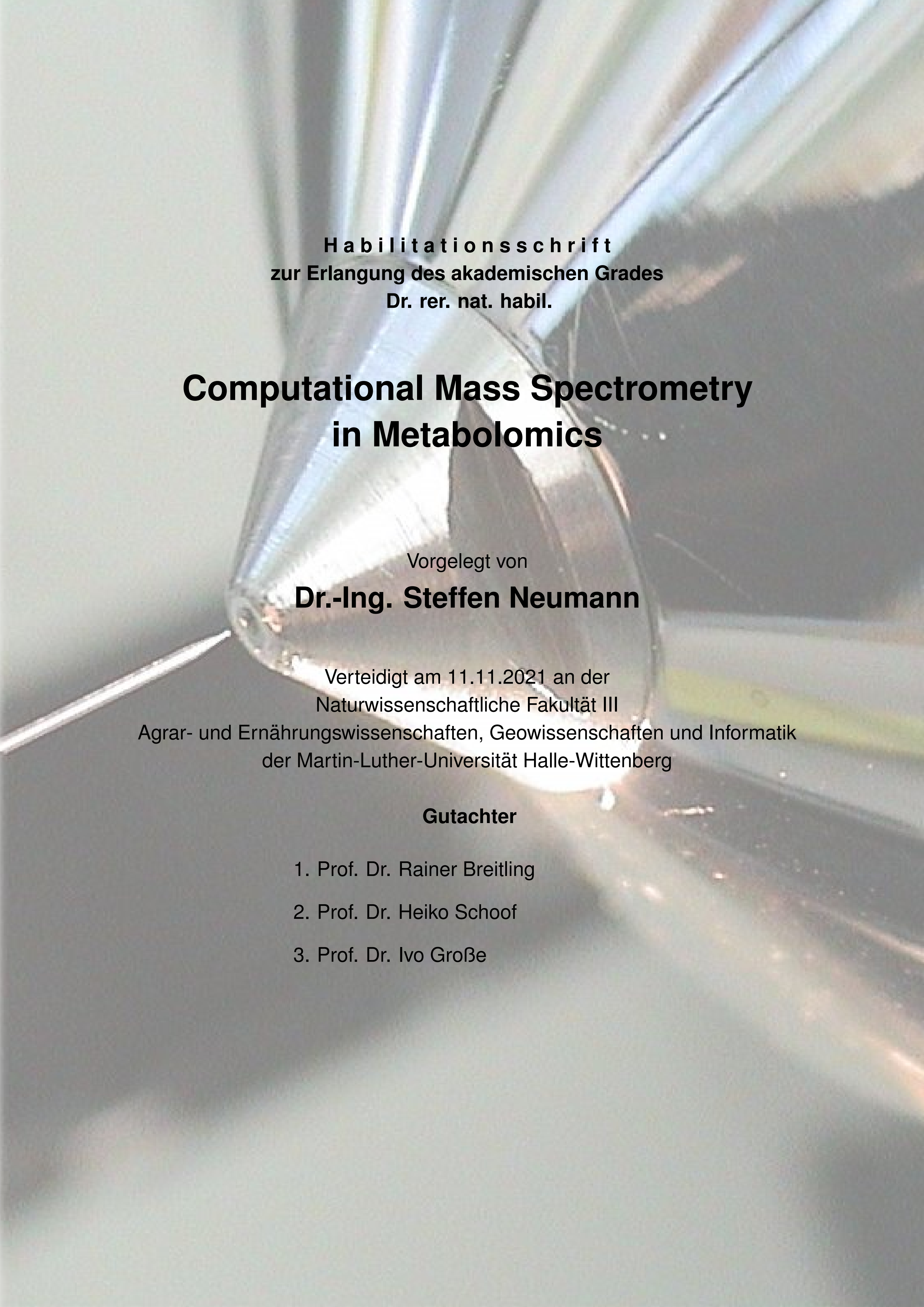


Habilitationsschrift
zur Erlangung des akademischen Grades
Dr. rer. nat. habil.

Computational Mass Spectrometry in Metabolomics

Vorgelegt von

Dr.-Ing. Steffen Neumann

Verteidigt am 11.11.2021 an der
Naturwissenschaftliche Fakultät III
Agrar- und Ernährungswissenschaften, Geowissenschaften und Informatik
der Martin-Luther-Universität Halle-Wittenberg

Gutachter

1. Prof. Dr. Rainer Breitling
2. Prof. Dr. Heiko Schoof
3. Prof. Dr. Ivo Große

Contents

I. Computational Mass Spectrometry in Metabolomics 5

1. Introduction 7

2. Metabolite Profiling 11
  2.1. Mass Spectrometry in Metabolomics 11
  2.2. Data Processing 13
    2.2.1. Feature Detection 15
    2.2.2. Deconvolution 16

3. Metabolite Identification 19
  3.1. Metabolite Identification with Reference Spectra 20
  3.2. *In silico* Identification 21
    3.2.1. Additional Structural Hints through Experimental Modification 23
    3.2.2. Integration and use of MetFrag in Metabolomics and Mass Spectrometry Software 23
  3.3. Integrated Identification with Spectral Libraries and *in silico* Approaches 23
  3.4. Critical Assessment of Small Molecule Identification Contest 24
  3.5. Integrating Multivariate Statistics and Metabolite Annotation 25
  3.6. Structure Elucidation with NMR 25

4. Data Sharing and Data Standards 27
  4.1. Data Standards and Data Repositories 27
  4.2. Metadata, Data Sharing and Reproducible Research 28

5. Software Environments and Biological Applications 31
  5.1. Metabolomics in R and Bioconductor 31
  5.2. Workflows for Metabolomics 32
  5.3. Biological Metabolomics Applications 33

6. Summary and Outlook 35

Bibliography 37

Articles in Journals 43

Preprints 51

Conference Proceedings 53

Books and Book Chapters 55

II. Reprints of Original Publications 57

7. Metabolite Profiling 59

8. Metabolite Identification 123

9. Data Sharing and Data Standards 233

10. Software environments and biological applications 271

Curriculum Vitae 345

Erklärung an Eides statt 347

Zusammenfassung

Unter Metabolomics versteht man Forschung an kleinen Molekülen z.B. in der Biologie mit dem Ziel, die Funktion von Stoffwechselprodukten (Metabolite) in biologischen Systemen zu beschreiben. Massenspektrometrie ist dabei eine Schlüsseltechnologie für die Messung der Metabolite. Durch den immensen technologischen Fortschritt in den letzten Jahren haben Menge und Komplexität der erzeugten Daten rasant zugenommen. Die Aufgabe der rechnergestützten Metabolomik ist es, Software und Datenbanken für Management und Analyse der Daten zu entwickeln.

Diese Arbeit beschreibt verschiedene Schritte einer typischen Analysepipeline von der Verarbeitung der Messungen und Kombination mehrerer Peaklisten aus verschiedenen Proben, mit dem Ziel, eine Datenmatrix zu erstellen. Die statistische Analyse solcher Matrizen hat dann das Ziel, interessante Metaboliten aufzudecken. Für deren biochemische Interpretation müssen die Metabolite identifiziert werden, einschließlich ihrer Molekularstrukturen.

Tandem-Massenspektren können dabei als Fingerabdruck der Moleküle genutzt und gegen Datenbanken mit Spektren bekannter Verbindungen verglichen werden. Informatische Ansätze ermöglichen die Identifizierung über Spektraldatenbanken hinaus. Diese Fortschritte in der rechnergestützten Metabolomik haben stark von der Entwicklung offener Datenformate und Repositorien profitiert, die die Grundlage der FAIR Prinzipien sind.

Wenngleich rechnergestützte Analysen die einzelnen Aufgaben wesentlich schneller erledigen als eine manuelle Interpretation, so geht der Nutzen über die reine Beschleunigung dieser Aufgaben hinaus. Die Effizienzsteigerung erlaubt die Bearbeitung gänzlich neuer, bis dato zu komplexer Analysen. Diese Herausforderungen treiben auch die Entwicklungen in der Informatik voran, von Datenbanken für die ständig wachsenden Datenmengen über effizientere Algorithmen bis hin zu Visualisierungen großer (biologischer) Netzwerke.

Stichworte: Computergestützte Massenspektrometrie, Metabolomik, Datenprozessierung, Metabolitenprofil, Metabolitenidentifikation, Datenpublikationen, Datenstandards

Abstract

Metabolomics is the modern term for the field of small molecule research in biology with the aim to capture the metabolites in biological systems and describe their biochemical role. Today, mass spectrometry is a key technology for metabolomics research. Due to immense technological advances in mass spectrometry over the last years, the amount and complexity of the data produced has been growing rapidly. The task of computational metabolomics is to develop tools and databases for the handling and analysis of mass spectrometry data.

This thesis describes the steps in a metabolomics data processing pipeline, from processing of signals and alignment of several peak lists from different samples into a data matrix. The statistical analysis of metabolomics experiments will reveal a number of "interesting" metabolites, but for the biochemical interpretation it is required to determine the metabolite identities including their molecular structure.

Tandem mass spectra can be considered a fingerprint of a molecule, and thus it is possible to create databases of spectra from known compounds for later comparison. Computational approaches allow identification beyond spectral databases. Most of the advances in computational metabolomics came along with the development of open data formats and repositories of open data. Together they are the basis of FAIR data.

While computational and integrated approaches are certainly faster than performing the individual tasks manually, the real benefit is beyond mere speed-up of such tasks. Ultimately they allow to answer biochemical questions that could not be tackled before, and also spur novel developments in computer science, ranging from databases for the ever growing amounts of data, to faster or more efficient data analysis algorithms or visualisation approaches for large (biological) networks and their dynamic behaviour.

Keywords: Computational mass spectrometry, metabolomics, data processing, metabolite profiling, metabolite identification, data publication, data standards

Preface

This habilitation thesis was written after several years of research on computational mass spectrometry and metabolomics at the Leibniz Institute of Plant Biochemistry (IPB Halle). During that time, I had many collaborations with excellent researchers at the IPB, in Germany and world-wide. Without these collaborations, the work would have been less exciting and less successful. For this reason I use the pronoun “we” in most places of this work.

The first part of the thesis gives an overview of the field of metabolomics and brings the individual publications into perspective. The bibliography at the end of Part I is split into several categories. All numeric citations like [5] refer to work in the scientific community. The citations with author abbreviation like [BAN+14] refer to own published work in refereed articles, preprints, conference proceedings and books. Part II provides reprints of selected original research articles grouped analogous to the chapters in Part I. The full final article is included where the publication license allows, or an author preprint otherwise.

Part I.

Computational Mass Spectrometry in Metabolomics

Introduction

1

Biology is the research of “living matter” and spurred the interest of bright minds for hundreds of years. Mendel described principles of inheritance [1] and modern molecular biology has been studied for more than 50 years [2], resulting in important discoveries about the relationship between genotype and phenotypes. Research in the life-sciences aims at the understanding of living organisms, where all processes between the genome and the phenotype are of interest. The subject of studies include gene regulation, protein synthesis, their post-translational modifications, and the biochemistry of proteins and small molecules – metabolites.

Metabolomics is the modern term for the field of small molecule research in biology, but the underlying questions have been addressed already for hundreds of years by physicians using the smell and colour (and hence metabolic state) of urine for diagnosis. In 1971, Pauling *et al.* [3] analysed more than 200 metabolites in breath and urine headspace, but the terms “metabolomics” or “metabonomics” only appeared in the scientific literature more than 25 years later [4, 5]. In the last two decades, huge progress has been made regarding the number of metabolites that can be (simultaneously) detected, lowering the limits of detection with modern analytical technologies, and the increased throughput of samples that can be processed.

Today, *mass spectrometry* is a key technology for metabolomics research. Due to immense technological advances in mass spectrometry over the last years, the amount and complexity of the data produced has been growing rapidly. These advances would not have been possible without the extensive use of computers throughout the data processing and -analysis steps of the experiments. While the first mass spectrometers used photo platters to record spectra, computers such as the setup shown in Figure 1.1 became an integral part of the instruments already in the 1970s [6].

The digital recording of mass spectra also allowed to couple the MS instruments to chromatographic separation, such as liquid (LC-MS) or gas chromatography (GC-MS). These separation processes greatly reduce the complexity of the individual mass spectra, which in turn allows to measure more complex samples, such as full methanolic extracts of plants or human body fluids. The amount of data from raw spectra is overwhelming, hence a *feature detection* step is typically applied to extract the chromatographic and spectral peaks into so called feature lists. These feature lists can be used as metabolic fingerprints, which represent a molecular phenotype. Typical metabolomics experimental designs include the comparison

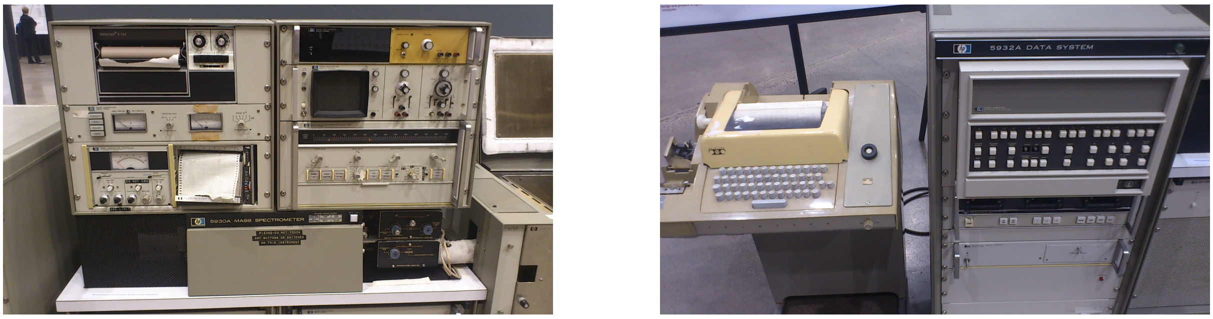


Figure 1.1.: The HP 5930A GC/MS (left) is coupled to a HP 5932A data station (right), which captures the spectra and stores them on magnetic tape. This system was exhibited at the 61$^{\text{st}}$ ASMS Conference on Mass Spectrometry and Allied Topics in Minneapolis in 2013. (Own photography)

of different genotypes, intervention studies or time-series experiments. These setups require the processing of dozens to hundreds, even thousands of samples. With microarrays it is possible to quantify the abundance of RNA and directly compare the gene expression across samples. In LC-MS and GC-MS however, the peak lists need to be matched across samples, and both chromatographic shifts and mass deviations have to be considered or even compensated for. A single metabolite will often give rise to more than one feature, and several metabolites can have very similar masses and/or chromatographic retention times. Thus, another data processing step is the grouping of features into compound spectra, and the annotation of ion species. Chapter 2 will describe the contributions to the data processing steps in metabolite profiling and give the required background on mass spectrometry.

For the biochemical interpretation, it is required to determine the metabolite identities, including the molecular structure. A main advantage here is that MS is independent of the availability of the genome sequence, and can be applied to any organism and tissue type. On the other hand, both analytical limitations and the chemical diversity of metabolites and biochemical processes prevents that all possible features are known *a priori*.

Thus, *metabolite identification* is an important task in computational metabolomics. For several organisms, including human and model organisms such as *Arabidopsis thaliana*, metabolite databases have been developed. If the compound is assumed to be known in that databases, it will be returned with a rather simple search for metabolites having a mass within an instrument-dependent error window. However, all compounds with a similar mass and of course all with the same molecular formula will be retrieved as false positive hits. Their number can be reduced if the molecular formula itself can be deduced from the accurate mass, isotopic pattern and further hints.

More structural information is available from higher-order mass spectra, such as tandem MS or MS^N. Here the analyte ions undergo fragmentation, and the fragmentation spectra provide a fingerprint of the molecular structure. Those spectra can be compared against reference data to identify the metabolite. Especially if no reference data are available, the spectra have to be interpreted, and structural hints can constrain the set of possible molecular structures. The topic and contributions to the metabolite identification task are described in Chapter 3.

The scientific discourse through letters among researchers and later articles in scientific journals has a long history going back centuries, but electronic data publications have emerged only in the last few decades. The amounts of data recorded in the life sciences mandate that they are available and enriched with experimental metadata. Contributions to open formats and structured data storage are described in Chapter 4.

All new methods developed show their value in real applications, and whether they can be applied by researchers worldwide. Most of our software implementations are available as Open Source software or easily accessible web applications. A holistic view of the resulting software environment and several biological applications are shown in Chapter 5. Chapter 6 concludes with a summary and an outlook.

Many of the challenges described here for metabolomics also apply to other – seemingly unrelated – disciplines. One task in environmental research is the monitoring of water quality which requires the profiling and comparison of samples across sites, time and in response to water treatment and the identification of unknown contaminants. Here, the environmental research questions will not be explicitly addressed unless the underlying problems or their solutions are markably different from the life sciences.

The task of *computational mass spectrometry* is to develop tools and databases for the handling and analysis of mass spectrometry data. While computational and integrated approaches are certainly faster than performing the individual tasks manually, the real benefit is beyond mere speed-up of such tasks. If determining the molecular formula of structure is fast, this enables to apply the analysis to entire experiments, rather than just one interesting metabolite. And if that data is available globally, this enables entirely novel data analysis strategies, and subsequently helps to answer biochemical questions that could not be tackled before [MRTN17]. These new applications also spur novel developments in computer science, ranging from databases for the ever growing amounts of data, to faster or more efficient data analysis algorithms or visualisation approaches for large (biological) networks and their dynamic behaviour.

Metabolite Profiling

2

In metabolomics, the aim is to capture the metabolite abundances in a biological system at a given point in time. In many experimental designs, the experimentalist will search for e.g. patterns in time series data, or differences between two or more sample classes representing wildtype and mutants, control versus treatment or healthy and diseased. This chapter will give an overview of mass spectrometry and the corresponding data processing steps for metabolite profiling.

2.1. Mass Spectrometry in Metabolomics

Due to the huge chemical diversity and limitations of today's analytical chemistry instrumentation, the full metabolome can not be obtained with a single technology. Mass spectrometry (MS) is a highly sensitive analytical method to characterise the composition even of complex samples. The samples can be in solution or in gas phase (e.g. in head space analysis). Although different types of MS instruments and configurations exist, the key principles remain the same. Figure 2.1 shows the schematic architecture of an LC-QqTOF mass spectrometer, and Figure 2.2 an actual instrument in operation at the IPB.

The molecules in the sample acquire a positive or negative charge in the *ion source*. In metabolomics, electron impact (EI) and electrospray ionisation (ESI) are the most widespread ionisation methods. The charged ions can then be accelerated with defined kinetic energy through an electric potential in the instrument. Multiply charged ions will consequently obtain a higher kinetic energy and travel at a higher speed. Therefore, mass spectrometry can only determine the *mass over charge ratio* m/z.

The actual m/z value is determined in the *analyser*. Common analysers include the quadrupole, which can be considered a bandpass filter where only ions within a specific m/z range can pass through. Cycling the filter mass from e.g. 50 to 600 in steps of one then allows to deduce the m/z or the detected ions based on the filter setting. Ion trap instruments can select (or trap) ions of a specified m/z value, before they induce an electric signal in the *detector*, where also the abundance is measured either in arbitrary units or sometimes as *counts per second*. The ion abundances across a specified m/z range is then called a *mass spectrum*. An example is shown in figure 2.3.

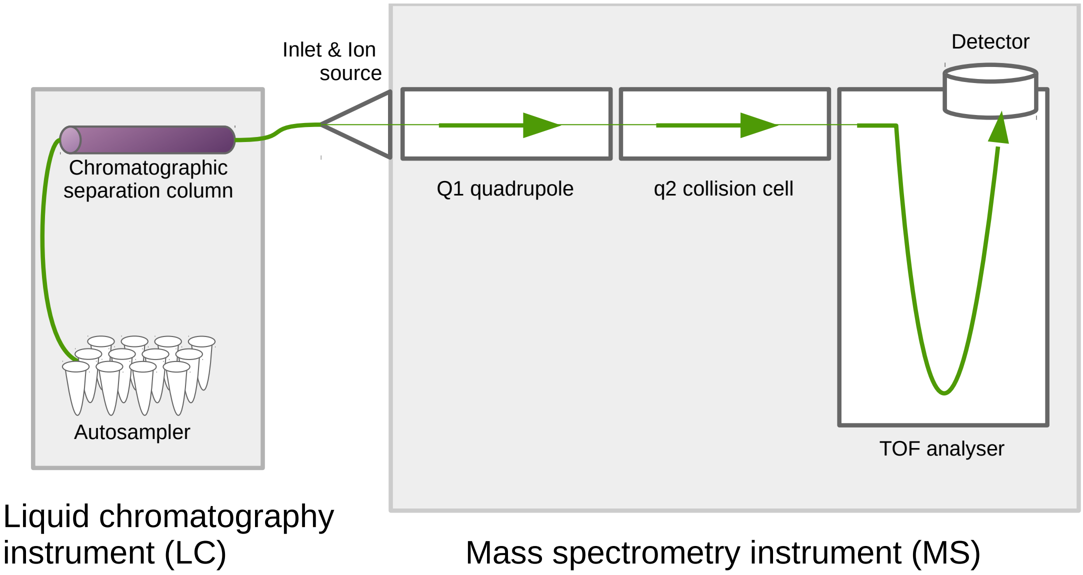

Inlet & Ion source
Detector
Chromatographic separation column
Q1 quadrupole
q2 collision cell
Autosampler
TOF analyser
Liquid chromatography instrument (LC)
Mass spectrometry instrument (MS)

Figure 2.1.: Schema and flow of sample and ions in an LC-QqTOF/MS instrument (left). Samples are selected in the autosampler holding e.g. 96 sample vials. One sample is subjected to separation in the chromatographic column, and transferred to the mass spectrometer. The ion source turns molecules into charged ions, which are then transferred to the time-of-flight (TOF) analyser and abundance is determined in the detector. Optionally, in case of MS/MS, ions can be selected in the quadrupole $Q1$, and undergo fragmentation in the collision cell $q2$ before transmission into the TOF analyser.

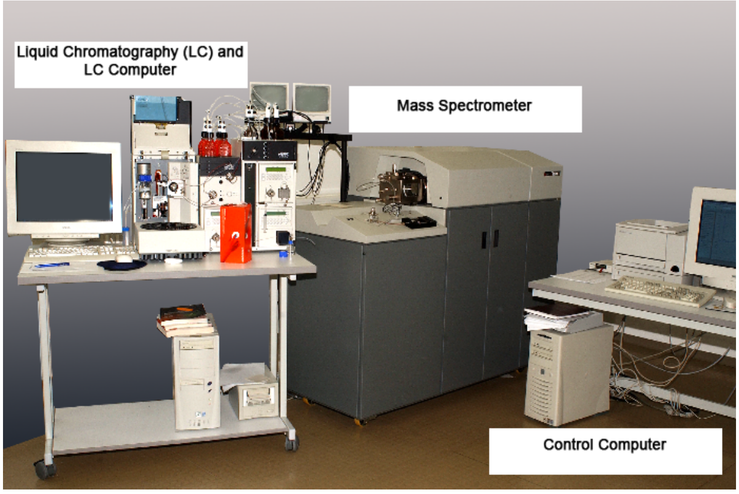


Figure 2.2.: A typical LC-QqTOF/MS in 2005 (Photo: Annett Kohlberg, IPB).

An optional step is the separation of the sample according to the physico-chemical properties with gas- or liquid chromatography, adding one (or more in the case of two-dimensional GC×GC or LC×LC separation) retention times to each spectrum. Recently, also ion mobility has been introduced into commercially available MS instruments, adding yet another type of separation. Here, the result is a *run*, where many spectra are measured at a rate of up to tens, rarely hundreds of scans per second.

During the ionisation process the uncharged molecules [M] obtain a charge, which is carried in positive mode by e.g. a proton H$^+$ to form a pseudo-molecular ion [M H]$^+$ or heavier adduct ions such as [M Na]$^+$, [M K]$^+$ or [M NH$_4$]$^+$, or in negative mode e.g. [M − H]$^-$ and [M Cl]$^-$. The ionisation is further complicated by dimerisation, where two or more molecules form a complex ion, e.g. [2 M H]$^+$, with corresponding increase of the observed m/z value. A doubly-charged ion, e.g. [M 2 H]$^{2+}$ with $z = 2$ will instead result in a m/z value of half the mass of the ion [M 2 H]$^{2+}$. Almost all of the atoms in the observed molecules [M] can occur in several isotopic variants with different numbers of neutrons in the nucleus and characteristic relative abundances. About 98.9% of carbon exists as ^{12}C with a mass of 12.0000, and 1.1% as ^{13}C with a mass of 13.0034. The natural relative abundance leads to the isotope patterns, i.e. groups of peaks with an average mass distance of ≈ 1.002 that can be observed in the example spectrum in figure 2.3. In general, all of these effects lead to complex mass spectra where a molecule gives rise to a multitude of peaks in the spectrum. On the one hand, this complicates the data analysis, on the other hand the defined mass differences can be exploited to annotate the spectrum.

Metabolite identification as discussed in Chapter 3 requires further structural hints in addition to the mass of the pseudo molecular ion. Multi-stage mass spectrometry introduces an additional *fragmentation* step in the MS instrument. In the schema in Figure 2.1 above, the $Q1$ quadrupole can be configured as a mass filter, which then selects ions of a specific mass. The collision cell $q2$ is filled with an inert gas such as nitrogen (N$_2$) or argon (Ar) at low pressure. An alternating electric field is applied to the quadrupole electrodes to apply energy to the ions, which dissociate into fragments upon collision with the gas, hence the term Collision Induced Dissociation (CID). Other fragmentation methods have been developed as well, such as infrared multiple photon dissociation (IRMPD) or higher energy collisional dissociation (HCD), which can result in different MS/MS spectra.

2.2. Data Processing

Regardless of the actual vendor and instrument category, a mass spectrometer will not measure the metabolome of an organism. Instead, it records the amount of ions arriving at the detector, while the biologists would be interested in concentrations of metabolites in a sample.

The HP 5930A/5932A system shown in Figure 1.1 was one of the first mass spectrometers that was sold together with a computer. In a “Scientific instrument selection guide” [7] they were advertised for $73 600, and were used e.g. to measure the abundance of altosid, which “is necessary for correlations of concentrations with biological response as well as to satisfy the requirements of regulatory agencies” [8]. Since then, computers are used in all stages of data processing and to facilitate the biological interpretation of the spectral data.

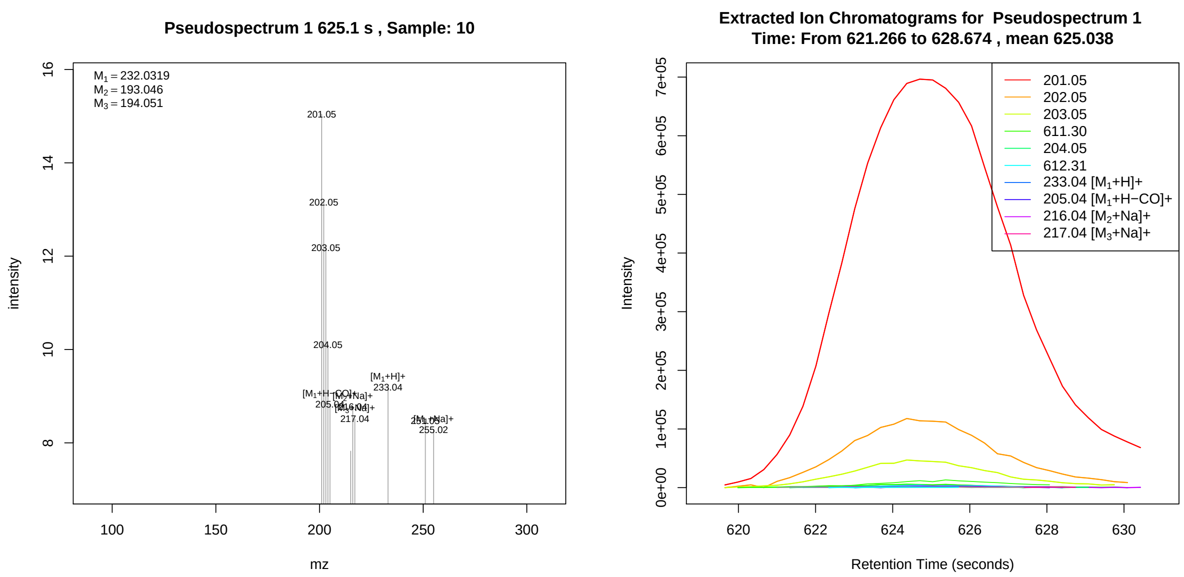


Figure 2.3.: Example mass spectrum of ions extracted by `CAMERA` (left), and corresponding extracted ion chromatograms (right). The metabolite causing the peaks is camalexin, measured on a Bruker micrOTOF-Q instrument from an *Arabidopsis thaliana* sample. Data is published in the MetaboLights repository as MTBLS2.

2.2.1. Feature Detection

The first step in a metabolomics data processing pipeline is the processing of signals, to reduce complex chromatographic data into peak lists, and align several peak lists from different samples into a data matrix. We are maintaining the successful Bioconductor package `xcms`, which is downloaded about 11 000 times per year[1] and was initially developed at the Scripps Institute [9].

The `xcms` package performs the steps 1) feature detection and quantification, 2) the retention time correction and alignment between several mass spectrometry runs and 3) the gap filling of features which were absent in some of the samples. The result is a rectangular $N \times M$ matrix with N columns for the samples and M rows for the detected features. The gap filling step is required because the downstream analysis often can not deal with matrices containing NA values, and imputing the missing values from the raw data is often a better choice than imputing the values based on the other samples or just random values.

Especially for high resolution LC-MS data we developed the feature detection algorithm centWave, which has become the algorithm of choice for high-resolution LC-MS data [TBN08]. For an evaluation of peak picking software (the new ccentWave and the original matchedFilter in `xcms` and mzMine [10]) we performed a thorough investigation of precision and recall of the algorithms. To do so, a known “ground truth” is needed to assess the quality of the algorithms. Instead of using an artificial mixture of a handful selected compounds and manual spectrum interpretation to determine “true” peaks, we used complex samples of Arabidopsis seed and leaf material, both pure at different concentrations and mixed at different ratios. The reliable peaks were then determined as the intersection of feature detected by all algorithms in several technical replication measurements. The rationale is that these peaks are the consensus features in the highest concentrations, but that their detection becomes more challenging at lower concentrations or even in mixtures of different samples, and that algorithms can be evaluated on the challenging samples. Using this reliable set of thousands of seed- and leaf specific peaks (removing those which occur in both samples), we were able to calculate the percentage of the “pure” features also found in the more complex mixture. We found that centWave has a better recall (i.e. finds more true peaks) and precision (i.e. less false positives) than both the matchedFilter approach and the peak picker in the open source tool mzMine. The evaluation data was published as MTBLS1430.

We also designed an evaluation protocol for the alignment of several LC-MS runs [LTNG08], which is the next step in the metabolomics workflow. The aim was again to create an unbiased ground truth with additional information that was not available to the alignment algorithms, instead of a (small) set of manually determined matching features. The ground truth was created from real-world samples, for which the alignment result was known with sufficient confidence from information not used for the actual alignment. This information included peptide identifications for the MS1 precursor ions obtained from tandem MS spectra, or ion species annotations such as isotopic patterns which have to occur also in the aligned data. This set consisted of 80.000 peaks. Several open source alignment algorithms were tested, again recording the statistically sound precision and recall measurements. As a side effect, the protocols developed for assessing the different peak picking and alignment tools are also useful to fine-tune the parameters to the chromatographic and mass spectrometry setup at the IPB. In both comparisons, the `xcms` algorithms were

---

[1] `http://bioconductor.org/packages/stats/bioc/xcms/`

top performers for the data acquired in the IPB metabolic profiling group, now available as MTBLS188. This excellent performance of the peak picking and alignment steps required a careful choice and optimisation of several parameters in the centWave algorithm, and relied on experience with both mass spectrometry and the data processing.

The authors of [11] describe an approach for a more objective optimisation of these parameters. A dilution series of a representative biological sample has to be measured in a pilot study on the given analytical setup prior to the actual experiment. The objective of the data processing is to detect as many features as possible with an intensity that correlates with the dilution steps. The optimal parameters were then found with a Design-of-Experiments strategy to avoid an exhaustive parameter scan. The drawback is the requirement for a separate pilot study.

To remove this requirement, the method described in [LDK15] implemented in the R package IPO relies only on the intrinsic properties of biological samples, namely the occurrence of isotopic peaks. The objective function can then be simplified to detect as many pairs of features as possible where one is the first isotopic peak.

The user defined parameters in the original centWave implementation was entirely untargeted and without any prior information about the sample and metabolites, even though some of that knowledge can be inferred from databases or previous experience. The feature detection algorithm apLCMS [12] used a database of expected or known metabolites. We implemented and evaluated a two-step strategy which uses a set of robust parameters in a first iteration of feature detection, and then extends the set of regions-of-interest (ROI) in centWave to cover expected peaks, e.g. isotopic peaks at predictable m/z values [TN16].

2.2.2. Deconvolution

In general there is a 1:n relation between the metabolites and the corresponding observed features in the MS data. Due to the characteristics of the mass spectrometry methods, the feature lists typically contain between 1 and 50 features for each single metabolite, where the additional features include isotope peaks, adducts and in-source fragmentation. It is desirable to assign features to metabolites for two reasons: 1) the mass differences between multiple features provide hints about the neutral molecular mass of the metabolite, which is of relevance for the biological interpretation, and 2) for the statistical analysis, the differences at the level of metabolites are of interest, rather than the underlying spectral features.

For the assignment of individual features to metabolites we first exploited their common intensity profile in the chromatographic domain, as shown in Figure 2.3. The underlying reason is that it is the metabolite that is traversing the chromatographic column, and only inside the mass spectrometer the individual ions are separated by their different masses.

We first created the ESI package [TBN07], which uses the retention time and pointwise correlation of the intensities across the chromatographic peak shape to assign peaks to *compound spectra*. Within these compound spectra, a fixed set of mass differences was used to annotate ions as e.g. [M+H] and [M+Na] if they have a mass difference of 21.9819 Da. This also creates a set of equations with one variable, and allows to calculate the mass of the neutral [M] molecule. The neutral mass can then be used to query metabolite databases, or to calculate the molecular formula, using e.g. our Rdisop package, which in turn is based on the decomp software library [13].

The ESI package was already very successful and formed the basis for the later developments on the `CAMERA` package [KTB12] to annotate ion species typically found in electrospray ionisation (ESI-MS). The **C**ollection of **A**lgorithms for **ME**tabolite p**R**ofile **A**nnotation in the `CAMERA` package represents the features as a graph, and multiple hints on the "relatedness" between pairs of features are assigned to the edges. In addition to the peak shape correlation, `CAMERA` also uses the intensity correlation across samples, which in turn is based on the assumption that the ratios between different ion species remains constant. This holds true for the majority of experiments, but in a few very specialised experimental designs the ratio may not be fixed. An example are salt stress experiments with different concentrations of Na in the samples. In `CAMERA`, the fixed set of mass differences was also replaced with a more dynamic rule set to cover a wider range of ion species.

A similar approach was used in the `RAMClustR` package, which is designed to assign precursor-product ion relationships in data-independent-acquisition (DIA) MS/MS data [BAN14]. This acquisition mode is a way to multiplex different instrumental parameters, but requires that corresponding features are linked for further interpretation.

Together, our developments formed the basis for automated mass spectrometry data processing pipelines, and have been used in a multitude of infrastructures and studies. These include e.g. integrated systems like MeltDB [14], MetaDB [14], XCMS online [15] or the workflows for Metabolomics [16] available for the Galaxy workflow environment.

Metabolite Identification

3

The statistical analysis of Metabolomics experiments will reveal a number of “interesting” metabolites. Although the MS profiling data often allows to determine the molecular formula, e.g. $C_9H_{15}O_6$ of the unknown features with reasonable reliability, no further identification of the molecular structure is possible from that information alone, because dozens or hundreds of compounds might be isomers of a single molecular formula. Mass spectrometry is also a key technology for the identification of small molecules.

In tandem mass spectrometry the ions are fragmented using e.g. collision induced dissociation (CID) as described in section 2.1. The resulting tandem mass spectra can provide additional structural hints. The acquisition of tandem mass spectra requires to specify the isolation window, i.e. specialised instrument parameters to isolate the precursor ion. This is either a tedious manual step, or it has to rely on data-dependent MS/MS methods (DDA, data-dependent acquisition) where a survey scan is acquired and the most intense top N ions are subsequently fragmented.. We have developed the MetShot approach [NTB12] to acquire tandem mass spectra of biologically relevant mass spectral features. After the acquisition, high-quality MS/MS spectra have to be extracted. To avoid the inclusion of background features, we combined the highly sensitive `xcms` feature finding and the compound spectra extraction from the `CAMERA` package.

Recently, developments in mass spectrometry instrumentation have introduced the concept of data *independent* acquisition, in short DIA. Here, the fragmentation spectra are acquired with either no precursor isolation at all (called MS^E for instruments manufactured by Waters), or in very broad, e.g. 25 Da wide, isolation windows which successively cover the whole mass range of interest. The latter has been termed **S**equential **W**indow **A**cquisition of all **TH**eoretical mass spectra (SWATH) for Sciex instruments. In both cases, the assignment between the precursor and the fragments is a challenge. For the case of DIA the RAMClustR approach was developed [BAN14], where a hierarchical clustering is used to group features and assign them to the precursor.

The metabolomics standards initiative (MSI) has defined four levels of identification [17]. These are *level one* for comparisons of two or more orthogonal properties (such as retention time and (tandem) mass spectrum) against authentic standards measured in-house on an identical analytical setup, *level two* where the same comparison is performed against external or literature data including spectral libraries. In *level three* only the compound class is known, while *level four* refers to the “known unknowns”, i.e. compounds where the identity is unknown, but which have been detected in other samples as well. The scientific discourse in this area has not stopped, and we are continuing to contribute to recommendations on metabolite identification [CDF14].

A more detailed summary on metabolite identification methods including several case studies has been reviewed in [NB10] and later with a different focus in [DEW13].

3.1. Metabolite Identification with Reference Spectra

The tandem mass spectra can be considered as a fingerprint of a molecule, and thus it is possible to create databases of spectra from known compounds for later comparison. Several spectral reference libraries have been created in the last decades for gas chromatography coupled to mass spectrometry, mostly using electron impact ionisation (GC/EI-MS). This particular type of instrumentation benefits from highly reproducible spectra, resulting in a good coverage of chemical substances in the libraries, and also very stable spectra even across different instrument vendors.

Reference libraries for LC/ESI-MS/MS were in contrast comparatively small, and the comparability of spectra across analytical setups was much lower. The MassBank consortium [HAK10] developed the first open database of reference spectra, accepting spectra from the community, but also spectra deposited as supplemental information for journal publications. The IPB Halle was the first European member of the MassBank consortium and is hosting a MassBank server[1]. We develop an ecosystem of tools and workflows around MassBank. This spectral database is an important resource for metabolomics researchers, but also the foundation for the development of computational mass spectrometry methods for metabolite identification. The spectra are used to train and validate computational models. MassBank records are now carrying explicit licensing terms, and in most cases the open Creative Commons license is used. In addition to online searches, the records are available for download and can be accessed through the version control system git. Since 2019, we prepare releases of the MassBank spectra. Releases are assigned a DOI and are archived on Zenodo (DOI 10.5281/zenodo.3570989).

With the multitude of spectral libraries available, the question arises how similar or different they are in their coverage. This can be evaluated comparing the InChiKeys (a hash value of a molecular structure) of the compounds contained in the libraries. The review in [VSN16] summarises the characteristics and analyses the content and overlap between different open and commercial libraries. Beyond the sheer number of unique or shared compounds or adducts covered, the next question is the coverage with regard to different biological questions. In [FSN18] the analysis was continued to determine the coverage between spectral libraries and genome scale metabolic networks (GSMN) for different model organisms. The GSMN were further analysed to determine distribution of neighbourhood coverage, whether known pathways are over-/underrepresented in the (non-)covered metabolites and occurrence in the scientific literature. Such an analysis also allows to declare a set of “most wanted” metabolites, for which reference spectra should be acquired to improve the coverage.

---

[1] `http://msbi.ipb-halle.de/MassBank/`

3.2. *In silico* Identification

Because reference spectra are often expensive to obtain (both in consumables and chemicals, but even more so in manpower), reference libraries will never be covering as many compounds as can be found in general purpose compound databases. Therefore, we are developing the MetFrag system [WSMHN10, RSW16, WRNSK17, RNP19]. The tool uses the tandem mass spectrum and the calculated mass of the neutral compound as input to search chemical structure databases such as KEGG, PubChem or ChemSpider for matching molecules. In some cases, it can be necessary to consider not only the known-unknowns, but also the unknown-unknowns where the structure has not yet been deposited in a chemical database. For that case, users can upload sets of structures as structure-data files (SDF format). Regardless of its origin, for each candidate every possible fragment is created using several heuristics. Because a mechanistic simulation of the process is computationally infeasible, we employ simplified *in silico* fragmentation methods, statistical models, and apply machine learning to a large set of training spectra. A user-friendly web application is available[2], but we also provide the source code under the LGPL open-source license for local deployments, a command line version and the R package `MetFragR` for inclusion into workflows.

The processing of a typical candidate structure might only take a few seconds or even less, but the candidate query in a large database like PubChem can return hundreds to thousands of candidates, which results in overall runtimes between minutes and several hours. Initially, the candidates were processed in an unspecified pseudo-random order as provided by the upstream structure database. If instead they could be retrieved in a pre-sorted order, where the promising candidates are processed first, the user could be presented with preliminary results, and might choose to not process the less promising candidates at all. Both the preliminary scoring and the sorted retrieval are required to be computationally very efficient. To that end, we created the MassStruct system, which is a relational database of the PubChem content, and includes an initial training step that allows to obtain a fast preliminary scoring for the candidate structures. In the training, we created a lookup table between observed tandem MS peaks and putative fragment structures, and stored them in a relational database. In our case, we chose the Open Source PostgreSQL with the chemistry extension pgchem::tigress, which in turn is based on the OpenBabel toolbox [18]. This combination allows to combine both the precursor information (exact mass or molecular formula) and structural properties like the “is-a-substructure-of” predicate in complex SQL statements. The performance and runtime of this approach were shown in [HWN11].

Because MetFrag allows both a precursor mass and a molecular formula based candidate selection, we also demonstrated in [NRWB13] the combined use of the SIRIUS tool [19] for the *de novo* calculation of fragmentation trees with MetFrag. In addition to a smaller candidate set, the calculated molecular formulae for the fragment peaks also allow to use the theoretical fragment masses instead of the measured ones, which can decrease the false positive rate of matched fragment peaks.

A problem arises if the correct compound is not contained in a compound database. Without the molecular structure, the MetFrag approach is not applicable. For small compound structures it is feasible to use structure generation to generate all plausible structures from a given molecular formula, which in turn can

---

[2] `http://msbi.ipb-halle.de/MetFrag/`

be deduced from the accurate mass and isotopic pattern of the unfragmented precursor. In [SGK12] we also used semiempirical quantum chemistry calculations to eliminate energetically unfeasible generated candidate structures.

A second approach to obtain candidate structures in cases where the chemical coverage for a given molecular formula in the chemical database is low was described in [GKN13], and included one case where the PubChem database had not a single candidate structure for the molecular formula of a novel metabolite discovered in *Nicotiana attenuata*. To overcome the limited coverage of metabolites for Nicotiana, we used structures of metabolites that are structurally related to the unknown compound: correlation networks of metabolites which show similar abundance behaviour across the samples can provide information which metabolites are co-regulated, and thus potentially originate from related biochemical processes. It had been proposed already in [20] that in a correlation network the identification of one node can support the identification of a connected node. Thus, we pooled all results from candidate queries using the precursor masses of nodes in the direct neighbourhood of an unknown metabolite to obtain chemically possibly related candidates. While these are guaranteed to be *not* the unknown structure if their precursor mass is different, especially those with a good MetFrag score can be expected to reveal structurally similar compounds. An experienced experimentalist can then use these together with the mass difference between the neighbouring nodes to deduce the possible structure of the unknown compound.

The results from MetFrag are the ranked candidate lists. They consider the score, but no chemical or structural information is used to navigate the results. To overcome this, we used a hierarchical clustering based on the chemical similarity (calculated as pairwise Tanimoto distance between the molecular fingerprints) as a postprocessing step. This visual representation helps to interpret which compound clusters include candidates with high scores. The next step is to calculate a representative structure for the individual candidate clusters, and obtain the maximum common substructure (MCSS) within clusters of chemically similar compounds. This approach was shown in [NRWB13], [GKN13] and later automated in [SGRN14].

A difficulty in the assessment of the ranked candidate list is that different compound classes with different structural properties can achieve different ranges of candidate scores. It is hence possible that a poor score of a candidate in the correct compound class is similar or even lower than a score in a different class. As a consequence, we developed a classification system that predicts for each candidate whether its score *within* its compound class is likely to be correct [WRNSK17].

In the original MetFrag scoring terms all fragment peaks in MS/MS spectra are treated alike (only the m/z, intensity and bond dissociation energy are used as weights in the final score) and no context information is considered. The increasing amount of available MS/MS spectra can be used as training data to model associations between fragment peaks and likely fragment structures. For the training set, it is thus possible to estimate the probability distribution of fragment structures given the MS/MS peaks $Pf|m$. For the test set and during application, they are used as weighting term for the assignment of fragments in the candidate structures [RNP19].

3.2.1. Additional Structural Hints through Experimental Modification

Mass spectrometry is superior to NMR for structure elucidation due to its higher sensitivity, especially for low-concentration compounds. On the other hand, the information content of MS/MS spectra is comparatively limited. One way to boost the structural hints is to modify the unknown compound in question to obtain additional spectral information. Deuterium (denoted D with an atomic mass of 2.0141 Da) is a stable, heavy isotope of hydrogen (denoted H with mass 1.0078) with two neutrons in the nucleus. **H**ydrogen **D**euterium e**X**change (HDX) has been applied in structure elucidation with electrospray ionisation mass spectrometry since the mid-90s [21, 22]. Deuterium can be used in either the mobile phase of chromatography (e.g. with D_2O instead of H_2O), or inside the mass spectrometer as part of the curtain gas (ND_3). The resulting MS/MS spectra differ by the mass of the additional neutrons in the fragments, which in turn allows to score and differentiate candidate structures with different numbers or positions of exchangeable hydrogens. The detailed method and an evaluation were shown in [RSS19].

3.2.2. Integration and use of MetFrag in Metabolomics and Mass Spectrometry Software

Due to the LGPL Open Source license, MetFrag can be used by and integrated into software developed by external collaborators in academia and industry alike. It is also easy to create an URL pointing to a landing page that passes all the query information to the MetFrag web application, saving the user manual copy&paste into the browser.

MetFrag has been included in the MolFind software [23] developed at University of Conneticut. For users of the Bruker Data Analysis software SmartFormula3D, a direct link to the MetFrag web application is available. In the Bruker MetaboScape software, MetFrag has been integrated as a module directly into the application. Nonlinear Dynamics (a Waters company) has ported[3] MetFrag to C# and integrated it into their Progenesis QI software. Other developments by third parties involving MetFrag are the tools for suspect and non-target analysis in the BMBF project FOR-IDENT[4], the Python wrapper of MetFrag for MS/MS based identification of LC/MS data[5] as part of the Eawag enviPy workflow project, and the integration of MetFrag into the Global Natural Products Social Molecular Networking (GNPS) resource [24] at UCSD.

3.3. Integrated Identification with Spectral Libraries and *in silico* Approaches

The approaches introduced above use either a spectral reference library or *in silico* approaches. Both have their benefits, i.e. the former contain actual experimental measurements, while the latter are backed by large chemical databases and usually a better chemical coverage. But both also have their drawbacks,

---

[3] `https://github.com/NonlinearDynamics/MetFrag.NET`

[4] `http://for-ident.hswt.de/pages/en/tasks.php?lang=EN`

[5] `https://pypi.python.org/pypi/pymetfrag/`

especially the limited chemical coverage for the spectral libraries on the one hand, and the imperfect *in silico* fragmentation and scoring on the other.

MetFusion [GN13] is a strategy and system to combine the compound hypotheses obtained by these complementary identification approaches. This strategy combines the best of both worlds: the identification using spectral libraries if similar spectra are available, and the huge chemical coverage of the compound databases queried by MetFrag. In the MetFusion software, the query spectrum for an unknown is passed simultaneously to both MetFrag and MassBank.

The core idea is that the candidates considered by MetFrag do include the correct solution, possibly with a low score. At the same time, MassBank will return structures with a similar mass spectrum. As MassBank does not restrict the search to compounds with the same precursor mass, these results can include structures that are chemically related to the unknown query, assuming that compounds with similar structures also have similar mass spectra. In the publication [GN13] we also included a specialised concept for cross validation of the performance: the results depend highly on the MassBank content, and whether the spectral library contains spectra from the correct or similar compounds. For the evaluation we thus "pruned" all results from MassBank above a defined chemical similarity to the correct solution.

Independent of the evaluation, an analysis of the chemical similarity to compound databases also allows to detect "blank spots" in the spectral library, and to prioritise which reference spectra should be acquired. With the given data and under certain assumptions MetFusion can be expected to identify 2 500 of the 15 000 KEGG compounds in the top 10 among all PubChem candidates.

3.4. Critical Assessment of Small Molecule Identification Contest

Since environmental research and metabolomics share many analytical and bioinformatics challenges, we initiated cooperations with Eawag, the Swiss Federal Institute of Aquatic Science and Technology and the Helmholtz Centre for Environmental Research (UFZ), especially in the area of metabolite and small molecule identification. Together with Dr. Emma Schymanski, we started the CASMI contest series: the *Critical Assessment of Small Molecule Identification* in 2012.

This contest was the first event where a set of spectral information of "unknown" compounds (i.e. unknown to the participants) was provided, and the community was called to submit hypotheses in Category 1 for the molecular formulae and the molecular structures for Category 2. Together with the challenges described in [SN13b] we have created a set of rules and an automatic evaluation pipeline. This allowed the rapid comparison of the available tools in an unbiased way after the contest submission deadline.

Because we were part of the organising team and possessed full knowledge about the "unknown" challenge compounds, we could only take part as internal participants [RGN13]. Together, all submissions from the four external participants had their strengths and weaknesses. The team of Rick Dunn (University of Birmingham) had the highest number of correct molecular formulae correct in Category 1, but their approach was to determine first the possible molecular structures, and then to submit their molecular formulae [25] to Category 1, while the team Dührkop et al. [26] used a *de novo* formula prediction without any database support, based on MS data alone. The final evaluation [SN13a] showed that while the molecular formula was found ranked first by one of the participants in 11 out of 14 cases, and always among the top 5, the correct structure was found only five times ranked first, and 8 times among the top 10.

After the initial contest CASMI was repeated, where the organisation was performed by different teams on different continents. The 2013 CASMI edition was organised by Prof. Takaaki Nishioka (Nara Institute of Science and Technology, Japan) and the 3^{rd} edition in 2014 was organised by Rick Dunn and members of the metabolite identification focus group of the Metabolomics Society. The 4^{th} edition was organised by Dr. Grégory Genta-Jouve (University of Paris Descartes, France), Prof. Olivier P. Thomas (University of Nice Sophia Antipolis, France) and Dr. Coralie Audoin (Laboratoires Clarins, France), and the 5^{th} by Dr. Dejan Nikolic (University of Illinois at Chicago, US), Dr. Nir Shahaf (Weizmann Institute of Science, Rehovot, Israel), Dr. Emma Schymanski (Eawag, CH) and Dr. Steffen Neumann [27].

3.5. Integrating Multivariate Statistics and Metabolite Annotation

So far, the workflow has been sequential, where first the raw data was processed, followed by statistical analysis and then identification of the interesting features. With the development of MS instruments capable of acquiring MS/MS spectra for most features, it has become possible to perform an integrated analysis of the LC-MS and MS/MS data as described in [TTP16].

The necessary innovation in that approach was to align all LC-MS features and their corresponding MS/MS spectra into two matrices, connected by the precursor ion information. The first matrix is derived from the LC-MS data and has quantification information, with samples in the columns, and features in the rows. This matrix can be subjected to a multivariate statistics, e.g. a principal component analysis (PCA), but other statistical methods can be applied as well. The second matrix is assembled from individual MS/MS spectra and contains the precursor features in the rows and the MS/MS fragment information in the columns. This matrix can be subjected to hierarchical clustering analysis (HCA), resulting in clusters of high spectral similarity. With the precursor as linking information, it is then possible to highlight spectral clusters within the PCA, or vice-versa show for a selection of LC-MS features into which spectral clusters they map.

Under the assumption that spectral similarity often translates into similar biochemistry, this enables the discovery of regulated metabolite families. In addition to a study performed at the IPB with data published as MTBLS297, we also re-used the GC-MS dataset MTBLS288 and reproduced some of their multivariate analyses now with added spectral clustering.

3.6. Structure Elucidation with NMR

Nuclear magnetic resonance (NMR) instruments measure the resonance of atoms, and provide an orthogonal method for both metabolite profiling and the elucidation of molecular structures. Analogous to the described reference libraries for mass spectrometry, several databases with 1D and 2D NMR spectra of pure compounds exist, e.g. [28, 29, 30].

If reference spectra are not available, several approaches exist to simulate NMR spectra. Then a large number of molecular structures can be scored based on the agreement with the measured spectrum. Again it is possible to use known molecular structures, or generate them *de novo* based on the molecular formula and possibly additional structural constraints. We have compared the performance of several machine learning techniques for the prediction of ^{1}H NMR spectra [KENS08]. Such a prediction allows to generate the spectra for a large number of candidate structures and rank structures based on spectral similarity to the simulated spectrum.

Data Sharing and Data Standards

4

The FORCE11 initiative has published a set of guidelines [31] that help to make data *FAIR*, which is an acronym for making data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. The details of how to implement these criteria are deliberately left to the individual scientific communities, but in metabolomics many of the required components to make data FAIR have been developed in the last years.

4.1. Data Standards and Data Repositories

Data standards are not required if the data acquisition, processing and later analysis are all performed in a single software, where no data or intermediate results are stored to disk, analysis is limited to a single experiment and where reproducibility is of no concern. Good scientific practice is the exact opposite and mandates the availability of data, preferably in (open) data standards. In the light of the FAIR principles, the metadata made available helps to make data sets **F**indable, and (raw) data standards help to make the data **I**nteroperable.

Storage and processing of mass spectrometry and metabolomics data can not be performed with simple text formats or *ad hoc* defined spread sheets. The complexity of the underlying data and requirements of data exchange and future-proof archival require a well-designed data model.

The early raw data exchange format mzXML [32] had been developed at the Seattle Proteome Center (SPC), while mzData [33] had been developed in the context of the Human Proteome Organisation (HUPO) and Proteomics Standards Initiative (PSI) communities. Several conversion tools exist to create mzData and mzXML from mass spectrometry instruments and other file formats. Since then, the developer communities of both mzData and mzXML collaborated to develop the joint successor mzML [MCS10]. The PSI also created a set of related data standards with similar design principles. These principles include a comparatively simple XML schema where flexible annotations in form of controlled vocabulary (CV) parameters are used. These annotations are taken from the separately maintained PSI-MS ontology. To constrain these generic tag-value pairs that characterise the CV terms, a *mapping file* describes which branch of ontology terms are allowed in which place of the XML schema. New terms can be added to the ontology without modifying the schema or the software parsers. Together, the schema, ontology and mapping file allow for a robust and future proof file format. We have supported the mzML format in the software package `xcms` and the data import package `mzR` [CMB12]. Later, the PSI developed the TraML format [DCN11] for the description of multiple reaction monitoring (MRM) and tandem mass spectrometry (MS/MS), which follows the same design principles. Nuclear magnetic resonance (NMR) is another important analytical technology in metabolomics, and the nmrML standard [SJW18] has been developed with a large number of international contributors, coordinated by the EU projects COSMOS and later PhenoMeNal.

While such data files can of course be stored on normal file systems, the development of databases for mass spectrometry data facilitates complex queries to retrieve a subset of current and archived data. The Model Driven Architecture paradigm (MDA) in software development allows to generate large parts of the required software and databases from a graphical model (Universal Modelling Language, UML) or from the XML Schema Definition (XSD). We have created graphical editors for the mzData raw data and ArMet (see Section 4.2 below) metadata standards, and developed a prototype for an infrastructure which allows experimentalists to edit, store and annotate their mass spectrometry data. It uses the Eclipse framework to generate Java objects, XML input/output bindings, database persistence and a user-friendly editor for both the XML files and database content. A prototype of a web frontend has been created to view, verify and upload to such a repository [KN06].

We also designed a data warehouse to store the results of the data processing described in section 2.2, i.e. the features detected in the MS data [GN07]. A data warehouse is a database optimised for online analytical processing (OLAP), and allows to retrieve subsets of data based on arbitrary filter criteria. The BioMart framework [34] has been used to establish the MetHouse database for preprocessed peaks as obtained e.g. by `xcms`. BioMart provides several frontends, including standalone and web interfaces, and a powerful command line and scripting client. Possible filters against the warehouse are e.g. plant genotype, growth conditions, treatment, ionisation mode or the model of the MS instrument used.

4.2. Metadata, Data Sharing and Reproducible Research

The term “reproducible research” refers to the ability to recreate and confirm a given analysis. For reproducible research it is a necessity that data underlying a publication is available as open data. If the used analysis software and scripts are published alongside with the data, together they allow to easily repeat individual steps or the whole analysis.

Once the data is in a vendor independent machine readable format, the next step is to publish various -omics data in a well annotated format, according to community accepted (at least minimal) information about the experiment [SRSF12].

The ArMet (*Ar*chitecture for *Met*abolomics) model [35] used in the databases described above has been one of the first implementations of a metadata format, and later served as the basis of the Core Information for Metabolomics Reporting (CIMR) described in [36]. Other -omics disciplines also resulted in checklists, which later on have been consolidated under the umbrella of the Minimum Information for Biological and Biomedical Investigations (MIBBI) community [KFS10].

But these individual checklists had not been created with a common, simple machine readable format in mind. This was later specified by the ISA-Tab consortium in form of the ISA-Tab format, which consists of several tab delimited spreadsheet-like files capturing information about the **I**nvestigation, one or more **S**tudies per investigation and one or more **A**ssays per study. In addition to the format itself, a whole set of related tools, databases and web applications have been created. We have contributed to the design of the ISA tools [RSBM10] and the related Risa package [GBNM14].

The MetaboLights repository [HSC13] at the European Bioinformatics Institute (EBI) is the first open access and long-term archive for metabolomics data. We started early to prepare data submissions for MetaboLights to support the design and testing of this repository, but also to have well-annotated data sets for the development of our software. These included MTBLS2 with MS data collected at the IPB as supplemental data for the MetShot publication [NTB12], MTBLS10 with MS data collected as supplemental data for [GKN13], MTBLS74 as supplemental data for [TSGN14] and MTBLS169 for [KTB12] and several more. It shows that leading-by-example can help to increase the adoption of the open data concept.

The metadata in ISA-Tab format serves not only as a description of the experiments performed, it also connects the experimental design to the data in the assays. The Bioconductor package Risa [GBNM14] can import an ISA-Tab description and calls analysis functions from e.g. `xcms` to create an `xcmsSet` object with the detected features. The subsequent statistical analysis can directly use the experimental design factors captured in the ISA-Tab information for the groupwise statistical analysis, visualisation of the data or for supervised machine learning algorithms.

Together, the open formats, data repositories, machine-readable metadata and integrated scripted analysis tools are the foundation of reproducible research in metabolomics. The computational tools such as those described in the previous chapters are becoming an integral part of conducting and analysing metabolomics experiments, and can be described in terms of standard operating procedures (SOPs), just like the SOPs for individual steps in the wetlab and analytical chemistry. All these aspects of why standards are important, which ones to use in metabolomics, what software and which libraries support them is summarised for both biologists and bioinformaticians in [RSSA16], together with examples that demonstrate the **R**eusability in cases where published data was re-analysed to showcase and compare novel data analysis strategies.

Software Environments and Biological Applications

5

Similarly to the data sharing efforts, it is of high importance to also pave the way for a software ecosystem to process and analyse metabolomics data. Finally, examples for the application of the previously described approaches in metabolomics experiments will be presented.

5.1. Metabolomics in R and Bioconductor

The R language was initiated by Ross Ihaka and Robert Gentleman under the umbrella of “The R Project for Statistical Computing” in 1992. The language is inspired and mostly compatible to the statistical language S, developed at the Bell Laboratories already in the 1970s. The package system in R and the Comprehensive R Archive Network (CRAN) spured literally thousands of contributed packages. On May 1st, 2002 the Bioconductor version 1.0 was released. Bioconductor started out as “an initiative for the collaborative creation of extensible software for computational biology and bioinformatics (CBB)” [37]. While the initial focus was on gene expression data and gene annotation, packages for other areas emerged soon after.

The organisation in individual packages allows to create a modularised data analysis pipeline for metabolomics data Starting from the signal processing tasks on mass spectrometry raw data. `xcms` has been initiated at the Scripps Institute as open source software and was extended and maintained at the IPB later. Figure 5.1 shows an overview of R packages for metabolomics and mass spectrometry that are (co-)developed at the IPB. These packages are just a subset of the impressive dependency network shown in Figure 2 of [SBH19].

Bioconductor enforces that all packages are tagged with different labels, to facilitate the organisation into BiocViews, collections of related packages. As of February 2020, the metabolomics biocView includes 63 packages related to metabolomics, ready to be used in research and education. Even more metabolomics packages are available on CRAN and in other package and source code repositories, but they are more difficult to find. Our recent review [SBH19] collected these and a much larger set of R packages for metabolomics into a review. In addition to the literature review, the publication also contains code to perform

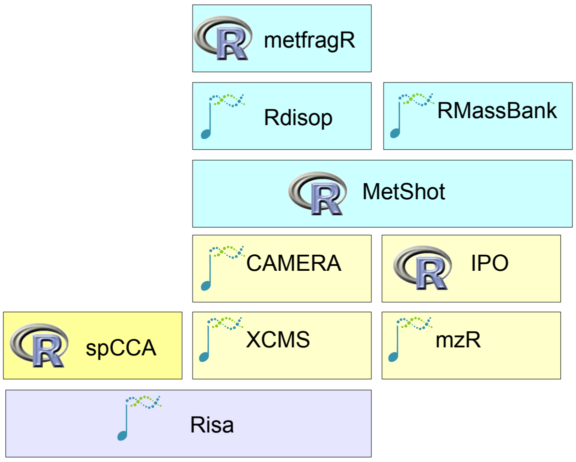


Figure 5.1.: R packages for metabolomics and mass spectrometry, developed, co-developed or maintained at the IPB. Some packages have been accepted into Bioconductor as indicated by the logo. Light yellow coloured packages are used for MS data processing, while the functions in the cyan coloured boxes are for metabolite identification. Risa deals with experimental metadata, and the purpose of spCCA is the combination of multiple -omics datasets.

a network analysis of the packages covered. The review was later turned into a live-book hosted at the RforMassSpectrometry project[1].

5.2. Workflows for Metabolomics

Not all metabolomics related data analysis software has been written in R, other projects have chosen Matlab, Python, Java, C++ or Python (to mention just a few) as the underlying programming language.

In computer science, workflow systems have been developed that usually allow to combine and integrate individual tools or modules from different sources and languages. The use of Galaxy [38] in metabolomics have been pioneered by the Workflows4metabolomics team in France [16], and the Galaxy-M team in Birmingham [39], with later contributions by the EU H2020 project PhenoMeNal [PBB19]. The KNIME workflow environment offers slightly different concepts, but in general a software module can be integrated in both the Galaxy and KNIME environment [40, 41].

Using a combination of public metabolomics data repositories and Galaxy instances with preinstalled data processing modules inside a docker container it is possible to provide a fully reproducible research workflow, as demonstrated in [PGBN18a].

---

[1] `https://rformassspectrometry.github.io/metaRbolomics-book/`

5.3. Biological Metabolomics Applications

One of the important areas of biological research is the elucidation of metabolic pathways. Before any deeper biochemical characterisation of the reactions can be performed, it is important to determine the “ingredients” of a pathway, i.e. the substrates, enzymes and resulting products of the biosynthesis steps. For a given enzyme, untargeted metabolite profiling can be used to obtain information about the putative substrates and products.

In [42] an experiment was performed to describe several steps in the biosynthesis of camalexin. Wild-type and *cyp79B2 cyp79B3* double knockout mutants were analysed with UPLC-ESI-QqTOF/MS. The hypothesis is that features that are missing in the knockout mutant are downstream of the enzyme, whereas features that accumulate are upstream of the enzyme. In [NTB12] we repeated the analysis on a subset of the samples and performed the data analysis with the R packages `xcms`, `CAMERA` and the newly developed `MetShot`.

In the first step, we used the `xcms` feature detection together with the `CAMERA` ion species annotation to obtain a metabolite profile matrix. Then we determined the features which are 1) differential, 2) have a higher intensity in the wild-type than in the mutant, and are 3) annotated as [M+H] or [M+Na] ions. For the subsequent acquisition of tandem MS spectra we included only features 4) above an intensity threshold that promised tandem MS spectra of reasonable quality. The following identification step used the molecular formula, retention time and MS/MS reference spectra to re-identify several metabolites also known from previous publications, but also revealed annotations for six metabolites previously uncharacterised.

A second plant metabolomics experiment was performed in [GKN13]. The biological question was the response of *Nicotiana attenuata* to simulated herbivory. The plant leaves were mechanically wounded and treated with the oral secretions of *Manduca sexta* larvae to induce large-scale changes especially in the secondary metabolic network as defence response. The leaf material was collected at six different time-points, and methanolic extracts were measured on a HPLC-ESI-TOF/MS platform at the Max Planck Institute for Chemical Ecology (Jena). The MS data processing was performed with `xcms` and `CAMERA`. Of the $\approx$1 000 reliably detected features across all samples, 324 were annotated as monoisotopic peaks, and for 326 an annotation of the ion species was possible. Almost 400 features were changed (t-test against 0h timepoint, $p \leq 0.05$, no multiple testing correction) after the herbivory feeding. We created a correlation network using the Pearson coefficient to obtain a hint towards the metabolic links between pairs or groups of metabolites. These connections can result from direct enzymatic conversions in a biochemical pathway, but also from indirect transcriptional controls.

Two properties of both the biological system and the analytical setup made the identification of metabolites challenging: first, only nine metabolites are known for *Nicotiana attenuata* and about 800 compounds for the entire Nicotiana genus (as of 02/2020, KNApSAcK [43]). Secondly, the analytical platform was not equipped to measure tandem MS spectra. For the identification we extracted therefore *in source* fragment peaks from the MS data and used MetFrag for the candidate scoring.

The molecular structures were obtained from PubChem, but in addition to the candidates of the compound of interest itself, we also retrieved the candidates of the immediate network neighbours. Here the assumption is that the network neighbours are biosynthetically related. This “guilt-by-association” approach can improve the structure elucidation of unknown metabolites based on their co-regulation with known pathways or sets of metabolites by providing additional candidates from the same or a related compound class.

Finally, in this paper we performed a hierarchical clustering with the chemical similarity as distance measure, and calculated the Maximum Common Substructure (MCS) of cluster members to represent their consensus structure. The MCS was later also used by [44] to calculate constraints for structure generation to resolve unknown-unknown compounds.

The approaches shown are applicable to all areas of metabolomics research, and are not limited to plant research. In nutrition experiments, the biomedical question is to determine the influence of a given diet on the human system. This influence can be monitored by metabolite profiling of e.g. blood serum. In [SGDN13] we applied the computational mass spectrometry pipeline to a dataset of 220 samples obtained in a nutrition study at the KU in Kopenhagen, Denmark. We then applied the `MetShot` approach to acquire tandem MS spectra for the most interesting features. Since the profiling measurements were performed in Kopenhagen, whereas the tandem MS spectra were acquired at the IPB Halle, it was required to map retention times between the two chromatographic systems. We used a predictive model that was iteratively trained with initially a few manually determined landmark features, and then refined with additional features that were automatically matched by `xcms` after the retention time correction with the initial model.

For the subsequent identification, we used the MetFusion system to obtain and score candidate structures having the correct exact mass (within a given error margin) from ChemSpider. In addition to the MetFusion score, we used the R package `Rdisop` [13] to calculate the similarity between the theoretical and the observed isotopic ratio for each candidate. While traditionally the molecular formula is determined from the spectra of an individual sample, we evaluated several alternatives to exploit the large ($\approx$220 samples) profiling dataset. We found that the averaged isotope ratios had a better accuracy, especially if a subset of 10% of the samples were used which had the highest intensities.

As another hint for the identification we used the retention time predicted for each candidate, also known as Quantitative Structure Retention time Relationshop (QSRR) [45]. Instead of directly training based on on the chemical fingerprints, we used the $\log D$ and experimental retention time for a small set of authentic standards to create a model for the prediction of retention times. The $\log D$ was calculated with cheminformatics software (ChemAxon Marvin).

In this study we demonstrated the benefit of using a mostly automatic data processing pipeline and multiple hints to streamline the metabolite identification task. In particular, fewer authentic standards had to be purchased to finally confirm the identification at the highest confidence level as defined by the Metabolomics Standards Initiative.

Summary and Outlook

6

The previous chapters covered multiple aspects of computational metabolomics, the work and progress that has been made over the last 15+ years in my group and the computational mass spectrometry community in general.

Successful (untargeted) metabolomics research first require automated, high-throughput algorithms for metabolite profiling. Early typical metabolomics experiments were conducted with just several dozen samples. Today, larger experiments include several thousands of samples. Metabolite profiling allowed to uncover interesting features, which in turn require the identification of their molecular structure. Nowadays, workflows for metabolite identification can use a range of spectral libraries and *in silico* algorithms.

In future work, we can expect to see improvements in all above mentioned areas, simplifying the data processing and initial data analysis in metabolomics experiments. Especially the renaissance of neural networks in form of deep learning approaches can be expected to dramatically change and likely improve several tasks along the data analysis pipeline. The coming challenges in biology and biochemistry will be to truly combine and interpret data from multiple -omics technologies and to obtain an understanding on the systems level.

Acknowledgements

I would like to thank all my previous and current PhD students and Post-Docs, who worked hard all the time and made the Bioinformatics and Mass Spectrometry group at the IPB such a success story: Thank you Yvonne Pöschl, Ralf Tautenhahn, Björn Egert, Stefan Kuhn, Sebastian Wolf, Michael Gerlich, Carsten Kuhl, Diana Trutschel, Daniel Schober, Susann Mönchgesang, Hendrik Treutler, Kristian Peters, Sarah Scharfenberg, Chris Ulpinnis, Susanne Marr, Oliver Duchrow and many Diploma, Bachelor and Master students not named individually. The best recognition of successful work is to see you in successful positions around the globe after your time in my group.

The IPB has always been an inspiring place for metabolomics, and I appreciate the collaborations with members from all departments, including Christoph Böttcher, Stephan Schmidt, Jürgen Schmidt, Gerd Balcke, Sophie Dietz, Stefanie Döll, Karin Gorzolka and Dierk Scheel as a very supportive department head until his retirement in 2018.

Since todays research does not take place in isolation, many thanks also go to the national and international collaborators, including all initial and current contributors to xcms, from Gary Siuzdak and Colin Smith to Paul Benton, Johannes Rainer, Jan Stanstrup, Michael Witting and others, the cheminformatics and metabolism team of Christoph Steinbeck at the EMBL-EBI, inclucding Keeva Cochrane, Ken Haug and Pablo Moreno, the ISA team including Susanna Assunta-Sansone, Phillipe Rocca-Serra, Alejandra Gonzalez-Beltran, David Johnson and Eammon Maguire, the PSI community including Andrew Jones, Henning Hermjakob, Eric Deutsch and many more, the MassBank community including the japanese team around Prof. Nishioka and Prof. Arita, the WANA team at the UFZ around Werner Brack, Tobias Schulze and Martin Krauss, and the entire PhenoMeNal-H2020 members, and the de.NBI network including Oliver Kohlbacher and his team and Nils Hoffmann. I am also honoured to have long-term collaborators who have been in touch through different parts of their career, including Emma Schymanski from UFZ, Eawag and now Uni Luxemburg, and Egon Willighagen from Wageningen, Uppsala, Cambridge, Stockholm and now Maastricht.

I also have to acknowledge the funding agencies DFG, BMBF and the European Commission, who supported the projects, as well as the Leibniz association for both the institutional support and funding through the SAW projects. Also thanks to William S. Lane, Harvard University, Cambridge MA, for permission to use the cover photograph of a mass spectrometer ion source inlet.

Bibliography

[1] G. Mendel, *Versuche über Pflanzen-Hybriden (1865)*. Arkana-Verlag, 1865.

[2] W. T. Astbury, “Molecular biology or ultrastructural biology?,” *Nature*, vol. 190, p. 1124, Jun 1961.

[3] L. Pauling, A. B. Robinson, R. Teranishi, and P. Cary, “Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography.,” *Proc Natl Acad Sci USA*, vol. 68, pp. 2374–2376, Oct 1971.

[4] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, “Systematic functional analysis of the yeast genome.,” *Trends Biotechnol*, vol. 16, pp. 373–378, Sep 1998.

[5] J. K. Nicholson, J. C. Lindon, and E. Holmes, “’Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.,” *Xenobiotica*, vol. 29, pp. 1181–1189, Nov 1999.

[6] H. Budzikiewicz and M. Schäfer, *Massenspektrometrie*. Wiley-VCH, 5 ed., 2005.

[7] HEWLET PACKARD, “Analytical and scientific instruments: a selection guide,” *Analytical Chemistry*, vol. 44, no. 10, p. 52LG–55LG, 1972.

[8] L. L. Dunham and R. J. Leibrand, “Residue Determination of an Insect Growth Regulator by Mass Fragmentography,” *British American Tobacco collection*, 1974.

[9] C. Smith, E. Want, G. O’Maille, R. Abagyan, and G. Siuzdak, “XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification,” *Anal Chem*, vol. 78, no. 3, pp. 779–787, 2006.

[10] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic, “MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data,” *BMC Bioinformatics*, vol. 11, no. 1, p. 395, 2010.

[11] M. Eliasson, S. Rännar, R. Madsen, M. A. Donten, E. Marsden-Edwards, T. Moritz, J. P. Shockcor, E. Johansson, and J. Trygg, “Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach.,” *Anal Chem*, vol. 84, pp. 6869–6876, Aug 2012.

[12] T. Yu and D. P. Jones, "Improving peak detection in high-resolution LC/M metabolomics data using preexisting knowledge and machine learning approach," *Bioinformatics*, vol. 30, no. 20, pp. 2941–2948, 2014.

[13] S. Böcker, M. Letzel, Zs. Lipták, and A. Pervukhin, "Decomposing metabolomic isotope patterns," in *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006)*, vol. 4175 of *Lect. Notes Comput. Sc.*, pp. 12–23, Springer, 2006.

[14] N. Kessler, H. Neuweger, A. Bonte, G. Langenkämper, K. Niehaus, T. W. Nattkemper, and A. Goesmann, "MeltDB 2.0-advances of the metabolomics software system.," *Bioinformatics (Oxford, England)*, vol. 29, pp. 2452–2459, Oct. 2013.

[15] R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, "Xcms online: a web-based platform to process untargeted metabolomic data.," *Analytical chemistry*, vol. 84, pp. 5035–5039, June 2012.

[16] F. Giacomoni, G. Le Corguillé, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E. A. Thévenot, and C. Caron, "Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics.," *Bioinformatics*, vol. 31, pp. 1493–1495, May 2015.

[17] L. W. Sumner, A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. Lane, J. C. Lindon, P. Marriott, A. Nicholls, M. Reily, J. Thaden, and M. R. Viant, "Proposed minimum reporting standards for chemical analysis," *Metabolomics*, vol. 3, no. 3, pp. 211–221, 2007.

[18] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox.," *J Cheminform*, vol. 3, p. 33, Oct 2011.

[19] F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher, and S. Böcker, "Computing Fragmentation Trees from Tandem Mass Spectrometry Data," *Analytical Chemistry*, vol. 83, no. 4, pp. 1243–1251, 2011.

[20] R. Breitling, D. Vitkup, and M. P. Barrett, "New surveyor tools for charting microbial metabolic maps.," *Nat Rev Microbiol*, vol. 6, pp. 156–161, Feb 2008.

[21] M. E. Hemling, J. J. Conboy, M. F. Bean, M. Mentzer, and S. A. Carr, "Gas phase hydrogen / deuterium exchange in electrospray ionization mass spectrometry as a practical tool for structure elucidation.," *Journal of the American Society for Mass Spectrometry*, vol. 5, pp. 434–442, May 1994.

[22] W. Lam and R. Ramanathan, "In electrospray ionization source hydrogen/deuterium exchange LC-MS and LC-MS/MS for characterization of metabolites.," *Journal of the American Society for Mass Spectrometry*, vol. 13, pp. 345–353, Apr. 2002.

[23] L. C. Menikarachchi, S. Cawley, D. W. Hill, L. M. Hall, L. Hall, S. Lai, J. Wilder, and D. F. Grant, "MolFind: a software package enabling HPLC/MS-based identification of unknown chemical structures.," *Anal Chem*, vol. 84, pp. 9388–9394, Nov 2012.

[24] D. D. Nguyen, C.-H. Wu, W. J. Moree, A. Lamsa, M. H. Medema, X. Zhao, R. G. Gavilan, M. Aparicio, L. Atencio, C. Jackson, J. Ballesteros, J. Sanchez, J. D. Watrous, V. V. Phelan, C. van de Wiel, R. D.

Kersten, S. Mehnaz, R. De Mot, E. A. Shank, P. Charusanti, H. Nagarajan, B. M. Duggan, B. S. Moore, N. Bandeira, B. Ø. Palsson, K. Pogliano, M. Gutiérrez, and P. C. Dorrestein, "MS/MS networking guided analysis of molecule and gene cluster families.," *Proc Natl Acad Sci USA*, vol. 110, pp. E2611–E2620, Jul 2013.

[25] J. W. Allwood, R. J. M. Weber, J. Zhou, S. He, M. R. Viant, and W. B. Dunn, "CASMI-The Small Molecule Identification Process from a Birmingham Perspective.," *Metabolites*, vol. 3, no. 2, pp. 397–411, 2013.

[26] K. Dührkop, K. Scheubert, and S. Böcker, "Molecular Formula Identification with SIRIUS," *Metabolites*, vol. 3, no. 2, pp. 506–516, 2013.

[27] **S. Neumann**, D. Nikolic, E. Schymanski, and N. Shahaf, "Critical assessment of small molecule identification: Looking at the fifth edition casmi 2017." http://www.metabonews.ca/Apr2018/MetaboNews_-Apr2018.pdf.

[28] C. Steinbeck and S. Kuhn, "NMRShiftDB – Compound identification and structure elucidation support through a free community-build web database," *Phytochemistry*, vol. 65, no. 19, pp. 2711–2717, 2004.

[29] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, and A. Scalbert, "HMDB 4.0: the human metabolome database for 2018.," *Nucleic acids research*, vol. 46, pp. D608–D617, Jan. 2018.

[30] E. Chikayama, Y. Sekiyama, M. Okamoto, Y. Nakanishi, Y. Tsuboi, K. Akiyama, K. Saito, K. Shinozaki, and J. Kikuchi, "Statistical indices for simultaneous large-scale metabolite detections for a single NMR spectrum.," *Anal Chem*, vol. 82, pp. 1653–1658, Mar 2010.

[31] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR guiding principles for scientific data management and stewardship.," *Scientific data*, vol. 3, p. 160018, Mar. 2016.

[32] P. G. A. 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, and R. Aebersold, "A common open representation of mass spectrometry data and its application to proteomics research.," *Nat Biotechnol*, vol. 22, no. 11, pp. 1459–66, 2004.

[33] S. Orchard, C. Taylor, H. Hermjakob, W. Zhu, R. Julian, and R. Apweiler, "Current status of proteomic standards development.," *Expert Rev Proteomics*, vol. 1, no. 2, pp. 179–83, 2004.

[34] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney, "EnsMart: A Generic System for Fast and Flexible Access to Biological Data," *Genome Res*, vol. 14, no. 1, pp. 160–169, 2004.

[35] 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, and D. B. Kell, "A proposed framework for the description of plant metabolomics experiments and their results," *Nat Biotechnol*, vol. 22, no. 12, pp. 1601–1606, 2004.

[36] J. C. Lindon, J. K. Nicholson, E. Holmes, H. C. Keun, A. Craig, J. T. M. Pearce, S. J. Bruce, N. Hardy, S.-A. Sansone, H. Antti, P. Jonsson, C. Daykin, M. Navarange, R. D. Beger, E. R. Verheij, A. Amberg, D. Baunsgaard, G. H. Cantor, L. Lehman-McKeeman, M. Earll, S. Wold, E. Johansson, J. N. Haselden, K. Kramer, C. Thomas, J. Lindberg, I. Schuppe-Koistinen, I. D. Wilson, M. D. Reily, D. G. Robertson, H. Senn, A. Krotzky, S. Kochhar, J. Powell, F. van der Ouderaa, R. Plumb, H. Schaefer, M. Spraul, and S. M. R. S. w. g. , "Summary recommendations for standardization and reporting of metabolic analyses," *Nat Biotechnol*, vol. 23, pp. 833–838, Jul 2005.

[37] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics.," *Genome Biol*, vol. 5, no. 10, p. R80, 2004.

[38] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, p. R86, 2010.

[39] R. L. Davidson, R. J. M. Weber, H. Liu, A. Sharma-Oates, and M. R. Viant, "Galaxy-M: a galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data.," *GigaScience*, vol. 5, p. 10, 2016.

[40] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Data Analysis, Machine Learning and Applications* (C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, eds.), Studies in Classification, Data Analysis, and Knowledge Organization, pp. 319–326, Springer Berlin Heidelberg, 2008.

[41] S. Aiche, T. Sachsenberg, E. Kenar, M. Walzer, B. Wiswedel, T. Kristl, M. Boyles, A. Duschl, C. G. Huber, M. R. Berthold, K. Reinert, and O. Kohlbacher, "Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry.," *Proteomics*, vol. 15, pp. 1443–1447, Apr 2015.

[42] C. Böttcher, L. Westphal, C. Schmotz, E. Prade, D. Scheel, and E. Glawischnig, "The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana.," *Plant Cell*, vol. 21, pp. 1830–1845, Jun 2009.

[43] F. M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin, L. K. Darusman, K. Saito, and S. Kanaya, “Knapsack family databases: integrated metabolite-plant species databases for multifaceted plant research.,” *Plant & cell physiology*, vol. 53, p. e1, 2 2012.

[44] J. E. Peironcely, M. Rojas-Chertó, A. Tas, R. Vreeken, T. Reijmers, L. Coulier, and T. Hankemeier, “Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics.,” *Anal Chem*, vol. 85, pp. 3576–3583, Apr 2013.

[45] J. Ghasemi and S. Saaidpour, “QSRR prediction of the chromatographic retention behavior of painkiller drugs.,” *Journal of chromatographic science*, vol. 47, pp. 156–163, Feb. 2009.

Articles in Journals

[AAAA15] R. Altenburger, S. Ait-Aissa, P. Antczak, T. Backhaus, D. Barceló, T.-B. Seiler, F. Brion, W. Busch, K. Chipman, M. López de Alda, G. d. A. Umbuzeiro, B. I. Escher, F. Falciani, M. Faust, A. Focks, K. Hilscherova, J. Hollender, H. Hollert, F. Jäger, A. Jahnke, A. Kortenkamp, M. Krauss, G. F. Lemkine, J. Munthe, **S. Neumann**, E. L. Schymanski, M. Scrimshaw, H. Segner, J. Slobodnik, F. Smedes, S. Kughathas, I. Teodorovic, A. J. Tindall, K. E. Tollefsen, K.-H. Walz, T. D. Williams, P. J. Van den Brink, J. van Gils, B. Vrana, X. Zhang, and W. Brack. Future water quality monitoring–adapting tools to deal with mixtures of pollutants in water resource management. *Sci Total Environ*, 512-513:540–551, Apr 2015.

[BAAB16] W. Brack, S. Ait-Aissa, R. M. Burgess, W. Busch, N. Creusot, C. Di Paolo, B. I. Escher, L. Mark Hewitt, K. Hilscherova, J. Hollender, H. Hollert, W. Jonker, J. Kool, M. Lamoree, M. Muschket, **S. Neumann**, P. Rostkowski, C. Ruttkies, J. Schollee, E. L. Schymanski, T. Schulze, T.-B. Seiler, A. J. Tindall, G. De Aragão Umbuzeiro, B. Vrana, and M. Krauss. Effect-directed analysis supporting monitoring of aquatic environments - An in-depth overview. *Sci Total Environ*, 544:1073–1118, Jan 2016.

[BAN14] C. D. Broeckling, F. A. Afsar, **S. Neumann**, A. Ben-Hur, and J. E. Prenni. RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Anal Chem*, 86(14):6812–6817, Jul 2014.

[BAS15] W. Brack, R. Altenburger, G. Schüürmann, M. Krauss, D. López Herráez, J. van Gils, J. Slobodnik, J. Munthe, B. M. Gawlik, A. van Wezel, M. Schriks, J. Hollender, K. E. Tollefsen, O. Mekenyan, S. Dimitrov, D. Bunke, I. Cousins, L. Posthuma, P. J. van den Brink, M. López de Alda, D. Barceló, M. Faust, A. Kortenkamp, M. Scrimshaw, S. Ignatova, G. Engelen, G. Massmann, G. Lemkine, I. Teodorovic, K.-H. Walz, V. Dulio, M. T. O. Jonker, F. Jäger, K. Chipman, F. Falciani, I. Liska, D. Rooke, X. Zhang, H. Hollert, B. Vrana, K. Hilscherova, K. Kramer, **S. Neumann**, R. Hammerbacher, T. Backhaus, J. Mack, H. Segner, B. Escher, and G. de Aragão Umbuzeiro. The SOLUTIONS project: challenges and responses for present and future emerging pollutants in land and water resources management. *Sci Total Environ*, 503-504:22–31, Jan 2015.

[BvRLS08] C. Böttcher, E. von Roepenack-Lahaye, J. Schmidt, C. Schmotz, **S. Neumann**, D. Scheel, and S. Clemens. Metabolome Analysis of Biosynthetic Mutants Reveals a Diversity of Metabolic Changes and Allows Identification of a Large Number of New Compounds in Arabidopsis. *Plant Physiol.*, 147(4):2107–2120, August 2008.

[CDF14] D. J. Creek, W. B. Dunn, O. Fiehn, J. L. Griffin, R. D. Hall, Z. Lei, R. Mistrik, **S. Neumann**, E. L. Schymanski, L. W. Sumner, and et al. Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics*, 10(3):350–353, Jun 2014.

[CMB12] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, **S. Neumann**, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, R. L. Moritz, J. Eckels, E. Deutsch, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick. A Cross-platform Toolkit for Mass Spectrometry and Proteomics. *Nat Biotechnol*, 2012.

[DCN11] E. W. Deutsch, M. Chambers, **S. Neumann**, F. Levander, P.-A. Binz, J. Shofstahl, D. S. Campbell, L. Mendoza, D. Ovelleiro, K. Helsens, L. Martens, R. Aebersold, R. L. Moritz, and M.-Y. Brusniak. TraML: a standard format for exchange of selected reaction monitoring transition lists. *Mol Cell Proteomics*, Dec 2011.

[DEW13] W. Dunn, A. Erban, R. Weber, D. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, **S. Neumann**, J. Kopka, and M. Viant. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9:44–66, 2013. 10.1007/s11306-012-0434-4.

[DPRC18] E. W. Deutsch, Y. Perez-Riverol, R. J. Chalkley, M. Wilhelm, S. Tate, T. Sachsenberg, M. Walzer, L. Käll, B. Delanghe, S. Böcker, E. L. Schymanski, P. Wilmes, V. Dorfer, B. Kuster, P.-J. Volders, N. Jehmlich, J. P. C. Vissers, D. W. Wolan, A. Y. Wang, L. Mendoza, J. Shofstahl, A. W. Dowsey, J. Griss, R. M. Salek, **S. Neumann**, P.-A. Binz, H. Lam, J. A. Vizcaíno, N. Bandeira, and H. Röst. Expanding the use of spectral libraries in proteomics. *Journal of proteome research*, 17:4051–4060, December 2018.

[EKMB19] P. Emami Khoonsari, P. Moreno, S. Bergmann, J. Burman, M. Capuccini, M. Carone, M. Cascante, P. de Atauri, C. Foguet, A. N. Gonzalez-Beltran, T. Hankemeier, K. Haug, S. He, S. Herman, D. Johnson, N. Kale, A. Larsson, **S. Neumann**, K. Peters, L. Pireddu, P. Rocca-Serra, P. Roger, R. Rueedi, C. Ruttkies, N. Sadawi, R. M. Salek, S.-A. Sansone, D. Schober, V. Selivanov, E. A. Thévenot, M. van Vliet, G. Zanetti, C. Steinbeck, K. Kultima, and O. Spjuth. Interoperable and scalable data analysis with microservices: applications in metabolomics. *Bioinformatics (Oxford, England)*, 35:3752–3760, October 2019.

[FCK08] A. Fernie, D. Centeno, J. Kopka, J. Freitag, K. Koehl, S. Trenkamp, N. Scahuer, E. von Roepenack-Lahaye, C. Boettcher, J. Kroymann, M. Pfalz, A. Matros, **S. Neumann**, R. Höfgen, and H.-P. Mock. Teaching (and learning from) Metabolomics: The 2006 PlantMetaNet ETNA Metabolomics Research School. *Physiol Plant*, 132:136–149, 2008.

[FSN18] C. Frainay, E. Schymanski, **S. Neumann**, B. Merlet, R. Salek, F. Jourdan, and O. Yanes. Mind the gap: Mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites*, 8(3):51, 2018.

[GBNM14] A. González-Beltrán, **S. Neumann**, E. Maguire, S.-A. Sansone, and P. Rocca-Serra. The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again. *BMC Bioinformatics*, 15:S11, 2014.

[GJS14] J. Griss, A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, **S. Neumann**, J. Fan, F. Reisinger, Q.-W. Xu, N. Del Toro, Y. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaíno, and H. Hermjakob. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics*, 13(10):2765–2775, Oct 2014.

[GKN13] E. Gaquerel, C. Kuhl, and **S. Neumann**. Computational annotation of plant metabolomics profiles via a novel network-assisted approach. *Metabolomics*, pages 1–15, 2013.

[GN13] M. Gerlich and **S. Neumann**. MetFusion: integration of compound identification strategies. *Journal of Mass Spectrometry*, 48(3):291–298, 2013.

[HAK10] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, **S. Neumann**, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, Jul 2010.

[HMN16] W. Hoehenwarter, S. Mönchgesang, **S. Neumann**, P. Majovsky, S. Abel, and J. Müller. Comparative expression profiling reveals a role of the root apoplast in local phosphate response. *BMC Plant Biol*, 16:106, 2016.

[HRS19] N. Hoffmann, J. Rein, T. Sachsenberg, J. Hartler, K. Haug, G. Mayer, O. Alka, S. Dayalan, J. T. M. Pearce, P. Rocca-Serra, D. Qi, M. Eisenacher, Y. Perez-Riverol, J. A. Vizcaíno, R. M. Salek, **S. Neumann**, and A. R. Jones. mzTab-M: A data standard for sharing quantitative results in mass spectrometry metabolomics. *Analytical chemistry*, 91:3302–3310, March 2019.

[HSC13] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, **S. Neumann**, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res*, 41(Database issue):D781–D786, Jan 2013.

[HWN11] C. Hildebrandt, S. Wolf, and **S. Neumann**. Database supported candidate search for Metabolite identification. *J Integr Bioinform*, 8(2):157, 2011.

[KENS08] S. Kuhn, B. Egert, **S. Neumann**, and C. Steinbeck. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, 9:400+, September 2008.

[KFS10] C. Kettner, D. Field, S.-A. Sansone, C. Taylor, J. Aerts, N. Binns, A. Blake, C. M. Britten, A. de Marco, J. Fostel, P. Gaudet, A. González-Beltrán, N. Hardy, J. Hellemans, H. Hermjakob, N. Juty, J. Leebens-Mack, E. Maguire, **S. Neumann**, S. Orchard, H. Parkinson, W. Piel, S. Ranganathan, P. Rocca-Serra, A. Santarsiero, D. Shotton, P. Sterk, A. Untergasser, and P. L. Whetzel. Meeting Report from the Second ”Minimum Information for Biological and Biomedical Investigations” (MIBBI) workshop. *Stand Genomic Sci*, 3(3):259–266, 2010.

[KTB12] C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, and **S. Neumann**. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*, 84(1):283–289, Jan 2012.

[KZN02] K. Koch, F. Zöllner, **S. Neumann**, F. Kummert, and G. Sagerer. Comparing bound and unbound protein structures using energy calculation and rotamer statistics. *In Silico Biology*, 2:32, 2002.

[LDK15] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, **S. Neumann**, G. Trausinger, F. Sinner, T. Pieber, and C. Magnes. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics*, 16:118, 2015.

[LTNG08] E. Lange, R. Tautenhahn, **S. Neumann**, and C. Gröpl. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, 9:375+, September 2008.

[MBH15] P. Moreno, S. Beisken, B. Harsha, V. Muthukrishnan, I. Tudose, A. Dekker, S. Dornfeldt, F. Taruttis, I. Grosse, J. Hastings, **S. Neumann**, and C. Steinbeck. BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics*, 16(1):56, 2015.

[MCS10] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, **S. Neumann**, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch. mzML - a Community Standard for Mass Spectrometry Data. *Mol Cell Proteomics*, Aug 2010.

[MRTN17] R. Meier, C. Ruttkies, H. Treutler, and **S. Neumann**. Bioinformatics can boost metabolomics research. *Journal of biotechnology*, May 2017.

[MSS16] S. Mönchgesang, N. Strehmel, S. Schmidt, L. Westphal, F. Taruttis, E. Müller, S. Herklotz, **S. Neumann**, and D. Scheel. Natural variation of root exudates in Arabidopsis thaliana – linking metabolomic and genomic data. *Sci Rep*, 6:29033, 2016.

[MST16] S. Mönchgesang, N. Strehmel, D. Trutschel, L. Westphal, **S. Neumann**, and D. Scheel. Plant-to-plant variability in root metabolite profiles of 19 arabidopsis thaliana accessions is substance-class-dependent. *International journal of molecular sciences*, 17, September 2016.

[NB10] **S. Neumann** and S. Böcker. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal Bioanal Chem*, 398(7-8):2779–2788, Dec 2010.

[NTB12] **S. Neumann**, A. Thum, and C. Böttcher. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics*, pages 1–8, 2012. 10.1007/s11306-012-0401-0.

[PBB19] K. Peters, J. Bradbury, S. Bergmann, M. Capuccini, M. Cascante, P. de Atauri, T. M. D. Ebbels, C. Foguet, R. Glen, A. Gonzalez-Beltran, U. L. Günther, E. Handakas, T. Hankemeier, K. Haug, S. Herman, P. Holub, M. Izzo, D. Jacob, D. Johnson, F. Jourdan, N. Kale, I. Karaman, B. Khalili, P. Emami Khonsari, K. Kultima, S. Lampa, A. Larsson, C. Ludwig, P. Moreno, **S. Neumann**, J. A. Novella, C. O'Donovan, J. T. M. Pearce, A. Peluso, M. E. Piras, L. Pireddu, M. A. C. Reed, P. Rocca-Serra, P. Roger, A. Rosato, R. Rueedi, C. Ruttkies, N. Sadawi, R. M. Salek, S.-A. Sansone, V. Selivanov, O. Spjuth, D. Schober, E. A. Thévenot, M. Tomasoni, M. van Rijswijk, M. van Vliet, M. R. Viant, R. J. M. Weber, G. Zanetti, and C. Steinbeck. PhenoMeNal: processing and analysis of metabolomics data in the cloud. *GigaScience*, 8, February 2019.

[PGBN18a] K. Peters, K. Gorzolka, H. Bruelheide, and **S. Neumann**. Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes. *Scientific data*, 5:180179, 2018.

[PGBN18b] K. Peters, K. Gorzolka, H. Bruelheide, and **S. Neumann**. Seasonal variation of secondary metabolites in nine different bryophytes. *Ecology and evolution*, 8(17):9105–9117, 2018.

[PTD19] K. Peters, H. Treutler, S. Döll, A. S. Kindt, T. Hankemeier, and **S. Neumann**. Chemical diversity and classification of secondary metabolites in nine bryophyte species. *Metabolites*, 9(10):222, 2019.

[PUM19] P. Püllmann, C. Ulpinnis, S. Marillonnet, R. Gruetzner, **S. Neumann**, and M. J. Weissenborn. Golden mutagenesis: An efficient multi-site-saturation mutagenesis approach by golden gate cloning with automated primer design. *Scientific reports*, 9:10932, July 2019.

[PWW18] K. Peters, A. Worrich, A. Weinhold, O. Alka, G. Balcke, C. Birkemeyer, H. Bruelheide, O. W. Calf, S. Dietz, K. Dührkop, E. Gaquerel, U. Heinig, M. Kücklich, M. Macel, C. Müller, Y. Poeschl, G. Pohnert, C. Ristok, V. M. Rodríguez, C. Ruttkies, M. Schuman, R. Schweiger, N. Shahaf, C. Steinbeck, M. Tortosa, H. Treutler, N. Ueberschaar, P. Velasco, B. M. Weiß, A. Widdig, **S. Neumann**, and N. M. v. Dam. Current challenges in plant eco-metabolomics. *International journal of molecular sciences*, 19, May 2018.

[RGN13] C. Ruttkies, M. Gerlich, and **S. Neumann**. Tackling CASMI 2012: Solutions from MetFrag and MetFusion. *Metabolites*, 3(3):623–636, 2013.

[RNP19] C. Ruttkies, **S. Neumann**, and S. Posch. Improving metfrag with statistical learning of fragment annotations. *BMC Bioinformatics*, 20(1):1–14, July 2019.

[RSBM10] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, **S. Neumann**, P. Sterk, W. Tong, and S.-A. Sansone. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356, Sep 2010.

[RSS19] C. Ruttkies, E. L. Schymanski, N. Strehmel, J. Hollender, **S. Neumann**, A. J. Williams, and M. Krauss. Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag. *Analytical and Bioanalytical Chemistry*, 411(19):4683–4700, July 2019.

[RSSA16] P. Rocca-Serra, R. M. Salek, M. Arita, E. Correa, S. Dayalan, A. Gonzalez-Beltran, T. Ebbels, R. Goodacre, J. Hastings, K. Haug, A. Koulman, M. Nikolski, M. Oresic, S.-A. Sansone, D. Schober, J. Smith, C. Steinbeck, M. R. Viant, and **S. Neumann**. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics*, 12(1):14, 2016.

[RSSN15a] C. Ruttkies, N. Strehmel, D. Scheel, and **S. Neumann**. Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an in silico generated compound database and MetFrag. *Rapid Commun Mass Spectrom*, 29(16):1521–1529, Aug 2015.

[RSSN15b] C. Ruttkies, N. Strehmel, D. Scheel, and **S. Neumann**. Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an in silico generated compound database and metfrag. *Rapid communications in mass spectrometry : RCM*, 29:1521–1529, August 2015.

[RSW16] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and **S. Neumann**. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*, 8:3, 2016.

[SBH19] J. Stanstrup, C. D. Broeckling, R. Helmus, N. Hoffmann, E. Mathé, T. Naake, L. Nicolotti, K. Peters, J. Rainer, R. M. Salek, T. Schulze, E. L. Schymanski, M. A. Stravs, E. A. Thévenot, H. Treutler, R. J. M. Weber, E. Willighagen, M. Witting, and **S. Neumann**. The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites*, 9, September 2019.

[SGDN13] J. Stanstrup, M. Gerlich, L. Dragsted, and **S. Neumann**. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. *Analytical and Bioanalytical Chemistry*, pages 1–12, 2013.

[SGK12] E. L. Schymanski, C. M. J. Gallampois, M. Krauss, M. Meringer, **S. Neumann**, T. Schulze, S. Wolf, and W. Brack. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Anal Chem*, 84(7):3287–3295, Apr 2012.

[SGRN14] E. L. Schymanski, M. Gerlich, C. Ruttkies, and **S. Neumann**. Solving CASMI 2013 with MetFrag, MetFusion and MOLGEN-MS/MS. *Mass Spectrometry*, 3(Special Issue 2):S0036–S0036, 2014.

[SJW18] D. Schober, D. Jacob, M. Wilson, J. A. Cruz, A. Marcu, J. R. Grant, A. Moing, C. Deborde, L. F. de Figueiredo, K. Haug, P. Rocca-Serra, J. Easton, T. M. D. Ebbels, J. Hao, C. Ludwig, U. L. Günther, A. Rosato, M. S. Klein, I. A. Lewis, C. Luchinat, A. R. Jones, A. Grauslys, M. Larralde, M. Yokochi, N. Kobayashi, A. Porzel, J. L. Griffin, M. R. Viant, D. S. Wishart, C. Steinbeck, R. M. Salek, and **S. Neumann**. nmrML: a community supported open data standard for the description, storage, and exchange of NMR data. *Analytical chemistry*, 90:649–656, January 2018.

[SN13a] E. L. Schymanski and **S. Neumann**. CASMI: And The Winner is .... *Metabolites*, 3(2):412–439, 2013.

[SN13b] E. L. Schymanski and **S. Neumann**. The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions. *Metabolites*, 3(3):517–538, 2013.

[SNMHB18] J.-A. Schüler, **S. Neumann**, M. Müller-Hannemann, and W. Brandt. Chemfrag: Chemically meaningful annotation of fragment ion mass spectra. *Journal of Mass Spectrometry*, 53(11):1104–1115, 2018.

[SNS15] R. M. Salek, **S. Neumann**, D. Schober, J. Hummel, K. Billiau, J. Kopka, E. Correa, T. Reijmers, A. Rosato, L. Tenori, P. Turano, S. Marin, C. Deborde, D. Jacob, D. Rolin, B. Dartigues, P. Conesa, K. Haug, P. Rocca-Serra, S. O'Hagan, J. Hao, M. van Vliet, M. Sysi-Aho, C. Ludwig, J. Bouwman, M. Cascante, T. Ebbels, J. L. Griffin, A. Moing, M. Nikolski, M. Oresic, S.-A. Sansone, M. R. Viant, R. Goodacre, U. L. Günther, T. Hankemeier, C. Luchinat, D. Walther, and C. Steinbeck. COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics*, 11(6):1587–1597, 2015.

[SNV15] J. Stanstrup, **S. Neumann**, and U. Vrhovšek. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal Chem*, 87(18):9421–9428, Sep 2015.

[SRK17] E. L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, and **S. Neumann**. Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9(1):22, 2017.

[SRSF12] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, **S. Neumann**, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide. Toward interoperable bioscience data. *Nat Genet*, 44(2):121–126, Feb 2012.

[TBN08] R. Tautenhahn, C. Böttcher, and **S. Neumann**. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9(1):504, 2008.

[TN16] H. Treutler and **S. Neumann**. Prediction, detection, and validation of isotope clusters in mass spectrometry data. *Metabolites*, 6, October 2016.

[TSGN14] D. Trutschel, S. Schmidt, I. Grosse, and **S. Neumann**. Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data. *Metabolomics*, Nov 2014.

[TSGN15] D. Trutschel, S. Schmidt, I. Grosse, and **S. Neumann**. Joint analysis of dependent features within compound spectra can improve detection of differential features. *Frontiers in bioengineering and biotechnology*, 3:129, 2015.

[TTP16] H. Treutler, H. Tsugawa, A. Porzel, K. Gorzolka, A. Tissier, **S. Neumann**, and G. U. Balcke. Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. *Anal Chem*, Aug 2016.

[vRBC17] M. van Rijswijk, C. Beirnaert, C. Caron, M. Cascante, V. Dominguez, W. B. Dunn, T. M. D. Ebbels, F. Giacomoni, A. Gonzalez-Beltran, T. Hankemeier, K. Haug, J. L. Izquierdo-Garcia, R. C. Jimenez, F. Jourdan, N. Kale, M. I. Klapa, O. Kohlbacher, K. Koort, K. Kultima, G. Le Corguillé, P. Moreno, N. K. Moschonas, **S. Neumann**, C. O'Donovan, M. Reczko, P. Rocca-Serra, A. Rosato, R. M. Salek, S.-A. Sansone, V. Satagopam, D. Schober, R. Shimmo, R. A. Spicer, O. Spjuth, E. A. Thévenot, M. R. Viant, R. J. M. Weber, E. L. Willighagen, G. Zanetti, and C. Steinbeck. The future of metabolomics in ELIXIR. *F1000Research*, 6, 2017.

[VSN16] M. Vinaixa, E. L. Schymanski, **S. Neumann**, M. Navarro, R. M. Salek, and O. Yanes. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry*, 78:23–35, Apr 2016.

[WMM16] G. Wohlgemuth, S. S. Mehta, R. F. Mejia, **S. Neumann**, D. Pedrosa, T. Pluskal, E. L. Schymanski, E. L. Willighagen, M. Wilson, D. S. Wishart, M. Arita, P. C. Dorrestein, N. Bandeira, M. Wang, T. Schulze, R. M. Salek, C. Steinbeck, V. C. Nainala, R. Mistrik, T. Nishioka, and O. Fiehn. Splash, a hashed identifier for mass spectra. *Nature biotechnology*, 34:1099–1101, November 2016.

[WRNSK17] M. Witting, C. Ruttkies, **S. Neumann**, and P. Schmitt-Kopplin. Lipidfrag: Improving reliability of in silico fragmentation of lipids and application to the caenorhabditis elegans lipidome. *PloS one*, 12:e0172311, 2017.

[WSMHN10] S. Wolf, S. Schmidt, M. Müller-Hannemann, and **S. Neumann**. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148, 2010.

[ZNKS05] F. Zöllner, **S. Neumann**, F. Kummert, and G. Sagerer. Database driven Test Case Generation for Protein-Protein Docking. *Bioinformatics*, 21(5):683–684, 2005.

Preprints

[MPR19] P. Moreno, L. Pireddu, P. Roger, N. Goonasekera, E. Afgan, M. van den Beek, S. He, A. Larsson, D. Schober, C. Ruttkies, D. Johnson, P. Rocca-Serra, R. J. Weber, B. Gruening, R. M. Salek, N. Kale, Y. Perez-Riverol, I. Papatheodorou, O. Spjuth, and **S. Neumann**. Galaxy-kubernetes integration: scaling bioinformatics workflows in the cloud. *bioRxiv*, 2019.

[TMW14] A. Thum, S. Mönchgesang, L. Westphal, T. Lübken, S. Rosahl, **S. Neumann**, and S. Posch. Supervised Penalized Canonical Correlation Analysis. *arXiv*, page arxiv.org/abs/1405.1534, May 2014.

Conference Proceedings

[ENH07] B. Egert, **S. Neumann**, and A. Hinneburg. Fast Approximate Duplicate Detection for 2D-NMR Spectra. In *Proceedings of DILS 2007*, volume 4, 2007.

[GFK06] I. Grosse, T. Funke, C. Kuenne, **S. Neumann**, A. Stephanik, T. Thiel, and S. Weise. Integrative Datenanalyse mit dem Plant Data Warehouse. In *Ausgewählte Vorträge aus GPZ-Arbeitsgemeinschaften*, volume 70 of *Vorträge für Pflanzenzüchtung*, pages 50–53, Göttingen, März 2006. Gesellschaft für Pflanzenzüchtung e. V. (GPZ).

[GN07] A. Gaida and **S. Neumann**. MetHouse: Raw and Preprocessed Mass Spectrometry Data. *Journal of Integrative Bioinformatics*, 4(1):457 – 464, March 2007.

[KN06] S. Klie and **S. Neumann**. Storage and Processing of Mass Spectrometry Data. In *Proc. of 17th Int. Conference on Databases and Expert Systems (DEXA 2006)*, pages 211–215. DEXA, IEEE, September 2006.

[Neu07] **S. Neumann**. Beyond Flat Files: Data Modeling, Editing, Archival and Interchange. In K. Kettner and M. G. Hicks, editors, *Proc. 2nd Workshop on Experimental Standard Conditions of Enzyme Characterizations*. Beilstein Institute, Logos-Verlag, 2007.

[PBC06] Y. Pöschl, C. Böttcher, S. Clemens, D. Scheel, S. Posch, and **S. Neumann**. Analysis of Metabolite relations in LCMS data using Bayesian Networks. In *German Conference on Bioinformatics, Short Papers*, pages 17–18, Tübingen, Germany, Sep 2006. ZBIT, Zentrum für BioInformatik Tübingen.

[SMM13] D. Schober, G. Mayer, A. Moing, M. Eisenacher, and **S. Neumann**. Ontological analysis of controlled vocabularies used in PSI/MSI supported XML standards. In M. Horbach, editor, *GI-Jahrestagung*, volume 220 of *LNI*, pages 1875–1888. GI, 2013.

[SWJ14] D. Schober, M. Wilson, D. Jacob, A. Moing, G. Mayer, M. Eisenacher, R. M. Salek, and **S. Neumann**. Ontology usage in Omics Standards Initiatives: Pros and Cons of enriching XML data formats with controlled vocabulary terms. *ONTOLOGIES AND DATA IN LIFE SCIENCES (ODLS 2014)*, 2014.

[TBN07] R. Tautenhahn, C. Böttcher, and **S. Neumann**. Annotation of LC/ESI-MS Mass Signals. *Lecture Notes in Computer Science*, 4414:371–380, 2007.

Books and Book Chapters

[LKG19] T. Lübken, M. Kopischke, K. Geissler, L. Westphal, D. Scheel, **S. Neumann**, and S. Rosahl. *Metabolomics: Practical Guide to Design and Analysis*, chapter Metabolic and transcriptional response of Arabidopsis thaliana wildtype and mutants to Phytophthora infestans. Chapman and Hall/CRC, 2019.

[Neu03] **S. Neumann**. *Soft volume models for protein-protein docking.* Dissertation, Universität Bielefeld, Technische Fakultät, 2003.

[NRWB13] **S. Neumann**, F. Rasche, S. Wolf, and S. Böcker. *The Handbook of Plant Metabolomics (Metabolite Profiling and Networking).*, chapter Metabolite identification and computational mass spectrometry. Wiley-Blackwell-VCH, 2013.

[NYMF19] **S. Neumann**, O. Yanes, R. Mumm, and P. Franceschi. *Metabolomics: Practical Guide to Design and Analysis*, chapter Mass Spectrometry Data Processing. 2019.

Part II.

Reprints of Original Publications

Metabolite Profiling

7

This part contains the explanation of my contributions to the topics introduced in part I and reprints of the original research articles for selected papers in peer-reviewed journals. The citations of articles included here are given in boldface. Permission to reprint was either obtained on a per-paper basis from the publishers, or is granted through an Open Access license. The footer includes the DOI and a link to the full publication.

My contributions in the area of metabolite profiling started with the supervision of the PhD student Ralf Tautenhahn and work on feature detection in `xcms` **[TBN08]** and feature alignment **[LTNG08]**. Later, I supervised the PhD student Carsten Kuhl, and we developed the graph-based feature annotation **[KTB12]**. Most recently, I supervised the PhD student Hendrik Treutler, and we incorporated additional prior information into the feature detection step in `xcms` **[TN16]**. For many years, I am now the maintainer of `xcms` and the `CAMERA` packages in Bioconductor.

My contributions in developments on feature clustering [BAN14] and automated parameter optimisation [LDK15] were limited to scientific discussions and creating packages for the R and Bioconductor environment from existing code.

The general principles in mass spectrometry data analysis were described in a textbook [NYMF19] where I coordinated the chapter on “Mass Spectrometry Data Processing”.

BMC Bioinformatics

BioMed Central

Research article

Open Access

Highly sensitive feature detection for high resolution LC/MS

Ralf Tautenhahn*, Christoph Böttcher and Steffen Neumann

Address: Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany

Email: Ralf Tautenhahn* - rtautenh@ipb-halle.de; Christoph Böttcher - cboettch@ipb-halle.de; Steffen Neumann - sneumann@ipb-halle.de
* Corresponding author

Published: 28 November 2008

BMC Bioinformatics 2008, **9**:504 doi:10.1186/1471-2105-9-504

Received: 9 July 2008
Accepted: 28 November 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/504

© 2008 Tautenhahn et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Liquid chromatography coupled to mass spectrometry (LC/MS) is an important analytical technology for e.g. metabolomics experiments. Determining the boundaries, centres and intensities of the two-dimensional signals in the LC/MS raw data is called feature detection. For the subsequent analysis of complex samples such as plant extracts, which may contain hundreds of compounds, corresponding to thousands of features – a reliable feature detection is mandatory.

Results: We developed a new feature detection algorithm *centWave* for high-resolution LC/MS data sets, which collects regions of interest (partial mass traces) in the raw-data, and applies continuous wavelet transformation and optionally Gauss-fitting in the chromatographic domain. We evaluated our feature detection algorithm on dilution series and mixtures of seed and leaf extracts, and estimated recall, precision and F-score of seed and leaf specific features in two experiments of different complexity.

Conclusion: The new feature detection algorithm meets the requirements of current metabolomics experiments. *centWave* can detect close-by and partially overlapping features and has the highest overall recall and precision values compared to the other algorithms, *matchedFilter* (the original algorithm of *XCMS*) and the centroidPicker from *MZmine*. The *centWave* algorithm was integrated into the Bioconductor R-package *XCMS* and is available from http://www.bioconductor.org/

Background

Metabolomics aims at the unbiased and comprehensive quantification of metabolite concentrations in organisms, tissues, or cells [1,2]. The combination of chromatographic separation with subsequent mass spectrometric detection has emerged as a key technology for multiparallel analysis of low molecular weight compounds in biological systems. Gas chromatography-mass spectrometry (GC/MS) based techniques are mature and well-established, but restricted to volatile compounds, often requiring chemical derivatisation. High-performance liquid chromatography-mass spectrometry (HPLC/MS) facilitates the analysis of compounds of higher polarity and lower volatility in a much wider mass range without derivatisation [3-5]. With LC/MS the injected sample is separated on the chromatographic column, resulting in the consecutive elution of different compounds. The mass spectrometer acquires mass spectra from the column output at a specified scan rate, so each compound can be measured in several consecutive scans. Due to the fact that each eluting compound gives rise to a number of mass signals (adducts, fragments and isotopic peaks), a metabolite induces several two-dimensional features.

DOI:10.1186/1471-2105-9-504

In the following, we use the term "feature" for a bounded, two-dimensional (*m/z* and retention time) LC/MS signal. The term "peak" is used for one-dimensional signals: both *m/z* peaks (centroids) in the mass spectrum and chromatographic peaks.

For complex metabolomics samples, the LC/MS data contains hundreds to thousands of metabolites. For the statistical analysis of biological experiments the feature intensity is of interest and has to be calculated from the raw data. Spectra can be acquired in profile mode or centroid mode. Vendor supplied centroidisation algorithms usually employs machine-specific models, which are superior to generic approaches. In addition, the centroid mode results in considerable size reduction of the LC/MS data set.

The processing pipeline for LC/MS based metabolomics can be divided into the following steps:

1. Signal preprocessing and centroidization in *m/z*,

2. Two-dimensional feature detection and integration

3. Alignment of corresponding features in multiple samples

4. Statistical analysis, chemical and biological interpretation.

Feature detection is a crucial step in the LC/MS data processing pipeline – it should be reliable, i.e. report as many as possible "real" features, while keeping the false positive rate low. The challenge for the algorithms is to detect features of low intensity induced by compounds with low abundance on the one hand, and to avoid feature-like signals caused by e.g. chemical noise on the other hand.

Several frameworks for feature detection (and alignment) of metabolomics LC/MS data have been developed in the last years, both commercial products such as MarkerLynx (Waters), the closed-source (but freely-available) MetAlign [6], or XCMS [7] and *MZmine* [8] which have open-source licenses. Other packages, some of them specific for LC/MS-based proteomics, have been reviewed in [9].

A widely used approach for the processing of LC/MS data is to transform the raw data into a matrix representation with the dimensions *m/z*, retention time and intensity. To convert high resolution mass spectra into this representation, it is necessary to divide the *m/z* axis into equidistant chunks depending on the resolution of the mass spectrometer, e. g. 0.1 *m/z* wide. This procedure is usually referred to as binning. Some drawbacks of this method were already mentioned in [7,10,11]. In particular, specifying the optimal bin size for the particular data set can be difficult. If the bin size is chosen too small, chromatographic peaks are alternating between bins and cannot be detected due to the loss of the chromatographic shape. If the bin size is too large, peaks can overlay each other and small features are rather buried by the increased chromatographic noise level. On the positive side it should be mentioned that the binning approach is all-purpose and allows for a fast data processing.

A density based LC/MS feature detection approach – an alternative to the common binning technique – was introduced by Stolt et al. [10]. The authors consider the emerging analyte as a region of data points with high density anked by a specific "data void". Based on these properties, they calculate a potential field which is then used to create a matrix of mass traces (runtime ~2 h/sample). Recently, the extraction of "pure ion chromatograms" using Kalman tracking was demonstrated in [11]. The applicability of Wavelet based techniques for peak picking in MALDI- and SELDI-TOF mass spectra was shown by e.g. [12-15]. Here we will discuss a new method for the reliable detection and integration of two-dimensional LC/MS signals, referred to as features. By using a combination of a density based technique to detect regions of interest in the *m/z* domain, and a Wavelet based approach to resolve chromatographic peaks, we achieve a high sensitivity even in very complex mixtures compared to two other algorithms, *matchedFilter* (the original algorithm of *XCMS*) and the *centroidPicker* from *MZmine*.

So far, there is no common method for evaluating the performance of feature detection algorithms. Even for the same feature detection algorithm, different parametrisation can lead to (vastly) different results, if e.g. many false positive noise signals are detected as features. Therefore the absolute number of detected features per sample is not suitable to characterise a feature detection algorithm. More elaborate approaches consider mixtures of known compounds spiked into complex samples [16]. To the best of our knowledge, no evaluation has been performed to assess recall and precision of feature detection algorithms for multiple complex samples.

The remainder of this paper is structured as follows: In section 2 we give a detailed description of the *centWave* algorithm, followed by the description of the experimental comparison between several feature detection algorithms. In section 3 we present the evaluation results and discuss the benefits of *centWave*, followed by a conclusion and outlook of expected future developments in section 4.

DOI:10.1186/1471-2105-9-504

Methods

This section describes the *centWave* method which combines density based detection of regions of interest in the *m/z* domain, and a Continuous Wavelet Transform (CWT) based approach for chromatographic peak resolution. The experimental setup is depicted as well as the layout of the evaluation procedure.

2.1 The centWave algorithm

2.1.1 Detecting regions of interest (ROI) in the m/z domain

To circumvent the mentioned problems of the binning technique, an alternative, fast computing approach was used which directly detects regions of interesting mass traces. Figure 1 shows the extracted ion chromatogram and the corresponding *m/z* centroids in the consecutive mass spectra for a typical LC/MS feature, recorded in centroid mode. With the chromatographic peak emerging, the consecutive centroids form a compact mass trace bounded in *m/z* and retention time. The *m/z* deviation is determined by the mass accuracy of the mass spectrometer and typically increases with lower signal intensities.

Due to the fact that the mass accuracy (μ, given in ppm) of the mass spectrometer and the minimum chromatographic peak width is known or can easily be assessed, it is possible to directly scan for regions where at least p_{min} centroids with a deviation less than μ ppm occur. This task is achieved by the following algorithm for samples in centroid mode, with scans numbered $s = 1,...,S$:

1. Initialisation:

(a) Initialise a list ROI using all *m/z* values mz_i^s from the first scan:

$$\forall\ i = 1,..., N,\ N = |mz^{s=1}|: \mathrm{ROI}(i).\mathrm{values}(1) = mz_i^{s=1}$$

(b) Initialise the *m/z* mean value for each actually processed region :

$$\mathrm{ROI}(i).\mathrm{mzmean} = mz_i^{s=1},\ i = 1,..., N,\ N = |mz^{s=1}|$$

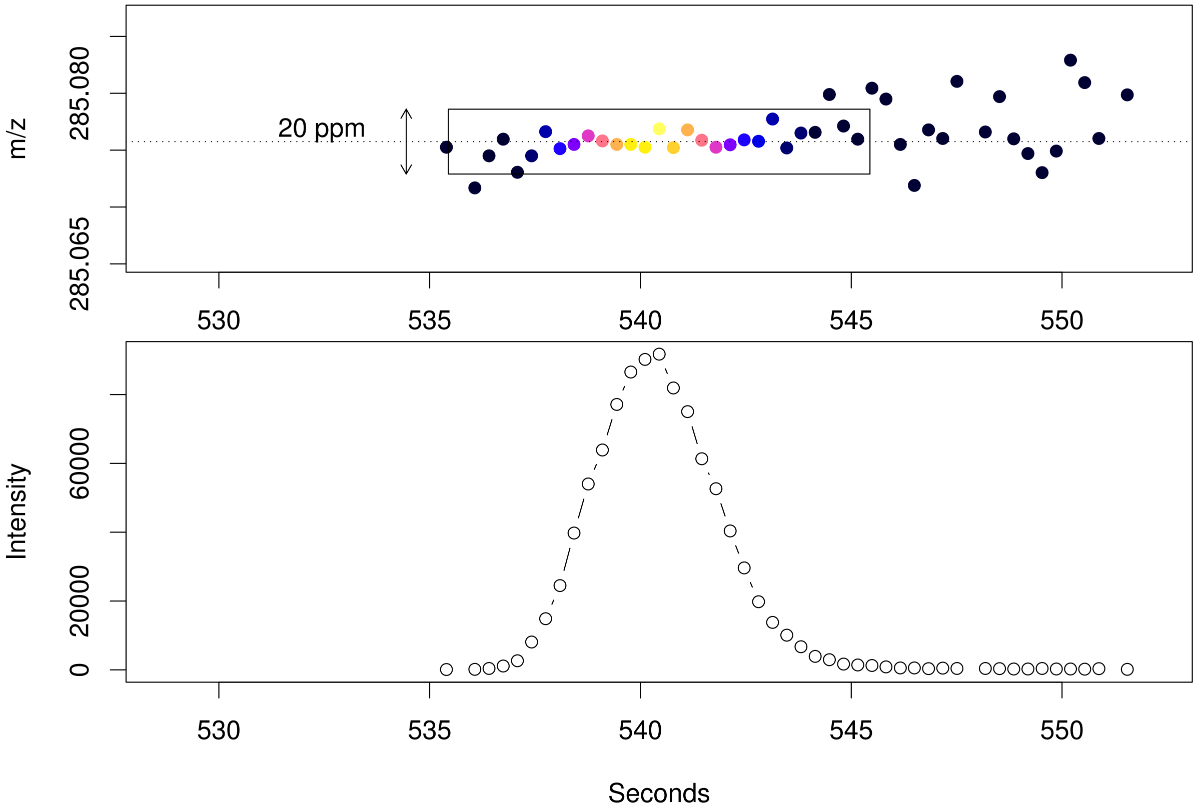


Figure 1

Mass trace and chromatographic peak of Biochanin A [*M* + *H*]⁺ mass signal. The upper panel shows the mass trace of the biochanin A [*M* + *H*]⁺ mass signal across 10 seconds with colour-coded intensities. The corresponding chromatographic peak is shown below.

DOI:10.1186/1471-2105-9-504

2. For each scan $s = 2,..., S$:

(a) For each *m/z* value mz_i^s, $i = 1,..., N$, $N = |\mathrm{mz}^s|$ in the current scan *s*:

Exists j, $j = 1,..., J$, $J = |\mathrm{ROI}|$ such that $|\mathrm{ROI}(j).\mathrm{mzmean} - \mathrm{mz}_i^s| <= \mu$?

- **Yes**: Append mz_i^s to ROI(*j*) and update the *m/z* mean value

$K = |\mathrm{ROI}(j).\mathrm{values}| + 1$, $\mathrm{ROI}(j).\mathrm{values}(K) = \mathrm{mz}_i^s$

$\mathrm{ROI}(j).\mathrm{mzmean} = \frac{1}{K}\sum_{k=1}^{K} ROI(j)\ .\mathrm{values}(k)$

- **No**: Initialise a new ROI and append it to the list

$J = |\mathrm{ROI}| + 1$, $\mathrm{ROI}(J).\mathrm{values}(1) = \mathrm{mz}_i^s$, $\mathrm{ROI}(J).\mathrm{mzmean} = \mathrm{mz}_i^s$

(b) Check & Cleanup:

- Remove all ROI which were not extended in step 2a *and* contain less than $\mathrm{p}_{\mathrm{min}}$ centroids

- Mark ROI that were not extended, but contain at least $\mathrm{p}_{\mathrm{min}}$ centroids as completed

Optionally an intensity filter (*prefilter* = (*k*, *I*), e. g. *prefilter* = (2, 100)) can be set to early discard regions of small intensity. Then only those ROI are retained (in step 2b) that contain at least *k* consecutive values with intensity ≥ *I*. This prefilter vastly speeds up the overall processing time.

Each *m/z* value needs to be considered only once, so the ROI algorithm is fast (approximately 10–20 seconds on a 2.5 GHz CPU for a measurement with 3000 scans). Figure 2 shows the result of the ROI detection algorithm for a small region of a complex LC/MS sample.

In some rare cases "gaps" are observed in the mass trace of features with low intensity. Due to the fact that each ROI is laterally extended for the following chromatographic peak detection, only a small contiguous region needs to be found for the successful detection of such features. To a certain extent, the algorithm is therefore able to detect features with such gaps. Otherwise, in case of samples which might show this phenomenon more often, the algorithm can easily be modified to be even more "gap-tolerant". In contrast to binning, this approach has the advantage that no fixed bin size has to be chosen. Each ROI is detected separately and the drawbacks of binning can be circumvented. Unlike binning the result is not a matrix but a list of mass traces with different lengths. Depending on the chromatography and the mass accuracy of the mass spectrometer, each ROI may contain none, exactly one or more than one distinct chromatographic peaks. Therefore it is necessary to subject each ROI to an extensive analysis in the chromatographic domain.

2.1.2 Detecting chromatographic peaks

Depending on the separation technique (e. g. HPLC/UPLC/CE) features can show considerable variations in their chromatographic width and shape. The matched filter approach makes use of a filter based on a model peak with defined shape and fixed width. This technique gives good results in most cases and was shown to work in principle also for peaks of differing width and shape (see [17,18]) but nevertheless some problems occur if the model peak width is not chosen appropriately. Figure 3 shows a mass trace from a HPLC/MS sample, containing three peaks of different width. The application of three independent matched filters with different width of the model peak (second derivative Gaussian) reveals the problem of assessing the perfect model peak width. Narrow peaks are found perfectly with a small model peak width (e. g. σ = 5–10 s) while broad peaks can only be properly detected with an increased model peak width (e. g. σ = 20 s).

Another aspect of this optimisation problem are chromatographic close-by peaks. Figure 4 shows the response of three independent matched filters with different σ on a chromatogram with many narrow, close-by peaks. It can be seen that only a matched filter with a very small model peak width (e. g. σ = 5 s) gives reasonable results in this case. Figure 3 and 4 are examples from the same LC/MS measurement. In this case, none of the three chosen model peak widths yields satisfying results for all occuring peaks. The enhancement of the matched filter approach is the peak detection on multiple scales using Continuous Wavelet Transform (CWT), which reliably detects chromatographic peaks of differing width. The CWT is widely used in signal processing and pattern recognition. The mathematical representation [19] is as follows:

$$(T^{\mathrm{wav}} f)(s,\tau) = \int_{-\infty}^{\infty} f(t)\psi_{s,\tau}(t)dt$$

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right), \quad s \in \mathbb{R}^+ - \{0\}, \quad \tau \in \mathbb{R}$$

where $f(t)$ is the signal, ψ the mother wavelet, s the scale and τ the translation. The result of the CWT is a two-

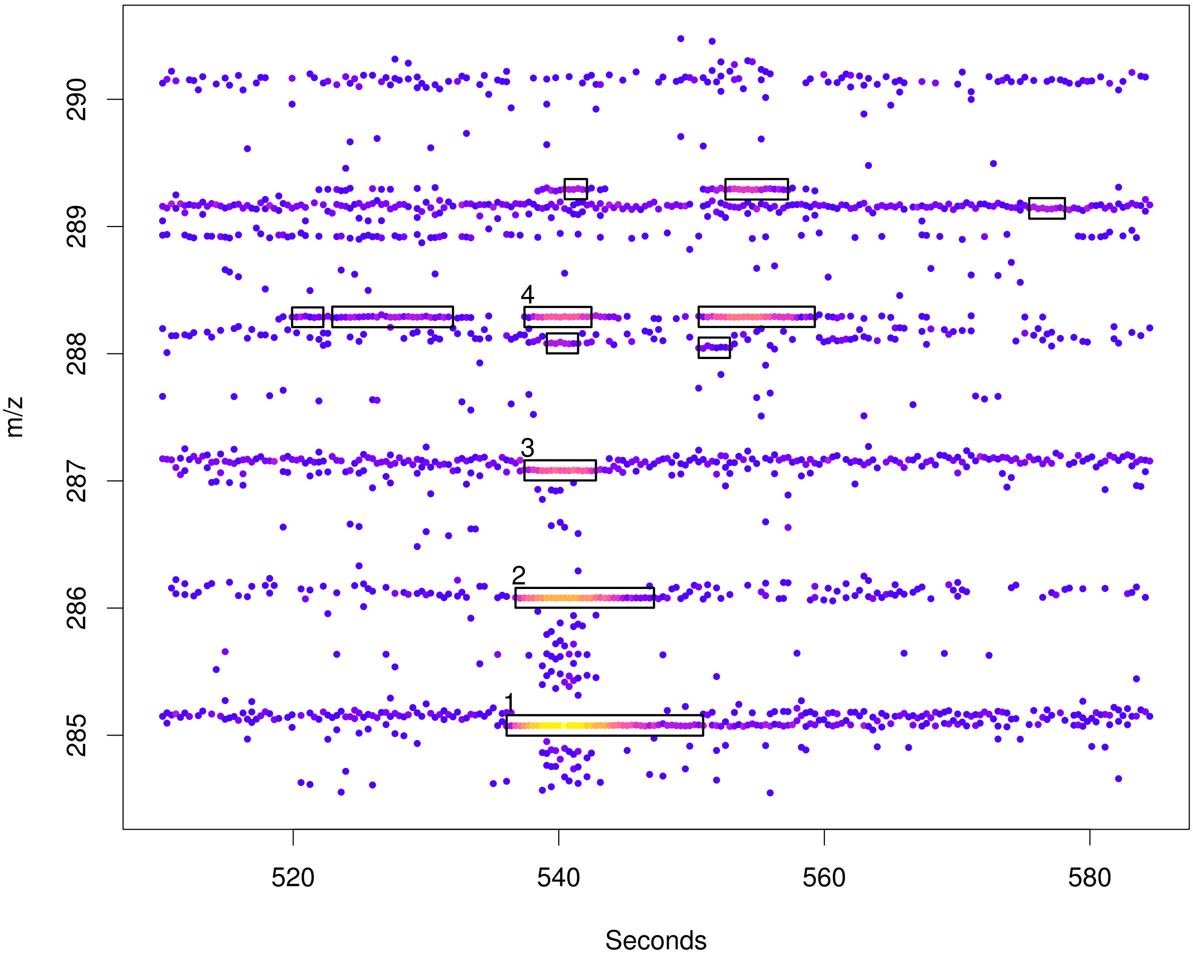


Figure 2

Region Of Interest (ROI) detection. Raw data in the chromatographic and *m*/*z* region around the $[M + H]^+$ mass signal (1) of biochanin A. In addition to the three isotopic peaks (2–4) other mass signals are marked as ROIs.

dimensional matrix of wavelet coefficients T^{wav}. Since the "Mexican Hat" wavelet (normalised second derivative of Gaussian $e^{-x^2/2}$, Figure 5) is used as the mother wavelet, the result of the CWT is comparable to the combined application of the matched filter technique with the second derivative Gaussian of different widths as model peak. The algorithms for CWT and CWT-coefficient analysis described and implemented in [13] for the peak detection in SELDI/TOF spectra were adapted for peak detection in the chromatographic domain.

2.1.3 The centWave workflow

The three relevant input parameters for the *centWave* algorithm are

1. Mass deviation μ in ppm, typically set to a generous multiple of the mass accuracy of the mass spectrometer. We use μ = 30 ppm for the Bruker MicrOTOF-Q, which is advertised with a mass accuracy of 3–5 ppm.

2. Chromatographic peak width range w_{min}, w_{max} in seconds, e. g. w_{min}, w_{max} = (5, 10) for UPLC separation as described in the experimental setup.

3. Signal to noise ratio threshold SNR_{Thr}, e.g. SNR_{Thr} = 10

DOI:10.1186/1471-2105-9-504

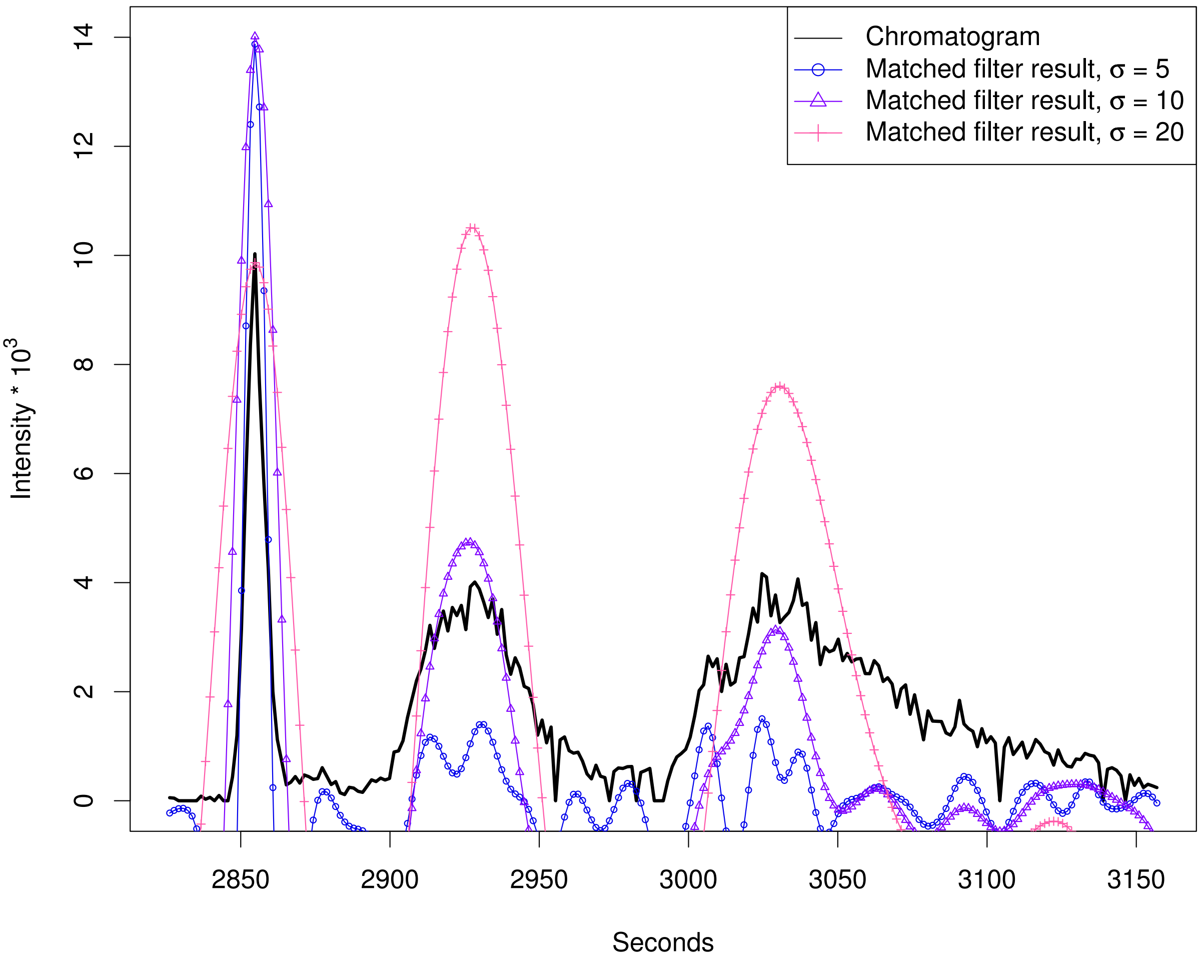


Figure 3
Matched filter effects, example region 1. HPLC/ESI-QTOF-MS of a *A. thaliana* leaf extract. Extracted ion chromatogram (277.213 – 277.221 *m*/z) and matched filter results using second derivative Gaussian with different filter widths. Negative filter values were omitted.

The following is the description of the most important steps of the *centWave* workflow:

- The scale range s_{min}, s_{max} for the CWT and the p_{min} parameter for the ROI detection are calculated from the input parameters w_{min}, w_{max} and the average inter-scan distance.

- ROI detection (see section 2.1.1) is performed using the parameters μ and p_{min}

- Chromatographic analysis of each detected ROI:

- To accommodate noise and baseline estimation, each ROI is laterally extended by a multiple of the expected chromatographic peak width

- Local noise and baseline estimation: Let x be the vector of intensity values of the actual (extended) ROI, and x_t the 10% trimmed x (5% of the smallest and 5% of the largest intensity values are discarded). Then the baseline BL is assessed as the mean value of x_t and the noise level NL as the standard devation of x_t.

DOI:10.1186/1471-2105-9-504

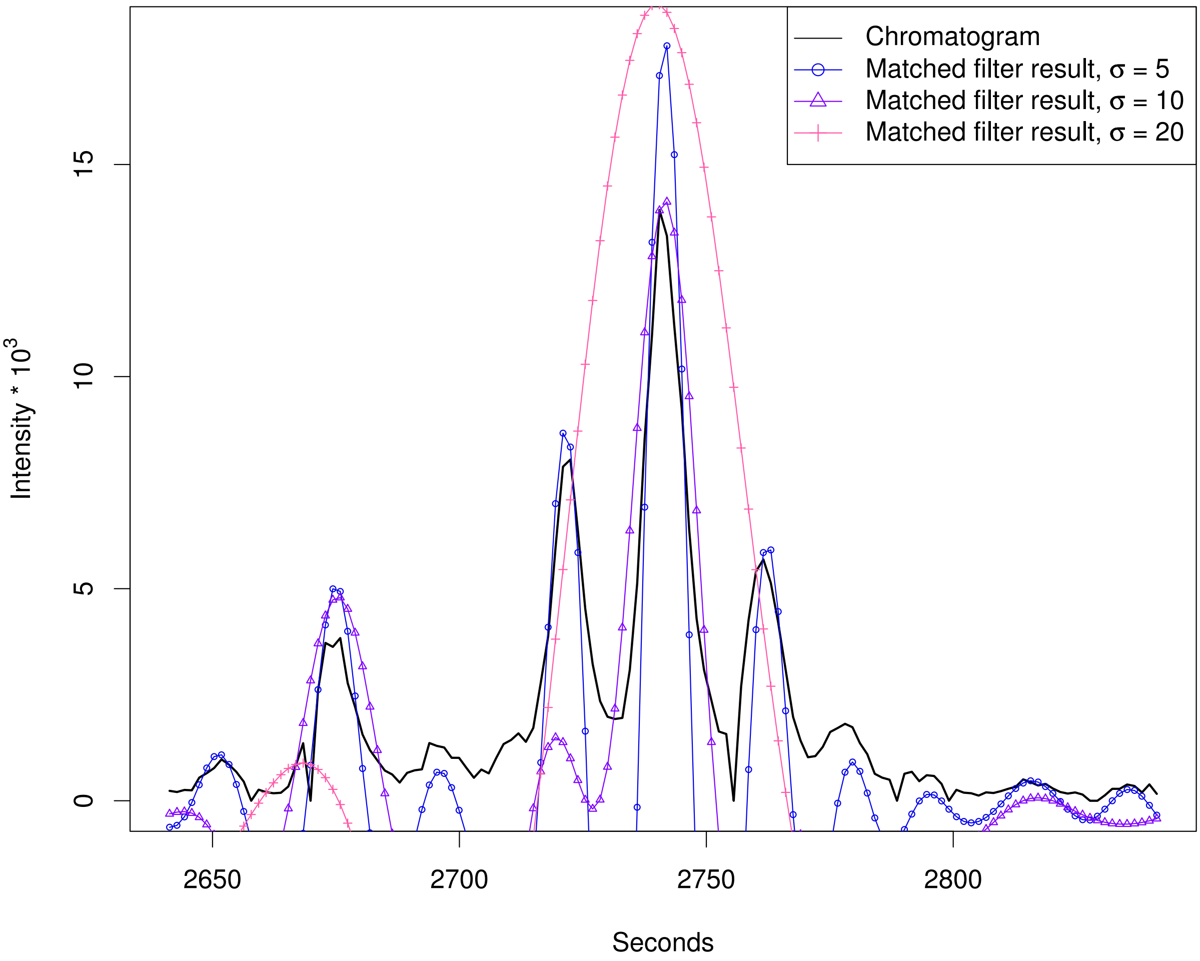


Figure 4

Matched filter effects, example region 2. HPLC/ESI-QTOF-MS of a *A. thaliana* leaf extract. Extracted ion chromatogram (967.53–967.56 *m*/*z*, same sample that was used for Figure 3) and matched filter results using second derivative Gaussian with different filter widths. Negative filter values were clipped.

- The Continuous Wavelet Transform (see 2.1.2) is applied to the intensity values of the ROI (the extracted ion chromatogram), using the scale range $s_{min}, \ldots, s_{max}$.

- Local maxima of the CWT coefficients at each scale are detected.

- "Ridges" can be identified by linking the detected local maxima (described in [13]). The ridges describe the scale range where the chromatographic peak was located. If more than one chromatographic peak was detected, the following steps are applied for each peak separately.

- Locate the chromatographic peak boundaries rt_{min} and rt_{max} by descent on the filtered peak data, i.e. the CWT coefficients of the scale where the peak was optimally located.

- Calculate the feature intensity I using the intensity values within rt_{min} and rt_{max}. I_{max} is defined as the maximal intensity value within this range.

- Compute the *m*/*z* centroid of the feature as the weighted mean of the *m*/*z* values within rt_{min} and rt_{max}.

DOI:10.1186/1471-2105-9-504

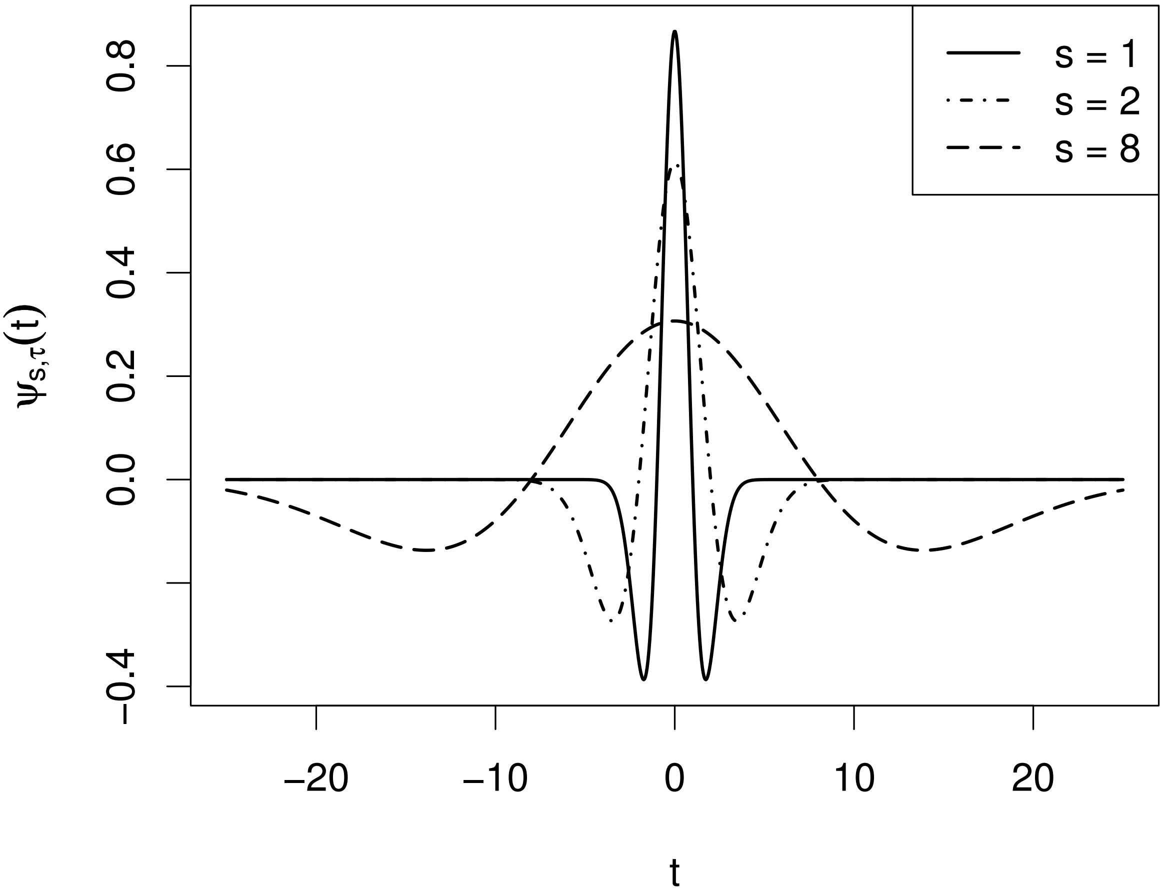


Figure 5
Mexican Hat Wavelet. Mexican hat wavelet at different scales.

- Calculate the signal to noise ratio SNR = (I_{max} – BL)/NL of the feature. Discard the feature if SNR < SNR_{Thr}.

- The deviation μ^* of *m*/*z* values within rt_{min} and rt_{max} is calculated in ppm. The value μ^* can be interpreted as the mass deviation value which would have been sufficient for the detection of this feature.

- Optionally, a Gaussian curve is fitted to the feature, using the Nonlinear Least Squares (NLS) implementation of R.

The result of the *centWave* algorithm for the regions shown in Figure 3 and 4 is depicted in Figure 6 and 7, respectively. The following experiments were designed to pose challenges with increasing complexity to the feature detection algorithms. We used complex mixtures with *Arabidopsis thaliana* leaf and seed extracts.

2.2 Experimental setup and Sample description

Arabidopsis thaliana (ecotype Col-0) was grown under controlled conditions and pooled after harvest. Methanolic extracts were prepared from ground seed and leaf tissue. o-Anisic acid, biochanin A, p-coumaric acid, ferulic acid, *N*-(3-indolylacetyl)-L-valine, kinetin, indole-3-acetonitrile, indole-3-carbaldehyde, kaempferol, phloretin, phlorizin and phenylglycine, rutin, and phenylalanine-d5 were used as marker compounds. The chromatographic separations were performed on an Acquity UPLC system (Waters) equipped with a modified C_{18} column with a 20 min water/acetonitrile gradient. The eluted compounds were detected by a Bruker MicrOTOF-Q in positive ion mode at a scan rate of 3 Hz. Mass calibration was performed against lithium formiate. The detailed experimental setup is available as Additional file 1.

DOI:10.1186/1471-2105-9-504

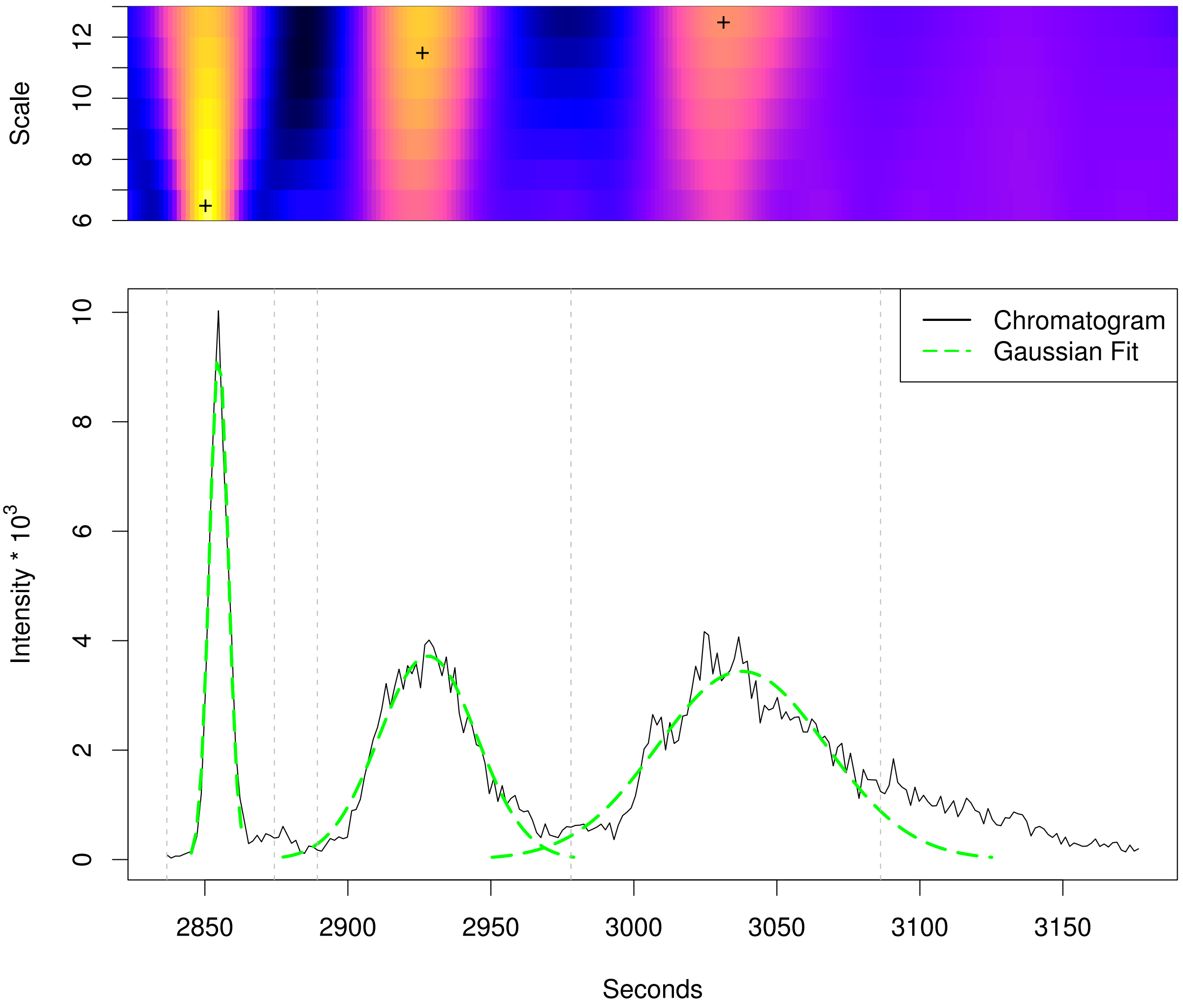


Figure 6
centWave results for example region 1. *centWave* results for example region 1. The lower part shows the same extracted ion chromatogram (277.213–277.221 *m/z*) as in Figure 3 and the detected chromatographic peaks from the *centWave* algorithm as Gaussian fits. The upper part shows the CWT coefficients on the different scales. A cross marks the scale where the peak was optimally localised. The vertical grey lines show the peak borders which were estimated from the coefficients of this scale.

Sample 1 A mixture containing each of the fourteen marker compounds (referred to as MM14) at a concentration of 20 μM was prepared and analysed by UPLC/ESI-QTOF-MS.

Sample set 2 Mixtures containing solvent and seed or leaf extracts were prepared with following volume portions (solvent/seed/leaf, v/v/v): 0/100/0, 25/75/0, 50/50/0, 75/25/0, 0/0/100, 25/0/75, 50/0/50, 75/0/25. The sample set (8 samples) was analysed by UPLC/ESI-QTOF-MS in ten technical replications.

Sample set 3 Mixtures containing solvent, seed, and leaf extracts were prepared with following volume portions (solvent/seed/leaf, v/v/v): 75/0/25, 0/75/25, 0/50/50. The sample set (3 samples) was analysed by UPLC/ESI-QTOF-MS in ten technical replications.

DOI:10.1186/1471-2105-9-504

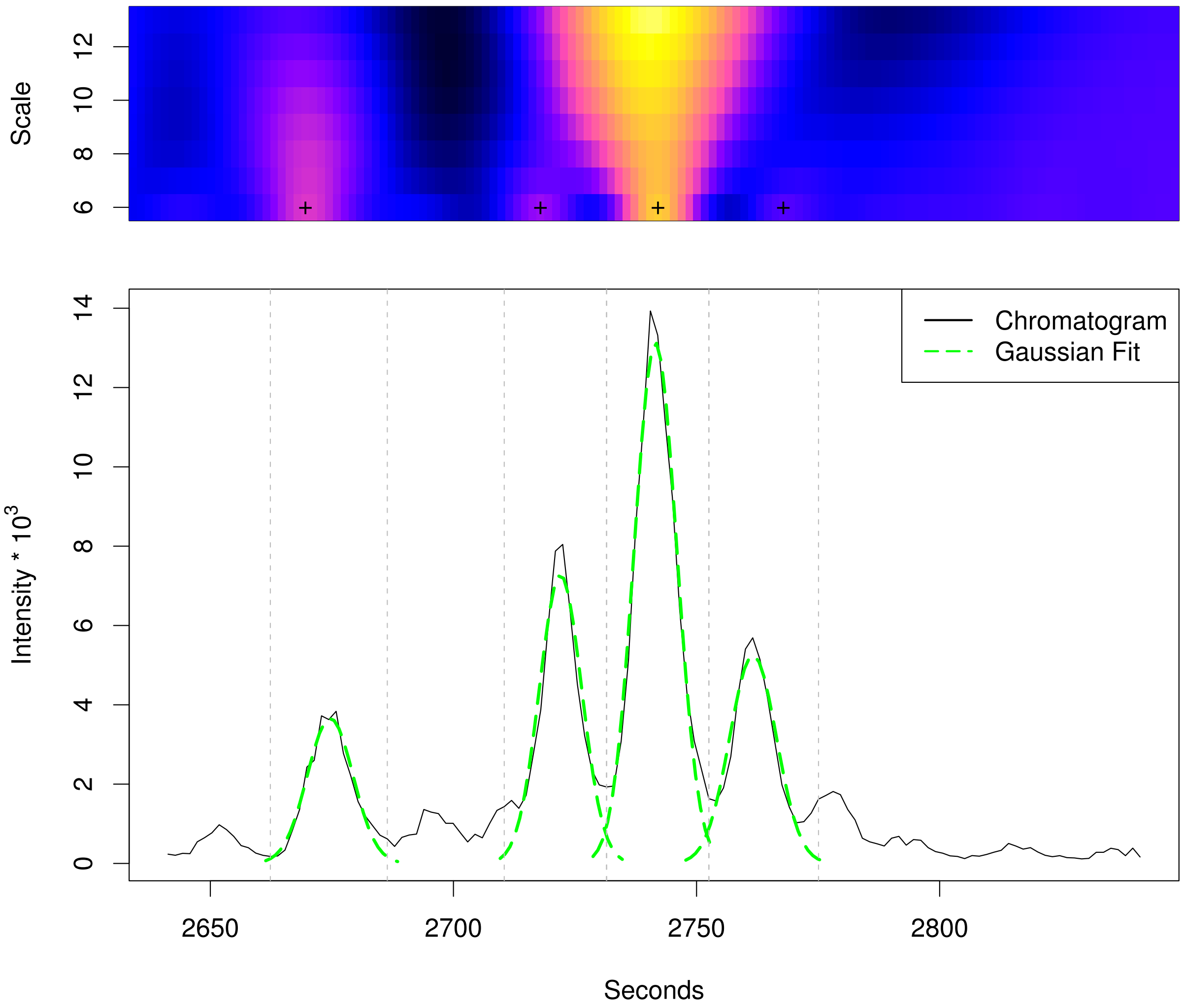


Figure 7
centWave results for example region 2. *centWave* results for example region 2. The lower part shows the same extracted ion chromatogram (967.53–967.56 *m/z*) as in Figure 4 and the detected chromatographic peaks from the *centWave* algorithm as Gaussian fits. The upper part shows the CWT coefficients on the different scales. A cross marks the scale where the peak was optimally localised. The vertical grey lines show the peak borders which were estimated from the coefficients of this scale.

All files were acquired in centroid mode and converted to mzData file format using Bruker CompassXport software. The data sets are available at http://msbi.ipb-halle.de/msbi/centwave/.

2.3 Parameter optimisation

Beside *centWave*, there are currently only two other feature detection algorithms available [9], which meet the following criteria: freely available, open source, and suited for feature detection in metabolomic LC/MS samples measured in centroid mode: *matchedFilter*- the originally implemented algorithm from XCMS and the *centroidPicker* from *MZmine* (Table 1).

The three algorithms tested have a number of parameters each, which have to be tuned to deliver good performance on the analytical setup. The *centWave* algorithm uses the *peakwidth* (= w_{min}, w_{max}) parameter to specify the chromatographic peak width range, the *ppm* parameter to set the tolerated mass deviation and *snthresh*, which defines the chromatographic signal-to-noise threshold. The *matchedFilter* algorithm has a similar parameter *snthresh*, the chro-

DOI:10.1186/1471-2105-9-504

Table 1: Overview of the evaluated feature detection algorithms

Algorithm	Framework	Version	Programming Language	Availability
centroidPicker	MZmine	0.60	Java	http://mzmine.sourceforge.net/
centWave matchedFilter	XCMS	1.12.1	C, R	http://bioconductor.org/packages/release/bioc/html/xcms.html

Overview of the compared feature detection algorithms for metabolomics data. MZmine has three feature detection algorithms implemented, but only the centroidPicker is suitable for centroided data and therefore was used for the evaluation.

matographic peak width is specified by the *fwhm* parameter, which defines the width of the model peak for matched filtering. The mass accuracy is indirectly defined by the bin size (parameter *step*).

The centroidPicker from *MZmine* also needs a bin size to be specified (*bin size*), and additionally the tolerated mass deviation (*m/z tolerance*). Moreover, there are five parameters that affect the chromatographic domain: *chromatographic threshold level*, *intensity tolerance*, *minimum peak duration*, *minimum peak height* and *noise level*. The first two of those are specified as a relative value, while the last three are set using absolute values.

The parameters of the three algorithms were tuned to detect as many of the real features, without allowing too many false positives. Based on known "good working" settings, we performed parameter sweeps and evaluated the number of real features and the number of other ("false") features for each setting. After initial optimisation of the other parameters, we found that for *centWave* and *matchedFilter* the *snthresh* parameter shows the highest influence on this ratio.

The centroidPicker from *MZmine* was more complex to optimise, due to its many parameters. Using settings from the authors as a starting point, a sweep was performed over a wide parameter range. Approximately 500 parameter settings were tried for *MZmine*, and about 50 for *matchedFilter* and *centWave*.

For the parameter optimisation we used the mixture of 14 compounds (MM14). Due to the electrospray ionisation, each compound gives rise to a number of features. A data set of known features was created using the separately measured substances. We annotated features that can be explained as adducts and fragments of the compound as well as their isotopic peaks. For all 14 compounds this results in a set of 296 features, about 21 features per compound. We observed up to eight in-source fragments per compound and also various cluster ions like $[2M+H]^+$ oder $[3M+Na]^+$. The annotations are available as Additional file 2. Manual verification shows, that 122 of the 296 known features are clearly visible in the MM14 mixture, while the other 174 features are hard to detect by the human eye. The 122 verified features are considered as required features, which should be detected by the algorithms.

All other features (beyond the 296) which were reported by the programs, but cannot be explained as features originating from the marker mixture, are considered as "false" features, e.g. (usually small) signals from solvents or chemical impurities, background noise etc.

As one result from the optimisation, we found that all algorithms are able to detect more than 100 from the 122 selected real features, but only if approximately 450 "false" features are tolerated. The total number of 122 real features are detected only with settings that give more false positives (see section 3.4). Therefore, as a trade-off between real and "false" features, we chose those parameter settings which detect a maximal number of real features, but return less than 450 "false" features.

Since the algorithms detect around 200–300 features in the separately measured blank solvent, these 450 "false" features can be explained as a "background", consisting of features originating from the solvents, tubes, vials, or impurities of the used marker compounds.

The result of the optimisation process can be seen in Table 2. These parameter settings were used for the following experiments.

2.4 Evaluation

Since feature detection can be seen as an information retrieval task, the performance can be assessed using the *precision* and *recall* values. The recall value (also referred to as *sensitivity*) measures the fraction of relevant items that are found by a query, while the precision value quantifies the relation of relevant items to the false positives. Denoting the total number of features that were detected by an algorithm by N, the number of real features that were found by TP, and the total number of real features by NP, we can measure Recall = $\frac{TP}{NP}$ and Precision = $\frac{TP}{N}$ of a fea-

Table 2: Parameter optimisation using the MM14 marker mixture

Algorithm	Number of detected MM14 features	Number of other reported features	Parameters
centWave	115	443	peakwidth = (5,10), ppm = 30, snthresh = 5, prefilter = (2,400)
matchedFilter	114	425	fwhm = 4, snthresh = 12, step = 0.02, mzdiff = 0, max = 50
MZmine	107	442	bin size = 0.05, chromatographic threshold level = 0.8, intensity tolerance = 0.7, minimum peak duration = 3, minimum peak height = 500, m/z tolerance = 0.03, noise level = 20

Number of features detected in the MM14 marker mixture and the parameter values that were chosen after the parameter optimisation step.

ture detection algorithm. A perfect feature detection algorithm will have both measures equal to 100%. False positives features lower the precision; false negatives (undetected real features) lower the recall.

For a compact representation of the results we used the *F-score* as a combined measure of precision and recall, which is defined as F-score = $\frac{2 \cdot R \cdot P}{R+P}$ [20]. A perfect feature detection will achieve a F-score of 100%, and both false positives and false negatives features lower its value. The F-score can be interpreted as a measure of the overall performance of a feature detection algorithm.

Results and discussion

We performed two experiments to assess the performance of the three algorithms. The experiments were designed to evaluate the sensitivity of the algorithms using complex biological samples at different concentrations.

First, the feature set representing the ground truth had to be created. For this purpose we used ten technical replicates of undiluted *Arabidopsis thaliana* seed and leaf extracts from Sample set 2 (solvent/seed/leaf): (0/100/0) and (0/0/100).

Since a manual annotation of the features was out of scope, we applied the following procedure to create a list of reliably detected features:

1. Feature detection on the 2 × 10 samples was performed using the three algorithms

2. We investigated the number of features which are found reproducibly in repeated measurements. The features detected in the ten technical replicates of undiluted seed and leaf extracts were separately aligned using XCMS *group* function (*mzwid* = 0.05, *bw* = 2). After the alignment only those features which were detected in at least seven out of the ten samples were retained. The resulting numbers of features are shown in Table 3.

3. We matched the aligned feature lists of all three algorithms (using 0.015 *m/z* and 5 s tolerance) and removed those features which had been found by only a single algorithm.

The resulting feature list contains 2281 features for the leaf- and 2345 features for the seed extract. 4076 features are unique, 550 features appear in both extracts. The filtering (step 2. & 3.) retained only the reliable features both across the replicates and detected by the majority of feature detection algorithms, see Figure 8. This data set was considered as ground truth feature data and used for the further evaluation.

3.1 Experiment 1

We evaluated the F-score (calculated from recall and precision values) for dilution series of the seed extract (Sample set 2 (solvent/seed/leaf): (25/75/0), (50/50/0), (75/25/0)). Feature detection was performed on the 3 × 10 samples with the three algorithms using the optimised parameters. Detected features that match the seed specific ground truth features were marked als true positives, while all other returned features were considered as false positives. The results are shown in the the left-most part of Figure 9. The same was done for the leaf specific features and different concentrations of the leaf extract (Sample set 2 (solvent/seed/leaf): (25/0/75), (50/0/50), (75/0/25)). The middle part of Figure 9 depicts the results. The *centWave* algorithm achieved up to 6% higher F-score values than *MZmine* and up to 14% more than *matchedFilter* in this experiment.

Table 3: Aligned features

Algorithm	Number of aligned features Seed	Leaf
centWave	2634	2423
matchedFilter	1568	1919
MZmine	2529	2699

Number of features that have been reliably detected in at least seven out of ten technical replicates from LC/MS analyses of seed and leaf extracts (Experiment 1).

DOI:10.1186/1471-2105-9-504

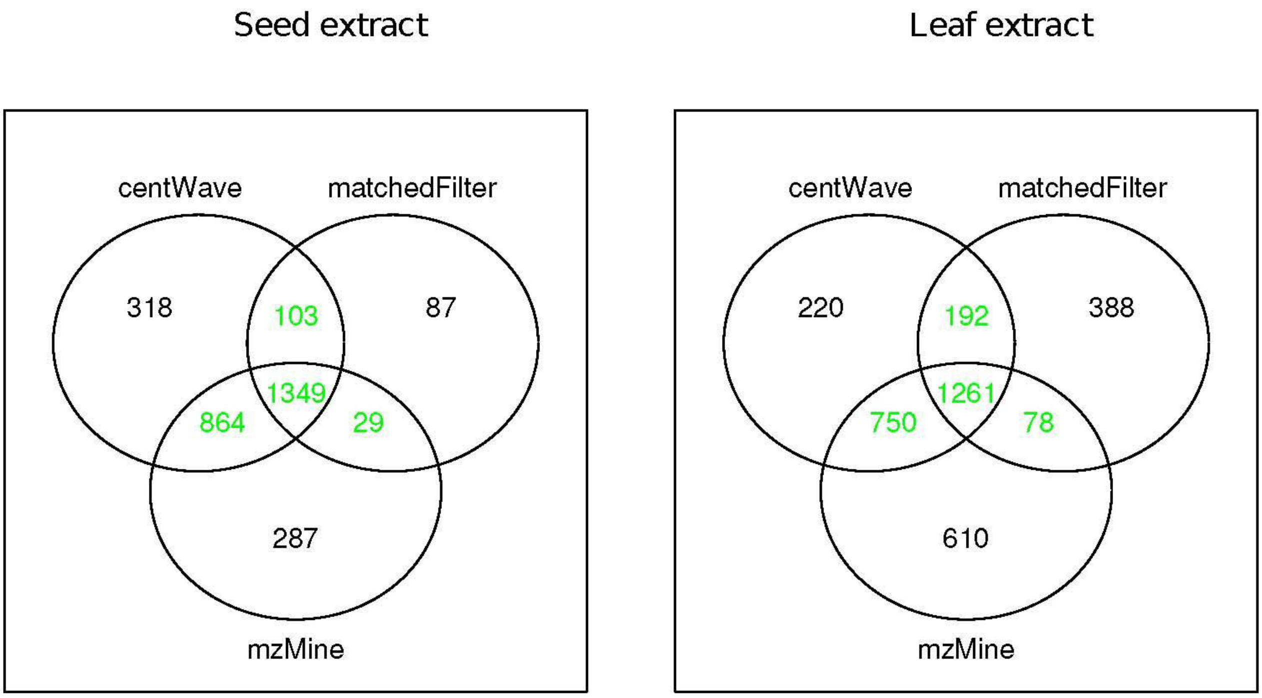


Figure 8
Venn Diagrams of Detected Features. Venn Diagrams showing the number of features in seed and leaf extracts that were found by the three different algorithms. Only the overlapping (green coloured) subsets were used as ground truth.

3.2 Experiment 2

For the second experiment we created mixtures of the seed and leaf extract at different concentrations (Sample set 3) and evaluated the F-score of the ground truth features. Again, feature detection was performed with the three algorithms. The ground truth seed and leaf specific features were considered together as true positives for this measurement. Thereby, the features which appear in both, seed and leaf extracts, were considered only once. All other features that were returned by the algorithms were considered as false positives. The right-most part of Figure 9 shows the result.

The detailed F-score, recall, and precision values of both experiments are available as Additional file 3. By manual inspection of the "true" features that were detected by *centWave*, but not by *MZmine* or *matchedFilter*, we found that these features were often close to other – in many cases larger – chromatographic peaks. This can be interpreted as a masking effect caused by noise level computation on the full chromatogram. The *centWave* algorithm uses local baseline and noise estimation to circumvent this problem.

Looking at the false positive features, we observed that *matchedFilter* frequently reports spikes (very narrow chromatographic peaks, consisting of 1–3 points) while *MZmine* tends to detect features in regions where only a high level noise can be seen.

3.3 Runtime

All three algorithms perform the feature detection for one sample in less than two minutes. *centWave* was the fastest algorithm in the test, with on average only one minute runtime per sample. The runtimes shown in Table 4 were measured as wall-clock time including all file input without other programs running. All measurements were done on an AMD Athlon 64 X2 Dual Core Processor 4200+ with 4GB RAM, running Linux (Ubuntu 6.06). Both frameworks can distribute the processing tasks, *MZmine* using Java RMI and XCMS using the Message Passing Interface (MPI) via Rmpi [21] on multicore architectures (and even cluster setups) to speed up the processing of many samples. This option was not used for the runtime measurements.

3.4 Alternative parameter settings

The optimisation strategy outlined in section 2.3 tried to balance the recall and precision, using a rough estimate of the potential chemical background signals. The settings in typical metabolomics profiling experiments of e.g. plant extracts usually will be tuned more aggressively towards

DOI:10.1186/1471-2105-9-504

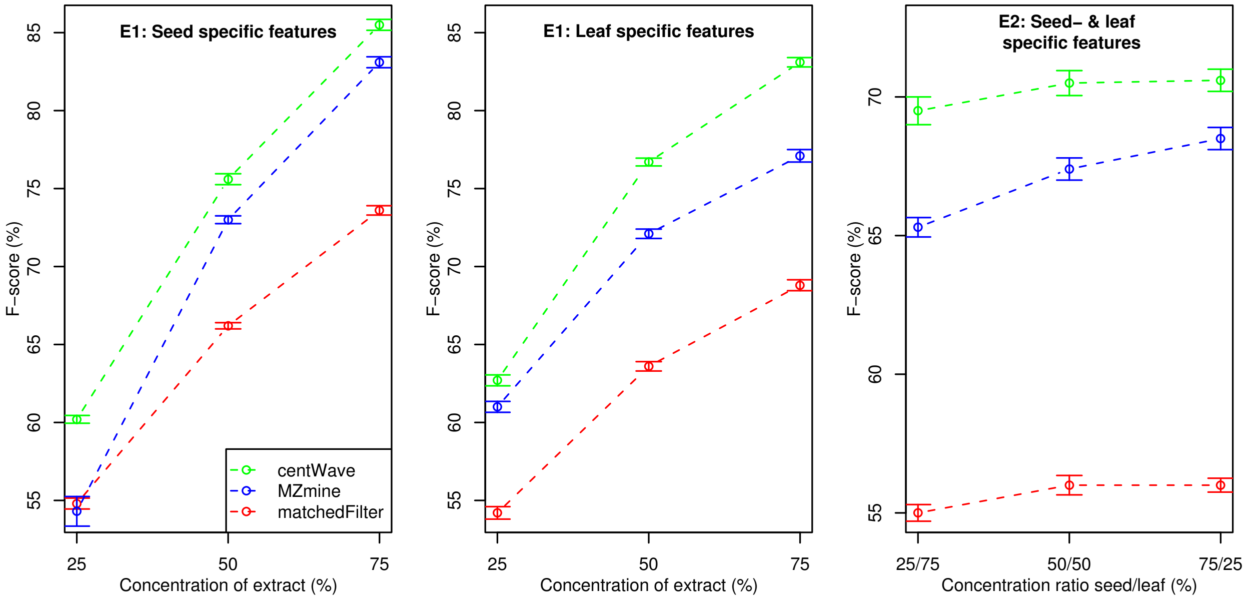


Figure 9

F-score values for Experiment 1 & 2. F-score (combined measure of recall and precision, calculated from the ground truth features) for dilution series of the seed and leaf extract (left-most and middle part) and for mixtures of the seed and leaf extract (right-most part of the figure). Detected features that match the respective ground truth features were counted als true positives, while all other features returned were considered as false positives. Higher F-score values represent better feature detection performance.

higher sensitivity. The resulting false positive features are often filtered by the downstream analysis, such as the alignment of replicate measurements and statistical tests for differential features.

We repeated the parameter optimisation, this time allowing up to 1000 background features. Essentially, the respective chromatographic threshold parameters were lowered, to achieve higher sensitivity. With these parameters we recreated the ground truth, and repeated both experiments. The results are depicted in Figure 10. The parameter settings and the number of features in the ground truth, as well as the detailed F-score, recall, and precision values are available as Additional file 4.

Table 4: Runtimes

	centWave	matchedFilter	MZmine
Runtime in minutes	1.02	1.85	1.54

Average wall-clock runtime in minutes for feature detection in one sample averaged across ten samples containing a 50/50 leaf/seed extract mixture.

With the second parameter set, we observed higher sensitivity for all algorithms. The number of aligned features almost doubled, the resulting ground truth contains 6649 unique features. The recall values of *matchedFilter* improved notably with the alternative parameter settings.

The results based on the second parameter set confirm the general trend shown above. The *centWave* algorithm achieved up to 6% and 15% higher F-score values than *MZmine* and *matchedFilter*, respectively.

Conclusion

We presented a new feature detection algorithm for high resolution LC/MS data called *centWave*. With the increasing deployment of high-resolution mass spectrometers such as QTOF or Orbitrap instruments, and high-throughput applications such as metabolomics experiments of highly complex samples, a reliable and sensitive feature detection is essential. *centWave* shows a high sensitivity, while trying to keep the false positive features low.

In the past, the Bioconductor project has attracted more and more development related to mass spectrometry and metabolite pathways. The implementation of *centWave* is

DOI:10.1186/1471-2105-9-504

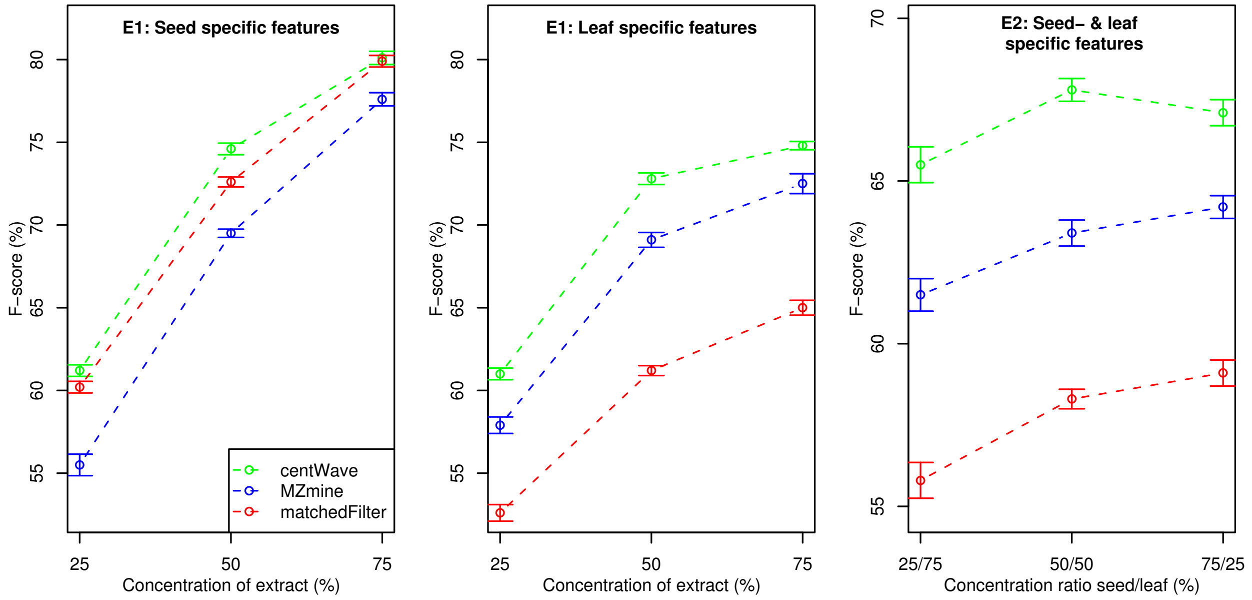


Figure 10

F-score values for Experiment 1 & 2 (alternative parameter settings). F-score (combined measure of recall and precision, calculated from the ground truth features) for dilution series of the seed and leaf extract (left-most and middle part) and for mixtures of the seed and leaf extract (right-most part of the figure). Detected features that match the respective ground truth features were counted als true positives, while all other features returned were considered as false positives. Higher F-score values represent better feature detection performance. Alternative parameter settings were used (see Additional file 4).

available in the R-package XCMS and can be obtained from http://bioconductor.org/packages/release/bioc/html/xcms.html. Integration with Bioconductor provides good support for the common file formats (netCDF, mzData and mzXML, with mzML currently under development) and allows for powerful downstream statistical analysis. The user feedback on the XCMS mailing list showed, that the *centWave* algorithm (introduced in 2007) is successfully used for LC-QTOF, LC-Orbitrap and even CE-MS or GC-MS data. For a more objective comparison we have evaluated *centWave* against two other open source algorithms. We performed two experiments to assess the performance of the algorithms, using complex chemical mixtures at different concentrations. The F-score, as a combined measure of recall and precision, was calculated using the ground truth data. The result was for *centWave* always higher than for *matchedFilter* and *MZmine*. The *centWave* algorithm is based on a sensitive detection of potentially interesting mass traces (ROIs), followed by an extensive chromatographic analysis, that reliably detects chromatographic peaks with different width via CWT. To allow for high sensitivity, baseline and noise are estimated locally. Some efforts are made to locate the exact chromatographic peak boundaries to provide accurate peak intensities. Feature quality can be assessed using numerous metrics, including signal to noise ratio, m/z fluctuation, and the residual of the Gaussian fit. Further development of the *centWave* algorithm will include an automatic estimation of the processing parameters.

In addition to *centWave* and the LC/MS data sets we have released the manual annotation of an LC/MS measurement of several pure compounds as a benchmark data set for both machine and software comparisons. The data sets are available at http://msbi.ipb-halle.de/msbi/centwave/.

Authors contributions

CB performed the LC/MS measurements and was involved in the development of the *centWave*. RT designed and implemented the *centWave* algorithm. RT and SN performed the evaluation of the algorithms. All authors contributed to, read and approved the fnal manuscript.

DOI:10.1186/1471-2105-9-504

Additional material

Additional file 1

***Experimental setup**. Detailed description of materials, chemicals, and protocols.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-504-S1.pdf]

Additional file 2

***MM14 annotations**. Feature annotations for the mixture of 14 compounds (MM14).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-504-S2.pdf]

Additional file 3

***Detailed recall, precision, and F-score values**. Tables containing the detailed F-score, recall, and precision values of both experiments.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-504-S3.pdf]

Additional file 4

***Results for the alternative parameter settings**. Venn diagrams of the ground truth data as well as the detailed F-score, recall, and precision values of both experiments using alternative parameter settings.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-504-S4.pdf]

Acknowledgements

Thanks to Colin Smith, who started the Bioconductor package XCMS and to Pan Du who wrote the MassSpecWavelet package for some of the CWT code.

References

1. Oliver S, Winson M, Kell D, Baganz F: **Systematic functional analysis of the yeast genome.** *Trends Biotechnol* 1998, **16(9):**373-378.
2. Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey R, Willmitzer L: **Metabolite profiling for plant functional genomics.** *Nature Biotechnology* 2000, **18:**115.
3. Dunn WB: **Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes.** *Physical Biology* 2008, **5:**24 [http://stacks.iop.org/1478-3975/5/011001].
4. Roepenack-Lahaye Ev, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D, Clemens S: **Profiling of Arabidopsis Secondary Metabolites by Capillary Liquid Chromatography Coupled to Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry.** *Plant Physiology* 2004, **134:**548-559.
5. Böttcher C, Roepenack-Lahaye Ev, Schmidt J, Schmotz C, Neumann S, Scheel D, Clemens S: **Metabolome Analysis of Biosynthetic Mutants Reveals Diversity of Metabolic Changes and Allows Identification of a Large Number of New Compounds in Arabidopsis thaliana.** *Plant Physiol* 2008:108.117754 [http://www.plantphysiol.org/cgi/content/abstract/pp.108.117754v1].
6. Tikunov Y, Lommen A, Vos Cd, Verhoeven H, Bino R, Hall R, Bovy A: **A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles.** *Plant Physiol* 2005, **139(3):**1125-37.
7. Smith C, Want E, O'Maille G, Abagyan R, Siuzdak G: **XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification.** *Analytical Chemistry* 2006, **78(3):**779-787.
8. Katajamaa M, Oresic M: **Processing methods for differential analysis of LC/MS profile data.** *BMC Bioinformatics* 2005, **6:**179.
9. Katajamaa M, Oresic M: **Data processing for mass spectrometry-based metabolomics.** 2007.
10. Stolt R, Torgrip R, Lindberg J, Csenki L, Kolmert J, Schuppe-Koistinen I, Jacobsson S: **Second-Order Peak Detection for Multicomponent High-Resolution LC/MS Data.** *Analytical Chemistry* 2006, **78(4):**975-983.
11. Aberg K, Torgrip R, Kolmert J, Schuppe-Koistinen I, Lindberg J: **Feature detection and alignment of hyphenated chromatographic-mass spectrometric data Extraction of pure ion chromatograms using Kalman tracking.** *J Chromatogr A* 2008, **1192:**139-146.
12. Lange E, Gröpl C, Reinert K, Kohlbacher O, Hildebrandt A: **High-accuracy peak picking of proteomics data using wavelet techniques.** *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 2006:243-254.
13. Du P, Kibbe WA, Lin SM: **Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.** *Bioinformatics* 2006, **22(17):**2059-2065.
14. Conrad TOF, Leichtle A, Hagehulsmann A, Diederichs E, Baumann S, Thiery J, Schütte C: **Beating the Noise: New Statistical Methods for Detecting Signals in MALDI-TOF Spectra Below Noise Level.** *CompLife* 2006, **4216:**119-128 [http://dblp.uni-trier.de/db/conf/complife/complife2006.html#ConradLHDBTS06]. Lecture Notes in Computer Science Springer
15. McLerran DF, Feng Z, Semmes OJ, Cazares L, Randolph TW: **Signal Detection in High-Resolution Mass Spectrometry Data.** *Journal of Proteome Research* 2008, **7:**276-285.
16. Nordström A, O'Maille G, Qin C, Siuzdak G: **Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum.** *Anal Chem* 2006, **78:**3289-3295.
17. Danielsson R, Bylund D, Markides K: **Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography – mass spectrometry.** *Analytica Chimica Acta* 2002:167-184.
18. Andreev V, Rejtar T, Chen HS, Moskovets E, Ivanov A, Karger B: **A Universal Denoising and Peak Picking Algorithm for LC-MS Based on Matched Filtration in the Chromatographic Time Domain.** *Analytical Chemistry* 2003, **75(22):**6314-6326.
19. Daubechies I: *Ten lectures on wavelets* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics; 1992.
20. Rijsbergen CJV: *Information Retrieval* Newton, MA, USA: Butterworth-Heinemann; 1979.
21. Yu H: **Rmpi: Parallel Statistical Computing in R.** *R News* 2002, **2(2**10-14 [http://CRAN.R-project.org/doc/Rnews/].

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

BMC Bioinformatics

BioMed Central

Research article

Open Access

Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements

Eva Lange*[1], Ralf Tautenhahn*[2], Steffen Neumann[2] and Clemens Gröpl[3]

Address: [1]Beatson Institute for Cancer Research, Proteomics and Mass Spectrometry Group, Scotland, UK, [2]Leibniz Institute of Plant Biochemistry, Bioinformatics and Mass Spectrometry, Halle, Germany and [3]Free University Berlin, Department of Mathematics and Computer Science, Berlin, Germany

Email: Eva Lange* - e.lange@beatson.gla.ac.uk; Ralf Tautenhahn* - rtautenh@ipb-halle.de; Steffen Neumann - sneumann@ipb-halle.de; Clemens Gröpl - groepl@inf.fu-berlin.de

* Corresponding authors

Published: 15 September 2008

BMC Bioinformatics 2008, **9**:375 doi:10.1186/1471-2105-9-375

Received: 31 March 2008

Accepted: 15 September 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/375

© 2008 Lange et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Liquid chromatography coupled to mass spectrometry (LC-MS) has become a prominent tool for the analysis of complex proteomics and metabolomics samples. In many applications multiple LC-MS measurements need to be compared, e. g. to improve reliability or to combine results from different samples in a statistical comparative analysis. As in all physical experiments, LC-MS data are affected by uncertainties, and variability of retention time is encountered in all data sets. It is therefore necessary to estimate and correct the underlying distortions of the retention time axis to search for corresponding compounds in different samples. To this end, a variety of so-called *LC-MS map alignment algorithms* have been developed during the last four years. Most of these approaches are well documented, but they are usually evaluated on very specific samples only. So far, no publication has been assessing different alignment algorithms using a standard LC-MS sample along with commonly used quality criteria.

Results: We propose two LC-MS proteomics as well as two LC-MS metabolomics data sets that represent typical alignment scenarios. Furthermore, we introduce a new quality measure for the evaluation of LC-MS alignment algorithms. Using the four data sets to compare six freely available alignment algorithms proposed for the alignment of metabolomics and proteomics LC-MS measurements, we found significant differences with respect to alignment quality, running time, and usability in general.

Conclusion: The multitude of available alignment methods necessitates the generation of standard data sets and quality measures that allow users as well as developers to benchmark and compare their map alignment tools on a fair basis. Our study represents a first step in this direction. Currently, the installation and evaluation of the "correct" parameter settings can be quite a time-consuming task, and the success of a particular method is still highly dependent on the experience of the user. Therefore, we propose to continue and extend this type of study to a community-wide competition. All data as well as our evaluation scripts are available at http://msbi.ipb-halle.de/msbi/caap.

DOI:10.1186/1471-2105-9-375

1 Background

Mass spectrometry (MS) has become the predominant technology for both proteomics and metabolomics experiments. In shotgun proteomics, proteins are first digested, then the resulting peptides are separated by liquid chromatography. The fractions of the mixture are transferred to the mass spectrometer. Soft ionization techniques like matrix-assisted laser desorption ionization (MALDI) or electrospray ionization (ESI) and high resolving mass analyzers are used to identify the individual compounds by peptide mass fingerprinting (PMF) or by tandem mass spectrometry. The latter uses another step of fragmentation and MS analysis (MS/MS). Multiple technologies also exist in metabolomics applications, where mass spectrometers are coupled to gas chromatography (GC), liquid chromatography (LC) or capillary electrophoresis (CE) for separation. For recent reviews see [1,2]. In this paper our focus is on LC-MS in proteomics and metabolomics applications.

The quantitative information in a proteomics LC-MS map can be used in numerous applications [3,4] ranging from additive series in analytical chemistry [5], analysis of time series in expression experiments [6,7], to applications in clinical diagnostics [8], where statistically significant markers detect certain states of diseases. Common applications in metabolomics are: The verification of substantial equivalence [9], or the profiling of, e.g., biosynthetic mutants to reveal cross-talk between pathways [10]. What applications have in common is that the same components in different measurements have to be related to each other. As with every laboratory experiment, chromatographic separation is stable and reproducible only to a certain extent. The retention time often shows large shifts, and distortions can be observed when different runs are compared. Even the m/z dimension might show (typically smaller) deviations. The overall change in RT and m/z is called *warp*. Pressure fluctuations, or changes in column temperature or mobile phase result in distorted elution patterns, and can even cause changes in the elution order of components. Elution order changes are not unlikely if their retention times are similar [11]. For example, in one of our data sets the ground truth contained 88 verified matching peptide signals, but no more than 66 of them can be aligned without elution order changes (see last figure in additional File 1 for further information). The correction of the shift in RT and m/z is called *dewarping* according to the time warping problem of Sakoe and Chiba [12] in speech processing. The advent of high-throughput quantitative proteomics and metabolomics makes an efficient solution to this problem an important task.

In general, the data processing pipeline for label-free LC/MS data proteomics and metabolomics applications can be divided into the following steps:

1. Signal preprocessing and centroidization,

2. Detection and extraction of two-dimensional signals, so-called *features*, which are caused by chemical entities,

3. Intensity normalization,

4. Compensation of retention time distortions by dewarping,

5. Computation of a consensus map by assigning corresponding features across multiple maps,

6. Statistical analysis, feature identification, and the biological interpretation.

A typical label-free quantification protocol might be the connection of the proposed analysis steps, but it can also consist of the comparison of LC-MS maps on the raw data level [13]. The comparison of LC-MS raw maps enables the search for differentially expressed peptides directly by using multiway data analysis methods (e.g., PARAFAC [14]). Hence, a typical analysis pipeline for this approach avoids the steps 2 and 5, and merely includes the preprocessing and intensity normalization of the LC-MS raw maps, the correction of the retention time distortion, as well as the statistical analysis, feature identification and the biological interpretation of the data. We call the dewarping and thereby superposition of multiple LC-MS raw maps the *LC-MS raw map alignment problem*. Several algorithms have been designed to deal with this problem [15-19]. They avoid errors introduced by centroidization and feature finding algorithms, but they tend to have high runtimes and are liable to time order changes. Moreover, the algorithms are usually described for pairwise alignment and do not easily generalize to a multiple alignment of *N* maps. In this paper we will concentrate on the typical label-free quantification analysis pipeline and focus on the so-called *LC-MS feature map alignment problem*, which comprises the dewarping of multiple feature maps as well as the grouping of corresponding features in different maps. Since feature maps have a much smaller data amount than raw maps, they allow for much faster dewarping algorithms. On the other hand, signal preprocessing, centroidization and feature finding may also introduce errors. Therefore, the quality of the feature maps strongly depends on the reliability of these processing steps.

Within the last four years several algorithms for LC-MS feature map alignment have been developed [20-27].

DOI:10.1186/1471-2105-9-375

These tools are either standalone tools or part of a whole framework for the analysis of MS based data. In this paper we concentrate on the comparison of the freely available feature map alignment algorithms implemented within the frameworks msInspect [25], MZmine [21], OpenMS [28] and XCMS [26], as well as the tools SpecArray [22] and XAlign [23] (see Table 1). Except for the alignment algorithm of MZmine, all methods estimate a linear (OpenMS, XAlign) or non-linear shift to correct the distortion of the RT dimension in all feature maps. The assignment of corresponding features and the determination of the consensus map is either done consecutively by processing the maps in a star-wise manner (MZmine, OpenMS, SpecArray, XAlign), or by a clustering approach (msInspect, XCMS). All algorithms take advantage of the more precisely measured m/z dimension to group corresponding features and to estimate the underlying warping function in RT. The general approach of the six different alignment methods compared by us will be described in the next section.

With the recent advent of LC-MS alignment reviews [13,29] it became obvious that a comprehensive unbiased performance study on a common benchmark set is needed to foster further competition and collaboration between the developers. In related fields, the Critical Assessment of Methods for Protein Structure Prediction (CASP) contests [30] and the Affycomp II Benchmark for Affymetrix GeneChip Expression Measures [31] have been quite fruitful in this respect. We have collected benchmark data sets from both proteomics and metabolomics experiments to compare *only the feature map alignment modules* of different software packages. We aim to minimize the influence of the preceding and subsequent processing steps. Therefore, we eliminated the influence of the individual signal processing modules by importing a common feature list. We furthermore abandoned the search for features in individual files based on features found in other measurements which is sometimes referred to as filling-in missing features. For proteomics, we have selected two data sets from the Open Proteomics Database [32], which have been used previously for the evaluation of the raw map alignment algorithm OBI-Warp [17]. For metabolomics data, no such public data repository currently exists, so we used two of our own data sets from a typical comparative metabolomics study. We are making these data sets available at http://msbi.ipb-halle.de/msbi/caap.

The remainder of this paper is structured as follows: In Sections 2.1 and 2.2 the benchmark data sets and the definition of ground truth are described. Section 2.3 introduces the MS software packages and how they were configured for the benchmark. The evaluation criteria are defined in Section 2.4. The results of our comparison are presented in Section 3, followed by a discussion of the merits of the underlying algorithms, and a conclusion of expected future developments in Section 4.

2 Methods

Before we describe the experimental setup and signal processing for the evaluation data sets we introduce some definitions that are used throughout the following sections. In our context, a *feature* is the two-dimensional (RT and m/z) signal caused by a single charge variant of a chemical entity. Feature detection involves identifying the signal region in the raw data (usually a union of convex sets) and fitting a theoretical model (e. g. elution profile, isotope distribution) to the observed data. The map align-

Table 1: Overview of alignment tools

framework *tool name*	input format	version	URL	programming language	operating system	source code available	modularity
msInspect *peptideMatch*	feature data in own tab-separated format	1.0.1	http://proteomics.fhcrc.org	Java, R	Windows Linux MaxOS	✓	✓
MZmine	raw data	0.60	http://mzmine.sourceforge.net	Java	Windows Linux MacOS	✓	-
OpenMS *MapAlignment*	feature data in featureXML or raw or peak data in mzData format	1.0	http://www.openms.de	C++	Linux MacOS (Windows)	✓	✓
SpecArray *PepMatch, PepArray*	feature data in own binary format	2.1	http://tools.proteomecenter.org	C	Linux	✓	✓
XAlign	feature data in own tabular separated format	03.09.2007	request from the author	C++	Windows	-	✓
XCMS	raw data	1.10.7	http://www.bioconductor.org	R, C	Windows Linux MacOS	✓	✓

ment problem has two aspects: (1) finding a suitable *transformation* of retention times, so that corresponding features will be mapped to nearby retention times, and (2) reporting the actual *groups* of corresponding features across multiple LC-MS feature maps. We will refer to these groups as *consensus features*, emphasizing that the individual features constituting a consensus feature should represent the same charge state of the same ionized compound. Referring to the consensus feature as a whole, one can then speak of an average retention time, mass charge ratio, etc. The collection of all consensus features constitutes a *consensus map*, which stores the correspondence information of all detected features in multiple LC-MS feature maps.

Ideally, each feature should be assigned to one consensus feature and each consensus feature should contain one feature from each map. However, limited dynamic range or large variation in the sample will lead to consensus features which do not extend across all LC-MS experiments. Artifacts of the feature detection phase, such as "broken" elution profiles, may also show up during the map alignment, resulting in consensus features which contain more than one feature from a particular map. As a special case, a consensus feature may consist of a single feature from a single map, if no other map contains the same charge state of the ionized compound. We will refer to these as *singletons*.

We consider the transformation of retention times as an intermediate step, because the downstream data analysis will mainly be concerned with groups of features and their average position, etc. rather than the distortions of retention times. The ultimate goal of multiple LC-MS feature map alignment is to derive a consensus map. This fact should be reflected by our quality metrics. An alignment method should create a "meaningful" partition of the feature maps: Corresponding features should be grouped in only one consensus feature instead of being split in multiple subsets, but the algorithm must also avoid grouping together unrelated features.

In Section 2.4 we introduce two measures that reflect the quality of a determined consensus map with respect to an optimal consensus map, the so-called *ground truth*. This is illustrated in Figure 1. The left part shows an optimal consensus map, representing the correspondence in four different feature maps. The right part shows a consensus map with various kinds of errors, which can occur in an alignment.

The quality of the transformation of retention times might also be assessed, but only after groups of corresponding features have been found. The transformation is often called a *warping function*, because original retention times x and transformed retention times y are related through a monotone increasing function $f(x) = y$. The difficulty with

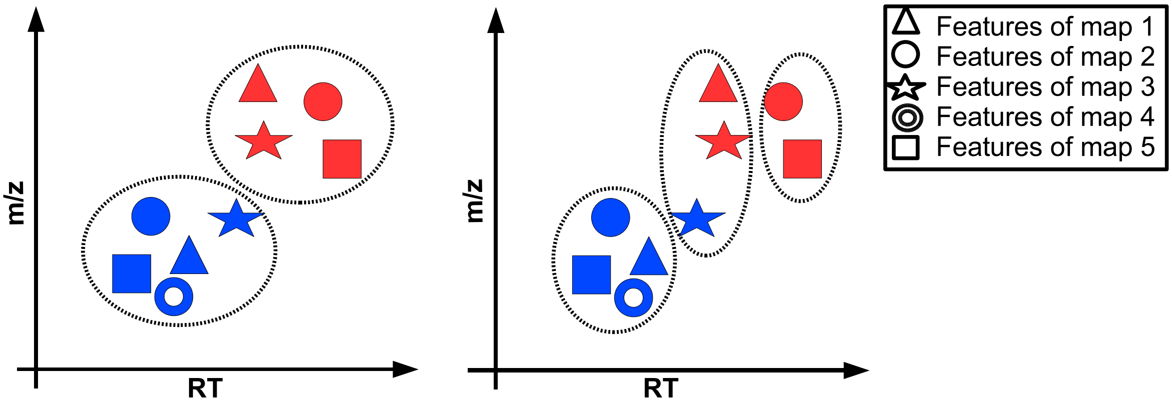


Figure 1

Consensus precision and recall. The left figure shows the two consensus features of a ground truth for the alignment of five feature maps. The features of the feature maps are distinguished by the five types of marker. Corresponding features in the different maps are illustrated by the same colour. The right figure shows three consensus features of a consensus map determined by an alignment algorithm. Note that the red features were assigned to separate consensus features, and the blue ones as well. The consensus feature in the middle even contains features from the same map. Thereby, the alignment results in a low recall value of $(1/2)\cdot(5/(2\cdot 5) + 4/(2\cdot 4)) = 0.5$. Since most of the determined consensus features are "relevant" the method achieved a precision of $(1/2)\cdot(5/7 + 4/5)\ \ 0.76$.

this approach is that the distance between corresponding features can be minimized by unrealistic, step-like warping functions. Hence in order to avoid overfitting, one has to include regularity (or "smoothing") conditions into the quality measure, which are hard to formalize.

In the following section we will describe the sample preparation of the complex biological proteomics and metabolomics data sets. Furthermore, we establish methods for the generation of proteomics and metabolomics ground truth consensus maps.

2.1 Proteomics data

We selected two proteomics data sets from the Open Proteomics Database (OPD) [32] resulting from two different experiments. The first data set originates from a dilution series of *Escherichia coli* and the other data set represents different cell states of *Mycobacterium smegmatis*. Both samples are of high complexity and provide typical alignment scenarios. They have previously been used for the evaluation of the LC-MS raw map alignment algorithm OBI-Warp [17].

We will briefly describe the sample preparation and the LC-LC-MS/MS analysis of the two experiments. Further information of the *E. coli* data set can be found on the OPD website and the *M. smegmatis* experiment is explicitly described in [33].

2.1.1 Experimental setup

Data set *P1*: LC-LC-ESI-IT-MS/MS

E. coli soluble protein extracts representing cells in exponential growth-phase were diluted in digestion buffer, denatured, and digested with trypsin. Tryptic peptide mixtures were separated by automated LC-LC-MS/MS. The injection quantity of the analyte was altered between two different runs: *021016_jp32A_10ul_3* (10 μL, [OPD: opd00005_ECOLI]) and *021010_jp32A_15ul_1* (15 μL, [OPD: opd00006 ECOLI]). We refer to these data sets as *P1_1* and *P1_2*, respectively. Chromatography salt step fractions were eluted from a strong cation exchange column (SCX) with a continuous 5% acetonitrile background and 10-min salt bumps of 0, 20, 40, 60, 80, and 100 mM ammonium chloride. Each salt bump was eluted directly onto a reverse-phase C_{18} column and washed free of salt. Reverse-phase chromatography was run in and peptides were analyzed online with an ESI ion trap mass spectrometer (ThermoFinnigan Dexa XP Plus). In each MS spectrum, the three tallest individual peaks, corresponding to peptides, were fragmented by collision-induced dissociation (CID) with helium gas to produce MS/MS spectra. Centroided mzXML data and corresponding SEQUEST identification results of P1_1 and P1_2 were downloaded from the OPD.

Data set *P2*: LC-LC-ESI-IT-MS/MS

M. smegmatis soluble protein extracts were diluted in digestion buffer, denatured, and digested with trypsin. Tryptic peptide mixtures were separated by automated LC-LC-MS/MS. The three different runs *6-17-03*, *7-17-03*, and *6-06-03* represent protein profiles of a *M. smegmatis* cell in middle exponential, early exponential and stationary phase [OPD: opd00009_MYCSM, opd00014_MYCSM, opd00028_MYCSM]. We refer to these data sets as *P2_1*, *P2_2*, and *P2_3*, respectively. The remaining setup is the same as above in *P1*. Centroided mzXML data and corresponding SEQUEST identification results of P2_1, P2_2, and P2_3 were downloaded from the OPD.

2.1.2 Data extraction

The raw data had been exported in centroided mode by the instrument. Preprocessing and data extraction was performed using TOPP tools [34]. We converted all data from *mzXML* to *mzData* format using FileConverter and transformed the data into a uniformly spaced matrix by bilinear resampling using Resampler. The spacing of the transformed matrix was 1 Th and 1 second. Afterwards we detected and extracted peptide signals in the resampled raw data maps using FeatureFinder ignoring the charge states to provide fair means of comparison for all alignment tools. The sizes of the feature maps from the *P1* and the *P2* alignment test set are available as additional File 2.

2.1.3 Ground truth

We established ground truth for the *P1* and the *P2* data sets by means of MS/MS information that was not available to the tested alignment procedures. As a consequence, our ground truth consist exclusively of features that can be annotated with a reliable peptide identification. This is discussed further below.

The reference method uses five steps: (1.) We establish an initial correspondence between MS/MS identifications and LC-MS features. (2.) We filter the peptide annotations based on the retention times of the features they are assigned to. The first two steps operate on each LC-MS/MS map individually. (3.) We compute an initial set of consensus features across multiple experiments. (4.) We reduce the list such that each feature is contained in at most one consensus feature. (5.) We filter the consensus features by comparing retention times across maps.

In the first step we scan through all peptide identifications. We disregard unreliable peptide identifications having a SEQUEST *XCorr* score less than 1.2. We check whether the RT and the m/z value of the precursor ion lies within the convex hull of a feature. In this case we assign the peptide identification to the feature. Each feature can be annotated with many peptide identifications originat-

DOI:10.1186/1471-2105-9-375

ing from many MS/MS scans within the experiment. The values in parentheses in additional File 2 are the number of annotated features.

In the second step we filter the peptide annotations with respect to the retention times of the features they are assigned to. If a peptide identification is assigned to two features with very different RTs in one map, it is likely that one or both features are falsely annotated. This observation is used to filter out dubious identifications which otherwise might give rise to incorrect consensus features in the ground truth. For each peptide identification, we compute the mean μ and standard deviation σ of the RT positions of the features to which it is assigned. If $\sigma > 100$ s, then the identification is considered dubious and removed from all features. Moreover, the identification is removed from all features, if any, whose RT positions deviate by more than 2σ from μ. These filters are applied for each experiment separately.

In the third step we compute an initial list of consensus features, in which features with identical identifications are grouped across maps. In the previous steps we have computed a set of associations between peptide identifications from MS/MS and LC-MS features. The consensus features in our ground truth should have unique peptide identifications. Therefore we start by compiling a complete list of all peptide identifications over all experiments. Then we step through this list and for each identification we find the best-scoring features associated with it, but at most one from each experiment, and add these features to the corresponding consensus feature. In this way we maximize the sum of XCorr values for the peptide identifications in a consensus feature. We discard dubious consensus feature whose m/z standard deviation is greater than 1.

Let the *total XCorr* score of a consensus feature be defined as the sum of XCorr values of all features contained in it. After step three, it is possible that a feature is contained in different consensus features from the initial list. In the fourth step we reduce the initial list such that each feature is contained in at most one consensus feature, whose total score is the largest among all consensus features containing it. We have developed a simple "greedy" strategy to achieve this goal. The purified list of candidate consensus features is sorted in order of decreasing total score. In each step we extract a consensus feature with maximum total XCorr score from the list. This consensus feature is added to the consensus map, and all consensus features having a non-empty intersection with it are also removed from the list. The process is iterated until no more consensus features can be found, i. e., the list has become empty.

In the fifth step, we apply a final filter for outliers and dubious identifications by comparing retention times across maps. We calculate the RT sample variance within all consensus features in the consensus map and discard consensus features whose standard deviation is greater than 2 times the sample standard deviation. Since this filter relies upon RT information and hence bears the risk of introducing bias into the ground truth, we confirmed that the removed consensus features are indeed outliers by visual inspection.

The numbers of consensus features in the ground truth are also shown in additional File 2. A ground truth is only considered if its number of consensus features corresponds to a least 10% of the number of annotated featues in the aligned feature maps.

As stated above, the assignment used as a ground truth is restricted to features in different feature maps that were annotated by a peptide identification. We believe that this will not introduce a bias toward any of the tools, based on the assumption that the features, which are selected for MS/MS fragmentation are chosen randomly and independently with the same probability p. For simplicity, consider the case of pairwise alignment. The extension to multiple map alignment will be discussed in Section 2.4. The classical *precision* value is defined as TP/(TP + FP). Note that the denominator does not depend on the ground truth, and the enumerator is expected to be a constant fraction $\mathrm{TP} = p \cdot \mathrm{TP}^*$ of the "real" true positive number TP*. Thus, it is still possible to *compare* the probability that a computed consensus feature is contained in the ground truth between the different tools, although the absolute precision values will be underestimated by a factor of p using the available ground truth. The *recall* value TP/(TP + FN) is not affected by such a bias, since both *TP* and *FN* will be underestimated by a factor of p, which cancels out. Hence, the classical recall value can still be used as an estimator for the probability that an "existing" consensus feature is actually computed by the tool.

2.2 Metabolomics data

We have selected a typical *Arabidopsis thaliana* metabolomics experiment, with different plant lines and treatments measured at multiple time points in triplicates. The same samples were measured on two different LC-MS setups as follows.

2.2.1 Experimental setup

Preparation of Extracts

Freshly ground *Arabidopsis thaliana* leaf tissue (130 ± 5 mg) was subjected twice to the following extraction procedure: mixing with 200 μL of methanol/water, 4/1 (v/v), sonication at 22°C for 15 min and centrifugation for 10

DOI:10.1186/1471-2105-9-375

min. Both extracts were combined and evaporated at reduce pressure in a vacuum centrifuge at ambient temperature. The remaining residue was redissolved in 400 μL methanol/water, 3/7 (v/v).

Data set M1: Capillary LC-ESI-QTOF-MS

1 μl of the extract was separated using an Ultimate capillary LC system (Dionex) on a modified C_{18} column (GROMSIL ODS 4 HE, 0.3 × 150 mm, particle size 3 μm, Alltech-Grom) applying a binary acetonitrile-water gradient at a flow rate of 5 $\mu Lmin^{-1}$. Eluted compounds were detected from m/z 75 to 1000 by an API QSTAR Pulsar i (Applied Biosystems/MDS Sciex) equipped with an Ionspray electrospray ion source in positive ion mode. Accumulation time was 2 s. Mass resolution for $[M + H]^+$ of a calibration peptide was RFWHM (resolution full width at half maximum) = 8500 at 829 m/z.

Data set M2: LC-ESI-QTOF-MS.

10 μl of the *A. thaliana* extract were separated using a Agilent 1100 Series HPLC system on a modified C_{18} column (Atlantis dC18, 2.1 × 150 mm, particle size 3 μm, Waters) applying the same binary gradient as above at a flow rate of 200 $\mu Lmin^{-1}$. Eluted compounds were detected from m/z 100–1000 by a MicrOTOF-Q (Bruker Daltonics) equipped with an Apollo II electrospray ion source in positive ion mode. Accumulation time was 1.5 s. Mass resolution for $[M + H]^+$ of a calibration peptide was RFWHM = 14000 at 829 m/z.

2.2.2 Data extraction

All data were exported in centroid mode by the converter software from Applied Biosystems and Bruker, respectively. The feature finding was done using XCMS [26] using the parameters *method* = "*centWave*", *peakwidth* = *c*(20, 50), *snthresh* = 5, *ppm* = 120 for the data set M1 and *ppm* = 30 for the data set M2, respectively. The number of features for each file is available as additional File 3.

2.2.3 Ground truth

In contrast to the proteomics data sets, usage of MS/MS information and SEQUEST annotation are not applicable. Compound spectra libraries exist for GC/EI-MS, but no extensive set of reference spectra is available for LC-ESI-MS. However, a relative annotation of "anonymous" substances is sufficient for the purpose of our alignment evaluation.

For soft ionization methods like LC-ESI-MS, different adducts (e.g. $[M + K]^+$, $[M + Na]^+$) and fragments (e.g., $[M - C_3H_9N]^+$, $[M + H - H_20]^+$) occur. Using these known mass differences and verification techniques such as peak shape comparison by correlation analysis, features which originate from the same substance can be grouped together as annotated feature groups. Even if the substances are unknown, their spectra can be reconstructed in this way. Details are described in [35].

We used features that do not only have the same retention time but also show high correlation (Pearson correlation coefficient > 0.9) in their chromatographic peak shapes to create annotated feature groups. The correlation verified feature annotations were created using the R-Package *ESI*, which can be downloaded from http://msbi.ipb-halle.de/msbi/esi.

Only those highly confident feature groups that were reproducible over at least four files and show limited deviation across the files (data set M1: ΔRT = 90 s, Δm/z = 0:02 Th, data set M2: ΔRT = 20 s, Δm/z = 0:01 Th) were used to create a verified alignment of these feature groups. Subsequently, the aligned feature groups were split up into their consensus features, which form the alignment ground truth. The number of features for each file and the size of the ground truth for each alignment are available in the additional File 3.

2.3 Computation of alignments

In the following subsections we will shortly describe the general approach of the six alignment methods as well as their most relevant parameters. Furthermore, we present our procedure to import the input feature lists into the different tools. Each program provides a consensus map in a proprietary file format which was parsed for the evaluation.

2.3.1 OpenMS

The open source framework OpenMS [36] offers a multiple LC-MS map alignment algorithm [28] for raw as well as feature maps.

The maps are aligned in a star-wise manner with the most complete map as the reference map. The correction of the warp in RT and m/z and the determination of a consensus map are performed in two steps called *superposition phase* and *consensus phase*. This modularization allows for the implementation of a general algorithm that either aligns multiple raw maps using just the superposition phase, or aligns multiple feature maps applying both phases. In the superposition phase the parameters of a suitable affine transformation are determined using a general paradigm for point pattern matching algorithms called *pose clustering*. The optimal transformation, which is defined as the transformation that maps as many elements of one map as possible close to elements in the other map, is determined by a so-called *voting schema*. The pose-clustering algorithm considers the different measuring accuracies of the RT and m/z dimension as well as the intensity information of the LC-MS map elements. After the estimation of the initial transformation by the pose-clustering approach, landmarks are searched in the two maps. These landmarks are used for the refinement of the affine warp by a linear regression step. The following consensus phase is based on a nearest neighbors search and determines the final consensus map given the dewarped feature maps. The OpenMS multiple feature map alignment algorithm is implemented in the TOPP tool MapAlignment. The most important parameter for the user are $precision_{RT}$, $precision_{m/z}$ and *mz_bucket_size*. The parameter *mz_bucket_size* is a parameter for the superposition phase. It restricts the computation of all possible transformations by mapping only features in both maps that have similar m/z positions. Whereas, $precision_{RT}$ and $precision_{m/z}$ are parameters of the consensus phase that define the maximal distance of corresponding features for the grouping process. The metabolomics feature lists were converted into the featureXML input format by the FileConverter TOPP tool.

2.3.2 msInspect

The multiple feature map alignment algorithm presented in [25] is part of the open source LC-MS analysis platform *msInspect*. The software package is written in the platform independent language Java and is freely available at http://proteomics.fhcrc.org.

Before a consensus map, the so-called *peptide array*, is determined the algorithm corrects the non-linear distortions of the RT dimension of all maps in a star-wise manner with respect to a certain reference map. It is assumed that the distortion in RT is explained by a global linear trend plus a remaining non-linear component. In the first step, the linear trend is estimated using the most intense features with similar m/z positions. This initial model of the RT transformation is used to iteratively determine a non-linear transformation using smoothing-spline regression methods from the previous model. After dewarping all maps, a global alignment is performed by applying divisive clustering, with user-supplied tolerances in RT and m/z of assigned features. The algorithm optionally offers the automatic choice of the optimal RT and m/z tolerances using the quality of clustering. The quality of the alignment is defined by the number of clusters that include at most one feature from each map.

msInspect uses various tsv (tab-separated values) files for input and output. We implemented utilities for converting data from our feature map format featureXML into the msInspect tsv format and to extract the resulting consensus map from the msInspect output files. The alignment algorithm of msInspect provides the setting of two parameters: scanWindow, which is the maximum size of a consensus feature in time space, and massWindow, the maximum size of a consensus feature in mass space. The option – optimize is used to determine the best choices for the two parameters with respect to the number of *perfect matches*, which contain exactly one feature of each map. We used the parameters suggested by the optimizer but also different parameters to evaluate msInspect's alignment algorithm.

2.3.3 SpecArray

Li et al. [22] developed a multiple feature map alignment algorithm embedded in the open source software suite *SpecArray* http://tools.proteomecenter.org.

The proposed algorithm computes all pairwise alignments and combines them to a final consensus map. To correct the distortion in RT a retention time calibration curve (RTCC) is iteratively computed for each pairwise alignment by pairing features with similar m/z values to construct an original feature pairs set. The RTCC curve is estimated by minimizing the root mean square distance of the features' RT positions to the monotonic function. Pairs with a small pairing score are removed and the reduced set of feature pairs is again used to estimate a RTCC. The two steps are repeated until only the pairs with a high pairing score remain and each feature in one map is paired with at most one feature in the other map. The final RTCC curve and the distance of peptides in m/z is used to select likely and unique feature pairs from the original set of feature pairs. The combination of all pairwise alignments yields the final consensus map, or the so-called *super list*. The parameters for the alignment algorithm are hard-coded and cannot be changed by the user. Calculating all pairwise alignments results in a high runtime and makes the algorithm inapplicable for the comparison of a large number of feature maps. SpecArray provides two tools for the alignment of feature maps. Whereas, PepMatch performs the actual alignment step, PepArray can be used for the postprocessing and filtering of the consensus map. We avoid the filtering step and use the unprocessed final consensus map for evaluation purposes.

We implemented software to convert our feature map format featureXML into the SpecArray's binary feature format pepBof. Furthermore, we forced SpecArray to directly export our consensus format by the addition of some lines of code to the sources of PepMatch.

2.3.4 XAlign

Zhang et al. [23] propose a stand-alone tool, called *XAlign*, for the alignment of multiple feature maps. The Xalign software for Windows is available upon request from the author.

XAlign computes in a first step a so-called gross-alignment, where the algorithm corrects a systematic shift in

DOI:10.1186/1471-2105-9-375

RT. In the second step, a final consensus map, the so-called *micro alignment*, is determined. The gross-alignment algorithm aligns multiple maps in a star-wise manner, where the reference map is chosen as follows: for all predefined RT and m/z windows the most intense features of each map are determined. If a window contains features from all maps, the features are called significant and their intensity weighted average mean RT position is calculated. The map with the minimal difference of all its significant features to the averaged RT positions is chosen as the reference map. Afterwards, all other maps are dewarped with respect to the reference by estimating a linear function that minimizes the mean absolute deviation of the RT positions of significant features. In the micro-alignment phase features yielding a high correlation coefficient are successively grouped together and establish the final consensus map. XAlign [23] is designed as a component of a data analysis pipeline for protein biomarker discovery. The stand-alone executable runs in the Windows command line. It reads tab-separated feature lists and generates several output files including the alignment table and peak statistics.

2.3.5 XCMS

The XCMS package presented in [26] is part of Bioconductor [37], a larger open source software project for bioinformatics written in the platform-independent programming language R. All Bioconductor packages can be obtained from http://www.bioconductor.org. XCMS is designed for both LC/MS and GC/MS data. It includes functionality for visualization, feature detection, non-linear retention time alignment and statistical methods to discover differentially expressed metabolites. We modified XCMS to skip the feature detection step and imported the featurelists directly from feature map format featureXML. XCMS' feature-matching algorithm makes use of fixed-interval bins (e.g., 0.1 Th wide) to match features in the mass domain. After this initial binning of features by mass, groups of features with different retention time in each bin are resolved. Kernel density estimation is used to calculate the distribution of features in chromatographic time and subsequently boundaries of regions where many features have similar retention times are identified.

XCMS supports an optional retention time correction step where "well-behaved" groups of features are used to calculate a nonlinear retention time deviation for each sample. The resulting deviation profiles are then used to correct the retention times of the original samples. The matching and retention time correction procedure can be repeated for an increasingly precise alignment. However, we observed that it is hard to predict whether the retention time correction will actually lead to a better consensus map and depends on the input. Therefore, we decided to report results both without and with the optional retention time correction step.

2.3.6 MZmine

The MZmine toolbox [38] for processing and visualization of LC/MS data is used via a graphical user interface. Due to its implementation in Java it is platform independent. MZmine is open source and can be downloaded from http://mzmine.sourceforge.net. We modified MZmine to skip the feature detection step and import featurelists instead.

MZmine's alignment approach does not estimate any dewarping transformations. The toolbox currently implements a simple alignment method utilizing a so-called *master feature list*, where features from each map are aligned against the master list. A score function is used to compute the similarity of a feature and a row of the master list, which represents the current consensus feature. If the score obtained between the best matching master list row and a feature is "good enough" (both the m/z and retention time difference are within tolerances) the feature is assigned to that row, otherwise it is appended to the master list. MZmine offers two alignment algorithms, "slow aligner" and "fast aligner", which differ in the implementation of the score function. We found only minimal differences in the alignment quality of both algorithms so we used the "fast aligner" due to the better runtime.

2.3.7 Parameters

We performed extensive test runs to optimize the parameters controlling the tolerance in RT and m/z for our test data. Using the known deviations of the data as a starting point we varied the parameters of each tool within reasonable ranges. The parameters which yielded the best results on the first experiment of each data set were choosen. The final settings are shown in Table 2.

2.4 Evaluation

The performance of an information retrieval system can be assessed using the *precision* and *recall* values. Our evaluation of the map alignment problem will follow these lines. As stated in the beginning of this section, the correction of retention times is a very important aspect of the LC-MS map alignment problem, and there is a trade-off between the smoothness of the warping function and the remaining distance among matched features. But at the end, the purpose of warping the retention times is to find groups of corresponding features that are reported as consensus features, which is why our analysis focuses on this aspect of the map alignment problem. That is, we will evaluate the quality of the *consensus map* rather than the *warping function*, because we consider the latter an intermediate step for the map alignment problem. Given a "query" feature in one map, the consensus map can serve

DOI:10.1186/1471-2105-9-375

Table 2: Alignment parameters

Tool	Parameter	Metabolomics Data		Metabolomics Data	
		Data Set P1	Data Set P2	Data Set M1	Data Set M2
msInspect	massWindow	1.5	1.5	0.1	0.05
	scanWindow	250	300	250	300
MZmine	m/z tolerance size	1.5	1.5	0.03	0.025
	RT tolerance size (absolute)	150	300	50	30
OpenMS	m/z bucket	0.5	0.5	0.1	0.01
	precision m/z	2	2	0.1	0.1
	precision RT	150	300	100	100
SpecArray	(hard coded parameters)	-	-	-	-
XAlign	m/z variation	2	2	0.04	0.03
	retention time variation	3	3	0.5	0.5
XCMS	mzwid	2.5	2.5	0.15	0.05
	bw	40	80	30	30
	retcor method	loess	linear	loess	loess
	span	0.75	-	0.75	0.75

to retrieve related "items" in the other maps. Consensus features are simply taken as sets of features; assigning an appropriate average position to these sets etc. is another problem and not addressed here.

In the frequentist interpretation, *precision* is the probability that a found item is relevant, whereas *recall* is the probability that a relevant item is found. In the special case of pairwise map alignment, the relevant items are matching features; an item is either found or not. In order to extend these concepts to the multiple map alignment problem, we need to deal with consensus features that do not contain features from all maps, as well as consensus features reported by tools, that overlap but are not identical to the ground truth.

Let us denote the consensus features in the ground truth by gt_i, where the index i runs from 1 to N. Likewise, the consensus features from the tool will be denoted by tool_j, for index $j = 1,...,M$. We consider the set of consensus features from the tool that contain at least two features (so that they can be used to retrieve items) and intersect with a given consensus feature from the ground truth. Thus, for each index i let us denote by M_i the set of all indices j such that $|\mathrm{tool}_j| \geq 2$ and $|\mathrm{gt}_i \cap \mathrm{tool}_j| > 0$. Now we can look at the cardinality of this index set, $|M_i|$. In some way, this is the number of "parts" into which consensus feature gt_i from the ground truth has been "split up" by the tool. But we can also look at the union of these consensus features, $\widetilde{\mathrm{tool}}_i := \bigcup_{j \in M_i} \mathrm{tool}_j$. Then $\widetilde{\mathrm{tool}}_i$ is the set of all items that can be retrieved if the query belongs to gt_i.

Therefore, following the classical definition of precision and recall, we define the *alignment precision*:

$$\mathrm{Precision}_{\mathrm{Align}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathrm{gt}_i \cap \widetilde{\mathrm{tool}}_i|}{|\widetilde{\mathrm{tool}}_i|}$$

and the *alignment recall*:

$$\mathrm{Recall}_{\mathrm{Align}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathrm{gt}_i \cap \widetilde{\mathrm{tool}}_i|}{|M_i| \cdot |\mathrm{gt}_i|}.$$

The factor $|M_i|$ in the denominator serves as a penalty for breaking up a consensus feature from the ground truth. Note that in the case of pairwise alignments, the summands in these definitions are either zero or one, and our definitions become equivalent to the classic precision and recall. Thus, their names are justified as generalizations. A perfect alignment will have both measures equal to one. False positives (erroneously grouped features) lower the alignment precision; false negatives (erroneously unaligned features) lower the alignment recall.

An example is shown and calculated in Figure 1.

An R script was written for the automated computation of the recall and precision values. The runtimes were measured as wall-clock time including all file input/output while no other programs were running. All measurements were done on an AMD Athlon 64 X2 Dual Core Processor 4800+ with 2 GB RAM running Linux (Ubuntu 6.06). Since XAlign does not run under Linux, we evaluated it under Windows XP running in a virtual machine using VMWare Workstation 5.5.3 on the same computer (native Windows XP should typically be 10–20% faster). The reported wall-clock runtimes are cumulative over all runs per data set.

DOI:10.1186/1471-2105-9-375

3 Results

The proteomics (P1, P2) and metabolomics (M1, M2) data sets pose different challenges for the alignment tools. Each tool has to correct the global trend of the retention time variation resulting from the flow rate variability from experiment to experiment. Furthermore, it has to overcome local distortions resulting from e. g. gradient noise or temperature changes and assign corresponding features across the different maps.

To illustrate the ground truth established for these data sets, we plot the retention time deviation versus the retention time. Figure 2 shows a significant shift between corresponding features in fraction 100 of P1_1 and P1_2, but almost no difference in scale. Figure 3 shows that fractions 20 of P2_2 and P2_3 are slightly scaled with respect to P2_1, but apart from that the retention times are in fact better correlated. While an average absolute retention time deviation of 57 s can be observed in the ground truth maps of P1, the average absolute retention time deviation for P2 is 131 s (before retention time correction). The retention time deviation plots for each single fraction of the data sets P1 and P2 are available as additional File 1.

The metabolomics data sets M1 and M2 contain a larger number of experiments (24 resp. 44). Therefore, we use box-whiskers plots for visualization. Figures 4 and 6 show that variation is higher in M1 than in M2, but still much smaller than in P1 or P2. The average absolute retention time deviation for the ground truth of the metabolomics data sets M1 and M2 is 5.4 s and 2.7 s respectively. Presumably, "large" deviations are the reason for most of the alignment errors. Loess regression curves for three randomly chosen files show that the global trends are not as pronounced as the local variation, see Figures 5 and 7.

Both proteomics data sets challenge the ability of the alignment tools to correct strong retention time variations. Especially the data of P2, which were measured during several weeks and show huge retention time deviations of around 13 minutes, confront the dewarping step of the tools with a serious problem. However, the highly complex metabolomics data sets reveal the capability of the alignment tools to assign the correct features across multiple maps. The maximum retention time deviations of feature maps in M1 and M2 are only 90 s and 20 s respectively, without an obvious global trend. The warps are mainly affected by local non-linear distortions of retention times similar to uncorrelated statistical noise. In M1 the high density of the feature maps complicates the determination of the correct consensus features. However, M2 challenges the grouping step of the tools by its large number of input maps.

Our evaluation of the tools' performance is based on alignment recall and alignment precision as defined in

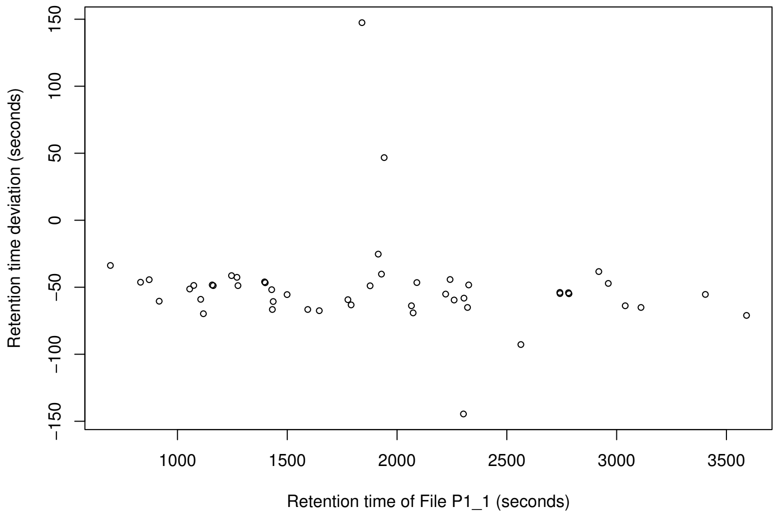


Figure 2
Retention time deviations of data set P1. Exemplary plot of retention time deviations in the ground truth of data set P1. Retention time deviation of File P1_2 is plotted against retention time of File P1_1 (fraction 100).

DOI:10.1186/1471-2105-9-375

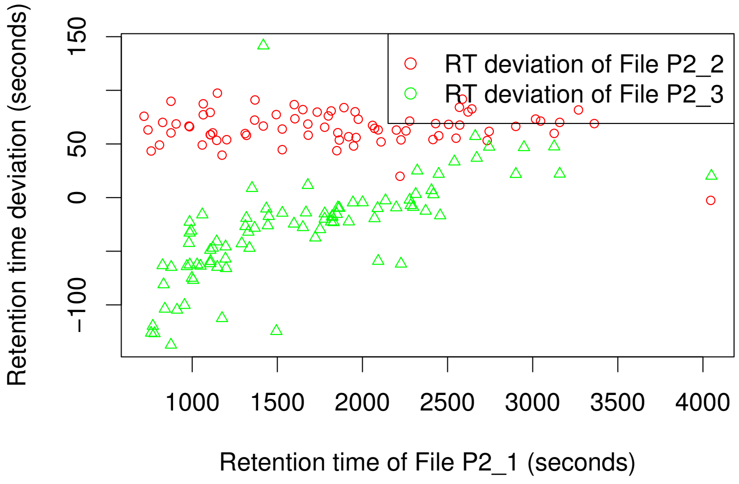


Figure 3
Retention time deviations of data set P2. Exemplary plot of retention time deviations in the ground truth of data set P2. Retention time deviations of File P2_2 and P2_3 are plotted against retention time of File P2_1 (fraction 20).

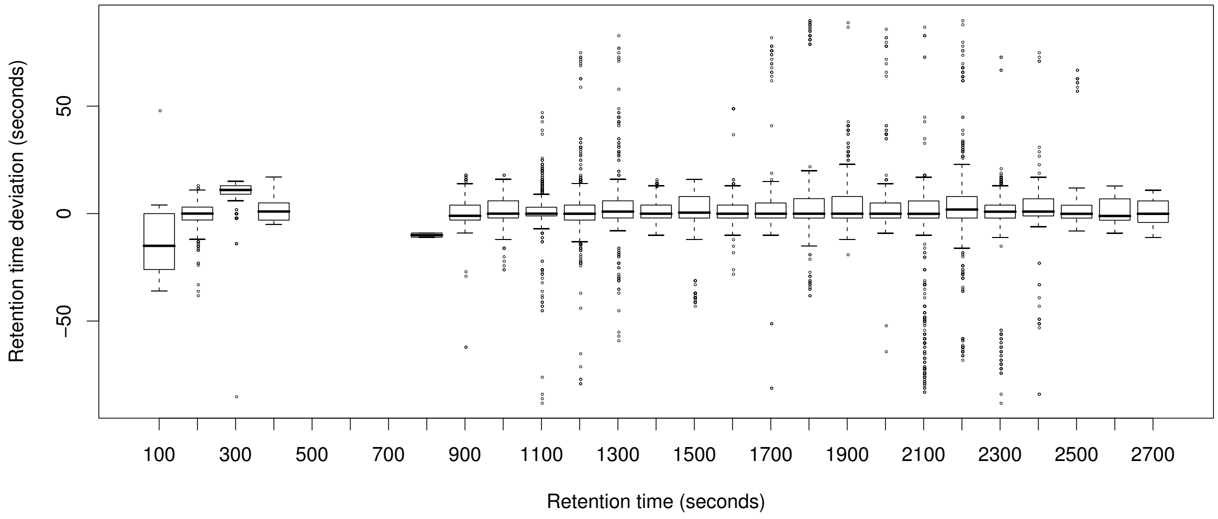


Figure 4
Retention time deviations of data set M1. Box-whiskers-plot showing the retention time deviations in the ground truth of data set M1.

DOI:10.1186/1471-2105-9-375

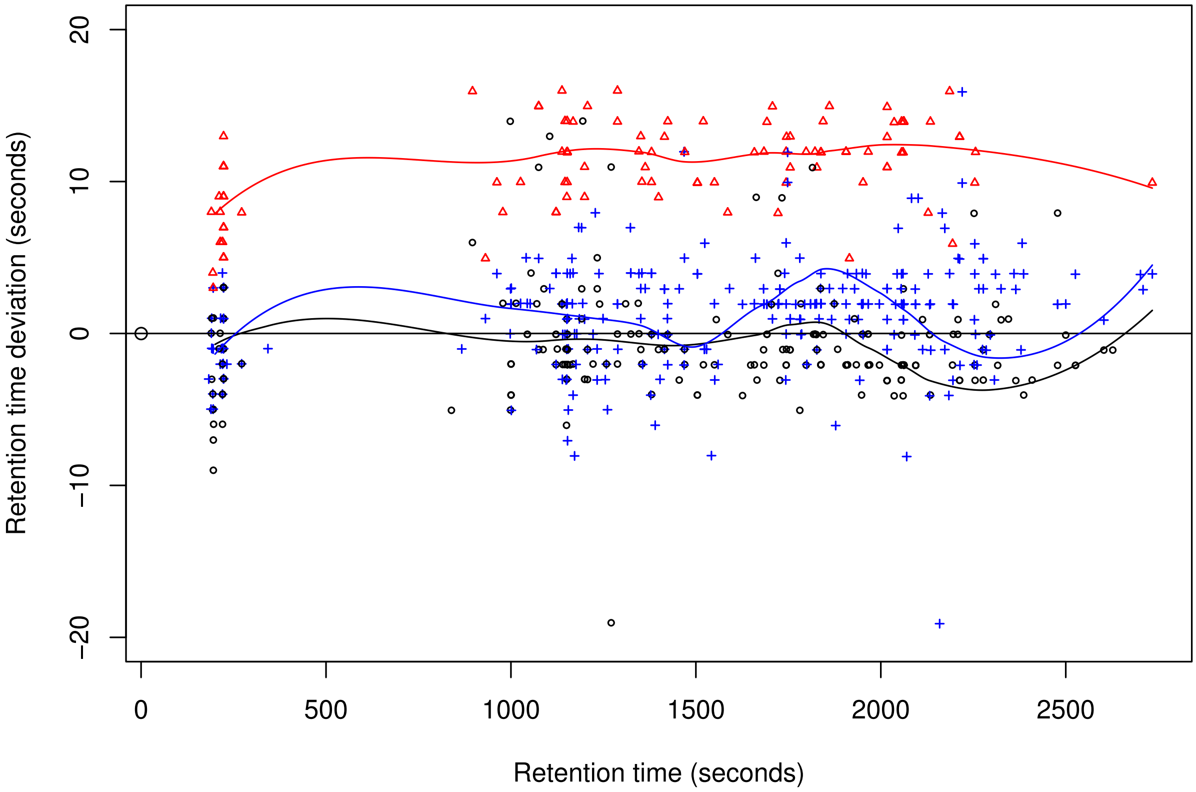


Figure 5
Retention time deviations in the ground truth of three randomly chosen files from data set M1. Loess regression curves were superimposed for better visualization.

Section 2.4, as well as their running times. Memory consumption was not a critical resource. Since the chromatographic separation steps for the metabolomics and the proteomics data sets resemble each other, we decided to test all tools on all data sets, even though most of them were originally designed for either metabolomics or proteomics data. Figure 8 shows a summary of the results on the different data sets.

The results for the proteomics data sets P1 and P2 are shown in Tables 3, 4, and 5. We found that OpenMS performs best on P1, closely followed by XAlign, XCMS and MZmine. All four tools achieved high recall as well as high precision values on this data set. However, SpecArray and msInspect result in slightly worse recall and precision values. The evaluation on the second proteomics data set shows a similar trend, despite the overall recall and precision of all tools is reduced on this more demanding data set. OpenMS again performs best on most fractions of P2 and is closely followed by XAlign, XCMS and MZmine. SpecArray and msInspect are closely ranked after these four tools. All programs completed within two minutes on the relatively small data sets of P1 and P2.

The results for the metabolomics data sets M1 and M2 are shown in Tables 6 and 7. Here, XCMS performs best on both data sets, and MZmine does equally well on M2, with OpenMS and XAlign not far behind. Alignment recall is much more discriminative than alignment precision, due to the penalty for breaking up a consensus feature from the ground truth. The running times were significantly different on these relatively large data sets, which

DOI:10.1186/1471-2105-9-375

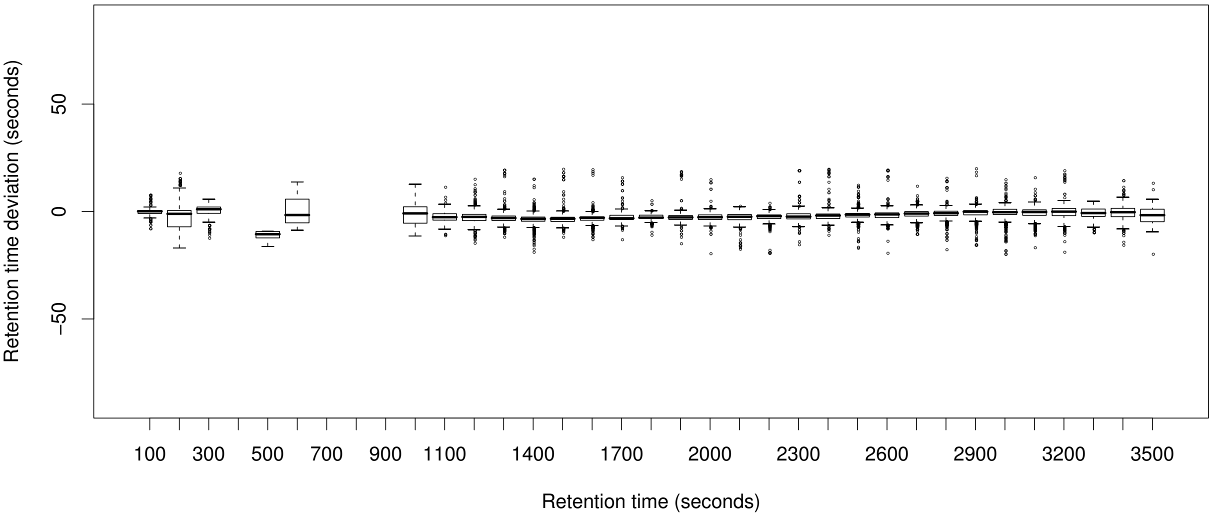


Figure 6
Retention time deviations of data set M2. Box-whiskers-plot showing the retention time deviations in the ground truth of data set M2.

contain more than 200 000 features in 24 (M1) respectively 44 (M2) feature maps. The alignments using SpecArray were canceled after 24 hours with an estimated remaining runtime of more than two weeks. SpecArray performs all pairwise map alignments and seems inapplicable to this kind of metabolomics data. In contrast, XCMS computes the alignment of the M1 and M2 in less than seven minutes. OpenMS requires 13 minutes for the determination of the metabolomics consensus maps. MZmine and XAlign both result in a high runtime of more than one hour for the quite complex metabolomics data sets.

msInspect has a runtime of only half an hour, but with very low recall and precision values. We were unable to obtain good results on the data sets M1 and M2 using msInspect with parameters suggested by the optimizer as well as different values chosen manually. In most cases the automatic choice of "optimized" parameters did not lead to better alignment results than manually chosen "good" values. Furthermore, we observed that a different order of the input files leads to different results with msInspect. Placing the feature list with the highest number of features on top of the list seems to give the best results.

Another outcome of our evaluation is that it is hard to predict whether XCMS map alignment should be used with or without retention time correction, and that the characteristics of the correction need to be checked.

4 Discussion and conclusion

The automatic alignment of LC-MS data sets is an important step in most analysis pipelines for metabolomics and proteomics high-throughput experiments. Algorithms that perform this task efficiently and accurately have a large impact not only on basic research in biology, but also on more applied questions such as biomarker discovery and drug research in general. Due to the importance of this step and the multitude of different approaches a meaningful standard data set and a sophisticated scoring method are needed. We offer both proteomics (P1, P2) and metabolomics (M1, M2) benchmark data sets, as well as proper quality measures ($Precision_{Align}$, $Recall_{Align}$) and an evaluation procedure. On the basis of these data sets we have assessed the performance of six freely available alignment tools.

Perhaps surprisingly, we observed that in many cases the largest part of the *systematic* deviation of retention time in our data sets could have been corrected by a simple shift without any further scaling or non-linear warping at all. The remaining error is very similar to statistical noise, not correlated among neighboring consensus features, and further scan-wise corrections of retention time will face the risk of overfitting. This suggests that the choice of the

DOI:10.1186/1471-2105-9-375

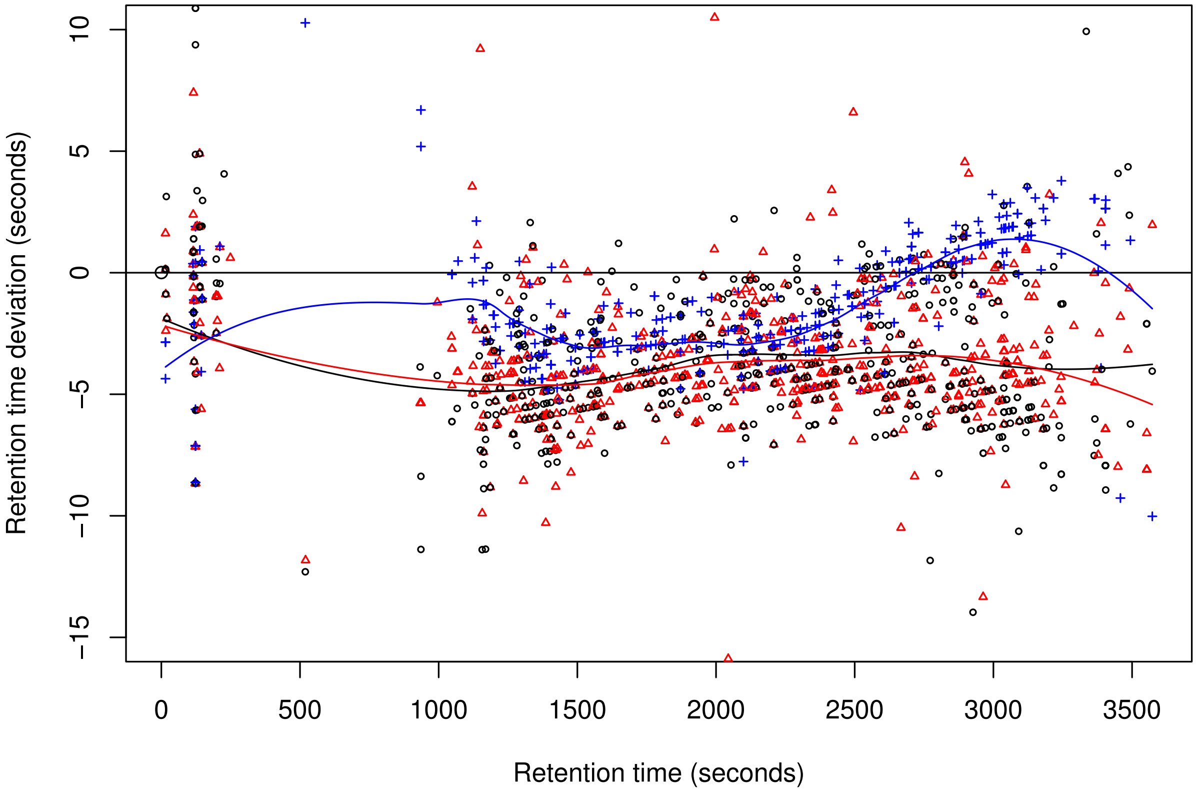


Figure 7

Retention time deviations in the ground truth of three randomly chosen files from data set M2. Loess regression curves were superimposed for better visualization.

warping function is less important than the following clustering step (i. e., the correction of the retention times of the individual features), as this will establish the actual consensus features.

The implemented methods are based on a variety of algorithmic principles with complementary strengths and weaknesses [13]. Combining them into "hybrid" approaches seems to be a promising direction for future research. However, such a project requires a long-term commitment and, if possible, several software developers are necessary. We expect to see a consolidation in the area in the future with a tendency toward open source frameworks such as Bioconductor or OpenMS.

Recently, the Association of Biomolecular Resource Facilities (ABRF) has organized a collaborative study focusing on evaluating the ability of proteomics laboratories to determine the identities of a complex mixture of proteins present in a single mass spectral *data set*, as a follow-up to an earlier study in which the actual *samples* were distributed [39]. This indicates the growing attention paid to data processing versus "wet-lab" techniques in the proteomics field. Similar competitions should be organized for all the other aspects of a typical LC-MS data processing pipeline, including the LC-MS map alignment problem. The experience from the plasma proteome project [40] has shown that it is difficult to assess the performance if many aspects change simultaneously.

We would like to encourage other MS software developers (including commercial vendors) to use our benchmark data for evaluation. Further benchmarks are also highly welcome, e. g. identical samples run at different laborato-

DOI:10.1186/1471-2105-9-375

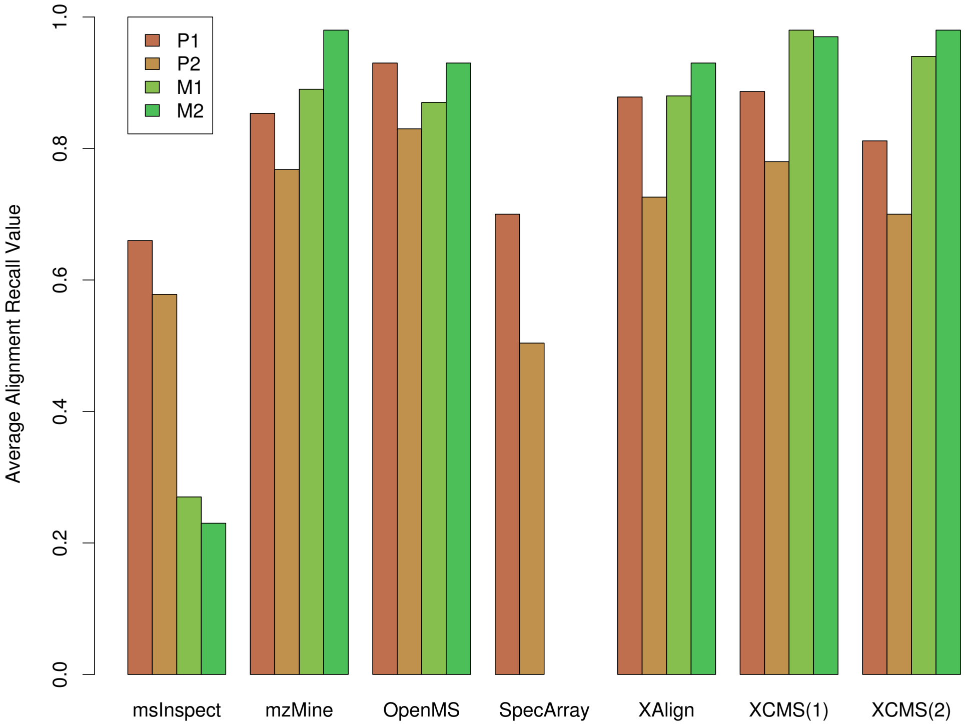


Figure 8
Result Overview. Average alignment recall values for the results on the four data sets P1, P2, M1 and M2. XCMS was evaluated without(1) and with(2) application of retention time correction. The detailed results are shown in Tables 3, 4 and 6.

Table 3: Alignment recall and precision results for the proteomics data set P1.

	msInspect	MZmine	OpenMS	SpecArray	XAlign	XCMS without retention time	XCMS with correction
fraction 00							
$\text{Recall}_{\text{Align}}$	0.52	0.75	**0.86**	0.61	0.82	0.72	0.62
$\text{Precision}_{\text{Align}}$	0.38	0.81	**0.86**	0.61	0.82	0.54	0.58
fraction 20							
$\text{Recall}_{\text{Align}}$	0.56	0.87	**0.92**	0.62	0.85	0.88	0.81
$\text{Precision}_{\text{Align}}$	0.45	0.88	**0.92**	0.62	0.85	0.84	0.80
fraction 40							
$\text{Recall}_{\text{Align}}$	0.63	0.87	**0.94**	0.75	0.87	0.92	0.81
$\text{Precision}_{\text{Align}}$	0.48	0.90	**0.94**	0.75	0.87	0.85	0.80
fraction 60							
$\text{Recall}_{\text{Align}}$	0.73	0.79	**0.96**	0.71	0.87	0.91	0.78
$\text{Precision}_{\text{Align}}$	0.54	0.84	**0.96**	0.71	0.87	0.80	0.75
fraction 80							
$\text{Recall}_{\text{Align}}$	0.70	0.92	**0.96**	0.74	0.90	0.94	0.89
$\text{Precision}_{\text{Align}}$	0.57	0.94	**0.96**	0.74	0.90	0.88	0.88
fraction 100							
$\text{Recall}_{\text{Align}}$	0.82	0.92	0.94	0.77	**0.96**	0.95	**0.96**
$\text{Precision}_{\text{Align}}$	0.56	0.94	0.94	0.77	**0.96**	0.89	**0.96**

DOI:10.1186/1471-2105-9-375

Table 4: Alignment recall and precision results for the proteomics data set P2.

	msInspect	MZmine	OpenMS	SpecArray	XAlign	XCMS	
						without retention time	with correction
fraction 00							
Recall$_{\text{Align}}$	0.23	**0.77**	**0.77**	0.07	0.65	0.70	0.58
Precision$_{\text{Align}}$	0.07	0.6	**0.65**	0.05	0.49	0.31	0.44
fraction 20							
Recall$_{\text{Align}}$	0.67	0.87	**0.92**	0.57	0.84	0.89	0.86
Precision$_{\text{Align}}$	0.24	0.71	**0.77**	0.42	0.70	0.55	0.66
fraction 40							
Recall$_{\text{Align}}$	0.44	**0.79**	0.76	0.60	0.71	0.72	0.72
Precision$_{\text{Align}}$	0.26	**0.76**	0.74	0.41	0.69	0.56	0.69
fraction 80							
Recall$_{\text{Align}}$	0.73	0.61	**0.80**	0.65	0.58	0.64	0.49
Precision$_{\text{Align}}$	0.34	0.56	**0.70**	0.44	0.56	0.50	0.45
fraction 100							
Recall$_{\text{Align}}$	0.82	0.80	0.90	0.63	0.85	**0.95**	0.85
Precision$_{\text{Align}}$	0.39	0.65	**0.75**	0.44	0.69	0.65	0.69

Table 5: Wall-clock runtime for the proteomics data sets P1 and P2 in minutes.

Data set	msInspect	MZmine	OpenMS	SpecArray	XAlign	XCMS	
						without retention time	with correction
P1	1	0.67	1.6	1.85	1.15	0.53	0.90
P2	0.75	1.22	0.36	5.19	0.29	0.33	0.49
Total	1.75	1.89	1.96	7.04	1.44	0.86	1.39

Table 6: Alignment recall and precision results for the metabolomics data sets M1 and M2

Data set	msInspect	MZmine	OpenMS	SpecArray	XAlign	XCMS	
						without retention time	with correction
M1							
Recall$_{\text{Align}}$	0.27	0.89	0.87	-	0.88	**0.98**	0.94
Precision$_{\text{Align}}$	0.46	**0.74**	0.69	-	0.70	0.60	0.70
M2							
Recall$_{\text{Align}}$	0.23	**0.98**	0.93	-	0.93	0.97	**0.98**
Precision$_{\text{Align}}$	0.47	**0.84**	0.79	-	0.79	0.58	0.78

Table 7: Wall-clock runtime for the metabolomics data sets M1 and M2 in minutes

Data set	msInspect	MZmine	OpenMS	SpecArray	XAlign	XCMS	
						without retention time	with correction
M1	12	20	4.4	-	51	0.9	1.4
M2	24	44	8.7	-	35	5.5	5.8
Total	36	64	13.1	-	86	6.4	7.2

DOI:10.1186/1471-2105-9-375

ries under "identical" conditions, or even on MS equipment from different vendors. We will collect future results and contributions upon request on http://msbi.ipb-halle.de/msbi/caap.

5 Authors' contributions

EL performed data extraction and established ground truth for the proteomics data sets, with methodology developed together with CG. RT and SN performed data extraction and established ground truth for the metabolomics data sets. EL performed parameter tuning and test runs for msInspect, SpecArray and OpenMS. RT performed parameter tuning and test runs for MZmine, XAlign and XCMS. All authors developed and decided upon the evaluation criteria. All authors read and approved the final manuscript.

Additional material

Additional file 1

The proteomics and metabolomics ground truth data sets as well as the evaluation script are available at http://msbi.ipb-halle.de/msbi/caap.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-375-S1.pdf]

6 Acknowledgements

We thank Christoph Böttcher for providing the data sets M1 and M2. EL performed the major part of her work during her PhD period at the Free University Berlin funded by the Berlin Center for Genome Based Bioinformatics (BCB). EL is now supported by Cancer Research UK. We thank Anja Kasseckert for proof-reading.

References

1. Colinge J, Bennett KL: **Introduction to Computational Proteomics.** *PLoS Computational Biology* 2007, **3(7):**e114.
2. Dunn WB: **Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes.** *Physical Biology* 2008, **5:**011001 [http://stacks.iop.org/1478-3975/5/011001]. (24pp)
3. Ong SE, Mann M: **Mass spectrometry-based proteomics turns quantitative.** *Nat Chem Biol* 2005, **1(5):**252-262.
4. Ong SE, Foster LJ, Mann M: **Mass spectrometric-based approaches in quantitative proteomics.** *Methods (San Diego, Calif.)* 2003, **29(2):**124-130.
5. Gröpl C, Lange E, Reinert K, Kohlbacher O, Sturm M, Huber CG, Mayr B, Klein C: **Algorithms for the automated absolute quantication of diagnostic markers in complex proteomics samples.** In *Procceedings of CompLife 2005, Lecture Notes in Bioinformatics* Edited by: Berthold M. Springer, Heidelberg; 2005:151-163.
6. Bisle B, Schmidt A, Scheibe B, Klein C, Tebbe A, Kellermann J, Siedler F, Pfeiffer F, Lottspeich F, Oesterhelt D: **Quantitative Profiling of the Membrane Proteome in a Halophilic Archaeon.** *Mol Cell Proteomics* 2006, **5(9):**1543-1558.
7. Niittylä T, Fuglsang AT, Palmgren MG, Frommer WB, Schulze WX: **Temporal analysis of sucrose-induced phosphorylation changes in plasma membrane proteins of Arabidopsis.** *Mol Cell Proteomics* 2007, **6(10):**1711-1726.
8. Vissers JPC, Langridge JI, Aerts JMFG: **Analysis and Quantification of Diagnostic Serum Markers and Protein Signatures for Gaucher Disease.** *Mol Cell Proteomics* 2007, **6(5):**755-766.
9. Catchpole GS, Beckmann M, Enot DP, Mondhe M, Zywicki B, Taylor J, Hardy N, Smith A, King RD, Kell DB, Fiehn O, Draper J: **Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops.** *Proc Natl Acad Sci U S A* 2005, **102(40):**14458-14462.
10. Böttcher C, v Roepenack-Lahaye E, Schmidt J, Schmotz C, Neumann S, Scheel D, Clemens S: **Metabolome Analysis of Biosynthetic Mutants Reveals Diversity of Metabolic Changes and Allows Identification of a Large Number of New Compounds in Arabidopsis thaliana.** *Plant Physiol* 2008, **147(4):**2107-2120.
11. Snyder LR, Dolan JW: *High-Performance Gradient Elution: The Practical Application of the Linear-Solvent-Strength Model* Wiley; 2007.
12. Sakoe H, Chiba S: **Dynamic programming algorithm optimization for spoken word recognition.** *IEEE Trans. Acoustics, Speech and Signal Processing* 1976, **26(11):**43-49.
13. Vandenbogaert M, Li-Thiao-Té S, Kaltenbach HM, Zhang R, Aittokallio T, Schwikowski B: **Alignment of LC-MS images, with applications to biomarker discovery and protein identification.** *Proteomics* 2008, **8(4):**650-672.
14. Bro R: **Parafac: tutorial and applications.** *Chemom Intell Lab Syst* 1997, **33:**149-171.
15. Bylund D, Danielsson R, Malmquist G, Markides KE: **Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography mass spectrometry data.** *J Chromatogr A* 2002, **961(2):**237-244.
16. Prakash A, Mallick P, Whiteaker J, Zhang H, Paulovich A, Flory M, Lee H, Aebersold R, Schwikowski B: **Signal Maps for Mass Spectrometry-based Comparative Proteomics.** *Molecular & cellular proteomics : MCP* 2006, **5(3):**423-432.
17. Prince J, Marcotte E: **Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping.** *Anal Chem* 2006, **78(17):**6140-6152.
18. Listgarten J, Neal RM, Roweis ST, Wong P, Emili A: **Difference detection in LC-MS data for protein biomarker discovery.** *Bioinformatics (Oxford, England)* 2007, **23(2):**e198-204.
19. Listgarten J, Emili A: **Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry.** *Molecular & cellular proteomics : MCP* 2005, **4:**419-434.
20. Radulovic D, Jelveh S, Ryu S, Hamilton T, Foss E, Mao Y, Emili A: **Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry.** *Molecular & cellular proteomics : MCP* 2004, **3(10):**984-997.
21. Katajamaa M, Miettinen J, Oresic M: **Processing methods for differential analysis of LC/MS profile data.** *BMC bioinformatics* 2005, **6:**179.
22. Li XJ, Yi EC, Kemp CJ, Zhang H, Aebersold R: **A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry.** *Molecular & cellular proteomics : MCP* 2005, **4(9):**1328-1340.
23. Zhang X, Asara J, Adamec J, Ouzzani M, Elmagarmid AK: **Data preprocessing in liquid chromatography/mass spectrometry-based proteomics.** *Bioinformatics (Oxford, England)* 2005, **21(21):**4054-4059.
24. Jaitly N, Monroe M, Petyuk V, Clauss T, Adkins J, Smith R: **Robust Algorithm for Alignment of Liquid Chromatography-Mass Spectrometry Analyses in an Accurate Mass and Time Tag Data Analysis Pipeline.** *Anal. Chem* 2006, **78(21):**7397-7409.
25. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng JK, Fang R, Lin C, Chen J, Goodlett D, Whiteaker J, Paulovich AG, McIntosh M: **A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.** *Bioinformatics (Oxford, England)* 2006, **22(15):**1902-1909.
26. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G: **XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.** *Anal Chem* 2006, **78(3):**779-787.
27. Wang P, Tang H, Fitzgibbon MP, Mcintosh M, Coram M, Zhang H, Yi E, Aebersold R: **A statistical method for chromatographic alignment of LC-MS data.** *Biostatistics (Oxford, England)* 2007, **8(2):**357-367.

DOI:10.1186/1471-2105-9-375

28. Lange E, Gröpl C, Schulz-Trieglaff O, Leinenbach A, Huber C, Reinert K: **A Geometric Approach for the Alignment of Liquid Chromatography-Mass Spectrometry Data.** *Bioinformatics* 2007, **23(13):**i273-i281.
29. America AHP, Cordewener JHG: **Comparative LC-MS: A landscape of peaks and valleys.** *Proteomics* 2008, **8(4):**731-749.
30. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A: **Critical assessment of methods of protein structure prediction – Round VII.** *Proteins* 2007, **69(Suppl 8):**3-9.
31. Irizarry RA, Wu Z, Jaffee HA: **Comparison of Affymetrix GeneChip expression measures.** *Bioinformatics* 2006, **22(7):**789-794.
32. Prince JT, Carlson MW, Lu RWP, Marcotte EM: **The need for a public proteomics repository.** *Nat Biotechnol* 2004, **22:**471-472.
33. Wang R, Prince JT, Marcotte EM: **Mass spectrometry of the M. smegmatis proteome: Protein expression levels correlate with function, operons, and codon bias.** *Genome Res* 2005, **15:**1118-1126.
34. Kohlbacher O, Reinert K, Gröpl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M: **TOPP-the OpenMS proteomics pipeline.** *Bioinformatics* 2007, **23(2):**191-197.
35. Tautenhahn R, Böttcher C, Neumann S: **Annotation of LC/ESI-MS Mass Signals.** *BIRD, Lecture Notes in Computer Science* 2007, **4414:**371-380 [http://dblp.uni-trier.de/db/conf/bird/bird2007.html#TautenhahnBN07]. Springer
36. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O: **OpenMS – An open-source framework for mass spectrometry.** *BMC bioinformatics* 2008, **9:**163 [http://www.openms.de].
37. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics.** *Genome biology* 2004, **5:**R80.
38. Katajamaa M, Miettinen J, Oresic M: **MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data.** *Bioinformatics (Oxford, England)* 2006, **22:**634-636.
39. The Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (ARGF): **iPRG2008 Study – Initial Results Presentation at ABRF2008.** *ABRF2008 Symposium, Salt Lake City, Utah* 2008 [http://abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm].
40. Omenn GS: **The HUPO Human Plasma Proteome Project.** *Expert Rev Proteomics* 2006, **3(2):**165-168.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

DOI:10.1186/1471-2105-9-375

analytical chemistry
Article
pubs.acs.org/ac

CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets

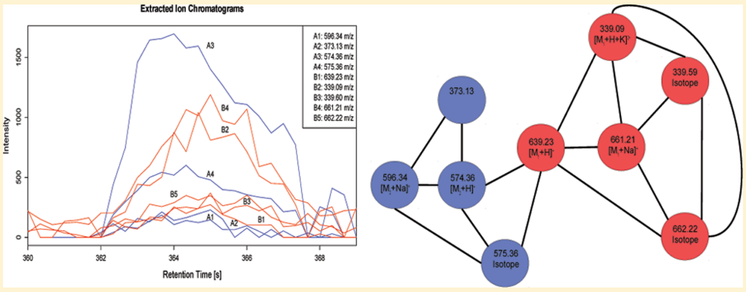
Carsten Kuhl,*[†] Ralf Tautenhahn,[‡] Christoph Böttcher,[†] Tony R. Larson,[§] and Steffen Neumann*[†]

[†]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany

[‡]Department of Chemistry and Molecular Biology, Center for Metabolomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[§]Centre for Novel Agricultural Products, Department of Biology, University of York, United Kingdom

Supporting Information



ABSTRACT: Liquid chromatography coupled to mass spectrometry is routinely used for metabolomics experiments. In contrast to the fairly routine and automated data acquisition steps, subsequent compound annotation and identification require extensive manual analysis and thus form a major bottleneck in data interpretation. Here we present CAMERA, a Bioconductor package integrating algorithms to extract compound spectra, annotate isotope and adduct peaks, and propose the accurate compound mass even in highly complex data. To evaluate the algorithms, we compared the annotation of CAMERA against a manually defined annotation for a mixture of known compounds spiked into a complex matrix at different concentrations. CAMERA successfully extracted accurate masses for 89.7% and 90.3% of the annotatable compounds in positive and negative ion modes, respectively. Furthermore, we present a novel annotation approach that combines spectral information of data acquired in opposite ion modes to further improve the annotation rate. We demonstrate the utility of CAMERA in two different, easily adoptable plant metabolomics experiments, where the application of CAMERA drastically reduced the amount of manual analysis.

Mass spectrometry (MS) is one of the dominant analysis methods for metabolomics experiments. In metabolite profiling studies, a large number of complex samples are analyzed. Typically, samples are separated prior to ionization and MS-based detection, mostly chromatographically either by gas chromatography (GC) or liquid chromatography (LC). An overview of techniques and applications was given by Dunn.[1] Depending on the sample preparation method and the analyzed organism, samples contain anywhere between dozens to thousands of compounds, e.g., the estimated number of metabolites in *Escherichia coli*[2] is just above 1000, in human serum[3] above 4000, and 5000 to 25000 for higher plants.[4] The coverage within an experiment is much lower due to analytical limitations.

The typical metabolomics data processing pipeline first performs a feature detection step. The term feature describes a two-dimensional bounded signal: a chromatographic peak (retention time) and a mass spectral peak (*m*/*z*). Several software packages exist for feature detection, for example, the closed-source but freely available MetAlign,[5] or frameworks with open-source licenses, such as OpenMS,[6] MZmine,[7] and XCMS.[8] Other packages, some of them specific for LC/MS-based proteomics, have been reviewed elsewhere.[9]

Upon ionization, an individual chemical compound gives rise to one or more ion species, which can be observed in the same mass spectrum. Those ion species include isotopologue ions, fragment ions, and, in particular for electrospray ionization (ESI), adduct and cluster ions. A summary can be found in Keller et al.[10]

Received: September 23, 2011
Accepted: November 23, 2011
Published: November 23, 2011

ACS Publications © 2011 American Chemical Society

DOI:10.1021/ac202450g
Reprinted with permission. © 2011 American Chemical Society.

For biological interpretation, users are mainly interested in the compounds, rather than the redundancy of the different ion species, which induce an undesired bloat in the number of observed features, e.g., for an *Arabidopsis thaliana* seed extract Bottcher et al.[11] reported 434 features for 180 compounds. The complexity of both the downstream statistical analysis and subsequent compound identification especially in untargeted metabolite profiling experiments is unduly increased.

To address these problems, two additional processing steps are desired for LC/MS data analysis: (1) grouping all features which are derived from the same analyte, and (2) annotation of the type of ion species. The first step alone achieves both a data reduction and a first estimation of the total number of detectable compounds in a MS analysis. Such an estimate can be used for the optimization of the analytical protocol, similar to Yanes et al.[12] where the authors used the feature number as optimization criterion. Both steps together can reveal quasi-molecular ions, whose annotation is essential for further metabolite identification, such as elemental composition calculation based on accurate mass and isotope pattern or tandem MS analysis.

The authors of Brown et al.[13] have developed a workflow using the retention time, m/z-difference, and intensity correlation across samples to group related features, both reducing the number of relevant features down to 50% and matched 60% of the remaining features against the Manchester Metabolome Database (MMD). Intensity correlation across samples is also used by Alonso et al.,[14] and a data-reduction of 86% is reported.

Alternatively, similarity across chromatographic peak shapes allows the grouping of related features. Ipsen et al.[15] use a χ^2 test to check for exact coelution. In case of LC/MS data acquired on TOF instruments with a time-to-digital converter, the test provides p-values for the (un)certainty of coelution. The test works best with low ion counts, and the instruments' detector saturation correction had been disabled for this evaluation. ACD/IntelliXtract[16] is a commercial software solution to cluster features based on their retention time and the annotation of ion species according to a given rule table.

Both correlation across samples and peak shape analysis techniques are used in the R package ESI.[17] A fixed m/z-difference rule table is used for annotation and detection of isotopic peaks. The same approach was later used by Scheltema et al.[18] for high-resolution LC/MS data. By explicitly removing features exhibiting both similar peak shapes and intensity correlation across samples, they achieved a 60% size-reduction of the feature list.

In this paper we present the CAMERA package, which integrates multiple methods for grouping related features and uses a dynamic rule table for the annotation of ion species. We evaluate the performance of CAMERA with several validation experiments and demonstrate the analysis of two metabolomics experiments.

■ THEORY, ARCHITECTURE, AND ALGORITHMS

The analysis workflow with CAMERA is shown in Figure 1 and numbering (1−5) describes the typical workflow order. In the next paragraphs the steps are explained in more detail.

Creating Compound Spectra Based on Retention Time ①. The initial creation of compound spectra has to be fast, if dozens to hundreds of samples with thousands of features have to be processed. We select the most intense feature from the feature table not yet assigned to a compound spectrum and calculate a feature specific retention time window, typically 60% of the chromatographic peak fwhm (full width at half-maximum) around the centroid. All features within this range are then included into a new compound spectrum. This step is repeated until all features are assigned to a compound spectrum. The most intense feature usually has the highest signal-to-noise (S/N) ratio and often provides the most accurate estimate of the centroid and retention time.

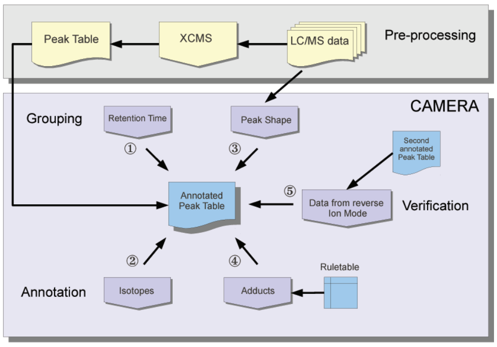


Figure 1. The CAMERA workflow for LC/MS data analysis. Raw data files are preprocessed with XCMS (upper part) and the resulting feature lists are passed to CAMERA. The feature grouping steps integrate retention time ① and chromatographic peak shape ③. Features are identified as isotopic peak ②, and adducts are annotated ④ using a dynamic rule table. Optionally, the annotation can be verified ⑤ with LC/MS data acquired in the opposite ion mode.

Isotopic Peak Detection and Charge State Calculation ②. The detection of isotopic patterns is required to deduce the charge states. Within each compound spectrum we calculate a pairwise m/z distance matrix and detect isotopes which exhibit a m/z-difference of $1.0033/z$[19] and also pass an additional intensity ratio check, described in detail in the Supporting Information, section S1.

Compound Spectrum Refinement Graph ③. Depending on the chromatographic separation, the resulting compound spectra might still encompass features of two or more closely coeluting compounds. We use a graph-based algorithm to integrate three more cues for an improved separation (see Figure 2 for an example).

First, we use the chromatographic peak shape similarity. CAMERA uses the raw data to obtain the extracted ion chromatograms (EIC) for each feature and calculates a pointwise pearson correlation of the intensities between the chromatographic peak boundaries for all pairs of features in a compound spectrum. CAMERA uses the EICs from the sample which had the most intense feature, often the one with the best S/N ratio. Alternatively, the peak shape correlation can be performed for all samples in the experiment. Second, we include the pearson correlation of intensities across all samples for each pair of features in a compound spectrum. Finally we encode the isotope relationship between two features detected in step ② as 1, and 0 otherwise. These three values are combined as shown in eq 1.

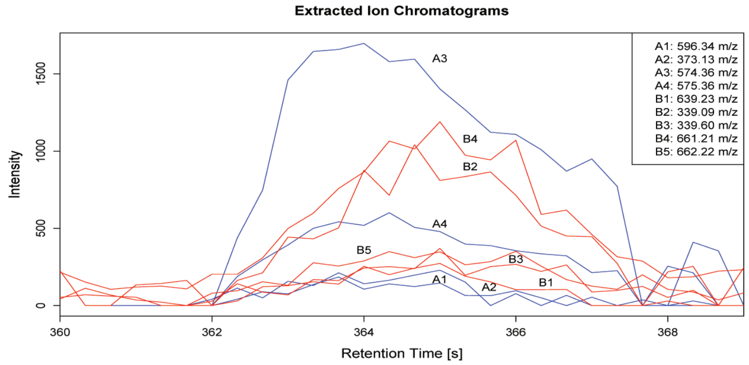
$$\mathrm{score}(x, y) = \mathrm{CAS}_{xy} + \mathrm{ISO}_{xy} + \frac{1}{N}\sum_{i=1}^{N} \mathrm{CPS}_{ixy} \qquad (1)$$

The score which represents the relationship between two features x and y is the combination of the intensity correlation across samples (CAS) for these two features, the binary

DOI:10.1021/ac202450g
Reprinted with permission. © 2011 American Chemical Society.

id	m/z	rt [s]
1	339.09	369.1
2	339.59	369.1
3	373.13	368.6
4	574.36	368.7
5	575.36	368.7
6	596.34	368.4
7	639.23	368.7
8	661.21	369.1
9	662.22	369.1

Initial compound spectrum



	1	2	3	4	5	6	7	8
2	1.0							
3	0.34	0.36						
4	0.72	0.67	0.90					
5	0.68	0.47	0.68	1				
6	0.38	0.38	0.87	0.86	0.75			
7	0.76	0.72	0.73	0.82	0.88	0.70		
8	0.94	0.79	0.35	0.70	0.67	0.51	0.79	
9	0.88	0.84	0.46	0.62	0.70	0.62	0.81	1.0

Relationship scoring table

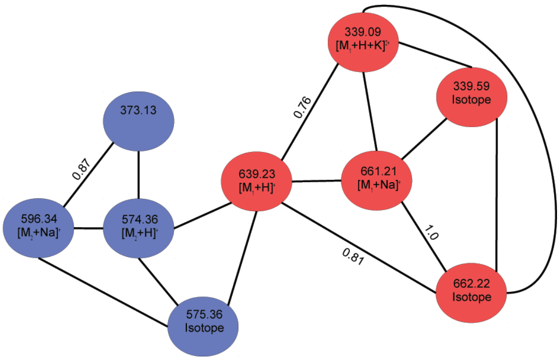


Figure 2. Schematic clustering of low-intensity features initially grouped by retention time into a single compound spectrum. Top left: the features, initially grouped by retention time. Top right: the EICs of all features. The labels A and B correspond to the result after graph clustering. Bottom left: the scoring matrix, used as edge weights in the graph. Bottom right: the relationship graph, where edges indicate an above-threshold score. The node labels include the ion species annotation, and the node color shows the graph separation after refinement with the LPC algorithm (A = blue, B = red).

encoded presence or absence of an isotope relationship, and the peak shape correlation (CPS_i) calculated for sample i.

In a graph, all features in a compound spectrum, which could still include features of two or more closely coeluting compounds, are represented as nodes, connected by edges with this score as edge weight. Several algorithms for graph separation have been developed, and we employ the "Highly-connected-subgraphs" (HCS[20]) from the R package RBGL or the "label propagation community" (LPC[21]) from the R package igraph. After the graph clustering, the initial compound spectrum is split into one refined compound spectrum for each subgraph. Figure 2 shows an example for a relationship graph before and after separation. Both coeluting compounds were separated completely.

Annotation of Adducts, Common Neutral Losses, and Cluster-Ions ④. For ESI, uncharged compounds are ionized through adduct formation with cations or anions or abstraction of protons. In addition, neutral losses occur leading to the formation of fragment ions. An annotation of these ion species reduces the number of features which have to be considered further in the downstream analysis. From at least two annotated ions, the molecular mass can be calculated, which is necessary to search in compound libraries or to calculate the elemental composition of the neutral compound.

CAMERA uses a dynamic rule set, which is created from the combination of lists of observable ions. Each rule describes a specific ion species with the mass difference to the molecular mass, ion charge, and the number of molecules the ion species contains. All m/z-differences within a compound spectrum are matched against the dynamic rule set. Matches with the same molecular mass hypothesis (below a given relative error) are combined into hypothesis groups. If no peaks can be explained via the rules, a reliable annotation is impossible. CAMERA does not use ad-hoc heuristics such as assuming that the most intense feature in a spectrum is the $[M + H]^+$-ion. Afterward, conflicting hypothesis groups are resolved as described in the Supporting Information, section S2.

Combining Data from Opposite Ion Modes for Verification ⑤. In metabolite profiling, samples are often measured in both positive and negative ion modes to increase the metabolite coverage. Although some compounds ionize in only one mode, many compounds are detectable in both. In these cases, the complementary ions provide further evidence for the quasi-molecular ion.

CAMERA includes a novel annotation verification algorithm using compound spectra measured in both ion modes. The algorithm calculates m/z-differences for all features of corresponding compound spectra from both modes within a retention time window. These differences are matched against a second, cross-polarity rule table. If a cross-polarity rule matches, it will either (1) annotate two previously unannotated ions, e.g., $[M + H]^+$ and $[M - H]^-$, or (2) verify an existing annotation, or (3) conflict with an existing annotation. In the latter case, the existing annotation is replaced. The cross-polarity rule table should only contain common and trusted combinations because these rules can override the single-polarity annotations.

Documentation and Availability. CAMERA is implemented in R, the packages for Windows (both 32 and 64 bit), Mac OS, and Linux are available from the Bioconductor repository[22] since release 2.4 in 2009.

DOI:10.1021/ac202450g

Reprinted with permission. © 2011 American Chemical Society.

EXPERIMENTAL SECTION

Reagents and Materials. All solvents used for sample preparation and analyses were of LC/MS-grade quality (CHROMASOLV, Fluka). A list of standard compounds used for the recovery experiment including sum formulas, molar masses, PubChem IDs and suppliers can be found in the Supporting Information, section S3. L-Tryptophan-2′,4′,5′,6′,7′-d_5 (98%) was purchased from Cambridge Isotope Laboratories. *Arabidopsis thaliana* (ecotype Col-0) was grown for 6 weeks on a soil/vermiculite mixture (3/2) in a growth cabinet with 8 h light (150 μE m^{-2}s^{-1}) at 22 °C and 16 h dark at 20 °C. Seeds of *Brassica napus*, *Brassica oleracera*, and *Brassica rapa* were kindly provided by D. Strack, Department of Secondary Metabolism, Leibniz Institute of Plant Biochemistry, Halle. All other seeds were obtained from local distributors. Procedures for extraction of leaf and seed material are provided in the Supporting Information, section S4.

Feeding Experiments. Plants were sprayed with 5 mM aqueous silver nitrate solution 1 h after the beginning of the light period. After 5 h, 25 rosette leaves originating from 5 individual plants were excised at the petiole and immersed in PCR tubes containing either 200 μL of water or 200 μL of an aqueous [ring-D_5]-Trp solution (1 mM), respectively. Leaves were incubated for an additional 2 days in a growth cabinet under the same conditions as described above. Individual leaves of the same treatment were pooled, frozen in liquid nitrogen, and stored at −80 °C until analysis.

Ultraperformance Liquid Chromatography (UPLC)/ESI-Quadrupole Time-of-Flight (QTOF)MS Analysis. Chromatographic separations were performed on an Acquity UPLC system (Waters) equipped with a HSS T3 column (100 mm × 1.0 mm, particle size 1.8 μm, Waters) applying the following binary gradient at a flow rate of 150 μL min^{-1}: 0−1 min, isocratic 95% A (water/formic acid, 99.9/0.1 (v/v)), 5% B (acetonitrile/formic acid, 99.9/0.1 (v/v)); 1−16 min, linear from 5 to 95% B; 16−18 min, isocratic 95% B; 18−20 min, isocratic 5% B. The injection volume was 2.7 μL (full loop injection). Eluted compounds were detected at a spectra rate of 3 Hz from m/z 100−1000 using a MicrOTOF-Q-I hybrid quadrupole time-of-flight mass spectrometer (Bruker Daltonics) equipped with an Apollo II electrospray ion source in positive and negative ion modes. We made sure that the concentration of the samples do not lead to saturation of the MS detector system, which is known to cause shifts of m/z, and retention time centroids of the features leads to truncated chromatographic peak profiles and distorted isotopic patterns. For detailed instrument settings and acquisition of collision-induced dissociation mass spectra see the Supporting Information, section S4.

LC/MS Data Preprocessing. Processing of raw data was performed with the XCMS package.[8] For the feature detection, we used the XCMS *centWave*[23] algorithm with the following parameters: snthresh = 6, ppm = 30, peakwidth = (5,12), prefilter = (2,200). The feature alignment was performed with the standard *group.density* algorithm from XCMS with bw = 3 and mzwid = 0.015. Afterward, each data set was processed with CAMERA functions in the following order *groupFWHM*, *findIsotopes*, *groupCorr*, and *findAdducts* using standard parameters. The Supporting Information, section S9 provides runtime measurements of CAMERA.

RESULTS AND DISCUSSION

We evaluated CAMERA with several experiments. First, using standards we analyzed the performance of compound spectrum creation and success rate of molecular mass annotation. Then, we processed the output from two different experiments, where the CAMERA results were used to perform targeted profiling of phenolic choline esters and tryptophan-derived metabolites, respectively.

Evaluation on Known Compound Mixture. For the evaluation we used a mixture of 39 known compounds (short, MM39), covering a broad mass range between 161 and 822 Da and different physicochemical properties (see the Supporting Information, section S3). The mixture was measured as pure solution and spiked in different concentrations (20, 5, 1, and 0.2 μM) into methanolic extracts of *Arabidopsis thaliana* leaves to simulate a realistically complex matrix.

The first evaluation focuses on the extraction of compound spectra, which requires a data set and a gold standard of true positive and true negative cases, i.e., pairs of peaks which should or should not be part of the same compound spectrum. Because it is very tedious to manually create a gold standard of a sufficiently large number of features from different compounds which coelute, we altered the retention times in the raw data files to artificially force “coelution” for this evaluation. We used only those peaks in the compound spectra of the MM39 for which a reliable annotation exists, to rule out false positives and randomly collected peaks from the remaining file with unrelated retention times to assemble a negative set. These data sets allowed us to calculate the precision and recall for the collection of compound spectra. The default peak shape correlation threshold of 0.75 results in a recall of 0.93, with a precision of 0.48. We also analyzed the influence of different acquisition parameters (scan rates varied from 0.5 to 6 Hz). Precision and recall had a standard deviation of 0.07 and 0.03, respectively, across the different conditions, see the Supporting Information, section S5 for details, including the ROC curves.

We then evaluated how successfully CAMERA could annotate the different ion species from a compound, which is required for the calculation of the molecular mass. We created baseline values for all annotatable compounds: we define an annotatable compound as observed to produce (1) the protonated molecular ion, (2) its first isotopic peak (required to calculate the charge state), and (3) the most prominent adduct ion (observed at 20 μM). This strategy serves as the gold standard to determine the number of annotatable compounds in the MM39 measurement; 35 out of 39 compounds pass the above requirements for the 20 μM positive mode measurement. If the mixture is diluted 2 orders of magnitude to 0.2 μM, many peaks drop below the detection limit and only 10 compounds remain annotatable.

CAMERA was able to detect the correct molecular mass in 90% of all annotatable compounds in either the positive or negative mode across all concentrations. After combining results from both ionization modes, CAMERA correctly determined molecular mass for all annotatable compounds and additionally for four compounds that were not on the gold standard list. Because the manual assignment of corresponding features in both positive/negative mode data is quite cumbersome, the combined annotation promises to annotate more compounds than a human operator could do on a routine basis.

DOI:10.1021/ac202450g

Reprinted with permission. © 2011 American Chemical Society.

Table 1. Calculation of Molecular Mass for the MM39 Compound Mixture Analyzed by UPLC/ESI-QTOFMS in Positive and Negative Ion Mode, Either in Pure Solvent or Spiked at Different Concentrations into a *Arabidopsis thaliana* Leaf Extract[a]

	in solvent	spiked into leaf extract				overall	
	20 µM	20 µM	5 µM	1 µM	0.2 µM		
ESI(+)	32 (35)	29 (32)	24 (28)	18 (21)	10 (10)	113 (126)	89.7%
ESI(−)	15 (19)	18 (18)	15 (16)	6 (6)	2 (3)	56 (62)	90.3%
ESI(±)	36	35	28	23	15	137	

[a] The number of annotatable compounds is shown in brackets. In the combined case, the annotations of the positive ion mode are verified and augmented with the negative ion mode data.

It is remarkable that the complex leaf matrix does not have an observable negative effect on the annotation performance. Table 1 shows the results for the individual concentrations. On closer inspection, the missing molecular mass annotations have few common causes: they occurred either because the compound spectrum did not contain enough explained features or, in other cases, several hypotheses had the same precedence scores and we did not count those as successful. For some compounds the compound spectra contained only the molecular ion and fragment ions but no further adducts. In this case, the compound cannot be annotated directly, unless the neutral loss is added to the rule set. The Supporting Information, section S10 shows an overview of the frequency of annotated adducts we observed.

In the measurements with different scan rates, we found that CAMERA only missed up to two correct annotations in those cases where either an essential (albeit low abundant) feature was not found by the feature detection algorithm or features were assigned to a different compound spectrum, especially in the case of lower scan rates where chromatographic peaks were covered by only a few scans. This suggests that CAMERA can also be used for LC/MS measurements with a low scan rate, e.g., on Orbitrap instruments at high resolution.

Case Study I: Screening for Phenolic Choline Esters in Brassicaceous Seeds. In this section we use untargeted LC/MS profiles of seeds from some *Brassicacea*, and demonstrate how CAMERA can be used to perform a neutral loss screen for phenolic choline esters as a targeted analysis strategy on a TOF instrument.

Phenolic choline esters accumulate in considerable amounts in seeds of many plant species within the *Brassicacea*[24] family. Representatives of this compound class structurally characterized so far include substituted cinammoyl and benzoyl cholines, which are further diversified by glycosylation or oxidative coupling to monolignols. A total of 30 phenolic choline esters could be identified in seeds of the model plant *Arabidopsis thaliana* and the oil crop *Brassica napus* using LC/ESI-tandem mass spectrometry.[25] A study of the fragmentation behavior of phenolic choline esters under positive-ion electrospray-CID conditions revealed a loss of trimethylamine as the initial fragmentation step (see the Supporting Information, section S6). The formation of the corresponding fragment ion $[M - C_3H_9N]^+$ requires different collision energies depending on the compound. However, it is also inducible by in-source CID allowing a systematic screening for phenolic choline esters even by single-stage MS. For that purpose, the neutral loss detection has to be performed *in silico* after data acquisition by searching for a given m/z-difference between pairs of peaks within a set of extracted compound spectra.

We prepared extracts from seeds of 12 different *Brassicacea* species and cultivars and analyzed each extract by UPLC/ESI(+)-QTOFMS at four different in-source CID voltages (0, 30, 60, and 90 V) to induce fragmentation of a broad range of phenolic choline esters. All 48 raw data files were preprocessed with XCMS with snthresh = 5 and ppm = 20, other parameters analogous to the evaluation section. Because of the large number of chromatographically unresolved compounds eluting near the void time, compounds with a retention time below 45 s were excluded from further analysis. Afterward CAMERA was used to create the compound spectra. Each compound spectrum was then screened for peak pairs displaying a m/z-difference of 59.074 ± 0.015 corresponding to a neutral loss of trimethylamine. In addition, we included a m/z-difference of 221.126 ± 0.015, related to a successive loss of trimethylamine and anhydrohexose (162.053 Da), because $[M - C_3H_9N]^+$-type ions formed from 4-*O*-hexosylated phenolic choline esters are known to readily eliminate their hexose moiety.[25] After alignment of positively screened peak pairs, the elimination of isotopic peak pairs, and application of a reasonable intensity threshold (1000 counts), we detected a total of 90 putative choline esters. A data matrix including m/z ratios of proposed molecular ions, retention times, and intensities can be found in the Supporting Information, section S6. It should be noted, that the number of putative candidates is rapidly increasing when tolerance thresholds for m/z-differences were increased. Therefore, use of mass analyzers providing adequate resolution and mass accuracy, such as TOFMS, is mandatory for this type of screening approach in order to ensure a highly specific neutral loss detection. In order to evaluate the obtained candidate list, previously published analytical data of choline esters from seeds of *Arabidopsis thaliana* and *Brassica napus* were used for compound annotation.[25] Out of 31 choline esters described recently, we were able to retrieve 22 from our list. Seven choline ester were consistently detected across all samples. Five of them could be annotated (Table 2), including sinapoylcholine, which is known to occur as a major phenolic choline ester in seeds of numerous *Brassicacea* species.[24] Although a rigorous evaluation is not possible because the choline ester composition of analyzed seeds is unknown, recovery of the majority of compounds described in the literature demonstrates the usability of CAMERA for such a

Table 2. Five Phenolic Choline Esters Found and Annotated in All 12 Brassicaceous Seeds[a]

m/z	t_r [s]	NL	elemental composition	annotation
280.15	275	−59	$C_{15}H_{22}NO_4^+$	FC
310.16	279	−59	$C_{16}H_{24}NO_5^+$	SC
458.21	403	−59	$C_{25}H_{32}NO_7^+$	FC(5-8′)G
472.21	221	−221	$C_{22}H_{34}NO_{10}^+$	SC 4-*O*-Hex
476.23	303	−59	$C_{25}H_{34}NO_8^+$	FC(4-*O*-8′)G

[a] NL, neutral loss; FC, ferulolycholine; SC, sinapoylcholine; G, guaiacyl; Hex, hexose.

DOI:10.1021/ac202450g

Reprinted with permission. © 2011 American Chemical Society.

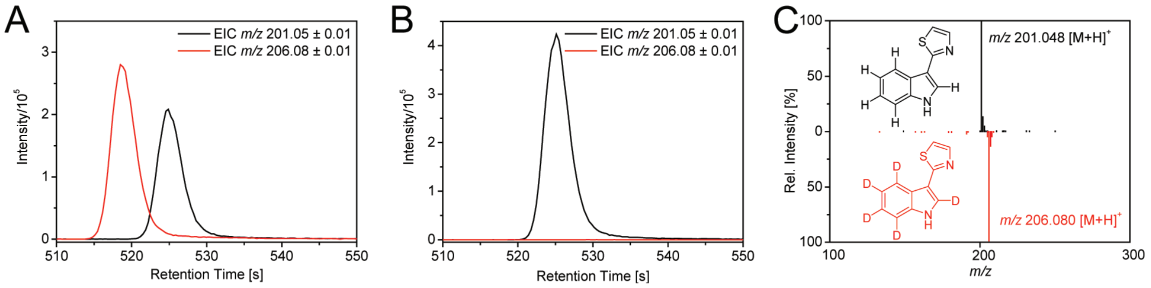


Figure 3. Identification of the phytoalexin camalexin as Trp-derived metabolite in silver nitrate-treated /*Arabidopsis thaliana*/ leaves using [ring-D_5]-Trp as the isotope-labeled tracer. Extracted ion chromatograms (EICs) corresponding to the protonoted molecular ions of camalexin (black) and D_5 -camalexin (red) obtained from UPLC/ESI(+)-QTOFMS analyses of extracts of [ring-D_5]-Trp-fed leaves (A) and control leaves (B). Extracted compound spectra of camalexin and its isotopologue are shown in the right picture (C).

screening approach. An additional advantage of this approach compared to triple quadrupole MS-based neutral loss scanning techniques is that any number of neutral losses can be simultaneously detected after data acquisition, allowing screening for a broad range of compound classes.

Case Study II: Identification of Trp-Derived Metabolites from *Arabidopsis thaliana* after [ring-D_5]-Trp Feeding. In vivo administration of isotope-labeled substrates combined with mass spectrometry-based analysis represents a powerful tool to investigate biochemical pathways. The detection of an isotope-labeled substrate incorporated into a known metabolite allows one to deduce a biosynthetic relationship between the fed precursor and the metabolite under study. Nontargeted screening for metabolites and their isotopologues after partial isotope-labeling of an endogenous precursor pool has been applied to explore unknown biosynthetic pathways and to discover novel intermediates and products.[26]

To demonstrate the applicability of the CAMERA package for such an analytical approach, the metabolic fate of the aromatic amino acid Trp was studied in the model plant *Arabidopsis thaliana* using [ring-D_5]-Trp as the isotope-labeled tracer. In *Arabidopsis*, Trp represents an important precursor for a variety of secondary metabolites including the phytoanticipin indol-3-ylmethyl glucosinolate and the phytoalexin camalexin (3-thiazol-2′-yl-indole).

Arabidopsis leaves were sprayed with silver nitrate to induce expression of Trp-metabolizing enzymes, detached from plants and fed with [ring-D_5]-Trp or water as the control. Methanolic extracts of label-fed and control leaves were analyzed in duplicate by UPLC/ESI-QTOF-MS in the positive and negative ion modes. In order to identify Trp-derived metabolites, the raw data was processed with XCMS and CAMERA to extract compound spectra and annotate isotopic peaks within these spectra. Afterward, deisotoped compound spectra extracted from data sets of label-fed leaves were screened for feature pairs that exhibit an m/z-difference of 5.031, reflecting the exchange of five hydrogen atoms by deuterium. Since deuterium labeling can slightly shift retention times, we searched for these feature pairs between compound spectra within a sliding retention time window of 8 s. For this purpose, we created a dedicated script using CAMERA functionality for the positive/negative polarity combination. We also included the m/z-difference of 4.025 because indole ring hydroxylation (a frequently observed transformation in Trp metabolism in *Arabidopsis*) results in a loss of one of the five deuterium labels. The retention time for Trp-candidates was restricted between 45 and 600 s. All features related to unlabeled Trp-metabolites have to be detectable in both label-fed and control samples whereas the labeled ones in label-fed samples only, see Figure 3. After those filtering steps, 46 putative Trp-derived metabolites could be identified in the positive ion mode and 34 in the negative ion mode. Corresponding candidate lists including compound annotation can be found in the Supporting Information, section S7.

To verify the obtained candidate lists, tandem mass spectra of quasimolecular ions of putative pairs of nonlabeled and labeled metabolites were acquired and compared (Supporting Information, section S8). Because of low peak intensities or low incorporation rates, only 19 candidate pairs could be rigorously verified following this strategy. Together with literature data, a total of 23 Trp-metabolites could be identified, of which 20 were already known from the literature. This case study clearly demonstrates applicability of CAMERA for such a screening approach, even in case of a retention time shifts when using deuterium labels.

■ CONCLUSIONS

The CAMERA package is designed to postprocess XCMS feature lists and to collect all features related to a compound into a compound spectrum. For this, a set of algorithms has been implemented in CAMERA, such as the fast retention time-based grouping but also a novel, graph-based algorithm to integrate the peak shape analysis, isotopic information, and intensity correlation across samples. The automatic sample selection avoids poor results if compounds have a low intensity (or are absent) in some samples. The ion species annotation uses a dynamic rule set and a new strategy to combine spectral information from samples measured in the positive and negative ion modes, resulting in both more and more reliable ion species annotation. We evaluated the reliability of the molecular mass calculation and found a 90% success rate for MM39 in different concentrations, both pure and after spiking the mixture at various concentrations into a complex *Arabidopsis thaliana* leaf extract.

Finally, we performed two experiments, demonstrating advanced analyses which can be performed with CAMERA. The first case study essentially performed a neutral loss screen for putative phenolic choline esters using multiple in-source voltages to induce fragmentation. In total, 90 putative choline esters were detected. The second case study demonstrated the search for mass differences as a result of [ring-D_5]-Trp feeding in *Arabidopsis thaliana* leaves. CAMERA was used to detect pairs of isotopologue features indicating 46 Trp-derived

DOI:10.1021/ac202450g
Reprinted with permission. © 2011 American Chemical Society.

metabolites. In addition to 20 already known compounds, 3 new ones were found and verified with tandem MS. Both studies can easily be adopted to other compound classes and metabolites. The CAMERA packages for Windows, Mac OS, and Linux, manuals, and tutorials are freely available from the Bioconductor repository and its mirrors under the open source GPL license.

ASSOCIATED CONTENT

Supporting Information

Experimental procedures and characterization data for all new compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ckuhl@ipb-halle.de (C.K.); sneumann@ipb-halle.de (S.N.).

ACKNOWLEDGMENTS

This work was supported by the California Institute of Regenerative Medicine (Grant TR1-01219), the National Institutes of Health (Grants R24 EY017540-04, P30 MH062261-10, and P01 DA026146-02), and the Department of Energy (Grants FG02-07ER64325 and DE-AC0205CH11231).

REFERENCES

(1) Dunn, W. B. *Phys. Biol.* **2008**, *5*, 011001.
(2) Feist, A. M.; Henry, C. S.; Reed, J. L.; Krummenacker, M.; Joyce, A. R.; Karp, P. D.; Broadbelt, L. J.; Hatzimanikatis, V.; Palsson, B. *Mol. Syst. Biol.* **2007**, *3*, 121.
(3) Psychogios, N.; et al. *PLoS One* **2011**, *6*, e16957.
(4) Trethewey, R. N. *Curr. Opin. Plant Biol.* **2004**, *7*, 196−201.
(5) Tikunov, Y.; Lommen, A.; Vos, C. d.; Verhoeven, H.; Bino, R.; Hall, R.; Bovy, A. *Plant Physiol.* **2005**, *139*, 1125−37.
(6) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinf.* **2008**, *9*, 163.
(7) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.
(8) Smith, C.; Want, E.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.
(9) Katajamaa, M.; Oresic, M. *J. Chromatogr., A* **2007**, *1158*, 318−328.
(10) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. *Anal. Chim. Acta* **2008**, *627*, 71−81.
(11) Böttcher, C.; von Roepenack-Lahaye, E.; Schmidt, J.; Schmotz, C.; Neumann, S.; Scheel, D.; Clemens, S. *Plant Physiol.* **2008**, *147*, 2107−2120.
(12) Yanes, O.; Tautenhahn, R.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2011**, *83*, 2152−2161.
(13) Brown, M.; Wedge, D. C.; Goodacre, R.; Kell, D. B.; Baker, P. N.; Kenny, L. C.; Mamas, M. A.; Neyses, L.; Dunn, W. B. *Bioinformatics* **2011**, *27*, 1108−1112.
(14) Alonso, A.; JuliÃ, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S. *Bioinformatics* **2011**, *27*, 1339−1340.
(15) Ipsen, A.; Want, E. J.; Lindon, J. C.; Ebbels, T. M. D. *Anal. Chem.* **2010**, *82*, 1766−1778.
(16) ACD/IntelliXtract, Advanced Chemistry Development, Inc. www.acdlabs.com/intellixtract, 2007.
(17) Tautenhahn, R.; Böttcher, C.; Neumann, S. *Lect. Notes Comput. Sci.* **2007**, *4414*, 371−380.
(18) Scheltema, R.; Decuypere, S.; Dujardin, J.; Watson, D.; Jansen, R.; Breitling, R. *Bioanalysis* **2009**, *1*, 1551−1557.
(19) Yergey, J. A. *Int. J. Mass Spectrom. Ion Phys.* **1983**, *52*, 337−349.
(20) Hartuv, E.; Shamir, R. *Inf. Process. Lett.* **2000**, *76*, 175 −181.
(21) Raghavan, U. N.; Albert, R.; Kumara, S. *Phys. Rev. E: Stat. Nonlinear, Soft Matter Phys.* **2007**, *76*, 036106.
(22) Gentleman; Rossini; Dudoit; Hornik. The Bioconductor FAQ. http://www.bioconductor.org,2003.
(23) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.
(24) Bouchereau, A.; Hamelin, J.; Lamour, I.; Renard, M.; Larher, F. *Phytochemistry* **1991**, *30*, 1873−1881.
(25) Böttcher, C.; von Roepenack-Lahaye, E.; Schmidt, J.; Clemens, S.; Scheel, D. *J. Mass Spectrom.* **2009**, *44*, 466−476.
(26) Feldberg, L.; Venger, I.; Malitsky, S.; Rogachev, I.; Aharoni, A. *Anal. Chem.* **2009**, *81*, 9257−9266.

DOI:10.1021/ac202450g
Reprinted with permission. © 2011 American Chemical Society.

metabolites

Article

Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data

Hendrik Treutler [1,2,*] and Steffen Neumann [1]

[1] Department of Stress and Developmental Biology, Leibniz Institute for Plant Biochemistry, Weinberg 3, Halle 06120, Germany; steffen.neumann@ipb-halle.de
[2] Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, Halle 06120, Germany
* Correspondence: hendrik.treutler@ipb-halle.de; Tel.: +49-345-5582-1472

Academic Editor: Peter D. Karp
Received: 31 August 2016; Accepted: 14 October 2016; Published: 20 October 2016

Abstract: Mass spectrometry is a key analytical platform for metabolomics. The precise quantification and identification of small molecules is a prerequisite for elucidating the metabolism and the detection, validation, and evaluation of isotope clusters in LC-MS data is important for this task. Here, we present an approach for the improved detection of isotope clusters using chemical prior knowledge and the validation of detected isotope clusters depending on the substance mass using database statistics. We find remarkable improvements regarding the number of detected isotope clusters and are able to predict the correct molecular formula in the top three ranks in 92% of the cases. We make our methodology freely available as part of the Bioconductor packages *xcms* version 1.50.0 and *CAMERA* version 1.30.0.

Keywords: isotope cluster; software; raw data

1. Introduction

The elucidation of the metabolism provides deep insights into complex processes in the cell such as responses to nutrition deficiency, pathogen exposure, and drought stress in plants or the implications of mutations, age, and tissue development in animals. Mass spectrometry is a key technology for the identification and quantification of metabolites in biological samples. After measurement using mass spectrometers, feature detection algorithms extract basic properties about peaks in the raw data such as retention time and peak height. The set of properties describing single peaks are called *features* and the exhaustive extraction of features is a prerequisite for downstream analyses such as metabolite identification and quantitative comparisons between samples.

The feature detection algorithm *centWave* in the R package *xcms* version 1.50.0 [1] adapts the following procedure. First, a set of *regions of interest* (ROIs) is identified in the ROI identification step, where ROIs are two-dimensional intervals in the mass-to-charge (m/z) dimension and the retention time dimension containing potential signals. The set of ROIs is examined in the ROI examination step in order to validate, localize, and quantify features. In the ROI identification step, a heuristic method is applied to the raw data to substantially reduce the processing time of the more computationally intensive ROI examination step. This heuristic method aims at a high specificity at the cost of sensitivity, especially in case of features with a low signal-to-noise ratio. Consequently, potentially important features in the raw data are not detected and the information behind these features cannot be used in downstream analyses.

Most chemical elements are present in different variants called isotopes. Though chemically almost equivalent, the isotopes of a particular chemical element differ in mass and are thus well distinguishable using mass spectrometry. The isotopes of each element have a known natural

DOI:10.3390/metabo6040037

abundance and the distribution of isotopes across all atoms of a molecule results in a set of related signals. The features extracted from these signals are called *isotopologue features* and the set of all isotopologue features from one analyte is called *isotope cluster* also known as isotope pattern. Unfortunately, many of these signals are below the detection limit which results in the underestimation of isotopologue features.

Based on isotope clusters, it is possible to determine the charge state, abundance, and elemental composition of the measured ion with high precision. The arrangement of isotopologue features to isotope clusters leads to a considerable reduction of data complexity facilitating the interpretation of data sets. It has been demonstrated that the analysis of isotope clusters leads to an increased confidence and precision of comparative analyses [2]. Isotope clusters from precursor ions and tandem mass spectrometry are pivotal for the determination of the molecular formula using software like SIRIUS [3], Rdisop [4], and others [5–12]. The molecular formula strongly facilitates the identification of molecules known as a major bottleneck in metabolomics [13,14] and has been demonstrated metabolome-scale [15]. There are approaches in metabolomics and proteomics which use isotope clusters to improve peak picking [16–18]. In addition, isotope clusters have been used as a valuable source for the assessment of the data quality [19] and for database searches with high precision [20].

The detection of isotope clusters is usually performed after peak picking by consideration of coeluting features separated by certain distances in the m/z dimension. However, a validation of putative isotope clusters in terms of the removal of leading peaks from hydrogen–losses and the decomposition of overlapping isotope clusters into individual isotope clusters is usually lacking in case of small molecules. The deconvolution of overlapping isotope clusters has been described in case of peptides and proteins, for isotope dilution experiments, and in case of substances with known molecular formula [17,21,22].

Aiming at the exhaustive detection and precise validation of isotope clusters, we propose the following approach for liquid chromatography–high resolution mass spectrometry data. We predict new ROIs for putative isotope peaks based on previously detected features and implement this approach in combination with the *centWave* algorithm as part of the R package *xcms* version 1.50.0 [23]. We validate putative isotope clusters depending on the mass of the substance based on database statistics and implement this approach as part of the R package *CAMERA* version 1.30.0 [24].

For evaluation purposes, we apply the modified *centWave* algorithm to different sets of mass spectrometry raw data and detect and validate isotope clusters as proposed. We evaluate the results using various performance measures and find remarkable improvements regarding the number of detected isotope clusters. The extended R packages *xcms* and *CAMERA* are available at Bioconductor [25].

2. Results

We demonstrate the performance of our approach for an enhanced isotope cluster detection and validation. First, we describe the workflow which includes our approach; Second, we evaluate the proposed targeted peak picking with predicted isotope ROIs compared to peak picking with random ROIs and traditional peak picking on basis of various performance measures; Third, we evaluate the proposed isotope detection routine with mass–specific isotope cluster validation compared to several isotope detection routines on basis of various performance measures; Fourth, we present the isotope ratio quantiles which are used for the validation of isotope clusters; Fifth, we exemplify the proposed isotope detection routine with and without mass–specific isotope cluster validation on six example substances.

2.1. Workflow of the Approach

We integrated the proposed methodology into an untargeted workflow which extracts annotated peak tables from LC-MS raw data as summarized in Figure 1. The user supplies the LC-MS raw data files in a *xcms*-supported format, namely one of AIA/ANDI NetCDF, mzXML, mzData, or mzML.

DOI:10.3390/metabo6040037

The workflow incorporates one function from the R package *xcms* [23], one function from the R package *CAMERA* [24], and two new function as follows.

First, we perform peak picking without any prior knowledge which we denote as *traditional peak picking*. Here, we use the *centWave* algorithm [1] which applies a heuristic for the detection of ROIs (ROI identification step). Given the set of detected ROIs, chromatographic peaks are extracted using continuous wavelet transformation (ROI examination step). This step results in a peak table with one row for each detected feature and one column for each feature property such as m/z, retention time, integrated peak area, and signal-to-noise ratio.

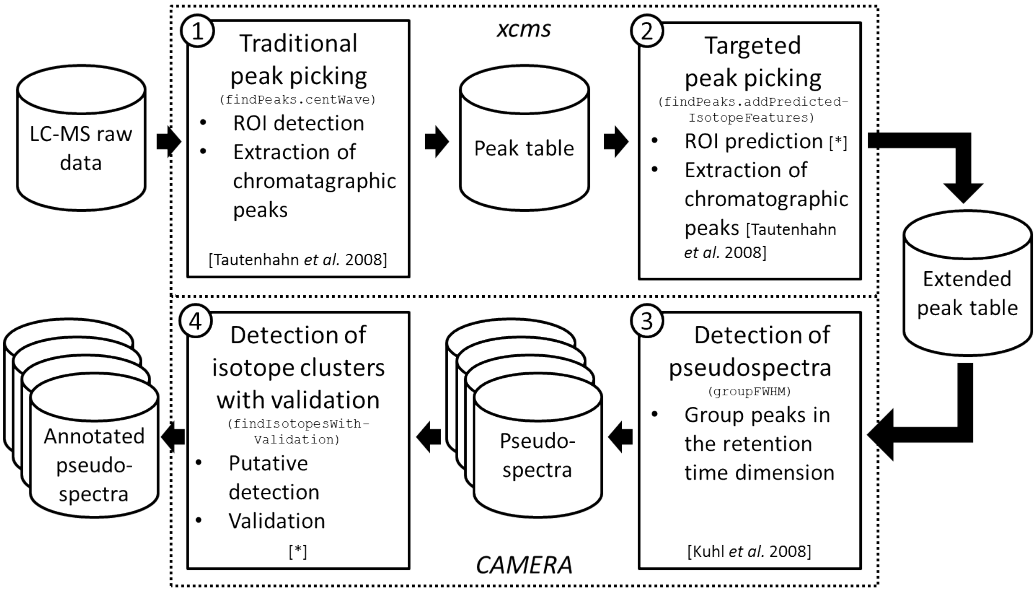

LC-MS raw data
xcms
① Traditional peak picking
(findPeaks.centWave)
• ROI detection
• Extraction of chromatographic peaks
[Tautenhahn et al. 2008]
Peak table
② Targeted peak picking
(findPeaks.addPredicted-IsotopeFeatures)
• ROI prediction [*]
• Extraction of chromatographic peaks [Tautenhahn et al. 2008]
Extended peak table
③ Detection of pseudospectra
(groupFWHM)
• Group peaks in the retention time dimension
[Kuhl et al. 2008]
Pseudo-spectra
④ Detection of isotope clusters with validation
(findIsotopesWith-Validation)
• Putative detection
• Validation
[*]
Annotated pseudo-spectra
CAMERA

Figure 1. Workflow of the proposed approach. We depict data sets with cylinders, algorithms with continuous rectangles, and R packages with dotted rectangles. Each algorithm rectangle comprises the step number (top left corner), the purpose of the algorithm (heading), the R function name (monospace font), algorithm steps (itemized), and a reference for the algorithm or the individual algorithm steps (in square brackets, asterisk stands for this manuscript). ① The workflow starts with traditional peak picking on LC-MS raw data to extract a peak table comprising features; ② This peak table is extended by a targeted peak picking which targets on isotope features; ③ The extended peak table is split into putative compound spectra denoted pseudospectra; ④ The detection and validation of isotope clusters is performed on each pseudospectrum resulting in annotated pseudospectra.

Second, we perform the proposed targeted peak picking as described in Section 4.1. Here, a set of isotope ROIs is predicted on basis of the previously extracted peak table. Given the set of predicted isotope ROIs, chromatographic peaks are extracted using continuous wavelet transformation (ROI examination step). Notably, this ROI examination step is identical to the ROI examination step in the traditional peak picking step with the exception that we use relaxed peak picking parameters this time. This step results in an extended peak table which is enriched with features corresponding to isotope isotope peaks as demonstrated in the second results section.

Third, we extract *pseudospectra* from the extended peak table [24]. This step aims at the extraction of compound spectra on basis of the retention times, but multiple coeluting compounds are potentially assigned to the same spectrum which is the reason for the usage of the term pseudospectrum. In case of multiple raw data files a retention time correction (*xmcs* function `retcor`) can be advisable prior to the extraction of pseudospectra. This step results in a set of pseudospectra. Each pseudospectrum is a peak table comprising all properties of a subset of the features from the extended peak table.

Fourth, we detect isotope clusters in each pseudospectrum using the proposed isotope detection routine with mass–specific isotope cluster validation as described in Section 4.2. Here, putative isotope clusters are detected and putative isotope clusters are validated based on database statistics as demonstrated in the third results section. This step results in a set of annotated pseudospectra, i.e., the given set of pseudospectra enriched with isotope annotations.

The presented workflow is implemented exemplarily in the vignette IsotopeDetectionVignette in R package *CAMERA* in version 1.30.0. In addition the R package *CAMERA* supports a number of further analyses given the set of annotated pseudospectra. This includes, amongst others, the annotation of adducts and neutral losses, the filling of missing values, and the combination of results from opposite ion modes.

2.2. Targeted Peak Picking Using Predicted Isotope ROIs

We examine whether the proposed prediction of isotope ROIs in combination with the *centWave* algorithm increases the number of detected isotope peaks. To verify the specificity of the predicted isotope ROIs to isotopes, we compare predicted isotope ROIs with the same number of random ROIs denoted *noise ROIs*. In addition, we compare our approach to the unmodified *centWave* algorithm with different signal-to-noise thresholds snthr. We evaluate our approach based on a dilution series experiment with 40 LC-MS measurements. These data sets comprise both strong and weak signals and constitute the basis to test the detection of weak signals like isotope peaks.

We evaluate the performance of predicted isotope ROIs detected with different relaxed signal-to-noise thresholds `snthr'` as described in Section 4.1 on 40 LC-MS measurements described in Section 4.4. We quantify the performance using the performance measures (i) number of detected peaks; (ii) number of detected isotope peaks; (iii) number of detected isotope clusters; (iv) *isotope coverage*; and (v) Peak Picking Score (*PPS*). The isotope coverage is the ratio between the number of detected isotope peaks and the number of detected peaks. The isotope coverage ranges from 0 to 1, where 0 means that no isotope clusters have been detected and 1 means that all peaks are part of isotope clusters. A higher isotope coverage indicates a higher peak picking quality as exploited in [19]. The PPS was proposed in [19] for the quantification of the peak picking quality and implemented in the R package *IPO*. The PPS is defined as the ratio between the number of reliable peaks squared and the number of non–reliable peaks. The number of reliable peaks is defined as the number of peaks in isotope clusters which are detected in the *IPO* package by a custom isotope detection routine. The number of non–reliable peaks is defined as the number of peaks which are not in a isotope cluster although it is to be expected based on different criteria. We compute each performance measure as a function of the relaxed signal-to-noise threshold `snthr'` $\in \{100, 95, ..., 5\}\% * \text{snthr}$, where snthr $=$ 25 is the signal-to-noise threshold used in the traditional peak picking step.

In Figure 2 we show the performance of the traditional peak picking in combination with targeted peak picking with isotope ROIs as well as traditional peak picking in combination with targeted peak picking with noise ROIs for varying signal-to-noise threshold `snthr'`. In addition, we show the performance of traditional peak picking with varying signal-to-noise threshold `snthr`. In case of predicted isotope ROIs, all five measures increase with decreasing `snthr'`. The isotope coverage appears to saturate for a relaxed signal-to-noise threshold `snthr'` of approximately 6.25. For this threshold, we find in case of predicted isotope ROIs an average increase of approximately +10% peaks, +37.6% isotope peaks, +33.5% isotope clusters, +25.2% isotope coverage, and +102.8% PPS in contrast to noise ROIs, suggesting an isotope-specific improvement of peak picking. More specifically, 20 isotope clusters could be extended and 37 isotope clusters could be newly detected. In addition, we find that the PPS decreases for a relaxed signal-to-noise threshold `snthr'` lower than 5. This finding confirms the general observation that peak picking with a too low signal-to-noise threshold results in unreliable peaks and is therefore not advisable. We also tested the performance of traditional peak picking with varying signal-to-noise threshold `snthr` and find that the number of peaks more

DOI:10.3390/metabo6040037

than doubles. However, the proportion of low–intensity peaks which are not part of isotope clusters increases disproportionately and there is no specificity for isotope peaks.

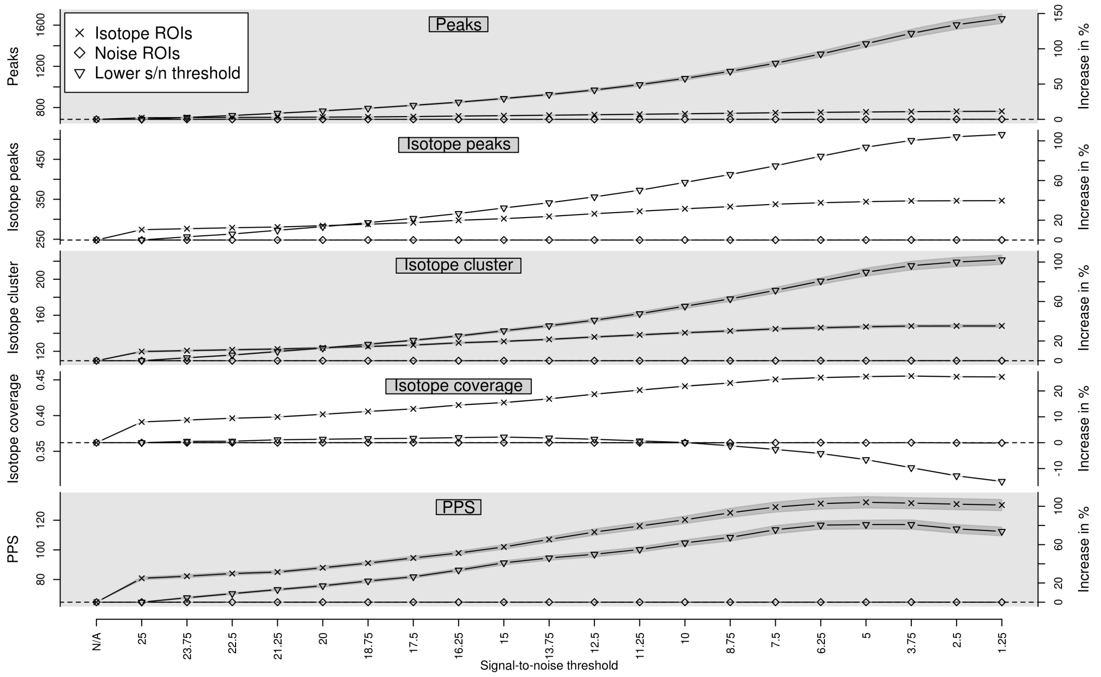
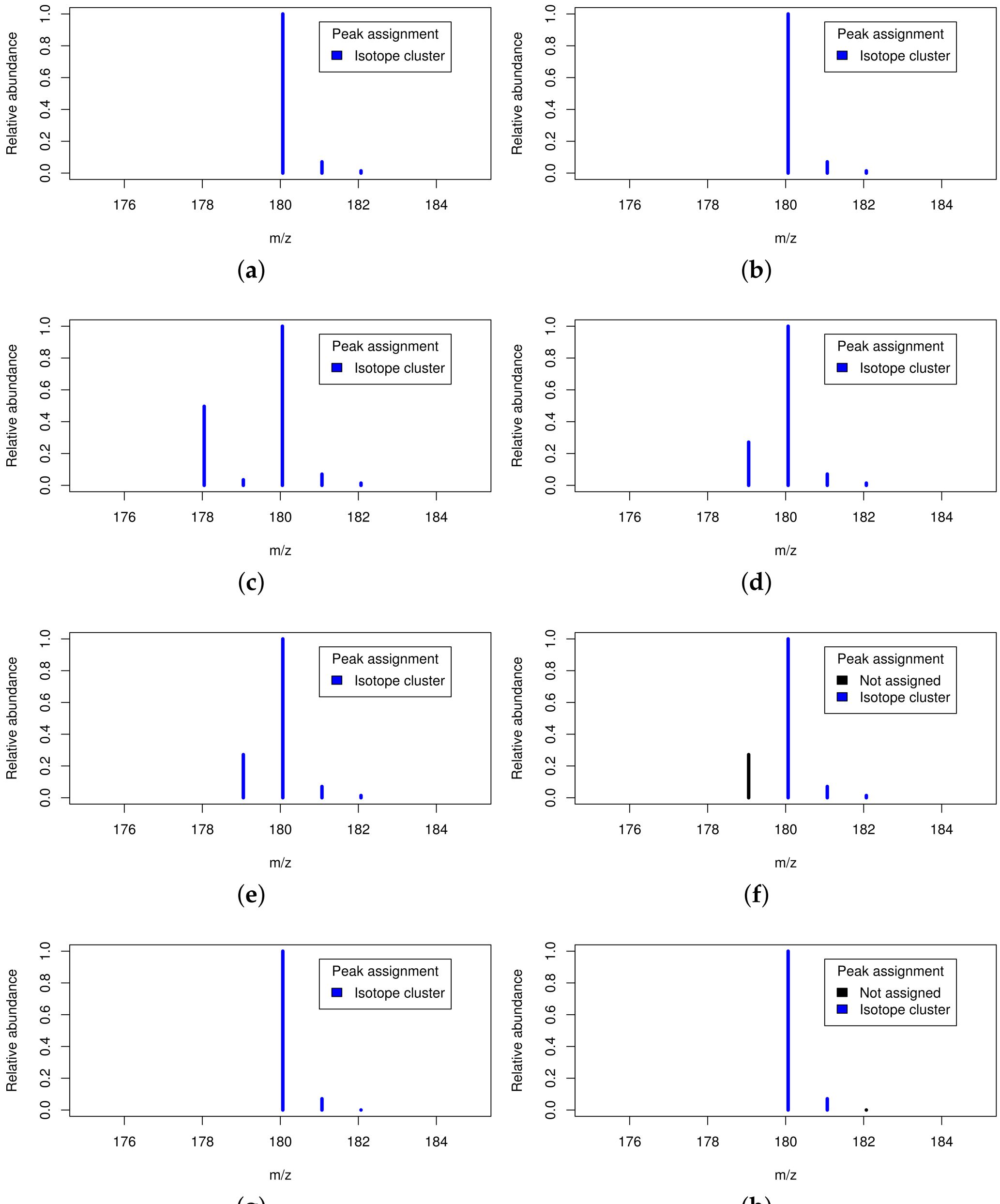


Figure 2. Evaluation of predicted isotope ROIs for varying relaxed signal-to-noise threshold `snthr'`. We show the mean (solid line) and the standard error of the mean (SEM, interval in dark grey) of the performance measures (**i**) number of detected peaks; (**ii**) number of detected isotope peaks; (**iii**) number of detected isotope clusters; (**iv**) isotope coverage; and (**v**) Peak Picking Score (PPS). In case of isotope ROIs and noise ROIs, we plot the performance of each measure without additional ROIs in the first column ("N/A") as reference value (horizontal dashed line) and in the subsequent columns with additional ROIs for decreasing relaxed signal-to-noise threshold `snthr'`. In case of "Lower S/N threshold", we plot the performance of each measure for decreasing signal-to-noise threshold `snthr` without additional ROIs. All four measures increase for predicted isotope ROIs with decreasing signal-to-noise threshold `snthr'` in contrast to noise ROIs.

2.3. Isotope Cluster Detection and Validation

There is a multitude of isotope detection routines for the recognition of isotope clusters. These detect coeluting features which are separated by certain distances in the m/z dimension and group these features to isotope clusters. However, a validation of detected isotope clusters is typically based on simple *ad hoc* rules. There are at least four cases for which the validation of isotope clusters can be beneficial as shown in Figure 3.

First, valid isotope clusters can be verified which strengthens the trust in the data; Second, multiple coeluting substances with mass differences of a few dalton can result in isobaric ion species and thus in overlapping isotope clusters [26]. These are potentially misinterpreted as a single isotope cluster affecting downstream analyses. This necessitates the deconvolution of the overlapping isotope cluster into at least two valid isotope clusters; Third, substances can be affected by hydrogen loss as reported in [27] and exploited in [28]. This leads to mass differences similar to isotope peaks ($\mathrm{mass}(^1\mathrm{H}) = 1.008 \approx 1.0034 = \mathrm{mass}(^{13}\mathrm{C}) - \mathrm{mass}(^{12}\mathrm{C})$) and results in a small trailing peak which is potentially misinterpreted as monoisotopic peak of the putative isotope cluster. This may result in the assumption of a wrong monoisotopic mass and may even lead to the rejection of the entire isotope cluster on the basis of failed intensity-checks [24]. Although this small trailing peak corresponds to the same substance, it needs to be removed from the isotope cluster in order to allow more precise molecular formula predictions. Fourth, the intensity of small peaks is systematically underestimated by some mass spectrometers which leads to distorted ratios between different isotope peaks as reported previously [3]. This intensity bias would lead to distorted molecular formula predictions and the removal of these underestimated peaks from the isotope cluster allows more precise molecular formula predictions.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3. Four cases necessitating the validation of putative isotope clusters. Figure 3**a**,**b**: Valid isotope cluster without and with isotope cluster validation; Figure 3**c**,**d**: Two overlapping isotope clusters without and with isotope cluster validation; Figure 3**e**,**f**: Hydrogen loss without and with isotope cluster validation; Figure 3**g**,**h**: Underestimated small peak without and with isotope cluster validation.

DOI:10.3390/metabo6040037

We compare the proposed isotope detection routine with mass–specific isotope cluster validation (IDR_{NewVal}) with the isotope detection routine without isotope cluster validation ($IDR_{NewNoVal}$), the isotope detection routine implemented in the *AStream* package ($IDR_{AStream}$) [29], the isotope detection routine implemented in the *CAMERA* package (IDR_{CAMERA}) [24], and the isotope detection routine implemented in the *mzMatch* package ($IDR_{mzMatch}$) [30]. The isotope detection routines from *AStream*, *CAMERA*, and *mzMatch* apply different requirements for the validation of isotope clusters. In $IDR_{AStream}$ it is required that the abundance of the monoisotopic peak, the first isotope peak, and the second isotope peak decreases strictly, which corresponds to a ratio <1 between consecutive isotope peaks. In IDR_{CAMERA} it is required that the ratio of the monoisotopic peak to the first isotopic peak is within an interval which is given by the ratios of the monoisotopic peak to the first isotopic peak of a substance consisting exactly one carbon atom and a substance consisting exactly $mass_{mono}/mass(^{12}C)$ carbon atoms, where $mass_{mono}$ is the assumed monoisotopic mass of the substance. In $IDR_{mzMatch}$ it is required that isotope peaks show a high correlation regarding coelution.

We evaluate the performance of the isotope cluster detection and validation described in Section 4.2 on a dilution series experiment with 40 LC-MS measurements described in Section 4.4. We quantify the performance using the performance measures (i) number of detected peaks; (ii) number of detected isotope peaks; (iii) number of detected isotope clusters; and (iv) isotope coverage, i.e., the proportion of detected isotope peaks versus all detected peaks. We compute each performance measure without predicted isotope ROIs as well as with predicted isotope ROIs for a relaxed signal-to-noise threshold `snthr'` of 6.25. We present the results with predicted isotope ROIs relative to the results without predicted isotope ROIs in Figure 4. These results are a subset of the results in Figure A1 in the Appendix A where we present the results for varying relaxed signal-to-noise threshold `snthr'`. We relate the results to the quality of the predicted molecular formulas presented in the Appendix B on a gold standard of 11 data sets with known content.

In Figure 4 we show the performance measures for IDR_{NewVal}, $IDR_{NewNoVal}$, $IDR_{AStream}$, IDR_{CAMERA}, and $IDR_{mzMatch}$. We find that all four measures increase with predicted isotope ROIs in case of all isotope detection routines. $IDR_{NewNoVal}$ detects the most isotopes which reflects the fact that there are no constraints regarding the shape of the isotope cluster. This indicates that a certain proportion of the detected isotope clusters might be invalid. We point out, that this highly sensitive algorithm can be useful in case of substances containing uncommon elements such as Cl, Br, Se, or B as scrutinized in [31]. $IDR_{mzMatch}$ detects by far the lowest number of isotopes which reflects that this algorithm requires a high degree of correlation between isotope peaks resulting in a high specificity at the cost of sensitivity. $IDR_{NewNoVal}$ and $IDR_{mzMatch}$ show the lowest number of correctly predicted molecular formulas as shown in Appendix B. We find comparable results for $IDR_{AStream}$, IDR_{CAMERA}, and IDR_{NewVal}. Also the numbers of correctly predicted molecular formulas are similar as shown in Appendix B. Interestingly, IDR_{NewVal} showed the highest number of correctly predicted molecular formulas and was also able to rank the highest number of correct molecular formulas to the first three ranks. Remarkably, in case of 85% to 92% of all tested ions the detected isotope clusters from all isotope detection routines with or without predicted isotope ROIs were sufficient for the prediction of the correct molecular formula to the first three ranks. This finding states, that the prediction of molecular formulas from isotope clusters works well in general and hence it is challenging to improve upon.

DOI:10.3390/metabo6040037

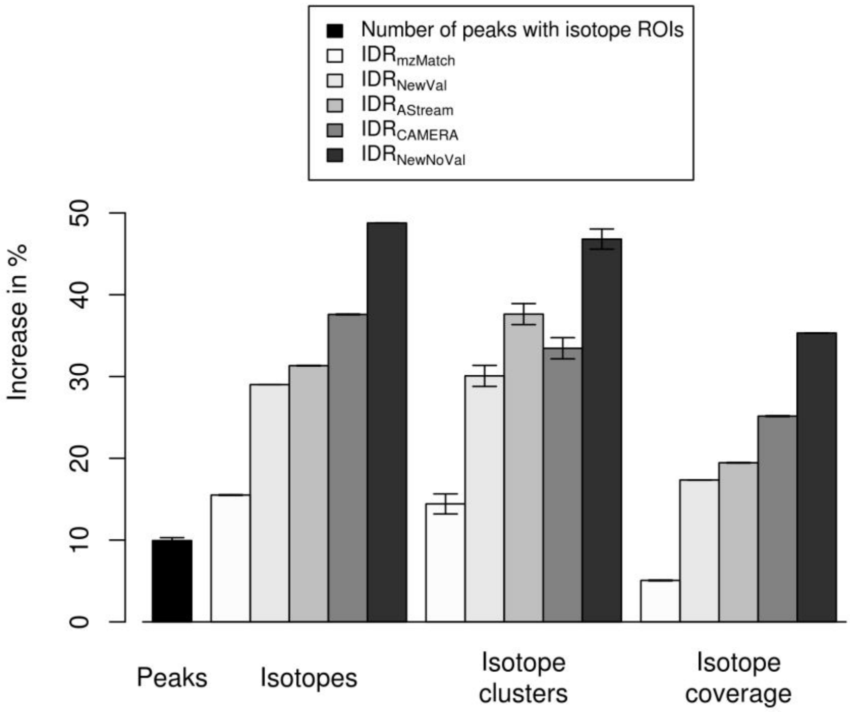


Figure 4. Evaluation of predicted isotope ROIs in combination with different isotope detection routines for a relaxed signal-to-noise threshold `snthr'` of 6.25. We plot the increase of the mean and the standard error of the mean (SEM, error bars) of the performance measures (**i**) number of detected peaks; (**ii**) number of detected isotope peaks; (**iii**) number of detected isotope clusters; and (**iv**) isotope coverage relative to the performance of the *CAMERA* isotope detection routine without predicted isotope ROIs. All four measures increase with predicted isotope ROIs.

2.4. Isotope Cluster Statistics

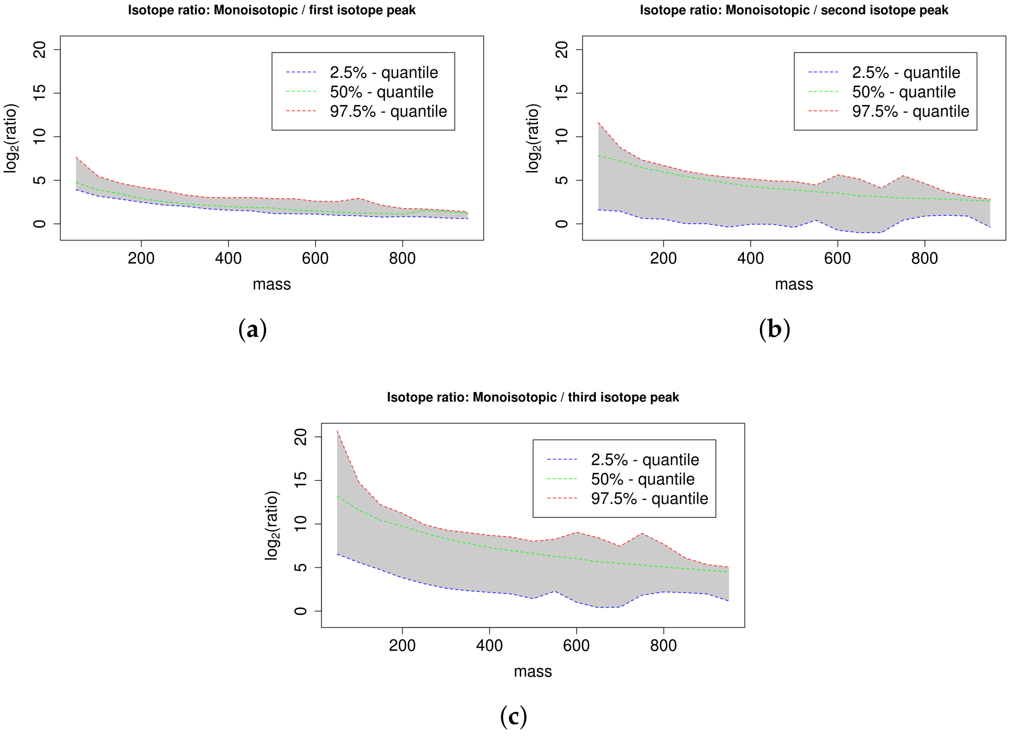
We examine the compounds of the publicly available databases ChEBI [32], KEGG [33], KNApSAcK [34], LIPID MAPS [35], and PubChem [36] in order to compute mass–specific confidence intervals for the abundance–ratio of the monoisotopic peak to the first to fifth isotope peak as described in Section 4.3. For each database and each isotope peak, we compute multiple quantiles in order to define confidence intervals with different confidence levels. We validate isotope clusters on basis of mass–specific confidence intervals of peak abundance–ratios as described in Section 4.2.

We exemplarily examine the interval size and magnitude of the computed confidence intervals of isotope ratios. A small interval size indicates a small range of observed isotope ratios for the analyzed substances and allows a precise definition of valid isotope ratios, whereas a large interval size indicates a diverse range of observed isotope ratios for the analyzed substances and requires a loose definition of valid isotope ratios. If the interval size and magnitude of the computed confidence intervals depends on the mass range, then mass–specific confidence intervals can increase the specificity of isotope cluster validation.

See Figure 5 for the 95% confidence interval of the ratios of the monoisotopic peak to the first; second, and third isotope peak for the database KEGG with a mass window size of 50 dalton. The ratio of the monoisotopic peak to the first isotope peak depends on the abundance of the first isotope peak, which is dominated by the proportion of ^{13}C. This results in a relatively narrow confidence interval, because the variation of the number of carbon atoms is limited within a 50 dalton mass window. The ratio of the monoisotopic peak to the second isotope peak depends on the abundance of the second isotope peak, which is dominated by the proportion of ^{13}C and ^{34}S. The 97.5%-quantile and the 50%-quantile are higher compared to the case of the first isotope peak because the second isotope

DOI:10.3390/metabo6040037

peak has typically a lower abundance than the first isotope peak. In contrast, the 2.5%-quantile is smaller compared to the case of the first isotope peak because a subset of compounds comprises at least one sulfur (partially also chlorine or bromine) with a high abundance of ^{34}S (or ^{37}Cl, ^{81}Br) causing a relatively high abundance of the second isotope peak and thus a small ratio of the monoisotopic peak to the second isotope peak. This results in a relatively large confidence interval. The ratio of the monoisotopic peak to the third isotope peak mainly depends on the abundance of the third isotope peak, which is dominated by the proportion of ^{13}C and ^{34}S (and ^{37}Cl, ^{81}Br). This results in a relatively large confidence interval analogous to the case of the second isotope peak. The quantiles are higher compared to the case of the second isotope peak because the third isotope peak has typically a lower abundance compared to the second isotope peak. We find that the magnitude of the quantiles substantially depends on the mass of the substances. Specifically, the quantiles are typically inversely proportional to the substance mass. For example, in case of the mass interval 200 to 250 dalton versus the mass interval 800 to 850 dalton the 50%-quantiles deviate by a factor of 3.5 in case of the ratio of the monoisotopic peak to the first isotope peak, by a factor of 8.4 in case of the ratio of the monoisotopic peak to the second isotope peak, and by a factor of 25.6 in case of the ratio of the monoisotopic peak to the third isotope peak. This finding suggests that mass–specific confidence intervals can indeed increase the specificity of isotope cluster validation. See Figure C1 in Appendix C for an overview of all computed quantiles and the resulting symmetric confidence intervals of the ratio of the monoisotopic peak to the first isotope peak for the database PubChem with a mass window size of 50 dalton.



(**a**) (**b**)

(**c**)

Figure 5. 95% confidence interval of the ratio of the monoisotopic peak to the first (**a**), second (**b**), and third isotopic peak (**c**) of all compounds in KEGG for different compound masses arranged in mass windows of size 50 dalton. We plot the 50%-quantile in green, the 2.5%-quantile in blue, and the 97.5%-quantile in red and we emphasize the enclosed 95% confidence interval in grey. The ratios decrease with increasing compound mass reflecting the increasing proportion of isotopic atoms.

DOI:10.3390/metabo6040037

2.5. Exemplary Isotope Cluster Detection

We exemplify the detection of isotope clusters for selected substances to demonstrate the proposed isotope detection routine without isotope cluster validation $\mathrm{IDR}_{\mathrm{NewNoVal}}$ and the isotope detection routine with mass–specific isotope cluster validation $\mathrm{IDR}_{\mathrm{NewVal}}$. We simulate the mass and relative intensity of the monoisotopic peak and the first five isotope peaks of six substances with enviPat [37] in centroid mode with a resolution of 10,000, namely (i) aspartic acid which has a low mass and comprises only the elements CHNO (see Table 1 for details); (ii) cysteine which has a low mass and comprises sulfur; (iii) chloramphenicol which has a low mass and comprises chlorine; (iv) digoxigenin monodigitoxoside which has a medium mass and comprises only the elements CHNO; (v) 2-Chloro-2′-deoxyadenosine-5′-triphosphate which has a medium mass and comprises chlorine; and (vi) autoinducer-2 which has a low mass and contains boron. The isotopic fine structure of these substances is not detectable at this resolution and hence each simulated peak is a mixture of multiple peaks from the isotopic fine structure. We only include isotope peaks with an abundance of at least 0.01% of the abundance of the monoisotopic peak which results in isotope clusters of size 4, 5, 6, 6, 6, and 6 respectively.

For each isotope cluster, we calculate the minimal absolute mass error Δm^{abs} in units of dalton and the minimal relative mass error Δm^{ppm} in units of PPM which are required for a successful isotope cluster detection. The incorporation of a mass error is necessary because the mass differences between individual isotope peaks depend on the elemental composition and hence deviates from the default mass difference of ^{13}C isotopes. It is possible to use only one of both parameters or a combination of both parameters to enable the detection of isotope clusters (see Equation (2) in Section 4.2).

We merge all six isotope clusters resulting in a single synthetic spectrum comprising 33 peaks. We apply the isotope detection routines $\mathrm{IDR}_{\mathrm{NewNoVal}}$ and $\mathrm{IDR}_{\mathrm{NewVal}}$ as described in Section 4.2 to the synthetic spectrum. We evaluate whether the isotope detection routines are able to assemble the original isotope clusters.

In Table 1 we show the results. We find that $\mathrm{IDR}_{\mathrm{NewNoVal}}$ is able to detect all six isotope clusters provided that a sufficiently large mass error is set (e.g., $\Delta m^{abs} = 0.01$). In case of a smaller mass error (e.g., $\Delta m^{abs} = 0.005$) we find that isotope clusters become split at isotope peaks which are dominated by the isotopes of sulfur, chlorine, or boron, i.e., the second isotope peak of substance (ii); the second and fourth isotope peak of substance (iii); the second isotope peak of substance (v); and the first isotope peak of substance (vi). We find that $\mathrm{IDR}_{\mathrm{NewVal}}$is able to validate all but one isotope cluster. The first peak of the boron-containing substance (vi) is not included in the isotope cluster, because the abundance of this peak is too small relative to the space of biological substances of this mass. Hence, the excluded peak is assumed to be a potential hydrogen-loss. However, this isotope cluster can be correctly identified without validation or with specialized approaches [31].

DOI:10.3390/metabo6040037

Table 1. Isotope cluster detection exemplified for six substances. We show the substance name, the sum formula, the mass of the monoisotopic peak and the first five isotope peaks (rounded to five digits), the mass difference to the monoisotopic peak (Δm, rounded to five digits), the relative peak intensity (Int., normalized to 100 and rounded to two digits), the absolute m/z error Δm^{abs} and the relative m/z error in ppm Δm^{ppm} for a successful isotope cluster detection (Δm^{abs} is rounded to five digits and Δm^{ppm} is rounded to one digit), whether the isotope cluster assignment using the isotope detection routine without isotope cluster validation $\text{IDR}_{\text{NewNoVal}}$ is successful or not (No val., "+"/"–"), and whether the isotope cluster assignment using the isotope detection routine with mass–specific isotope cluster validation $\text{IDR}_{\text{NewVal}}$ is successful or not (Val., "+"/"–"). $\text{IDR}_{\text{NewNoVal}}$ is able to detect the isotope clusters of all substances and $\text{IDR}_{\text{NewVal}}$ successfully validates the isotope clusters of all but one substance.

Substance Name	Sum Formula	Mass	Δm	Int.	Δm^{abs}	Δm^{ppm}	No Val.	Val.
Aspartic acid	$C_4H_7NO_4$	133.037508		100.00	0.00191	14.3	+	+
		134.040468	1.00296	4.96			+	+
		135.041918	2.00441	0.93			+	+
		136.044728	3.00722	0.04			+	+
Cysteine	$C_3H_7NO_2S$	121.019749		100.00	0.00895	73.9	+	+
		122.021976	1.00223	4.59			+	+
		123.016385	1.99664	5.05			+	+
		124.019165	2.99942	0.19			+	+
		125.018404	3.99866	0.03			+	+
Chloramphenicol	$C_{11}H_{12}Cl_2N_2O_5$	322.012327		100.00	0.00913	28.4	+	+
		323.015369	1.00304	13.00			+	+
		324.009595	1.99727	66.20			+	+
		325.012562	3.00024	8.53			+	+
		326.007250	3.99492	11.54			+	+
		327.010016	4.99769	1.45			+	+
Digoxigenin monodigitoxoside	$C_{29}H_{44}O_8$	520.303618		100.00	0.00078	1.5	+	+
		521.307027	1.00341	32.24			+	+
		522.309803	2.00619	6.70			+	+
		523.312531	3.00891	1.04			+	+
		524.315166	4.01155	0.13			+	+
		525.317742	5.01412	0.01			+	+
2-Chloro-2′-deoxyadenosine-5′-triphosphate	$C_{10}H_{15}ClN_5O_{12}P_3$	524.961858		100.00	0.00817	15.6	+	+
		525.964411	1.00255	13.30			+	+
		526.959596	1.99774	35.41			+	+
		527.962023	3.00017	4.63			+	+
		528.963673	4.00182	1.11			+	+
		529.966017	5.00416	0.12			+	+
Autoinducer-2	$C_5H_{10}BO_7$	192.055590		24.37	0.00689	35.9	+	–
		193.052059	0.99647	100.00			+	+
		194.055706	2.00012	6.13			+	+
		195.056530	3.00094	1.59			+	+
		196.059851	4.00426	0.09			+	+
		197.060963	5.00537	0.01			+	+

3. Discussion

Aiming at the exhaustive detection and precise validation of isotope clusters we propose an additional targeted peak picking step with predicted isotope ROIs and the mass–specific validation of putative isotope clusters based on database statistics. Compromising between peak reliability and

DOI:10.3390/metabo6040037

exhaustive detection we use a relaxed signal-to-noise of 6.25 threshold for predicted isotope ROIs and achieve an increase of +37.6% isotope peaks and +102.8% PPS. We use this relaxed signal-to-noise threshold by default in the freely available implementation of this algorithms in the R package xcms. The targeted peak picking with predicted isotope ROIs can easily be adapted in other tools such as *MZmine2* [38], *apLCMS* [39], and related approaches [40]. The validation of putative isotope clusters in combination with predicted isotope ROIs results in the highest number of correctly predicted molecular formulas and also the highest number of correct molecular formulas among the first three ranks. However, the ranks of correctly predicted molecular formulas were robust with respect to different approaches for peak picking and isotope cluster detection and it is challenging to improve upon. We exemplify the use of the proposed isotope detection routine with and without mass–specific isotope cluster validation and find that it is possible to detect substances with and without biologically unusual elements using an absolute mass error of 0.01 dalton. Consequently, we use this absolute mass error by default in the freely available implementation of these algorithms in the R package *CAMERA*.

The enhanced isotope cluster detection and validation presented in this work could improve the accuracy of substance quantification. All isotope peaks of one isotope cluster originate from the same substance and we point out that the consideration of a greater number of features from a certain substance—although small and noisy—reduces the technical variance in the data. In turn, this would enhance the precision and yield of comparative analyses, because a reduced data variance would not only improve calculated fold changes but would enable the statistically valid detection of smaller effect sizes. The slight improvement in molecular formula prediction could affect a considerable number of substances in case of metabolome-scale metabolite identification studies. Especially in untargeted metabolomics reliable hints for metabolite identification are urgently needed.

4. Materials and Methods

We present the methodology of the proposed approach and the used data for evaluation. Specifically, we describe (i) the targeted peak picking with predicted isotope ROIs; (ii) the detection and mass–specific validation of isotope clusters; (iii) the computation of isotope ratio quantiles; and (iv) two sets of mass spectrometry raw data.

4.1. Targeted Peak Picking with Predicted Isotope ROIs

A requirement for the prediction of isotope ROIs is a set of peaks that have been detected previously. This initial peak picking can be accomplished by one of the numerous peak picker which are available [1,18,38]. In untargeted approaches, these peak picker typically do not use any prior knowledge and we refer to this kind of peak picking as *traditional peak picking*. We propose the following approach for the targeted detection of isotope peaks. This approach is designed for liquid chromatography–high resolution mass spectrometry data and does not consider the isotopic fine structure available with ultrahigh resolution mass spectrometry.

Given a set of detected peaks from traditional peak picking, a maximum charge $Z = 3$, and a maximum number of isotopes $I = 5$ we predict putative isotope ROIs as follows. For each charge state $z \in \{1, ..., Z\}$ and for each isotope number $i \in \{1, ..., I\}$, we compute the theoretical m/z distance to the monoisotopic peak

$$d_{z,i} = \frac{i * \Delta m}{z}, \tag{1}$$

where $\Delta m = \text{mass}(^{13}\text{C}) - \text{mass}(^{12}\text{C}) \approx 1.003355$. We use Δm as an approximation for the mass difference between successive peaks in isotope clusters because the isotopic nuclide ^{13}C has usually the largest impact on isotope clusters in biological samples. Other isotopic nuclides such as ^{15}N, ^{18}O, and ^{34}S cause isotope peaks with mass differences which can only be discriminated from ^{13}C-isotope peaks using mass spectrometers with resolution above 40,000 (in case of ions with an m/z of 500 dalton). For each peak detected by traditional peak picking we predict for each charge state z and for each isotope number i one putative isotope ROI. Each putative isotope ROI is composed of the retention time

DOI:10.3390/metabo6040037

interval of the detected peak and the m/z interval of the detected peak shifted by $d_{z,i}$ as exemplified in Figure 6. An additional targeted peak picking is performed based on the set of predicted isotope ROIs using a relaxed signal-to-noise threshold `snthr'` $=$ `snthr` $* r/100$, where `snthr` is the signal-to-noise threshold for traditional peak picking and $r \in \{100, 95, ..., 5\}$. Subsequently, the peak table from traditional peak picking and the peak table from the targeted peak picking on basis of putative isotope ROIs are merged and redundant peaks are removed.

For control purposes, we generate a set of noise ROIs given the set of predicted isotope ROIs as follows. To approximate the distribution of the predicted isotope ROIs in the m/z dimension and the retention time (RT) dimension, we calculate the minimum and maximum m/z and RT of the predicted isotope ROIs and use a uniform distribution in the calculated intervals of both dimensions. To approximate the distribution of peak widths in m/z and RT we calculate a histogram of peak widths in m/z relative to the peak m/z and a histogram of peak widths in RT. For each predicted isotope ROI we sample one new noise ROI which m/z and RT is uniformly drawn within the calculated ranges in m/z and RT and which peak width in m/z and RT is drawn from the calculated histograms. Subsequently, targeted peak picking is applied to the set of noise ROIs using a relaxed signal-to-noise threshold `snthr'` analog to predicted isotope ROIs and the results from traditional peak picking and targeted peak picking on basis of noise ROIs are merged as before.

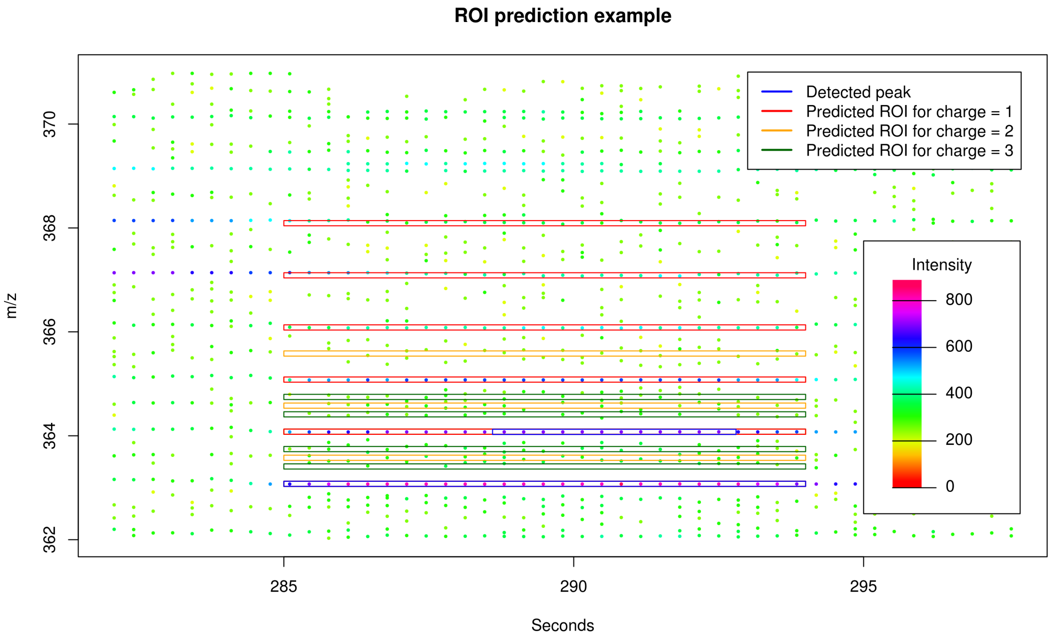


Figure 6. Exemplary section of LC-MS raw data. We mark two detected peaks from traditional peak picking in blue and 12 predicted isotope ROIs in red, orange, and green calculated on basis of the (monoisotopic) peak (apex $m/z \approx 363.075$ dalton / retention time ≈ 291 seconds) given a maximum isotope number $I = 5$ and a maximum charge state $Z = 3$. Via prediction of isotope ROIs, we are able to expand the region of the already detected first isotope peak and to encompass the signals of the second, third, fourth, and fifth isotope peak. Here, the subsequent peak picking procedure will not find relevant signals for the predicted isotope ROIs corresponding to the charge states 2 (orange) and 3 (green) and will reject these accordingly.

4.2. Detection and Mass–Specific Validation of Isotope Clusters

We propose an approach for the detection and validation of isotope clusters in liquid chromatography–high resolution mass spectrometry data which does not resolve the isotopic fine structure. In this approach we detect putative isotope clusters based on characteristic distances in the m/z dimension. We validate putative isotope clusters depending on the substance mass and we refer to this validation as *mass–specific validation*. We detect and validate isotope clusters given a set of coeluting features, a maximum charge $Z = 3$, a relative m/z error in ppm Δm^{ppm}, and an absolute m/z error Δm^{abs} as follows.

First, we detect putative isotope clusters. For each charge state $z \in [1, Z]$, we mark all pairs of peaks (p_1, p_2) for which

$$\delta_{z,p_1,p_2} = ||mass(p_1) - mass(p_2)| - \Delta m/z| \leq \max\left(\frac{mass(p_1) * \Delta m^{ppm}}{10^6}, \Delta m^{abs}\right) \quad (2)$$

holds, where $\Delta m = \text{mass}(^{13}C) - \text{mass}(^{12}C) \approx 1.003355$ is the expected distance between two isotope peaks (cf. Section 4.1). For each charge state and for each peak p, we compute all putative isotope clusters $(p_1, p_2, ..., p_n)$ for which $\delta_{c,p',p''}$ holds for each successive pair of peaks (p', p''). We retain the putative isotope cluster with the maximum number of peaks and remove the peaks of this putative isotope cluster from the set of available peaks. We iteratively perform the last steps with the remaining peaks until there are no putative isotope clusters with at least two peaks left.

Second, we validate the set of putative isotope clusters which have been extracted previously depending on the monoisotopic mass. See Figure 3 for four cases which necessitate the following validation of putative isotope clusters. For each putative isotope cluster $(p_1, p_2, ..., p_n)$ we examine the second to last peak $p' \in (p_2, ..., p_n)$. For each peak p' we compute the ratio of the abundance of the monoisotopic peak p_1 and the abundance of peak p'. Specifically, we compute the minimum and maximum ratio considering that the abundance estimates of both peaks are affected by the ubiquitous noise using an estimate of the signal-to-noise ratio of both peaks. If the computed interval of ratios does not overlap with the 99% confidence interval derived from the KEGG database for the current monoisotopic mass (mass window size 50) we split the putative isotope cluster. In this case we turn the peak p' into the new monoisotopic peak resulting in a new putative isotope cluster $(p', ..., p_n)$ which is validated as well. We retain all putative isotope clusters which comprise at least two peaks and consider these as validated isotope clusters.

4.3. Isotope Ratio Quantiles

We perform isotope statistics for each of the databases ChEBI, KEGG, KNApSAcK, LIPID MAPS, and PubChem as follows [32–36]. We iterate all compounds, compute the exact mass and the theoretical isotope cluster from the molecular formula, and record the ratio of the monoisotopic peak to the first to fifth isotope peak. We group all compounds by the exact mass in consecutive mass windows for each of the mass window sizes 10, 25, 50, 100, and 250 dalton to support different compromises between mass specificity and quantile robustness. For each mass window size, each mass window, and each isotope peak (1st–5th) we compute the isotope ratio for several p-quantiles, where $p \in \{5.0 \times 10^{-6}$, 0.999995, 1.0×10^{-5}, 0.99999, 5.0×10^{-5}, 0.99995, 1.0×10^{-4}, 0.9999, 5.0×10^{-4}, 0.9995, 0.001, 0.999, 0.005, 0.995, 0.01, 0.99, 0.025, 0.975, 0.05, 0.95, 0.1, 0.9, 0.5$\}$. For each mass window size and each isotope peak we record the isotope ratio in a matrix with one row for each p-quantile and one column for each mass window. We encapsulate the resulting data for each database, each mass window size, and each isotope peak in an R object of class *S4* named `compoundQuantiles`. This implementation supports a simple API for convenient retrieval of the data (see documentation of package *CAMERA* version 1.50.0 for details). Based on this implementation, it is also possible to compute isotope ratios amongst isotope peaks, e.g., the confidence interval of the isotope ratio between the third isotope peak and the fifth isotope peak for a given mass range.

DOI:10.3390/metabo6040037

4.4. Data Sets

4.4.1. MM48

We perform a case study based on a gold standard data set comprising 11 LC-MS measurements (UPLC-ESI-QTOF-MS, positive mode) each of a solution of 48 known reference substances denoted as *MM48*. The raw data is available in MetaboLights [41] accession MTBLS381 in Supplementary Materials link. This set of compounds was also used in [24] and the measurements have been deposited in MetaboLights accession MTBLS188. We compile a ground truth of detectable ions as follows. First, we assume a set of three expected ions ($[M]^+$, $[M + H]^+$, $[M + Na]^+$) as well as isotope peaks up to the fifth isotope peak (i.e., $[M + 1]^+$, $[M + 2]^+$, $[M + 3]^+$, $[M + 4]^+$, and $[M + 5]^+$ in case of the $[M]^+$ ion) for each compound and calculate the exact mass of these 18 molecular formulas (three ions each with an isotope cluster with six peaks); Second, we check the abundance of these ions in the 11 data sets and define all ions with a peak area of at least 1000 counts within a retention time interval of at most five seconds as measurable ions constituting the ground truth. Considering the set of ions which are measurable in at least six of 11 data sets, we detect 72 monoisotopic ions (see Figure 7), 63 isotope clusters with at least two ions, and 190 ions in total.

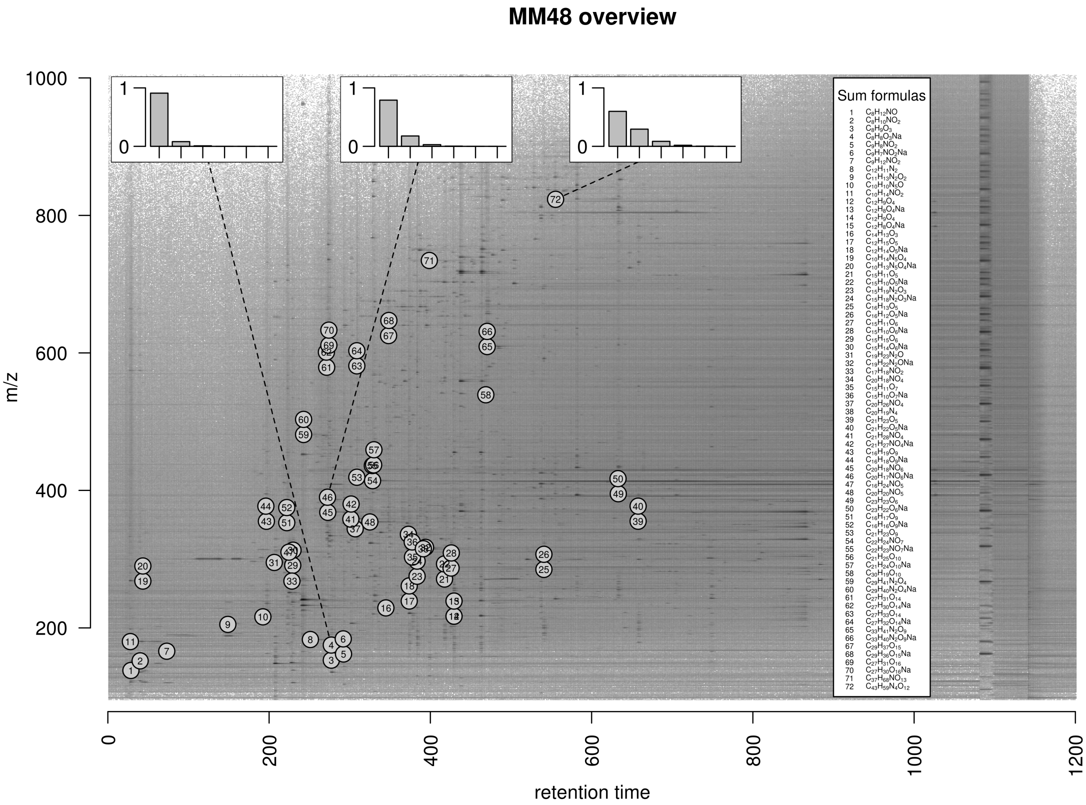


Figure 7. Overview of monoisotopic measurable ions in the MM48 data set. We plot the logarithmic raw data intensities in the dimensions mass-to-charge ratio (m/z) and retention time and mark the location of 72 monoisotopic ions which are measurable in at least six of eleven data sets. In case of three ions with exact mass 175.037, 390.095, and 823.413 dalton, we exemplarily plot the theoretical relative intensities of the monoisotopic peak and the first to fifth isotope peak in the insets at the top. The set of measurable ions spans a huge range in both dimensions with different isotope clusters constituting a diverse basis for validation purposes.

DOI:10.3390/metabo6040037

4.4.2. Dilution Series

We perform a case study based on 40 LC-MS measurements (UPLC-ESI-QTOF-MS, positive mode), which is a subset of the data used in [24] and is available from the MetaboLights repository with accession MTBLS188. This set of measurements is composed of a dilution series varying the ratio of solution and leaf sample. Specifically, the ratio of solution and leaf sample is 0:100, 25:75, 50:50, and 75:25 in 10 data sets each. This experimental design implies a diverse range of cases in the data regarding the signal-to-noise ratio of peaks and constitutes the basis to test the detection of weak signals like isotope peaks.

5. Conclusions

We implemented the targeted peak picking with predicted isotope ROIs in combination with the *centWave* algorithm as part of the R package *xcms* in version 1.50.0 (functions `findPeaks.centWaveWithPredictedIsotopeROIs` and `findPeaks.addPredictedIsotopeFeatures`). We implemented the mass–specific validation of putative isotope clusters as part of the R package *CAMERA* in version 1.30.0 (function `findIsotopesWithValidation`).

Supplementary Materials: The following are available online at www.ebi.ac.uk/metabolights/MTBLS381, 11 MM48 raw data files used for performance evaluation in the manuscript.

Acknowledgments: The open access fee was funded by the Deutsche Forschungsgemeinschaft (DFG funding No. NE 1396/5-1). The authors would like to thank Carsten Kuhl and Christoph Böttcher for providing the UPLC-ESI-QTOF-MS data and Sarah Scharfenberg for refining the manuscript.

Author Contributions: Hendrik Treutler, and Steffen Neumann conceived and designed the methodology; Hendrik Treutler performed the case studies; Hendrik Treutler wrote the paper. Hendrik Treutler, and Steffen Neumann read and approved the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Isotope Cluster Detection and Validation: Extended Results

We compare the proposed isotope detection routine with mass–specific isotope cluster validation ($\text{IDR}_{\text{NewVal}}$) against the isotope detection routine without isotope cluster validation ($\text{IDR}_{\text{NewNoVal}}$), the isotope detection routine implemented in the *AStream* package ($\text{IDR}_{\text{AStream}}$) [29], the isotope detection routine implemented in the *CAMERA* package ($\text{IDR}_{\text{CAMERA}}$) [24], and the isotope detection routine implemented in the *mzMatch* package ($\text{IDR}_{\text{mzMatch}}$) [30].

We evaluate the performance of the isotope cluster detection and validation described in Section 4.2 on a dilution series experiment with 40 LC-MS measurements described in Section 4.4. We quantify the performance using the performance measures (i) number of detected peaks; (ii) number of detected isotope peaks; (iii) number of detected isotope clusters; and (iv) isotope coverage, i.e., the ratio of the number of detected isotope peaks and the number of all detected peaks. We compute each performance measure as a function of the relaxed signal-to-noise threshold `snthr'` $\in \{100, 95, ..., 5\}$ % $*$ `snthr`, where `snthr` $=$ 25 is the signal-to-noise threshold of the traditional peak picking step. In the Section 2.3 we show an excerpt of these results, i.e., we present the results for each isotope detection routine with predicted isotope ROIs relative to the results of $\text{IDR}_{\text{CAMERA}}$ without predicted isotope ROIs in Figure 4.

In Figure A1 we show the performance measures for $\text{IDR}_{\text{NewVal}}$, $\text{IDR}_{\text{NewNoVal}}$, $\text{IDR}_{\text{AStream}}$, $\text{IDR}_{\text{CAMERA}}$, and $\text{IDR}_{\text{mzMatch}}$.

DOI:10.3390/metabo6040037

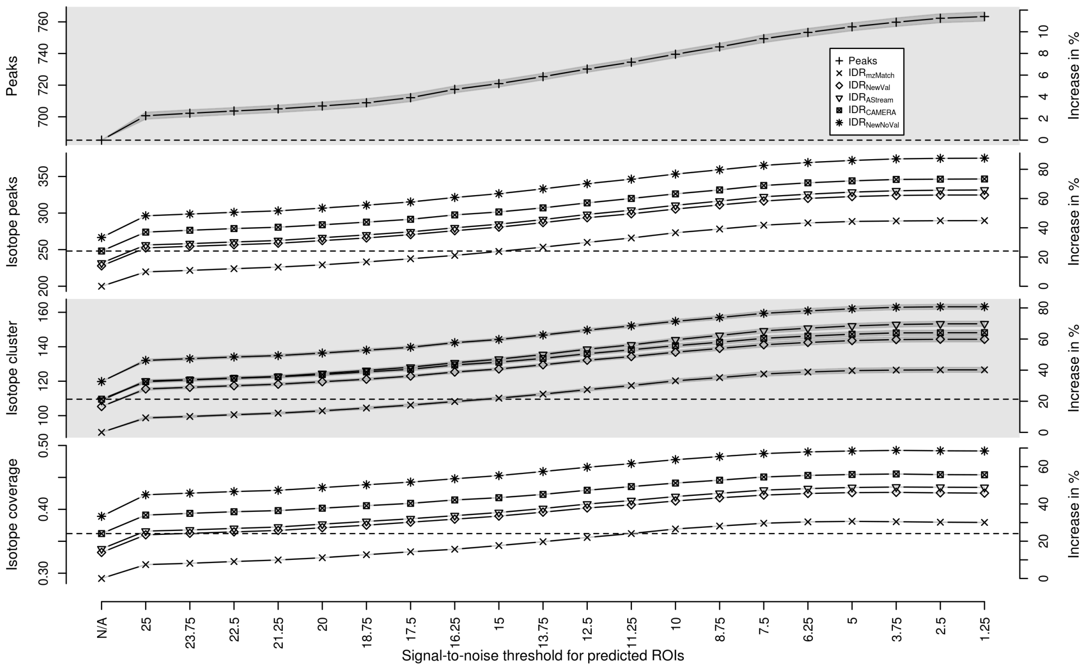


Figure A1. Evaluation of predicted isotope ROIs in combination with validated isotope clusters for varying relaxed signal-to-noise threshold `snthr'`. We plot the mean (solid line) and the standard error of the mean (SEM, interval in dark grey) of the performance measures (**i**) number of detected peaks; (**ii**) number of detected isotope peaks; (**iii**) number of detected isotope clusters; and (**iv**) isotope coverage. We plot the performance of each measure without additional ROIs in the first column ("N/A") as reference value (horizontal dashed line). All four measures of all isotope detection routines increase with decreasing signal-to-noise threshold `snthr'`.

Appendix B. Prediction of Molecular Formulas From Isotope Clusters

In order to study to which degree the proposed approach is capable of improving the detection and validation of isotope clusters, we test the quality of predicted molecular formulas. The prediction of molecular formulas is an important step towards the identification of substances and can be done automatically on the basis of isotope clusters. We use 11 LC-MS measurements with 48 known compounds and select a set of 72 ions. We predict for each ion a list of ranked molecular formula candidates using SIRIUS and evaluate the rank of the correct molecular formula [3].

We evaluate the performance of predicted isotope ROIs described in Section 4.1 and the isotope detection routine with mass–specific isotope cluster validation described in Section 4.2 on 11 LC-MS measurements of known compounds described in Section 4.4 using predicted molecular formulas from SIRIUS as described in the Appendix D.4. We quantify the performance using the number of compounds with a certain rank averaged over all measurements. If the proposed approaches increase the quality of detected isotope clusters, then the rank of the predicted molecular formulas should decrease and be ranked first in the ideal case. We compare different combinations of two peak picking approaches and five isotope detection routines, namely (iA) the traditional peak picking and (iB) the traditional peak picking in combination with targeted peak picking with predicted isotope ROIs (see Section 4.1) and (iiA) the isotope detection algorithm from *AStream*; (iiB) the isotope detection algorithm from *mzMatch*; (iiC) the isotope detection algorithm from *CAMERA*; (iiD) the proposed isotope detection algorithm without isotope cluster validation; and (iiE) the proposed isotope detection algorithm with mass–specific isotope cluster validation resulting in ten combinations of algorithms (see Section 4.2 and the Appendix D). In Table B1 we show the ranks of the predicted molecular formulas for ten algorithms averaged over 11 data sets.

Table B1. Molecular formula prediction from isotope clusters. Using SIRIUS we predict molecular formulas from isotope clusters which have been detected using different algorithms. In the first column we indicate whether we use targeted peak picking with predicted isotope ROIs ('+') or not ('−') and in the second column we indicate the isotope detection algorithm ($\mathrm{IDR}_{\mathrm{AStream}}$ for the algorithm implemented in R package *AStream*, $\mathrm{IDR}_{\mathrm{CAMERA}}$ for the algorithm implemented in R package *CAMERA*, $\mathrm{IDR}_{\mathrm{mzMatch}}$ for the algorithm implemented in R package *mzMatch*, $\mathrm{IDR}_{\mathrm{NewNoVal}}$ for the proposed isotope detection algorithm without isotope cluster validation, and $\mathrm{IDR}_{\mathrm{NewVal}}$ for the proposed isotope detection algorithm with mass–specific isotope cluster validation). We specify the number of ions with a molecular formula on rank 1, on rank 2, on rank 3, between rank 4 and rank 10, on a rank above 10, the number of ions which molecular formula is not among the top 1000 candidates ('No rank'), and the number of ions which have not been detected during peak picking ('No peak'). We arranged the isotope detection algorithms by the number of ions with molecular formula on rank 1.

Predicted Isotope ROIs	Isotope Detection Algorithm	Rank 1	Rank 2	Rank 3	$3 < \text{Rank} \leq 10$	Rank > 10	No Rank	No Peak
−	$\mathrm{IDR}_{\mathrm{mzMatch}}$	48.82	11.55	1.18	3.36	0	4.64	2.45
+	$\mathrm{IDR}_{\mathrm{mzMatch}}$	48.18	12	1.18	3.36	0	4.82	2.45
−	$\mathrm{IDR}_{\mathrm{NewNoVal}}$	49.09	10.91	0.91	1.55	0	7.09	2.45
+	$\mathrm{IDR}_{\mathrm{NewNoVal}}$	49.36	11.18	0.73	1.64	0	6.73	2.36
−	$\mathrm{IDR}_{\mathrm{AStream}}$	52.82	11.27	1.09	1.82	0	2.55	2.45
+	$\mathrm{IDR}_{\mathrm{AStream}}$	53.27	11.55	0.55	1.91	0	2.36	2.36
−	$\mathrm{IDR}_{\mathrm{CAMERA}}$	53.73	10.27	0.82	1.55	0	3.18	2.45
+	$\mathrm{IDR}_{\mathrm{CAMERA}}$	52.82	11	0.64	1.64	0	3.55	2.36
−	$\mathrm{IDR}_{\mathrm{NewVal}}$	53.82	11.09	1	1.55	0	2.09	2.45
+	$\mathrm{IDR}_{\mathrm{NewVal}}$	54.09	11.36	0.73	1.64	0	1.82	2.36

Appendix C. Isotope Cluster Statistics: Full Quantile Set for PubChem

In Figure C1 we depict all computed quantiles and the resulting symmetric confidence intervals of the isotope ratio of the monoisotopic peak to the first isotope peak for the database PubChem with a mass window size equal to 50 dalton. See Section 4.3 for a detailed description of the database statistics.

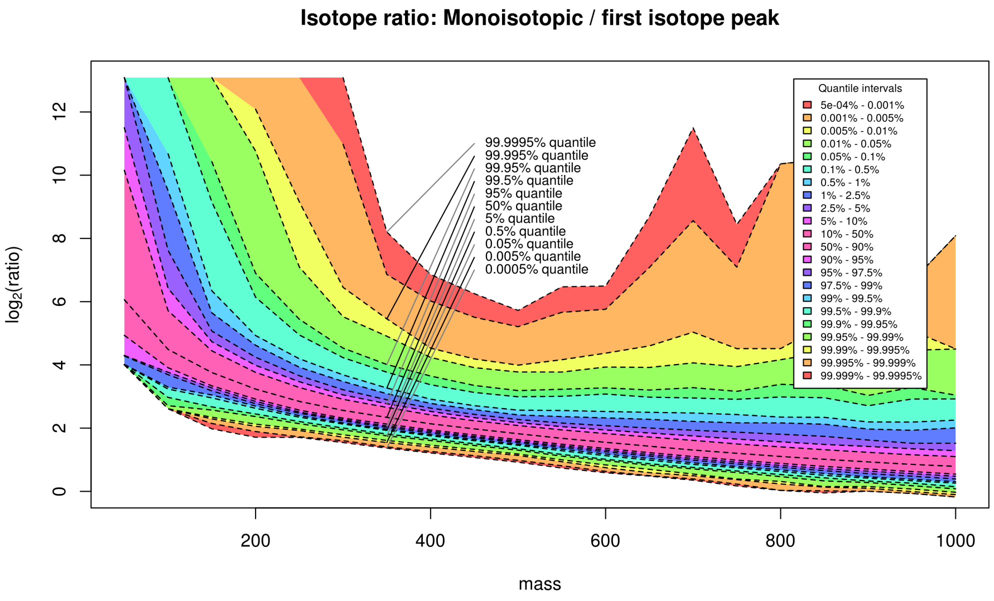


Figure C1. The full set of 23 quantiles of the monoisotopic peak versus the first isotopic peak for the PubChem database for different compound masses arranged in mass windows of size 50 dalton. We emphasize the enclosed confidence intervals with different colors.

DOI:10.3390/metabo6040037

Appendix D. Software Versions and Processing Parameters

Tools versions, used functions, and parameters of *xcms/CAMERA*, *AStream*, *mzMatch*, and *SIRIUS* are given subsequently.

Appendix D.1. xcms/CAMERA

We use the R package *xcms* version 1.44.0 [23] and the R package *CAMERA* version 1.27.0 [24] for peak picking using *centWave* [1], the grouping of features into pseudospectra, and the detection of isotope clusters. We processed the raw data of each LC-MS measurement individually as follows. We performed peak picking with the *centWave* algorithm with parameters `peakwidth` $= (5, 12)$, `prefilter` $= (2, 200)$, `ppm` $= 10$, and `snthr` $= 25$. We use a signal-to-noise ratio of 25, because it has been shown that this ratio yields reliable molecular formula predictions from mass spectrometry data [42]. Subsequently, we group detected peaks by retention time into pseudospectra-groups using function *groupFWHM* with `perfwhm` $= 1$ and standard parameters and detect isotope clusters using function *findIsotopes* with `intensityValue` = 'intb' and standard parameters.

Appendix D.2. AStream

We use the R package *AStream* version 2.0 [29] for the detection of isotope clusters. We import the peaks which have been detected using *xcms* into the *AStream* datalist structure. We apply the function `data.norm` with the parameters `mz.tol` $= 0.005$ (the mean m/z error for `ppm` $= 10$ as used in *xcms* and *mzMatch*) and we detect isotope clusters using function `isotope.search` with the parameter `mz.tol` $= 0.005$. In a postprocessing step we remove contradictory isotope annotations, i.e., if (i) peak B is annotated as [M + 1] isotope peak of peak A and (ii) peak C is annotated as [M + 2] isotope peak of peak A and (iii) peak C is annotated as [M + 1] isotope peak of peak B; then we remove annotation (iii).

Appendix D.3. mzMatch

We use the R package *mzmatch.R* version 2.0-13 [30] for the detection of isotope clusters. We import the peaks which have been detected using *xcms* via the peakML file format used by *mzMatch* using the function `PeakML.xcms.write.SingleMeasurement` with the parameters `writeRejected` = TRUE, `ppm` = 10, `addscans` = 0, and `ApodisationFilter` = FALSE. We convert this data using function `mzmatch.ipeak.Combine` and we detect isotope clusters using function `mzmatch.ipeak.sort.RelatedPeaks` with the parameters `ppm` = 10 and `rtwindow` = 50. In a postprocessing step we remove all isotope clusters with gaps, i.e., the isotope cluster with monoisotopic peak [M] and isotope peak [M + 2] without the [M + 1] isotope peak is considered non-evaluable and removed from the output. Approximately 10% of the isotope annotations are removed in this way.

Appendix D.4. Prediction of Molecular Formulas Using SIRIUS

We predict ranked candidate lists from isotope clusters using command–line SIRIUS [3] version 3.1.3. We use the parameters `-elements` = *CHNOPS*, `-isotope` = *score*, `-candidates` = 1000, `-ppm-max` = 10, and `-profile` = $qtof$ and give the ion species (`-ion`), the monoisotopic m/z (`-mz`), the (m/z, intensity) pairs (`into` intensity from *xcms*; `-ms1`), and an empty MS/MS spectrum (`-ms2`) as input. We rank the resulting candidate lists according to the `tree` score and select the rank of the correct molecular formula.

References

1. Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinform.* **2008**, *9*, 504, doi:10.1186/1471-2105-9-504.
2. Trutschel, D.; Schmidt, S.; Grosse, I.; Neumann, S. Joint Analysis of Dependent Features within Compound Spectra Can Improve Detection of Differential Features. *Front. Bioeng. Biotechnol.* **2015**, *3*, doi:10.3389/fbioe.2015.00129.

DOI:10.3390/metabo6040037

3. Böcker, S.; Letzel, M.C.; Lipták, Z.; Pervukhin, A. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics* **2009**, *25*, 218–224.
4. Dührkop, K.; Hufsky, F.; Böcker, S. Molecular Formula Identification Using Isotope Pattern Analysis and Calculation of Fragmentation Trees. *Mass Spectrom.* **2014**, *3*, doi:10.5702/massspectrometry.S0037
5. Stoll, N.; Schmidt, E.; Thurow, K. Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1692–1699.
6. Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinform.* **2006**, *7*, 234, doi:10.1186/1471-2105-7-234.
7. Zhang, J.; Gao, W.; Cai, J.; He, S.; Zeng, R.; Chen, R. Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2005**, *2*, 217–230.
8. Ipsen, A.; Want, E.J.; Ebbels, T.M.D. Construction of Confidence Regions for Isotopic Abundance Patterns in LC/MS Data Sets for Rigorous Determination of Molecular Formulas. *Anal. Chem.* **2010**, *82*, 7319–7328.
9. Pluskal, T.; Uehara, T.; Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal. Chem.* **2012**, *84*, 4396–4403.
10. Jarussophon, S.; Acoca, S.; Gao, J.M.; Deprez, C.; Kiyota, T.; Draghici, C.; Purisima, E.; Konishi, Y. Automated molecular formula determination by tandem mass spectrometry (MS/MS). *Analyst* **2009**, *134*, 690–700.
11. Meringer, M.; Reinker, S.; Zhang, J.; Muller, A. MS/MS Data Improves Automated Determination of Molecular Formulas by Mass Spectrometry. *MATCH Commun. Math. Comput. Chem.* **2011**, *2011*, 259–290.
12. Snider, R.K. Efficient calculation of exact mass isotopic distributions. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1511–1515.
13. McLafferty, F.W.; Turecek, F. Interpretation of Mass Spectra, 4th ed. *J. Chem. Educ.* **1994**, *71*, doi:10.1021/ed071pA54.5.
14. Clendinen, C.S.; Stupp, G.S.; Ajredini, R.; Lee-McMullen, B.; Beecher, C.; Edison, A.S. An overview of methods using (13)C for improved compound identification in metabolomics and natural products. *Front. Plant Sci.* **2015**, *6*, doi:10.3389/fpls.2015.00611.
15. Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K.E.; Breitling, R. MetAssign: Probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **2014**, *30*, 2764–2771.
16. Hussong, R.; Tholey, A.; Hildebrandt, A. Efficient Analysis of Mass Spectrometry Data Using the Isotope Wavelet. In Proceedings of the 3rd International Symposium on Computational Life Science (COMPLIFE 2007), Utrecht, The Netherlands, 4–5 October 2007; Volume 940, pp. 139–149.
17. Slawski, M.; Hussong, R.; Tholey, A.; Jakoby, T.; Gregorius, B.; Hildebrandt, A.; Hein, M. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinform.* **2012**, *13*, doi:10.1186/1471-2105-13-291.
18. Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.U.U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. *Mol. Cell. Proteom. MCP* **2014**, *13*, 348–359.
19. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; et al. IPO: A tool for automated optimization of XCMS parameters. *BMC Bioinform.* **2015**, *16*, doi:10.1186/s12859-015-0562-8.
20. Ojanperä, S.; Pelander, A.; Pelzing, M.; Krebs, I.; Vuori, E.; Ojanperä, I. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1161–1167.
21. Fabregat-Cabello, N.; Sancho, J.V.; Vidal, A.; González, F.V.; Roig-Navarro, A.F.F. Development and validation of a liquid chromatography isotope dilution mass spectrometry method for the reliable quantification of alkylphenols in environmental water samples by isotope pattern deconvolution. *J. Chromatogr. A* **2014**, *1328*, 43–51.
22. Haimi, P.; Uphoff, A.; Hermansson, M.; Somerharju, P. Software tools for analysis of mass spectrometric lipidome data. *Anal. Chem.* **2006**, *78*, 8324–8331.
23. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78*, 779–787.

24. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.
25. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80, doi:10.1186/gb-2004-5-10-r80.
26. Meija, J.; Caruso, J.A. Deconvolution of isobaric interferences in mass spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 654–658.
27. Johnstone, R.A.W.; Rose, M.E. *Mass Spectrometry for Chemists and Biochemists*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1996.
28. Yamagaki, T.; Watanabe, T. Hydrogen radical removal causes complex overlapping isotope patterns of aromatic carboxylic acids in negative-ion matrix-assisted laser desorption/ionization mass spectrometry. *Mass Spectrom.* **2012**, *1*, doi:10.5702/massspectrometry.A0005.
29. Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S. AStream: An R package for annotating LC/MS metabolomic data. *Bioinformatics* **2011**, *27*, 1339–1340.
30. Scheltema, R.A.; Jankevics, A.; Jansen, R.C.; Swertz, M.A.; Breitling, R. PeakML/mzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Anal. Chem.* **2011**, *83*, 2786–2793.
31. Meusel, M.; Hufsky, F.; Panter, F.; Krug, D.; Müller, R.; Böcker, S. Predicting the Presence of Uncommon Elements in Unknown Biomolecules from Isotope Patterns. *Anal. Chem.* **2016**, *88*, 7556–7566.
32. Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **2008**, *36*, D344–D350.
33. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **1999**, *27*, 29–34.
34. Afendi, F.M.M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L.K.; et al. KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **2012**, *53*, doi:10.1093/pcp/pcr165.
35. Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E.A.; Glass, C.K.; Merrill, A.H.; Murphy, R.C.; Raetz, C.R.; Russell, D.W.; et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **2007**, *35*, D527–D532.
36. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2015**, *44*, D1202–D1213.
37. Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees. *Anal. Chem.* **2015**, *87*, 5738–5744.
38. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, doi:10.1186/1471-2105-11-395.
39. Yu, T.; Park, Y.; Johnson, J.M.; Jones, D.P. apLCMS—Adaptive processing of high-resolution LC/MS data. *Bioinformatics* **2009**, *25*, 1930–1936.
40. Woldegebriel, M.; Vivó-Truyols, G. Probabilistic Model for Untargeted Peak Detection in LC–MS Using Bayesian Statistics. *Anal. Chem.* **2015**, *87*, 7345–7355.
41. Haug, K.; Salek, R.M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. MetaboLights—An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013**, *41*, D781–D786.
42. Koch, B.P.; Dittmar, T.; Witt, M.; Kattner, G. Fundamentals of Molecular Formula Assignment to Ultrahigh Resolution Mass Data of Natural Organic Matter. *Anal. Chem.* **2007**, *79*, 1758–1763.

© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).

DOI:10.3390/metabo6040037

Metabolite Identification

8

My contributions in the area of metabolite profiling started with supervising the student Stefan Kuhn and PhD student Björn Egert, and joint work with Christoph Steinbeck on prediction of NMR spectra [KENS08].

Databases of reference spectra are the foundation of metabolite identification approaches. In 2008 I initiated a collaboration with the MassBank consortium and IPB Halle became the first MassBank server **[HAK10]** outside of Japan.

Later, I designed metabolite annotation approaches. I supervised four PhD students working on metabolite identification: Sebastian Wolf implemented the initial MetFrag system **[WSMHN10]**, and the master student Christian Hildebrandt **[HWN11]** implemented the MassStruct system. The PhD student Michael Gerlich integrated MetFrag and MassBank into MetFusion **[GN13]**. I designed and implemented the `MetShot` package **[NTB12]** for improved automated LC-MS/MS data acquisition. The PhD student Hendrik Treutler developed the MetFamily system **[TTP16]**, with strong analytical chemistry insights by Gerd Balcke. I am supervising the PhD student Christoph Ruttkies, who improved the MetFrag system **[RSW16, RNP19]**, [WRNSK17], with additional developments by Emma L. Schymanski, Michael Witting and Stefan Posch. MetFrag was evaluated in conjunction with structure generation [SGK12] and for derivatised GC-APCI-MS or deuterium exchange analytical setups [RSSN15b, RSS19] and in lipidomics [WRNSK17].

I am also supervising the PhD student Sarah Scharfenberg, who developed an approach for post-processing of MetFrag result lists to obtain information on compound classes [MBH15] with the group of Chris Steinbeck. I am mentoring the PhD student Jördis Ann-Schüler, who developed an approach similar to MetFrag that uses energy calculations from MOPAC to explain the fragmentation in MS/MS [SNMHB18].

Since 2012 I am co-organising the CASMI contest with Emma L. Schymanski, the most recent edition was CASMI 2016 **[SRK17]**. I developed the automatic setup and evaluation of the contest. In several cases tools developed by my PhD students participated in CASMI [RGN13, SGRN14], sometimes as inofficial participants.

With this expertise, I also wrote both a review on methods [NB10] and a book chapter on computational methods [NRWB13] with Sebastian Böcker and recommendations as part of the Metabolomics Society task group on metabolite identification [DEW13].

Special Feature: Tutorial

Journal of MASS SPECTROMETRY

Received: 18 March 2010 Accepted: 1 June 2010

Published online in Wiley Interscience:

(www.interscience.com) DOI 10.1002/jms.1777

MassBank: a public repository for sharing mass spectral data for life sciences

Hisayuki Horai,[a] Masanori Arita,[a–c†] Shigehiko Kanaya,[d] Yoshito Nihei,[a] Tasuku Ikeda,[a] Kazuhiro Suwa,[b] Yuya Ojima,[a] Kenichi Tanaka,[d] Satoshi Tanaka,[e,f] Ken Aoshima,[e,f] Yoshiya Oda,[e,f] Yuji Kakazu,[a] Miyako Kusano,[c] Takayuki Tohge,[c] Fumio Matsuda,[c] Yuji Sawada,[c,f] Masami Yokota Hirai,[c,f] Hiroki Nakanishi,[f,g] Kazutaka Ikeda,[f,g] Naoshige Akimoto,[h] Takashi Maoka,[i] Hiroki Takahashi,[d] Takeshi Ara,[j] Nozomu Sakurai,[j] Hideyuki Suzuki,[j] Daisuke Shibata,[j] Steffen Neumann,[k] Takashi Iida,[l] Ken Tanaka,[m] Kimito Funatsu,[n] Fumito Matsuura,[o] Tomoyoshi Soga,[a] Ryo Taguchi,[f,g] Kazuki Saito[c] and Takaaki Nishioka[a*]

MassBank is the first public repository of mass spectra of small chemical compounds for life sciences ($<$3000 Da). The database contains 605 electron-ionization mass spectrometry(EI-MS), 137 fast atom bombardment MS and 9276 electrospray ionization (ESI)-MS^n data of 2337 authentic compounds of metabolites, 11 545 EI-MS and 834 other-MS data of 10 286 volatile natural and synthetic compounds, and 3045 ESI-MS^2 data of 679 synthetic drugs contributed by 16 research groups (January 2010). ESI-MS^2 data were analyzed under nonstandardized, independent experimental conditions. MassBank is a distributed database. Each research group provides data from its own MassBank data servers distributed on the Internet. MassBank users can access either all of the MassBank data or a subset of the data by specifying one or more experimental conditions. In a spectral search to retrieve mass spectra similar to a query mass spectrum, the similarity score is calculated by a weighted cosine correlation in which weighting exponents on peak intensity and the mass-to-charge ratio are optimized to the ESI-MS^2 data. MassBank also provides a merged spectrum for each compound prepared by merging the analyzed ESI-MS^2 data on an identical compound under different collision-induced dissociation conditions. Data merging has significantly improved the precision of the identification of a chemical compound by 21–23% at a similarity score of 0.6. Thus, MassBank is useful for the identification of chemical compounds and the publication of experimental data.

Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: MassBank; public database; distributed database; metabolite; spectral similarity

Introduction

Mass spectral data are important experimental data for supporting life science research. Researchers are encouraged to annotate/describe every detail of their experimental data, especially metadata, available to the public at publication of their studies. Full disclosure of supporting experimental data is required for other scientists to confirm the quality of experimental data.[1] However, most mass spectral or supplementary data in journal articles are not fully disclosed because they are published only as figures showing the mass-to-charge ratio (*m/z*) and the relative intensity values of major peaks.

Although published mass spectral data are valuable research products that should be shared as reference data for the identification of chemical compounds detected by mass spectrometry, their retrieval from journal archives is extremely time consuming. Therefore, mass spectral data as supporting experimental data and as useful research products should be publicly accessible not in figures but in digital format. However, at present there is no public repository for mass spectral data of small chemical compounds except for those of proteomics data. Before considering

---

* *Correspondence to: Takaaki Nishioka, Institute for Advanced Biosciences, Keio University, 14-1 Banba-cho, Tsuruoka, Yamagata 997-0035, Japan. E-mail: takaaki@sfc.keio.ac.jp*

† *Current address: Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan.*

a *Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan*

b *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan*

c *RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan*

d *Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan*

e *Biomarkers and Personalized Medicine Core Function Unit, Eisai Product Creation Systems, Eisai Co. Ltd, Tsukuba, Ibaraki 300-2635, Japan*

f *JST, CREST, Kawaguchi, Saitama 332-0012, Japan*

g *Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan*

h *Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan*

i *Research Institute for Production Development, Kyoto 606-0805, Japan*

j *Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan*

k *Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, 06120 Halle, Germany*

l *College of Humanities and Sciences, Nihon University, Tokyo 156-8550, Japan*

m *Institute of Natural Medicine, University of Toyama, Toyama 930-0194, Japan*

n *Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*

o *Faculty of Life Science and Biotechnology, Fukuyama University, Fukuyama, Hiroshima 729-0292, Japan*

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777 Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

the reasons for this, we will briefly review a few currently available mass spectral databases.

Several small-scale databases of mass spectral data of small chemical compounds provide reference mass spectral libraries for metabolite identification. The Golm Metabolome Database (GMD@CSB.DB), established by the Max Planck Institute of Molecular Plant Physiology (Golm, Germany), is a library of GC-MS data of plant metabolites.[2] The METLIN database of the Scripps Research Institute (San Diego, CA, USA) provides 8800 MS^2 data on 1662 metabolites and drugs[3] and the Glycan Mass Spectral Database (GMDB), created by the Research Center for Medical Glycoscience of the National Institute of Advanced Industrial Science and Technology (AIST), Japan, is a library of MS^n data of polysaccharide chains.[4] The Human Metabolome Database (HMDB) of the University of Alberta (Edmonton, Canada) contains liquid chromatography (LC)- and GC-MS data (as PNG images) of 799 and 279 endogenous metabolites reported in the literature that were found in biofluids, respectively.[5] All the electrospray ionization (ESI)-MS^2 data were collected at three different collision energy levels. Two major mass spectral databases, the Mass Spectral Library[6] [the National Institute of Standards and Technology (NIST)/Environmental Protection Agency (EPA)/National Institutes of Health (NIH), USA] and the Spectral Database System (SDBS)[7] of AIST provide 220 000 and 24 000 official mass spectral data, respectively. These national laboratories analyze purified natural and synthetic chemical compounds by electron-ionization mass spectrometry (EI-MS).

In those six databases, all mass spectra were analyzed under fixed, well-controlled experimental conditions. To retain the quality of the data as reference data for the identification of chemical compounds, curators do not mix data in their databases with data analyzed by other research groups.

In the life sciences, different types of mass spectrometers are used to analyze chemical compounds in biological samples because their diverse chemical structure results in different physicochemical properties.[8,9] For example, in most metabolomics studies, GC and LC are coupled to EI-MS and ESI-MS^n, respectively. EI-MS, which applies a standardized analytical method, yields reproducible data for an identical chemical compound. On the other hand, no standard experimental protocol is available for ESI-MS^n. Individual researchers optimized their experimental methods of ESI-MS^n depending on the physicochemical properties of their target chemical compounds. However, slight differences in the experimental methods of ESI-MS^n may yield different mass spectra for an identical chemical compound. Therefore, if a public repository were available, the mass spectral data analyzed by different experimental methods would be mixed. This raises concerns about the suitability of a public repository for sharing mass spectral data as reference data for the identification of chemical compounds detected by mass spectrometry. This may be the main reason for the continuing absence of a public repository of mass spectral data.

Although standardization of experimental methods of mass spectrometry is thought to be essential for sharing the mass spectral data of chemical compounds and standardized procedures to unify experimental protocols have been proposed, the metabolomics research community has not reached consensus on those proposals.[10,11] As research groups individually optimized their experimental methods based on their projects and the physicochemical properties of their target compounds, switching to other analytical methods would be almost impossible. Consequently, each group prepared its own reference mass spectral library by analyzing commercially available standard reagents. However, commercially available standard reagents, especially those of secondary metabolites produced by plants and microorganisms, are limited in number. Because this limited availability restricts the ratio of identified metabolites to those detected on LC-MS and -MS^2, it remains as low as 3–5% (48/1233) in plant[12] and 20–30% (175/626) in human tissues.[13]

Usually, metabolites are identified by comparing two data, retention index of chromatographic separation and mass spectrum, with authentic compounds analyzed under identical experimental conditions. New technologies such as single-cell mass spectrometry using matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry[14] and direct nano-ESI mass spectrometry[15] do not employ chromatographic separation but rather, they ionize all chemical compounds in a cell at once. Therefore, metabolite identification in new technologies depends solely on the reference library of the MS^n data.

In summary, although we must not expect the standardization of experimental protocols or platforms, this does not justify the absence of a public repository for mass spectral data.

Here, we report MassBank, the first public repository of mass spectral database of small chemical compounds ($<$3000 Da) for life sciences. Research groups contributing to the repository make their mass spectral data available to the public as supporting experimental data for other researchers. MassBank accepts mass spectral data analyzed on chemical compounds using optimized, up-to-date analytical methods. It is also the first internationally allied spectral database. As contributors deposit their mass spectral data not on a centralized repository, but on their own MassBank data servers, the contributed data and their quality are not mixed but independent from those of other contributors. Users of MassBank are provided with informatics tools to search the distributed data for identification of chemical compounds detected by mass spectrometry.

Experimental

Concepts underlying MassBank

We designed the software architecture and record format of MassBank based on three concepts. First, MassBank should be a public repository for sharing mass spectral data. Contributors should prepare their data in a common record format that defines the data field for the experimental methods, details the analytical parameters of the mass spectrometry and provides peak data. Second, data should be distributed on the Internet. Ideally, each contributor should have a local data server for publication of the formatted data. A contributor may have multiple databases to facilitate the separate management of data analyzed on different instruments, and (s)he could specify which data servers are and are not open to the public. Third, the query interface of MassBank functions as an access point to data servers distributed on the Internet.

Software architecture of MassBank servers

Despite its distributed design, from the user's point of view, MassBank should appear and function as a normal centralized database. Users should be able to access MassBank data without knowing where the data are or what data are involved and contributors should be able to update and manage their data independently.

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777 Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

Table 1. MassBank record

Tag	Description of record field
Summary section	
ACCESSION	Accession number
RECORD_TITLE	Short summary of the record, including the chemical name of the compound analyzed and the analytical method
DATE	Date of contribution
AUTHORS	Contributors and their affiliations
COPYRIGHT	Copyright notice
Chemical section	
CH$NAME	Chemical name of the compound analyzed
CH$COMPOUND_CLASS	Chemical class of the compound
CH$FORMULA	Chemical formula of the compound
CH$EXACT_MASS	Exact mass of the compound
CH$SMILES	SMILES code of the chemical structure of the compound
CH$IUPAC	InChI code of the chemical structure of the compound
Analytical section	
AC$INSTRUMENT	Mass spectrometer and name of manufacturer
AC$INSTRUMENT_TYPE	Type of ion analyzer
AC$ANALYTICAL_CONDITION/MODE	Ionization mode
Spectral section	
PK$NUM_PEAK	Total number of peaks
PK$PEAK	Peak data: *m*/*z*, intensity and relative intensity
Others	
MOLFILE_NAME	File name of the molfile that defines the chemical structure of the compound analyzed

Each data field is labeled by the tag specifying the data item. The 16 tags listed in the table are mandatory; they are shown on Record Editor.

To satisfy these requirements, we adopted a three-tier architecture for the MassBank system; it is comprised of database, application and presentation layers. The database layer stores the mass spectral data in text format in the relational MySQL database. The application layer is a search engine for the data stored in the database layer. The presentation layer is the user interface that specifies servers to be accessed. The application and presentation layers are implemented in Java on the Apache Tomcat web server.

Software distribution and maintenance

The MassBank system software is distributed free-of-charge under the GNU General Public License. The latest source codes are downloadable from SourceForge.net and they are provided for both Linux and Microsoft Windows operating systems (OS). MassBank Installer is a single archive file that includes precompiled object files and a script for the installation of required free software such as Apache, Tomcat and MySQL. As the MassBank Installer is not updated as often as the frequently updated MassBank system, we recommend that users install the MassBank system by means of the MassBank Installer first and then perform updates using the latest source codes from SourceForge.net.

An update service is provided to make maintenance of MassBank easy. The version of each component of the MassBank system is checked automatically using the http access to the MassBank.jp website. When an old component is found, the latest version is transferred and installed automatically.

MassBank record format

MassBank data must be prepared in the MassBank record format. Each record contains one mass spectrum attributable to one chemical compound with a specific chemical formula and each record consists of four sections: a summary, chemical, analytical and spectral section. Each data field carries a tag that specifies the data item (Table 1). For example, for the chemical, analytical and spectral sections the tags are CH$, AC$ and PK$, respectively.

The summary section contains the accession number that uniquely defines the record and summary information of the analytical and chemical sections, authors and copyright. The first three letters of the accession number specify the contributor.

The chemical section, CH$, defines the chemical information of the compound analyzed, including chemical names, the CAS number, compound category and IDs with links to available chemical compound databases such as KEGG,[16] PubChem,[17] KNApSAcK,[18] LipidBank,[19] and LipidMaps,[20] if available. The chemical structure is given in SMILES[21] and InChI code[22] and is defined separately by an MDL molfile.

The analytical section, AC$, describes the instrument types and analytical parameters used for mass spectrometry, including the instrument manufacturer, the catalog number of the mass spectrometer, the method of ionization, the type of ion analyzer, ionization voltage, matrix for MALDI ionization and the collision-induced dissociation (CID) conditions for MS^n measurement. For chemical compounds in biological samples that were separated and purified by LC, GC or capillary electrophoresis (CE) coupled to a mass spectrometer, the chromatographic column used, the chromatographic separation conditions and the retention index should be described in detail. These data are helpful for the identification of chemical compounds.

The spectral section, PK$, lists peak data with *m*/*z* and intensity and relative intensity values in integral or real numbers.

DOI:10.1002/jms.1777 Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

Evaluation of the precision of compound identification by spectral search

The query and target datasets (QSs, TSs) were prepared by extracting ESI-MS^2 data from MassBank data. The two datasets consisted of ESI-MS^2 data in which identical metabolites were analyzed under different analytical conditions. Using the QS spectrum as the query, a spectral search against TSs retrieved a list of similar spectra with corresponding similarity scores. If the metabolite of a similar spectrum was the same as the metabolite of the query spectrum, the search result was considered correct; if not, it was considered incorrect. Each search result was recorded with the similarity score. We repeated the spectral search for all QS spectra.

Considering the search results with a similarity score higher than the threshold, say s, to be true, we counted the number of true positives, TP(s), false negatives, FN(s) and false positives, FP(s), as follows.

> TP(s) = Total number of correct results with a similarity score higher than the threshold value s,
> FN(s) = Total number of correct results with a similarity score lower than the threshold value s,
> FP(s) = Total number of incorrect results with a similarity score higher than the threshold value s.

We then calculated the precision, recall and *F*-value at threshold s as follows.

$$\mathrm{Precision}(s) = \mathrm{TP}(s)/[\mathrm{TP}(s) + \mathrm{FP}(s)] \quad (1)$$

$$\mathrm{Recall}(s) = \mathrm{TP}(s)/[\mathrm{TP}(s) + \mathrm{FN}(s)] \quad (2)$$

$$F\text{-value}(s) = \text{Harmonic means between Precision(s) and Recall(s)} \quad (3)$$

Results

Tools for contributors

Contributors to MassBank must prepare the mass spectral data in the MassBank record format and deposit the formatted data on their own MassBank data servers. Previously, data preparation involved tedious manual work. For example, for the analytical section, contributors had to manually detail the experimental methods and analytical parameters of mass spectrometry. Additionally, experience with MySQL and the Linux OS was essential for data management on their data servers. To reduce the workload and the experience requirement, we developed two tools: Record Editor and Administration Tool.

Generally, mass spectrometers output mass spectral data in the form of binary raw data readable only by the specific software provided by the instrument manufacturer. Binary raw data contain the peak data and the analytical parameters used to control the mass spectrometers. Previously, contributors had to manually extract the peak data and the analytical method, including parameters from the binary raw data with appropriate software. Then they manually prepared the data of the analytical and spectral sections in the MassBank record format.

The Mass++ program can directly import the binary raw data of major instrument companies and output the data in mzML and other data formats.[23,24] Mass++ has newly incorporated functionality that imports binary raw data and automatically outputs the spectral data and the analytical methods in the MassBank record format. The formatted data output from Mass++ is then combined with the molfile that defines the structure of the chemical compound in the Record Editor. This tool automatically calculates the chemical formula and the exact mass of the molecule, and generates SMILES and InChI codes to complete the chemical data section. After the accession number of the record, the authors and other necessary data are manually input in the summary section, and the Record Editor outputs a complete MassBank record as shown in Fig. 1.

Finally, using Administration Tool on a web browser, contributors can upload and manage their data on their MassBank data servers. Thus, contributors no longer need to have experience with either Linux or MySQL commands for data management.

Manuals are available from the manual page of the MassBank site (http://www.massbank.jp/en/manual.html) for contributors wanting to know more about Record Editor and Administration Tool.

Statistics of MassBank data

As of January 2010, 16 research groups, 12 in Japan, 3 in the United States and 1 in Germany, are contributing data to MassBank (Table 2). Mass spectral data, chemical compounds and analytical methods are summarized for each research group on the website (http://www.massbank.jp/en/published.html). These data are distributed on eight MassBank data servers, one of which is located in the Leibniz Institute of Plant Biochemistry (Halle, Germany). Eight small research groups currently without their own data servers contribute their data to the MassBank data servers in Japan or Germany. In January 2010, MassBank data included 10 294 mass spectra [9276 ESI-MS^n, 605 EI-MS, 137 fast atom bombardment (FAB)-MS] of 2337 chemical compounds, 3045 ESI-MS^2 data of 679 synthetic drugs and 11 545 EI-, 795 CI-, 38 FD- and 1 FI-MS data of 10 286 volatile natural and synthetic compounds. The MassBank data consist of data analyzed on 21 different instrument types.

MassBank data are composed of the mass spectra of primary metabolites, flavonoids, gibberellins, saponins, carotenoids, phospholipids and oligosaccharides. Most of these were analyzed on ESI-MS^2, and some on FAB-MS. In their analysis on ESI-MS^n, different CID energies were applied to obtain as many product ions as possible. This resulted in 9276 ESI-MS^n data of 1889 chemical compounds, an average of 4.9 ESI-MS^n data per chemical compound. EI-MS data are for bile acids and volatile chemical compounds such as terpenoids, alkyl alcohols, aldehydes and carboxylic acids. Since standard experimental conditions are available for EI-MS, each chemical compound has only one spectral datum.

In collaboration with LipidBank (http://www.lipidbank.jp/), the official database of the Japanese Conference on the Biochemistry of Lipids (JCBL), MassBank also collects the mass spectra of lipids from the literature. As of June 2008, MassBank is the official database of the Mass Spectral Society of Japan.

Users can access MassBank data from two access points, one in Japan[25] and the other in Germany.[26] Monthly access to MassBank data originating from Japan, USA, UK, Germany, Spain and other countries has reached 7800 hits on average, more than half originated from countries other than Japan.

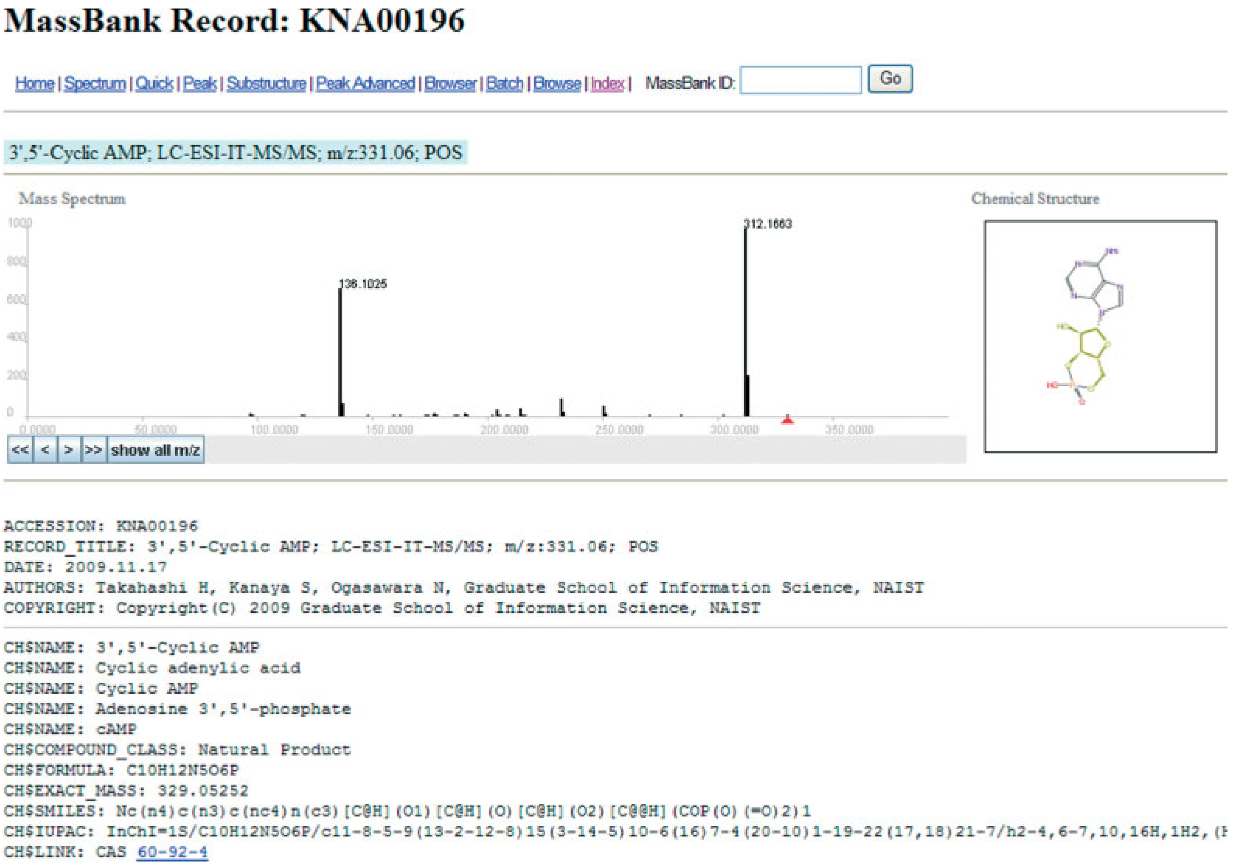
Tools for users

Here we briefly introduce the tools developed for users to access MassBank data and their functions. Users wanting to know more details about the functions can consult a user manual available as a pdf file from the MassBank website (http://www.massbank.jp/en/manual.html).

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.



MassBank Record: KNA00196

Home | Spectrum | Quick | Peak | Substructure | Peak Advanced | Browser | Batch | Browse | Index |
MassBank ID:
Go

3',5'-Cyclic AMP; LC-ESI-IT-MS/MS; m/z:331.06; POS

```
ACCESSION: KNA00196
RECORD_TITLE: 3',5'-Cyclic AMP; LC-ESI-IT-MS/MS; m/z:331.06; POS
DATE: 2009.11.17
AUTHORS: Takahashi H, Kanaya S, Ogasawara N, Graduate School of Information Science, NAIST
COPYRIGHT: Copyright(C) 2009 Graduate School of Information Science, NAIST
```

```
CH$NAME: 3',5'-Cyclic AMP
CH$NAME: Cyclic adenylic acid
CH$NAME: Cyclic AMP
CH$NAME: Adenosine 3',5'-phosphate
CH$NAME: cAMP
CH$COMPOUND_CLASS: Natural Product
CH$FORMULA: C10H12N5O6P
CH$EXACT_MASS: 329.05252
CH$SMILES: Nc(n4)c(n3)c(nc4)n(c3)[C@H](O1)[C@H](O)[C@H](O2)[C@@H](COP(O)(=O)2)1
CH$IUPAC: InChI=1S/C10H12N5O6P/c11-8-5-9(13-2-12-8)15(3-14-5)10-6(16)7-4(20-10)1-19-22(17,18)21-7/h2-4,6-7,10,16H,1H2,(
CH$LINK: CAS 60-92-4
```

Figure 1. Example of a MassBank record.

To obtain suitable search results, users should specify search conditions using the Search Parameter Setting applet before their first search. The users should first specify the search tolerance, that is the experimental error allowance in the *m/z* value, the cutoff threshold for lower intensity peaks and the precursor ion by the *m/z* value. Then, the users select the instrument type identical with or similar to the type of the query mass spectrum and the ionization mode (Fig. 2(a)). Currently, the applet displays 21 instrument types.

Spectral Search

Spectral Search retrieves MS^n data identical with or similar to the query data. The search results are output in the order of the similarity score together with the number of identical product ions.

MassBank currently adopts the database search algorithm that calculates the similarity score based on a modified cosine correlation proposed by Stein and Scott.[27] The intensity of the *i*th peak is weighed by a factor, *W*i, as follows:

$$W_\mathrm{i} = [\text{Intensity of peak}_\mathrm{i}]^m[m/z \text{ of peak}_\mathrm{i}]^n \qquad (4)$$

Stein and Scott empirically determined the optimal exponents as $m = 0.6$ and $n = 3$ by analyzing *ca* 12 000 EI-MS data of 8000 organic compounds in the NIST Mass Spectral Library. Similar to their method, we optimized the exponents as $m = 0.5$ and $n = 2$ by analyzing 8785 ESI-MS^2 data of *ca* 700 authentic compounds of primary metabolites.[28] The difference between the present exponents and those determined by Stein and Scott is primarily attributable to the smaller number of peaks and the higher intensity of higher *m/z* peaks in the ESI-MS^2 data analyzed.

By displaying the search results peak-by-peak on the three-dimensional display, users can identify peaks in a database mass spectrum that are common to peaks in the query mass spectrum (Fig. 2(b)). MassBank provides a batch service for heavy users who submit many MS^n data as queries to the search service.

Quick Search and Substructure Search

MassBank features two tools to search for chemical compounds in its repository: Quick Search and Substructure Search. Quick Search retrieves chemical compounds by the chemical name, chemical formula and a list of the *m/z* and relative intensity values. The search results show the chemical compounds with their chemical names, spectral data and chemical structure (Fig. 3). Substructure Search retrieves chemical compounds containing a specified chemical substructure as a part of their chemical structure (Fig. 4). Users can select three different search options depending on how many π electrons in the query substructure are included in the target structures. The number of π electrons should be (1) the same, (2) higher in the target data or (3) ignored.

Peak Search and Peak Difference Search

Peak Search retrieves MS^n data containing the peaks specified by the *m/z* values within a specified error allowance. Peak Difference Search shows chemical compounds containing one or more peak pairs whose *m/z* values are different from each other by the specified *m/z* values.

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

Table 2. Statistics of MassBank data as of January 2010

Research group	Group ID	Analytical method	Num of spectra	Num of compounds
Institute for Advanced Biosciences, Keio University	KO	ESI-QqTOF-MS/MS	914[a]	695
		ESI-QqQ-MS/MS	4 275	
		ESI-IT-(MS)^n	515	
PSC, RIKEN	PR	GC-EI-TOF-MS	241	767
		LC-ESI-TOF-MS	85	
		LC-ESI-QqQ-MS/MS	87	
		CE-ESI-TOF-MS	20	
		LC-ESI-QTOF-MS/MS	1 290	
Waters	WA	LC-ESI-Q-MS	2 721	577
		ESI-QqQ-MS/MS	273	
Akimoto, Graduate School of Pharmaceutical Sciences, Kyoto and Maoka, Research Institute for Production Development	CA	FAB-CID-EBEB-MS/MS	106	106
Taguchi, Graduate School of Medicine, The University of Tokyo	UT	ESI-QqIT-MS/MS	378	42
Kazusa DNA Research Institute	KZ	GC-EI-TOF-MS	273	163
Iida, College of Humanities and Sciences, Nihon University	NU	EI-MS	75	74
Tanaka, Institute of Natural Medicine, University of Toyama	TY	LC-ESI-IT-TOF-MS	91	69
Kimura, Faculty of Agriculture, Tottori University	TT	EI-MS	11	11
		FAB-MS	5	
Funatsu, Graduate School of Engineering, The University of Tokyo	JP	EI-MS	11 545	10 286
		CI-MS	795	
		FD-MS	38	
		FI-MS	1	
Leibniz Institute of Plant Biochemistry	PB	ESI-QqTOF-MS/MS	297	90
		ESI-QqQ-MS/MS	63	
Matsuura, Fukuyama University	FU	LC-ESI-QqQ-MS/MS	285	71
Metabolon, Inc.	MT	ESI-IT-MS/MS	149	149
Morii, University of Occupational and Environmental Health	UO	FAB-MS	26	25
		EI-MS	5	
		FD-MS	3	
		CI-MS	1	
Kanaya, Graduate School of Information Science, Nara Institute of Science and Technology	KNA	LC-ESI-IT-MS/MS	619	75
		LC-ESI-FT-MS	208	
Grant, University of Connecticut	CO	ESI-QqTOF-MS	510	102

[a] Number of merged spectra.

Peak Search Advanced

Peak Search Advanced is similar to Peak Search and Peak Difference Search in function, but it is different in that it specifies the peaks with the molecular formulae of the ions. Peaks in the merged data (see the next section for details) are annotated by the chemical formula within an error range of 50 ppm (the threshold is adjustable). Currently, there are 817 positive and 797 negative ESI-QqTOF-MS^2 merged data available as the target for Peak Search Advanced.

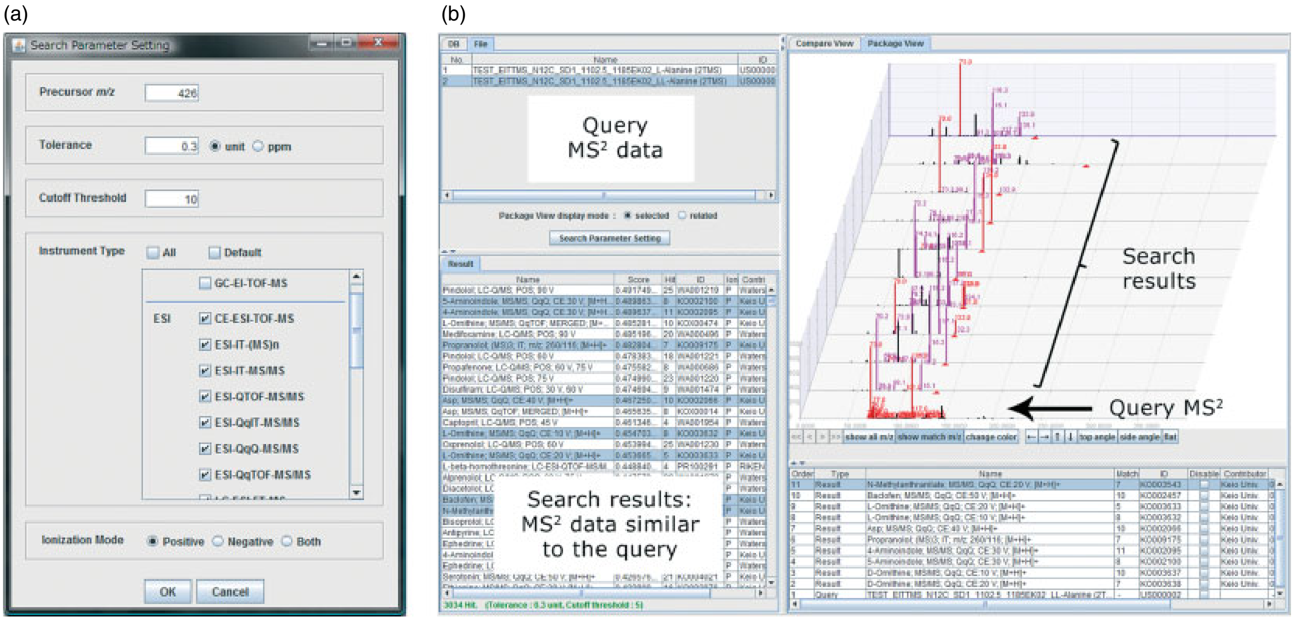
Merged mass spectra as artificial reference mass spectra for metabolite identification

One of the most important applications of MassBank data in the life sciences is metabolite identification. Generally, ESI-MS^2 data of chemical compounds are useful as reference data for metabolite identification when the analytical conditions of the query ESI-MS^2 data are the same as or very similar to those of the reference mass spectra. When the query and the reference chemical compounds are the same, the spectral search retrieves the reference mass spectrum with higher similarity scores. In other cases, the query and reference mass spectra are less similar or different even when the two chemical compounds are the same. As most MassBank users may encounter the latter situation, MassBank provides an artificial reference, that is the 'merged' mass spectrum.

As the reproducibility of the ESI-MS^2 data is reportedly low,[29,30] we evaluated the degree of reproducibility of MassBank ESI-MS^2 data for use as reference data in the metabolite identification. We took two datasets of common metabolites extracted from MassBank: datasets [QqQ] and [QqTOF] consisting of 4205 ESI-QqQ-MS^2 and 4431 ESI-QqTOF-MS^2 data of 856 common chemical compounds, respectively. Each chemical compound in each dataset has four or five spectral data. In the first experiment,

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.



(a)

Search Parameter Setting

Precursor m/z 426

Tolerance 0.3 ● unit ○ ppm

Cutoff Threshold 10

Instrument Type ☐ All ☐ Default
☐ GC-EI-TOF-MS
ESI ☑ CE-ESI-TOF-MS
☑ ESI-IT-(MS)n
☑ ESI-IT-MS/MS
☑ ESI-QTOF-MS/MS
☑ ESI-QqIT-MS/MS
☑ ESI-QqQ-MS/MS
☑ ESI-QqTOF-MS/MS

Ionization Mode ● Positive ○ Negative ○ Both

OK Cancel

(b)

Package View display mode : ● selected ○ related

Search Parameter Setting

3034 Hit. (Tolerance : 0.3 unit, Cutoff threshold : 5)

Figure 2. Search Parameter Setting and Spectral Search. (a) Search parameters are selected and input on the applet. The 'Precursor ion' is specified by the *m/z* value. 'Tolerance' is the error allowance of *m/z* values. When a peak in the query data and the corresponding peak in the target data have different *m/z* values but are within the tolerance, the two peaks are treated as identical. 'Cutoff threshold' is used to distinguish real peaks from noise peaks. (b) The left upper and lower panels show the QS and the search results in the order of the similarity score, respectively. When some of the search results are selected in the left lower panel, the three-dimensional display in the right upper panel shows the spectral search results in peak-by-peak mode.

the query dataset (QS) was [QqQ] and the TS was [QqTOF]. In the second experiment, QS and TS were [QqTOF] and [QqQ], respectively. We performed two spectral searches and evaluated precision (see Experimental section, Eqn (1)), recall (Eqn (2)) and the *F*-value (Eqn (3)) at various threshold similarity scores for each QS and TS pair. When the threshold of the similarity score was 0.6, the precision, recall and the *F*-value for TS = [QqQ] and [QqTOF] were [0.222, 0.327, 0.264] and [0.276, 0.292, 0.284], respectively. Thus, in their original form, ESI-MS^2 data in MassBank are not likely to serve as reference data.

ESI-MS^2 data using CID reflect the employed collision energy (Fig. 5(a)); smaller product ion nonlinearly increase with the collision energy. This is one of the major reasons for the low reproducibility of ESI-MS^2 data analyzed under different analytical conditions. Therefore we expect that merged mass spectra, that is superposition of spectra in different collision energies, would better serve as the reference mass spectra for metabolite identification.

In fact, metabolomics groups at the Institute for Advanced Biosciences, Keio University, Tokyo, Japan ('Keio group') and the RIKEN Plant Science Center, Yokohama, Japan ('RIKEN group') measured the ESI-MS^2 data of chemical compounds at five different CID collision energies in both positive and negative modes. The Keio group assessed 4570 ESI-QTQF-MS^2 data of 695 chemical compounds under five different collision energies at 10–50 V. For each chemical compound, the ESI-QTQF-MS^2 data were overlaid and merged into a single artificially merged MS^2 spectrum (Fig. 5(b)). Each of the chemical compounds has one merged mass spectrum. The Keio group contributed 914 merged ESI-QTQF-MS^2 data of 695 chemical compounds to MassBank. The RIKEN group measured 535 chemical compounds on LC-ESI-QTOF-MS^2 under the ramp mode, which we regard as merged mass spectra, in the range of 5–60 V collision energies in both positive and negative modes, contributing to a total of 1290 ESI-MS^2 data. Merged mass spectral data have the character 'X' in the third position of the record number, e.g. KOX000031. These merged ESI-QTOF-MS^2 data contain most of the product ions observed under the commonly adopted CID conditions for measuring ESI-MS.[2] Therefore, for each chemical compound, the merged data yield a representative fragmentation pattern.

Evaluation of compound identification using merged ESI-MS^2 data as reference data

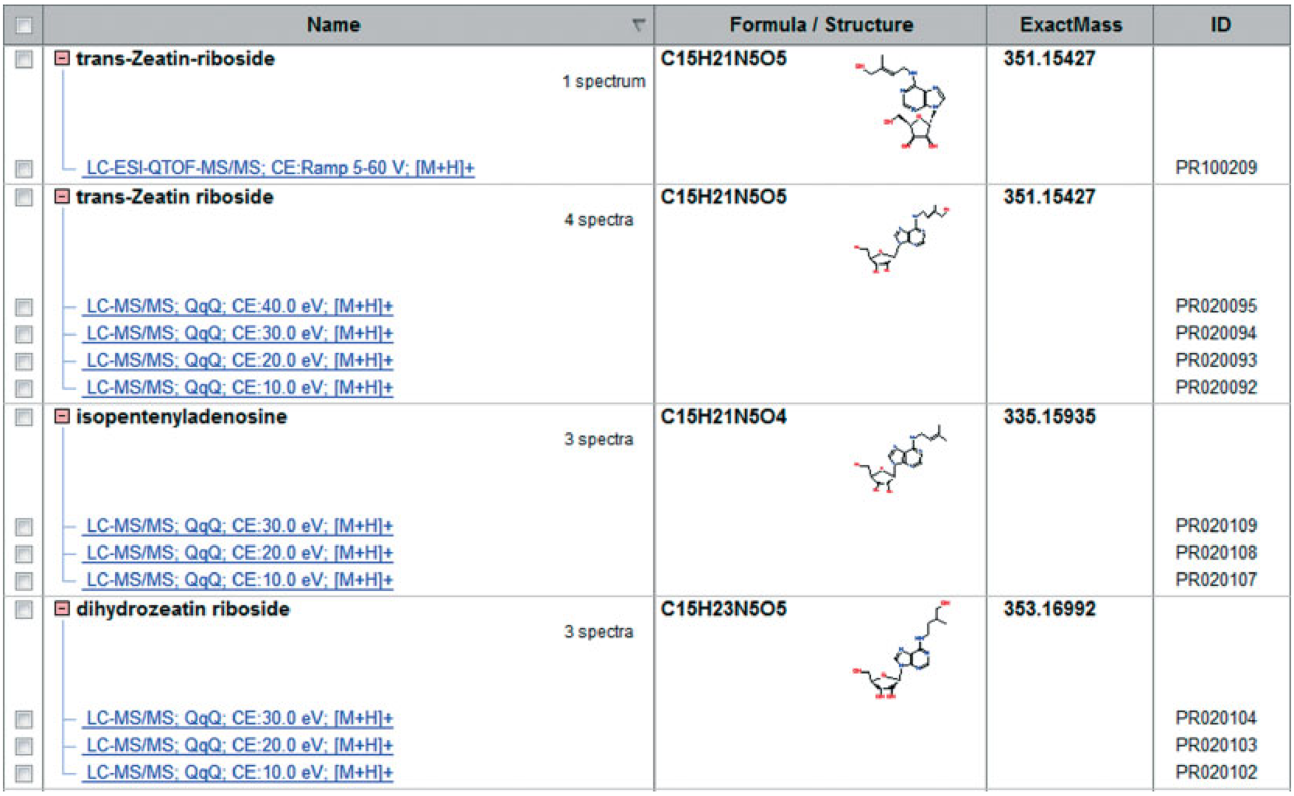
We evaluated the quality of merged ESI-MS^2 data as reference data vis-à-vis the original ESI-MS^2 data. The TSs [Merged QqQ] and [Merged QqTOF] were prepared by merging [QqQ] and [QqTOF] for each chemical compound. This yielded 856 merged data for each dataset. In the first experiment, QS was [QqQ] and TS was [Merged QqTOF], and in the second, QS and TS were [QqTOF] and [Merged QqQ], respectively. We performed two spectral searches and evaluated precision, recall and the *F*-value at various threshold similarity scores for each QS and TS pair. When the threshold of the similarity score was 0.6, precision, recall and *F*-value observed for TS = [Merged QqQ] and [Merged QqTOF] were [0.454, 0.307, 0.366] and [0.490, 0.299, 0.371], respectively. Therefore, merging the ESI-QqQ and QqTOF-MS^2 data improved the precision of the spectral searches by 23% and 21%, respectively, at similarity scores higher than 0.6. Merging the data did not significantly affect recall. The merged data improved metabolite identification using ESI-QIT-MS data as queries (data not shown). Therefore, a spectral search with weighting parameters optimized against the merged mass spectra yields satisfactory results for metabolite identification.

We recommend that contributors of ESI-MS^2 data deposit multiple data for each chemical compound analyzed under at least a few different levels of collision energy in both positive and negative mode.

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.



☐	Name	Formula / Structure	ExactMass	ID
☐	⊟ **trans-Zeatin-riboside** 1 spectrum	C15H21N5O5	351.15427	
☐	LC-ESI-QTOF-MS/MS; CE:Ramp 5-60 V; [M+H]+			PR100209
☐	⊟ **trans-Zeatin riboside** 4 spectra	C15H21N5O5	351.15427	
☐	LC-MS/MS; QqQ; CE:40.0 eV; [M+H]+			PR020095
☐	LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			PR020094
☐	LC-MS/MS; QqQ; CE:20.0 eV; [M+H]+			PR020093
☐	LC-MS/MS; QqQ; CE:10.0 eV; [M+H]+			PR020092
☐	⊟ **isopentenyladenosine** 3 spectra	C15H21N5O4	335.15935	
☐	LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			PR020109
☐	LC-MS/MS; QqQ; CE:20.0 eV; [M+H]+			PR020108
☐	LC-MS/MS; QqQ; CE:10.0 eV; [M+H]+			PR020107
☐	⊟ **dihydrozeatin riboside** 3 spectra	C15H23N5O5	353.16992	
☐	LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			PR020104
☐	LC-MS/MS; QqQ; CE:20.0 eV; [M+H]+			PR020103
☐	LC-MS/MS; QqQ; CE:10.0 eV; [M+H]+			PR020102

Figure 3. Quick Search. When, for example, the search involves chemical compounds containing 'adenine' in the name, Quick Search displays the chemical compounds matching the search together with the spectral data and chemical structure.

API services

The MassBank Application Programming Interface (API), the Simple Object Access Protocol (SOAP) interface to MassBank, allows users to write their own programs for accessing, customizing and utilizing MassBank. Currently available methods, downloadable from http://www.massbank.jp/en/download.html and described by a schema in Web Service Definition Language (WSDL) (http://www.massbank.jp/api/services/MassBankAPI?wsdl), are Spectral Search, Peak Search and Peak Difference Search.

We show an example using MassBank API. As described above, mass spectrometers output spectral data as binary raw data. Because binary raw data are not accepted as a query for a spectral search in MassBank, they must first be converted into text data format. Conducting a spectral search query for several hundred binary raw data outputs with a single run of LC-MS^n was a time-consuming task in metabolomics studies. The Mass++ program frees users from this burden with a new function that imports binary raw data for submission as a spectral search query using MassBank API and shows the search results in its own display mode. In the near future, MassBank will provide the WSDL batch service method for spectral searches.

Program source codes and tool manuals

MassBank is currently available in Linux and Microsoft Windows versions. Typically, the Windows version is released more than 6 months after the Linux version. The source codes of the MassBank system are freely available from SourceForge[31] with the GNU General Public License. Manuals for using the search tools, preparing the data in the MassBank record format, installing the MassBank system and for managing data on MassBank servers are available from the MassBank Manual download site.[32]

Discussion

Merged mass spectra for the identification of chemical compounds

Public mass spectral databases accept mass spectral data analyzed by nonstandardized analytical methods. Among different analytical methods, ESI-MS^n data are of low reproducibility; therefore, these data were not thought to be useful as reference data. However, Volná *et al.*[30] found that the fragmentation patterns are almost identical for all tandem mass analyzers and that only the ratios of the product ions differ somewhat. They recommend analyzing ESI-MS^n at three different CID collision energy levels. Our present analysis of MassBank data supports their findings. In fact, most contributors of ESI-MS^n data to MassBank analyzed each chemical compound under five collision energy levels ranging from 5 to 50 V to observe all possible product ions. Additionally, MassBank provides a merged mass spectrum for each compound. Although merging ESI-MS^n data statistically improved the precision of metabolite identification without decreasing recall, we encountered two problems with the merged data. First, the total number of product ions in the merged data tended to be much larger than the number of product ions in the original ESI-MS^n data. For example, merging five data increased the total number of product ions by 3.82 times (an average of 870 merged data). This resulted in an increase in the number of false-positive hits and a consequent decrease in precision. Second, the base

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777 Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

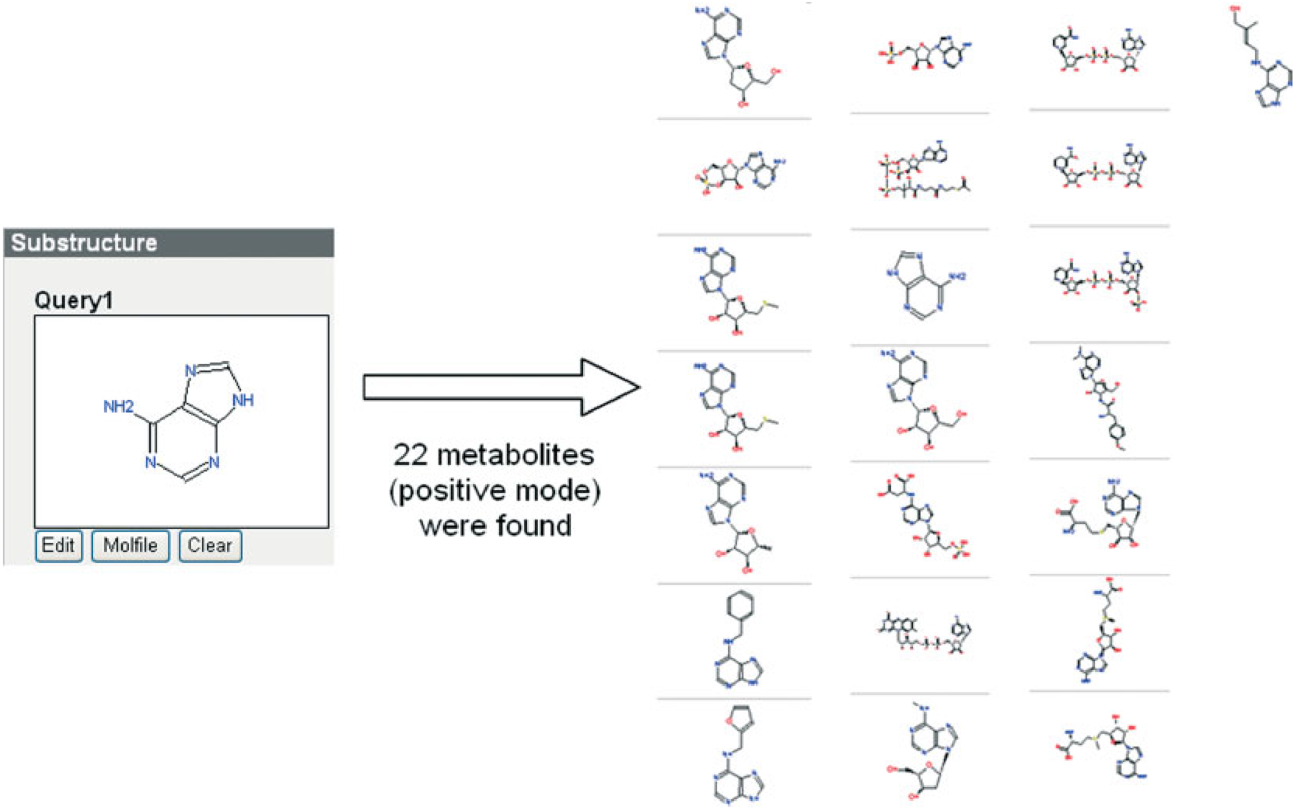


Figure 4. Substructure Search. When a substructure is submitted as a query, all chemical structures containing the query substructure are listed.

peak in the merged data was different from the base peak in the original data. The development of a better merging method and a new database-searching algorithm will solve these problems and improve metabolite identification in MassBank.

Cost of publication of a distributed database

In MassBank, contributing research groups openly avail their data to the public from their own data servers. From this aspect, MassBank is similar to the currently available mass spectral databases discussed in the Introduction (GMD@CSB.DB, METLIN, GMDB, HMDB, NIST/EPA/NIH Mass Spectral Library, SDBS). However, MassBank is different because it accepts data contributions from researchers and groups; the repository contains data analyzed with a wide range of mass spectrometry methods. Via the Search Parameter Setting interface, MassBank allows users to select datasets obtained with different analytical methods as the search target.

In other databases, only the owning research groups or laboratories contribute to their databases and the data in each database are prepared in different record formats. Consequently, the (owning) users of a database cannot access other (nonowned) databases in parallel. In MassBank, contributors must prepare their data in the specified record format. This includes not only the peak data but also the analytical method and conditions, and the chemical structure information on the analyzed chemical compounds. In addition, contributors must manage their data on their own local data servers. As the preparation of formatted data and data management on owned servers was time consuming, at the request of contributors we made efforts to reduce their workload. Our efforts resulted in an increase in the data deposited in MassBank in 2009.

The cost incurred by contributors in the preparation and management of their data in the MassBank-distributed database system is proportional to the amount of data deposited. Contributors of larger quantities of data need high-performance computers and large storage capacity. This is one of the rationales behind a distributed database system. In grant applications, contributors should include costs involved in the publication of experimental data as a necessary expense for the sharing of their data as a research product. Funding organizations should judge the performance of researchers not only based on publications but also on products made available to the wider research community.[1]

The freely available source code is also useful for an independent database project outside of the MassBank consortium. An example is MS/MS spectral tag (MS2T) viewer[12,33] where the data are prepared in the MassBank record format and whose database server is the MassBank clone. The users cannot access the viewer from the common MassBank interface, but only from its original website (http://prime.psc.riken.jp/lcms/ms2tview/ms2tview.html).

Retaining the quality of mass spectral data in MassBank

Some users of MassBank are concerned with the quality of MassBank data with respect to the technical quality of the mass spectrometry and the chemical purity and identification levels of the samples. At present, we cannot offer a practical method for evaluating the technical quality of contributed data. However, before data submission, contributors can easily look for experimental mistakes on Record Editor. Thus, mistakes such

DOI:10.1002/jms.1777 Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

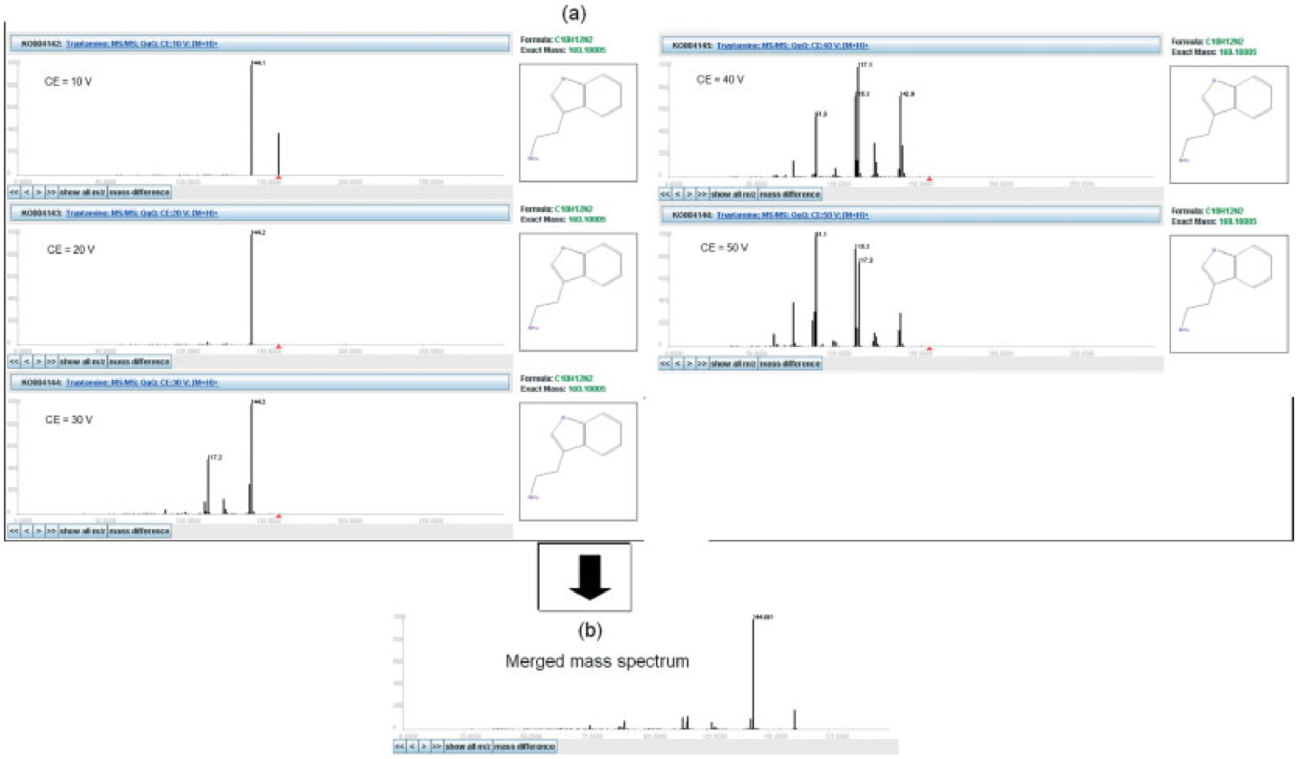


Figure 5. Merged mass spectral data. (a) The mass spectra of tryptamine analyzed on ESI-QqQ-MS^2 by different collision energies (CE), 10–50 V. (b) The five ESI-QqQ-MS^2 data of tryptamine were overlaid and merged into one 'merged mass spectrum'.

as the mislabeling of a test tube are caught by comparing the observed mass of the molecular ions with the calculated mass from the molfiles. For higher resolution MS^n data of known chemical compounds, chemical formulae may be uniquely assignable to most of the product ions in a higher *m/z* range within an error range of 50 ppm. Contributors are advised to add the chemical annotation of as many product ions as possible in an optional data field, PK$ANNOTATION, of the MassBank record format. Such chemical annotations are useful for the removal of mass spectral data that contain ions from contaminants. Annotations are also helpful to MassBank contributors who evaluate the mass accuracy of the data.

At present, MassBank data are the mass spectra of specific chemical compounds commercially available as purified reagents of metabolites. In the near future, we will accept the mass spectral data of metabolites detected and identified by LC-MS^n analysis of biological cell and tissue samples. In such cases, contributors must provide satisfactory experimental evidence for the identification of the chemical compounds in the chemical section.[11] We will also accept LC-, GC- and CE-coupled or direct MS^n data analysis of tissue pieces or single cells. Such data will include the mass spectra of identified and unidentified chemical compounds. Identified chemical compounds are indicated by their chemical names or structures and unidentified or unknown chemical compounds by their MS^n data. MS^n data are used as the tag of unidentified chemical compounds. By comparing the MS^n data analyzed on different biological samples, the intersample similarity or difference of the chemical compounds can be determined.

Sharing mass spectral data among research communities

Beginning in June 2008, the Mass Spectrometry Society of Japan supported MassBank as the official database of the society. In the near future, the society's journal will recommend the authors to register their mass spectral data in MassBank at the time they submit their manuscripts. MassBank will provide the authors with accession numbers for citation of the data in the manuscript. This will make it possible for readers to lookup data details on MassBank and to search for related articles with Spectral Search and other search tools available on MassBank. We plan to advocate the registration of mass spectral data in MassBank among contributors to other academic journals. In 2009 we started collaboration with LipidBank, the official database of the JCBL and organized joint special lectures on MassBank and LipidBank at annual meetings. The society and the conference will work jointly to seek continuous academic funding to support both MassBank and LipidBank.

MassBank provides a record field for copyright, the default holders of which are contributors, but none for data distribution. Because the distribution of mass spectral data is another method of data sharing, users and contributors propose to prepare a record field for data distribution in which contributors express under the terms of the Creative Commons Attribution Licenses.[34] We will prepare the record field and an FTP site for the download of data. Additionally, we consider augmenting the record documentation of MassBank by conforming to the guidelines for the controlled vocabularies from Proteomics Standards Initiative (PSI).[35]

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

Conclusions

MassBank is based on the three concepts. First, it is a public database of mass spectral data analyzed under nonstandardized experimental conditions. Second, it is a distributed database in which contributors prepare and provide their data from their own data servers on the Internet. Third, it develops and provides free tools for contributors to prepare and manage data on their sites. To improve the metabolite identification from mass spectra, we merged ESI-MS^2 data of identical chemical compounds analyzed under different experimental conditions. Merged data as a TS of spectral search were significantly improved precision without decreasing recall of the spectral search when compared with the unmerged original data set. This showed that merging spectral data is useful for generating reference data for metabolite identification.

Acknowledgements

We thank Ms Rie Matsuzawa and Ms Michi Kittaka for their input of data, Dr Toshiaki Katayama for his advice on the development of Web API on MassBank, Dr Zenzaburo Tozuka and Dr Yoshinao Wada for their support on introducing MassBank as the official database of the Mass Spectrometry Society of Japan and Dr Masaru Tomita for his financial support. This work was supported by a grant for the advancement and standardization of biological databases (2006-2010) from the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency (to T. N., M. A. and S. K.), a grant-in-aid for Scientific Research on the Priority Area from the Ministry of Education, Culture, Sports, Science and Technology of Japan (grant number 18016028 to T. N., M. A. and S. K.), research grants from Yamagata Prefecture and Tsuruoka City (to T. N. and M. A.) and a grant from the New Energy and Industrial Technology Development Organization (NEDO) of Japan as part of the 'Development of Fundamental Technologies for Controlling the Material Production Process of Plants' (to T. A., N. S., H. S. and D. S.).

References

[1] P. N. Schofield, T. Bubela, T. Weaver, L. Portilla, S. D. Brown, J. M. Hancock, D. Einhorn, G. Tocchini-Valentini, M. Hrabe de Angelis, N. Rosenthal. Post-publication sharing of data and tools. *Nature* **2009**, *461*, 171.

[2] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser. GMD@CSB.DB: the Golm Metabolome database. *Bioinformatics* **2005**, *21*, 1635.

[3] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747.

[4] A. Kameyama, N. Kikuchi, S. Nakaya, H. Ito, T. Sato, T. Shikanai, Y. Takahashi, K. Takahashi, H. Narimatsu. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. *Anal. Chem.* **2005**, *77*, 4719.

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

[6] National Institute of Standards and Technology, NIST Standard Reference Database 1A, NIST/EPA/NIH Mass Spectral Library with Search Program: (Data Version: NIST 08, Software Version 2.0f). http://www.nist.gov/srd/nist1a.htm. [Last accessed: March 2010].

[7] National Institute of Advanced Industrial Science and Technology, Japan. Spectral Database for Organic Compounds, SDBS. http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi. [Last accessed: March 2010].

[8] C. M. Dobson. Chemical space and biology. *Nature* **2004**, *432*, 824.

[9] J. Clardy, C. Walsh. Lessons from natural molecules. *Nature* **2004**, *432*, 829.

[10] S. A. Sansone, T. Fan, R. Goodacre, J. L. Griffin, N. W. Hardy, R. Kaddurah-Daouk, B. S. Kristal, J. Lindon, P. Mendes, N. Morrison, B. Nikolau, D. Robertson, L. W. Sumner, C. Taylor, M. van der Werf, B. van Ommen, O. Fiehn. The metabolomics standards initiative. *Nat. Biotechnol.*. **2007**, *25*, 846.

[11] L. Sumner, A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, J. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. Lane, J. Lindon, P. Marriott, A. Nicholls, M. Reily, J. Thaden, M. Viant. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*, 211.

[12] F. Matsuda, K. Yonekura-Sakakibara, R. Niida, T. Kuromori, K. Shinozaki, K. Saito. MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. *Plant J.* **2009**, *57*, 555.

[13] A. Sreekumar, M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher, A. M. Chinnaiyan. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **2009**, *457*, 910.

[14] L. A. McDonnell, R. M. A. Heeren. Imaging mass spectrometry. *Mass Spectrom. Rev.* **2007**, *26*, 606.

[15] T. Masujima. Live single-cell mass spectrometry. *Anal. Sci.* **2009**, *25*, 953.

[16] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **2010**, *38*, D355.

[17] United States National Library of Medicine, National Institutes of Health, National Center for Biotechnology Information. PubChem Compounds Database. http://pubchem.ncbi.nlm.nih.gov/. [Last accessed: March 2010].

[18] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, S. Kanaya. KNApSAcK: A comprehensive species-metabolite relationship database. In *Plant Metabolomics*, K. Saito, R. A. Dixon, L. Willmitzer(Eds). Springer-Verlag Berlin: NY, **2006**, 165.

[19] Japanese Conference on the Biochemistry of Lipids. Database of natural lipids. http://www.lipidbank.jp/. [Last accessed: March 2010].

[20] E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, E. A. Dennis. A comprehensive classification system for lipids. *J. Lipid Res.* **2005**, *46*, 839.

[21] D. J. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

[22] International Union of Pure and Applied Chemistry. The IUPAC International Chemical Identifier. http://www.iupac.org/inchi/. [Last accessed: March 2010].

[23] S. Tanaka, K. Aoshima, Y. Miura, Y. Oda. 57th ASMS Conference on Mass Spectrometry and Allied Topics (American Society for Mass Spectrometry), Philadelphia, PA, 31 May to 04 June, **2009**.

[24] Biomarkers and Personalized Medicine Core Function Unit, Eisai Product Creation Systems, Eisai Co. Ltd. Mass++. http://groups.google.com/group/massplusplus. [Last accessed: March 2010].

[25] Institute for Advanced Biosciences, Keio University. MassBank. http://www.massbank.jp. [Last accessed: March 2010].

[26] Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology. MassBank. http://msbi.ipb-halle.de/MassBank/. [Last accessed: March 2010].

[27] S. E. Stein, D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859.

[28] H. Horai, M. Arita, T. Nishioka. Comparison of ESI-MS in MassBank Database. 1st International Conference on BioMedical Engineering and Informatics, Sanya, Hainan, China, 28–30

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

May, **2008**. (The abstract is downloadable from the site http://www.massbank.jp/en/document.html).

[29] C. Hopley, T. Bristow, A. Lubben, A. Simpson, E. Bull, K. Klagkou, J. Herniman, J. Langley. Towards a universal product ion mass spectral library – reproducibility of product ion spectra across eleven different mass spectrometers. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1779.

[30] K. Volná, M. Holcapek, L. Kolárová, K. Lemr, J. Cáslavsky, P. Kacer, J. Poustka, M. Hubálek. Comparison of negative ion electrospray mass spectra measured by seven tandem mass analyzers towards library formation. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 101.

[31] Geeknet, Inc. SourceForge.net. http://sourceforge.net/projects/massbank/. [Last accessed: March 2010].

[32] Institute for Advanced Biosciences, Keio University. Mass++ Manual. http://www.massbank.jp/en/manual.html. [Last accessed: March 2010].

[33] RIKEN, Plant Science Center. Platform for RIKEN Metabolomics MS/MS spectral tag (MS2T) viewer. http://prime.psc.riken.jp/lcms/ms2tview/ms2tview.html. [Last accessed: March 2010].

[34] Creative Commons. Creative Commons Attribution Licenses. http://creativecommons.org/. [Last accessed March 2010].

[35] R. G. Cote, P. Jones, L. Martens, R. Apweiler, H. Hermjakob. The ontology lookup service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.* **2008**, *36*, W372.

Copyright © 2010 John Wiley & Sons, Ltd.

DOI:10.1002/jms.1777

Reprinted with permission. © 2010 John Wiley & Sons, Inc. All Rights Reserved.

Wolf *et al. BMC Bioinformatics* 2010, **11**:148
http://www.biomedcentral.com/1471-2105/11/148

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

In silico fragmentation for computer assisted identification of metabolite mass spectra

Sebastian Wolf[1*], Stephan Schmidt[1], Matthias Müller-Hannemann[2], Steffen Neumann[1]

Abstract

Background: Mass spectrometry has become the analytical method of choice in metabolomics research. The identification of unknown compounds is the main bottleneck. In addition to the precursor mass, tandem MS spectra carry informative fragment peaks, but the coverage of spectral libraries of measured reference compounds are far from covering the complete chemical space. Compound libraries such as PubChem or KEGG describe a larger number of compounds, which can be used to compare their in silico fragmentation with spectra of unknown metabolites.

Results: We created the MetFrag suite to obtain a candidate list from compound libraries based on the precursor mass, subsequently ranked by the agreement between measured and in silico fragments. In the evaluation MetFrag was able to rank most of the correct compounds within the top 3 candidates returned by an exact mass query in KEGG. Compared to a previously published study, MetFrag obtained better results than the commercial MassFrontier software. Especially for large compound libraries, the candidates with a good score show a high structural similarity or just different stereochemistry, a subsequent clustering based on chemical distances reduces this redundancy. The in silico fragmentation requires less than a second to process a molecule, and MetFrag performs a search in KEGG or PubChem on average within 30 to 300 seconds, respectively, on an average desktop PC.

Conclusions: We presented a method that is able to identify small molecules from tandem MS measurements, even without spectral reference data or a large set of fragmentation rules. With today's massive general purpose compound libraries we obtain dozens of very similar candidates, which still allows a confident estimate of the correct compound class. Our tool MetFrag improves the identification of unknown substances from tandem MS spectra and delivers better results than comparable commercial software. MetFrag is available through a web application, web services and as java library. The web frontend allows the end-user to analyse single spectra and browse the results, whereas the web service and console application are aimed to perform batch searches and evaluation.

Background

Mass spectrometry has become the analytical method of choice in metabolomics research [1]. Various ionisation methods are commonly used, such as electron impact (EI) used with gas chromatography (GC/MS), or the soft electrospray ionisation (ESI), which is employed in LC/ESI-MS systems. The main bottleneck in the interpretation of metabolomics experiments is the identification of compounds. In addition to the exact mass, tandem MS spectra provide additional structural hints, providing a fingerprint of the measured molecule. In tandem MS, the molecules are interacting with a collision gas at specified kinetic energies, hence the name *collision induced dissociation*. Large spectral libraries of measured reference spectra exist for GC/MS, such as the commercial NIST library '08 (Gaithersburg, MD) or the GMD [2], but for ESI-tandem MS spectral libraries are still few and comparably small [3,4]. A different approach towards identification is the interpretation of the measured spectra, usually with regard to the known (or hypothetical) molecular structure.

Fragmenter with a rule set like the commercial tools ACD Fragmenter [5] and Mass Frontier [6] generate

---

* Correspondence: swolf@ipb-halle.de
[1]Leibniz Institute of Plant Biochemistry- Department of Stress- and Developmental Biology, Weinberg 3, 06120 Halle(Saale), Germany

© 2010 Wolf et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1186/1471-2105-11-148

fragments based on cleavage rules known from the literature, in both cases the algorithmic details are not published. For some compounds, MassFrontier 5 is not able to identify any fragments in negative mode [7]. Hill et al. used Mass Frontier 4 to predict the tandem MS spectra of 102 test compounds, which were analysed using a Micromass Q-TOF II in positive mode, to identify the measured compound and its structure. Candidate compounds were retrieved from PubChem using the exact mass. MassFrontier used those structures as input and generated spectra which were compared to the measured spectra. Finally, the compounds were ranked according to the peaks common to both the predicted and measured spectra [8]. *Combinatorial Fragmenter* such as Fragment Identificator (FiD) proposed by Heinonen et al. [9] try to predict the fragmentation tree given both a metabolite's molecular structure and its tandem mass spectrum. Due to high computational complexity, even for a single medium sized compound (around 300 Da) runtimes can reach several hours. Another approach is the systematic bond disconnection method without a rule set as described in [10]. The resulting product ions from a single precursor structure are matched against the peaks measured with a high-resolution mass spectrometer. The software EPIC was tested against two hand annotated spectra from the literature and is not publicly available. The runtime was reported to be around 1 minute to process 1-(3-(5-(1,2,4-triazol-4-yl)-1H-indol-3-yl)propyl)-4-(2-(3-fluorophenyl)ethyl)piperazine (432 Da).

MetFrag is a combinatorial fragmenter using the bond disconnection approach, which is fast enough to screen dozens to thousands of candidates retrieved from e.g. KEGG, PubChem or ChemSpider compound databases. We do not attempt to create a mechanistically correct prediction of the fragmentation processes. Instead, we want to perform a search in compound libraries using the measured fragments as additional structural hints.

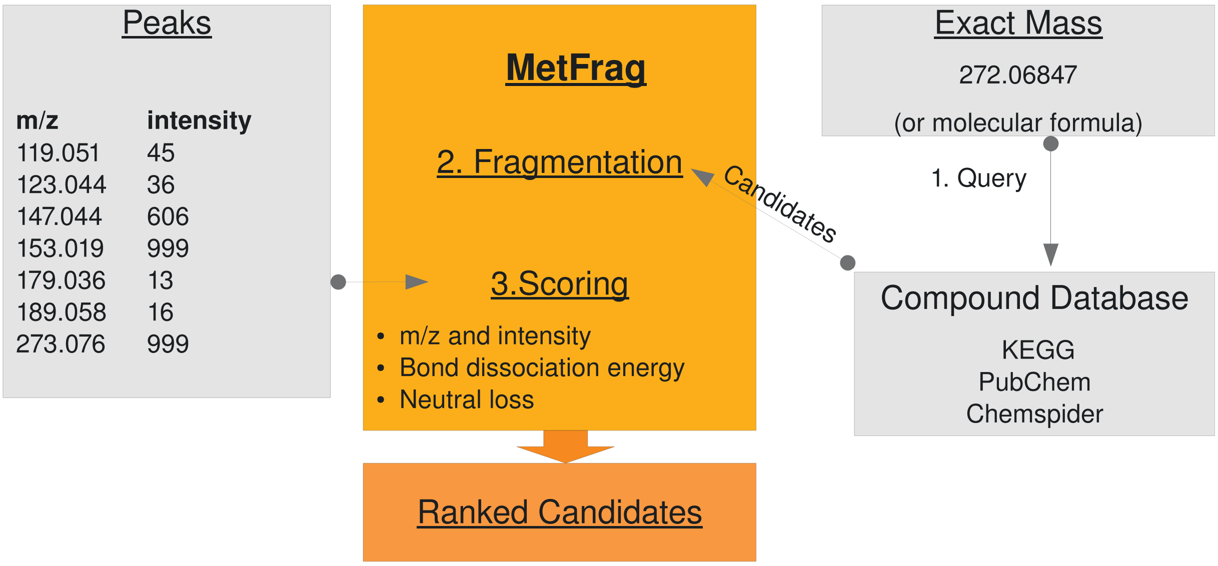
The paper is structured as follows: in the next section we describe the architecture and the in silico fragmentation algorithm, including heuristics to speed up calculations and to account for molecular rearrangements upon fragmentation. Afterwards, we explain the scoring function. In the results section we evaluate MetFrag on a set of 710 spectra from 151 compounds. The paper finishes with our conclusions. All detailed results are available as additional files.

Implementation

The workflow implemented in MetFrag is shown in Figure 1, and covered in detail in the following sections. MetFrag is implemented in Java and uses the Chemistry Development Kit [11], an open source Java library. The CDK provides algorithms and data structures for structural Chemo- and Bioinformatics and is able to read and write common formats such as MDL, CML, InChI, and many more.

Retrieval of candidates from compound libraries

First we perform a search in a general purpose compound database for candidate molecules based on the



m/z	intensity
119.051	45
123.044	36
147.044	606
153.019	999
179.036	13
189.058	16
273.076	999

Figure 1 Workflow of a search based on exact mass and tandem MS spectrum. First the upstream compound library is searched using their respective web service API. The scoring ranks the measured peaks against the in silico fragments.

exact mass (within an error range given in ppm) of the neutral and intact molecule. Currently three compound databases can be queried: KEGG Compound (about 16 021 entries, October 2009) [12], PubChem (37 million, June 2009) [13] and ChemSpider (23 million, October 2009) [14]. Optionally, the search can be restricted to compounds containing only the elements CHNOPS, commonly occurring in natural products.

Alternatively, the compound databases can be searched with the elemental composition if this has been derived from e.g. exact mass and isotopic pattern of the precursor. Finally, the set of candidates can be supplied by simply enumerating all database IDs to be processed, e.g. obtained by an independent search for metabolites of a pathway. To query other (local) libraries, a custom wrapper can be added which contains the search logic.

The results usually contain dozens to thousands of hits with a similar (or identical in case of isomeric compounds) mass. The databases are accessed via their webservice interface and the resulting candidate compounds are downloaded automatically. Hydrogens are added explicitly to the structure where necessary.

In silico fragmentation of candidates

MetFrag generates all possible topological fragments of a candidate compound in order to match the fragment mass with the measured peaks. The problem of enumerating all possible molecular fragments can be solved by creating a fragmentation tree. The root consists of the intact molecule, and each node represents a fragment, obtained by splitting the molecule at a given bond. We implemented this as an iterative, breadth-first algorithm. One major speed determining factor is the number of fragments generated, because of the combinatorial nature of the algorithm. Thus, the *maximum tree depth* was introduced to improve the performance and specificity. We perform additional application-specific steps to prune the search space and take care of molecular rearrangements, see below. For each candidate structure the fragments are generated in the following way (Figure 2):

Initially the candidate structure is pushed into an "unprocessed" queue. The candidate structure is preprocessed using a (small) set of rules, which describe molecular rearrangements during the CID fragmentation that can not be accounted for by the simple bond disconnection approach. Each application of these rules results in one or more derived fragments which are added to the

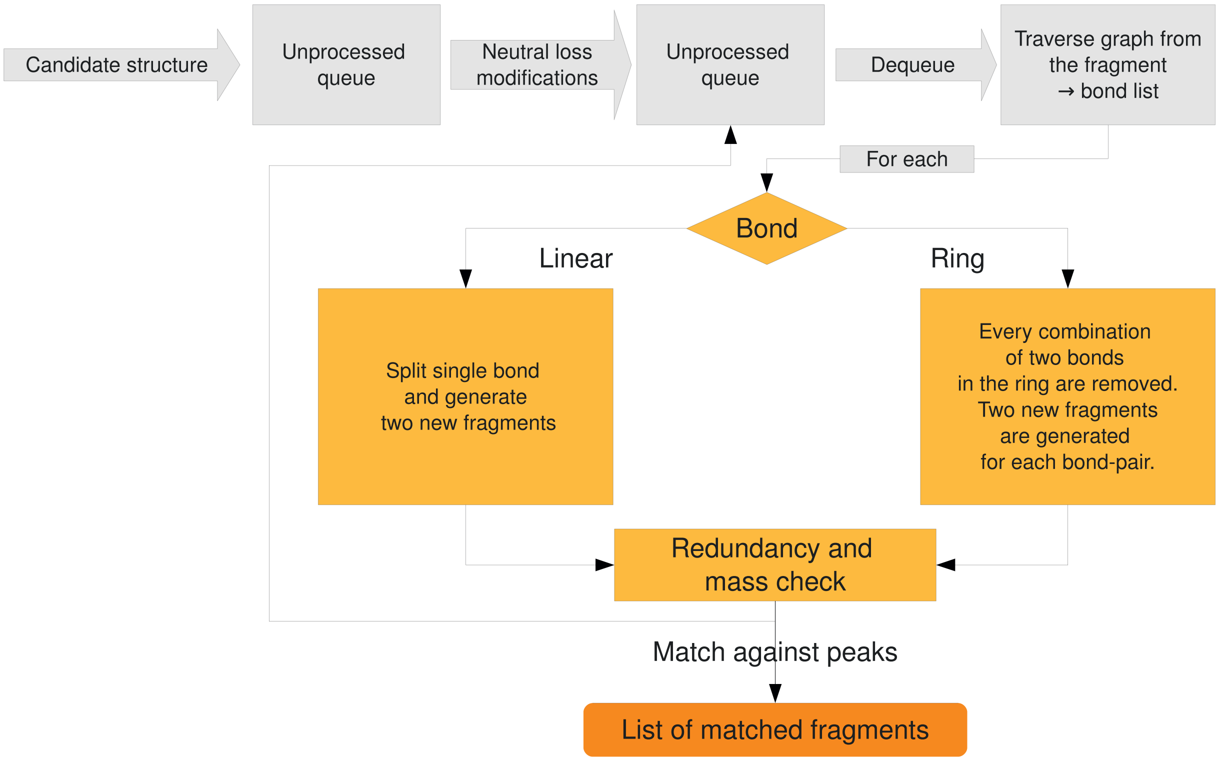

Candidate structure
Unprocessed queue
Neutral loss modifications
Unprocessed queue
Dequeue
Traverse graph from the fragment → bond list
For each
Bond
Linear
Ring
Split single bond and generate two new fragments
Every combination of two bonds in the ring are removed. Two new fragments are generated for each bond-pair.
Redundancy and mass check
Match against peaks
List of matched fragments

Figure 2 Algorithm for in silico fragmentation. Each compound is fragmented using the bond dissociation approach. Bonds in ring systems need special treatment. Every possible structure is generated until a given tree depth is reached. The redundancy heuristics and mass checks reduce the search space.

DOI:10.1186/1471-2105-11-148

"unprocessed" queue. The actual rules will be described later in this paper.

Then a structure is dequeued and its molecular graph is traversed to collect all bonds to be split. A linear bond (which is not part of a ring system) only needs to be cleaved and results in two new fragments. Within a ring system two bonds have to be split simultaneously, to create the new fragments. Only the fragments larger than the peak with the smallest mass are created, since smaller fragments can not explain an experimental peak.

Before proceeding to the next fragment, a redundancy check is performed to eliminate duplicate fragments. Redundancy occurs if a fragment *A* is part of both parent fragments *AB* and *ABC*, or the fragment *A* appears in different places of the molecule, as in *ABA*. In both cases the redundant structures would cause longer runtimes and higher memory consumption without gaining any information. In addition to full (and time consuming) graph isomorphism checks we describe simpler heuristics later in this paper.

Finally, the in silico fragments are matched against the query peaklist. The measured peaks correspond to the charged fragments, so the matching function adds (positive mode) or removes (negative mode) a proton (1.007 Da) to the fragment mass. In a few cases, fragment ions can have an intrinsic charge, where one of the heteroatoms is charged. In this case the fragment mass is used as-is, but a penalty is added to the bond dissociation energy of this fragment (see below).

The accuracy of a mass measured by an MS instrument is typically expressed relatively in ppm. In practice we found that especially for low masses, an additional (absolute) deviation has to be considered. Hence MetFrag uses two values mzppm and mzabs respectively, to calculate the mass error used for fragment matching.

Peaks that have such an explanation are subsequently removed from the query peaklist and the fragment-peak pair is saved for the final scoring. If the peak with the smallest mass has been explained, this will raise the minimal-mass cut-off, resulting in even fewer fragments that need to be considered. The "unprocessed" queue is then populated with the created and filtered fragments and processed as described above. The fragmentation terminates if the queue is empty or the maximum tree depth has been reached. The candidate is then scored based on all matched fragment-peak pairs as explained in the following section.

Scoring candidates based on fragments explaining the measured peaks

The score is an extension of a simple peak count: S_i of a candidate compound i is calculated based on all fragments F_i that explain peaks in the measured spectrum and the bond dissociation energy (BDE) calculated during the in silico fragmentation:

$$S_i = \frac{1}{\max(w)} w_i - \frac{1}{2\max(e)} e_i \tag{1}$$

where

$$\begin{aligned} w_i &= \sum_{f \in F_i} (\mathrm{int}_f)^{0.6} \cdot (\mathrm{mass}_f)^3 \\ e_i &= \frac{1}{|F_i|} \sum_{f \in F_i} \sum_{b \in B_f} \mathrm{BDE}_b \end{aligned}$$

In general a peak with a high mass and intensity is more characteristic than peaks with lower mass and intensity. This is reflected by the weighted peak count w_i, as already proposed by [3,15]. The exponents $m = 0.6$ and $n = 3$ we use are taken from the literature [15]. The weights w_i are scaled by max(w) such that it is between 0 and 1. We also take the bond dissociation energy (BDE) into account, the higher the BDE, the less likely we consider a fragment. We use the standard enthalpy change upon bond fragmentation from literature, see e.g. [16]. For each candidate f we sum up BDE_b for all bonds B_f cleaved along the fragmentation tree for the explained fragments F_i. Afterwards, for each candidate the arithmetic mean e_i of these BDEs is scaled by 2 max(e) such that it is between 0 and 0.5.

Neutral loss rules account for rearrangements

The ionised molecules typically have a single charge. After the fragmentation, the charge remains with either of the resulting fragments, the other is neutral. Because only charged ions can be measured, the mass difference between the two charged ions before and after the fragmentation is referred to as the "neutral loss" [17].

One example of a common neutral loss is H_2O, which is not a true substructure of any molecule. Instead, H_2O is formed after a hydroxyl group (OH) and a single H are split off at *different* (though usually nearby) positions (see Figure 3, where the distance is three). Because individual H atoms are not considered during the in silico fragmentation, the resulting fragment would never be found without special treatment. MetFrag is checking for structural patterns that can lead to such a non-topological fragmentation. We check within a specified topological distance of the OH-group for another hydrogen and remove both OH and H.

This non-topological fragmentation is handled by the rules shown in Table 1, other neutral losses are covered by the bond-disconnection approach. Rules can be added easily, e.g. if the compounds measured belong to

DOI:10.1186/1471-2105-11-148

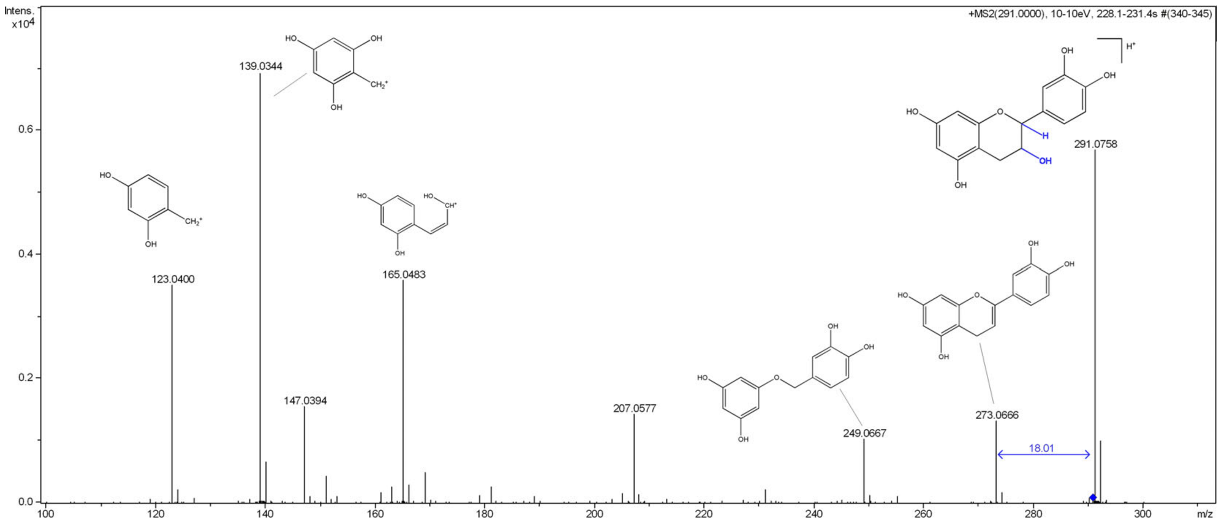


Figure 3 Annotated tandem MS spectrum of Epicatechin. This spectrum for Epicatechin was measured on a Bruker-micrOTOFQ mass spectrometer and manually annotated by an expert. The measured peaks and corresponding fragments for the major signals are depicted. In addition, the non-topological water loss is highlighted in blue.

Table 1 Neutral loss rules

Ion Mode[a]	Exact Mass[b]	Topological Fragment[c]	Neutral Loss[d]	Maximum Distance[e]
+ -	18.0106	OH	H2O	3
+ -	27.0109	CN	HCN	3
+ -	17.0266	NH2	NH3	3
+ -	30.0106	COH	CH2O	3
+	46.0055	COOH	HCOOH	3

These rules are applied to the initial candidate structures to account for rearrangements during the tandem MS fragmentation, i.e. neutral losses of unconnected fragments: [a]ionisation mode where this rule can be applied, [b]exact mass in Da of the neutral loss, [c]molecular formula of the characteristic fragment, [d]all atoms that are removed, e maximum number of bonds traversed to match neutral loss.

unusual compound classes. MetFrag reads these during start up and applies the rules to the initial candidates, resulting in new (derived) candidate molecules.

Elimination of redundant fragments

We implemented three alternative *structure redundancy checks*. Intuitively, a proper graph isomorphism check is the best approach to eliminate structures with the same molecular connectivity. In practice, graph isomorphism checks are not fast enough to process thousands of structures in reasonable time.

Alternatively we implemented an *atom based* redundancy check: each atom is labelled with a unique identifier and resulting fragments are compared to others based on atom IDs. This method will not detect the redundancy as in *ABA* mentioned above, because the atoms in the two identical substructures *A* carry different IDs. This method showed the same identification rate at much lower runtime requirements. To reduce the complexity of the test even further, the *molecular formula* redundancy check was introduced, which compares fragments based only on their elemental composition. This check will detect the *ABA* redundancy, but will produce false positives if two structures have the same elemental composition, but with different bond structure, i.e. connectivity. If two fragments have the same molecular formula, the one that requires the lower bond dissociation energy is chosen. This way the fragments which are more likely to occur are considered. The molecular formula redundancy check is used by default, because the results are comparable at considerably reduced runtime.

Structure clustering

Depending on the upstream database, the MetFrag result list can contain many similar structures or stereo isomers which have identical MetFrag scores. Therefore, we cluster the hits with tied ranks using the pairwise Tanimoto [18] distance of the molecular fingerprints, as implemented in the CDK [11]. All hits with a pairwise similarity ≥ 0.95 are collapsed into one cluster.

DOI:10.1186/1471-2105-11-148

User interface and available APIs

Our MetFrag application features an user friendly web interface, http://msbi.ipb-halle.de/MetFrag/. The required input includes the tandem MS peaklist with intensities (Figure 4, top left), selection of the upstream compound database and respective search parameters (top right). Alternatively, a list of database IDs can be provided explicitly. This allows e.g. to select the candidates based on their occurrence in specific pathways. Figure 4 also shows the results browser. A feedback form allows to store all input data, user rating of the hypotheses, and further comments. This helps to collect user-provided test- and training data. Spectra will *not* be saved unless explicitly granted. The web interface is based on Java Server Faces (JSF) [19], using the Apache MyFaces [20] implementation, ICEfaces [21] (a component library with AJAX capabilities) in an Apache Tomcat [22] servlet container. Thus, MetFrag is platform independent and accessible using most javascript enabled browsers.

We also provide a BioMoby [23] web service, which can be called from other software, including the Taverna workflow engine. Finally, the actual MetFrag algorithms are available as Java library, which can be used to perform batch searches and evaluation.

Results and Discussion

In this section we give an example of MetFrag results for an exemplary compound, and describe the full test data sets and evaluation criteria. We evaluate MetFrag on two data sets, measured on different instruments, using either KEGG or PubChem as compound library.

For the evaluation we use the merged spectra from different collision energies of compounds where the database id is known. If MetFrag returns multiple hypotheses with tied ranks, we report the most pessimistic position: even if the correct solution has the highest observed score, if 9 other candidates also have the same score, then we assign rank 10.

In addition to the worst case rank we report the *cluster rank*. Clusters of compounds having a structural Tanimoto similarity ≥ 0.95 are collapsed and treated as one *compound cluster*. Again, this measure is quite conservative, because ranks are collapsed only within results having identical scores, and still the worst case cluster rank is reported. The standard deviation of both the raw and cluster ranks for a larger benchmark data set can be quite high, therefore we report not only the average rank, but also the median and 75% quantile.

Example: Spectrum of Naringenin

As an example we show the analysis of a tandem MS spectrum of Naringenin ($C_{15}H_{12}O_5$, KEGG C00509) with MetFrag. Using KEGG as compound library with a realistic 10 ppm window around the exact mass of 272.068 Da will return 15 hits. Each candidate structure is retrieved and fragmented as described in the previous section.

After scoring each structure, the first three results can be seen in Figure 4. The details window shows the fragments that can be explained by the spectrum. The same query in PubChem yields 736 candidates, and Figure 5 shows the 9 top ranked solutions, including the correct compound at worst case rank 8. The similarity would collapse the isomers into two clusters, resulting in a cluster rank 5.

Benchmark data sets

Two data sets were used for evaluation, together consisting of 710 spectra of 151 known compounds. Current instruments allow the acquisition of so called *ramp* spectra, which combine a range of collision energies in one measurement. In both data sets the compounds were measured at different collision energies. Depending on the compound, informative fragmentation might occur only at higher energies. For other compounds, even low collision energies can lead to a very high degree of fragmentation. For this reason we use *composite* spectra: two peaks p_1 and p_2 from different collision energies are merged $\overline{\mathrm{mz}} = \mathrm{avg}(\mathrm{mz}_1, \mathrm{mz}_2)$ if $|\mathrm{mz}_1 - \mathrm{mz}_2| \leq 0.01$ Th, retaining the higher intensity $\max(int_1, int_2)$.

Data set I with compound library KEGG

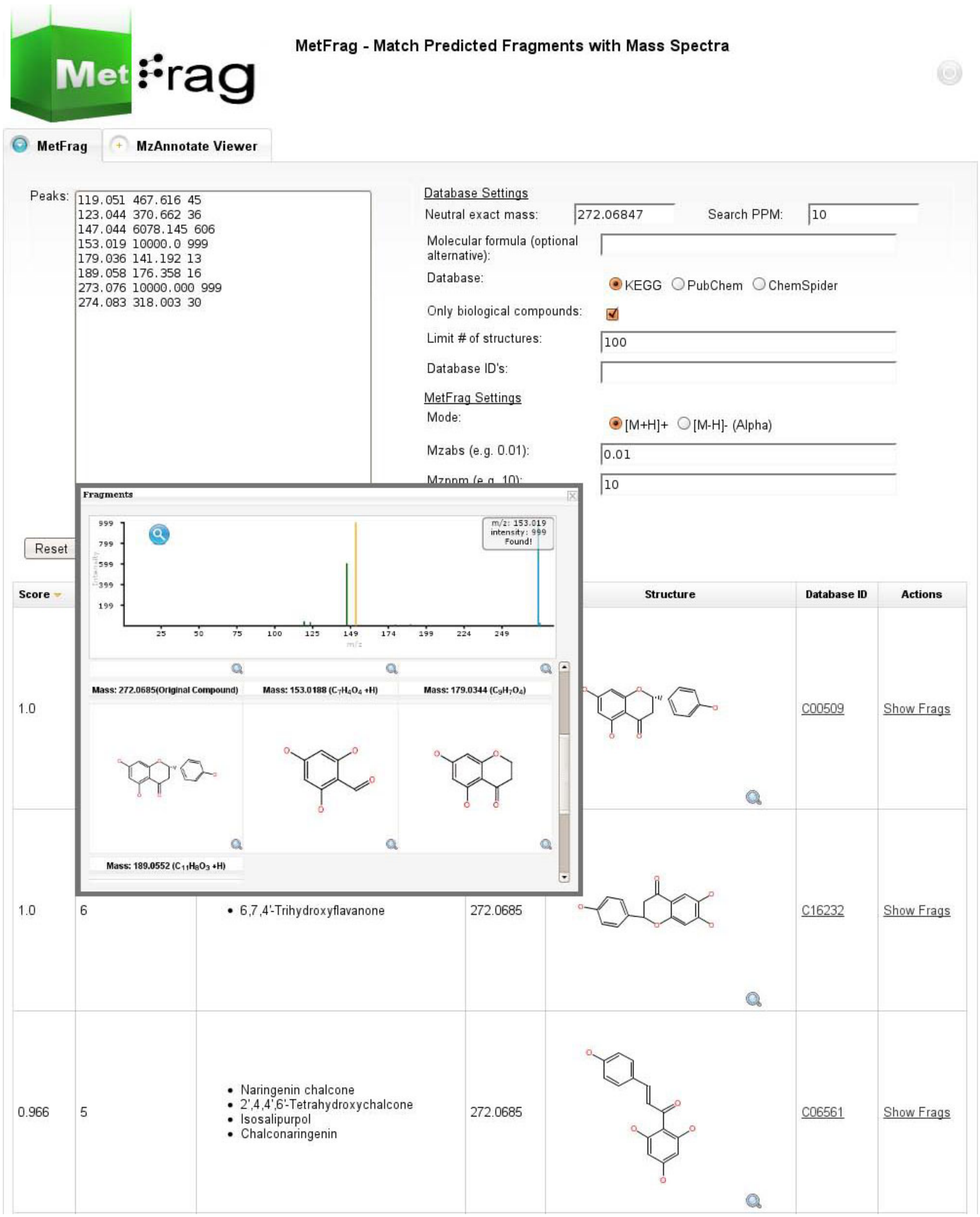
The first data set consists of 200 spectra from 49 compounds obtained on the API QSTAR Pulsar I in positive mode at several different collision energies, e.g. 10, 20, 30 and 40 eV. The spectra were measured at the IPB and are publicly available in the MassBank database http://msbi.ipb-halle.de/MassBank/, see additional file 1 for a list of accession numbers.

MetFrag was used to identify the compounds using the 49 composite spectra within KEGG. Fragments are generated until a tree depth of two is reached. The instrument specific deviation was set to mzabs = 0.01 and mzppm = 50.

The initial list of candidates obtained from KEGG contained on average 10.3 compounds. The correct compound has a median of 3 in the MetFrag result list. 25 of the correct compounds were ranked in the top 3 hits and 11 of these are ranked first. MetFrag is a great improvement over a mass-only library search. With 16 021 entries KEGG is a comparably small library. However, the compounds are highly relevant to metabolomics research.

Data set II searched against PubChem

For the second data set we used the PubChem database, with a much larger collection of natural and synthetic compounds. A collection of 102 compounds with an average mass of 372.5 Da has been measured on a



Peaks:

```
119.051 467.616 45
123.044 370.662 36
147.044 6078.145 606
153.019 10000.0 999
179.036 141.192 13
189.058 176.358 16
273.076 10000.000 999
274.083 318.003 30
```

Database Settings

Neutral exact mass: 272.06847
Search PPM: 10
Molecular formula (optional alternative):
Database: ● KEGG ○ PubChem ○ ChemSpider
Only biological compounds: ☑
Limit # of structures: 100
Database ID's:

MetFrag Settings

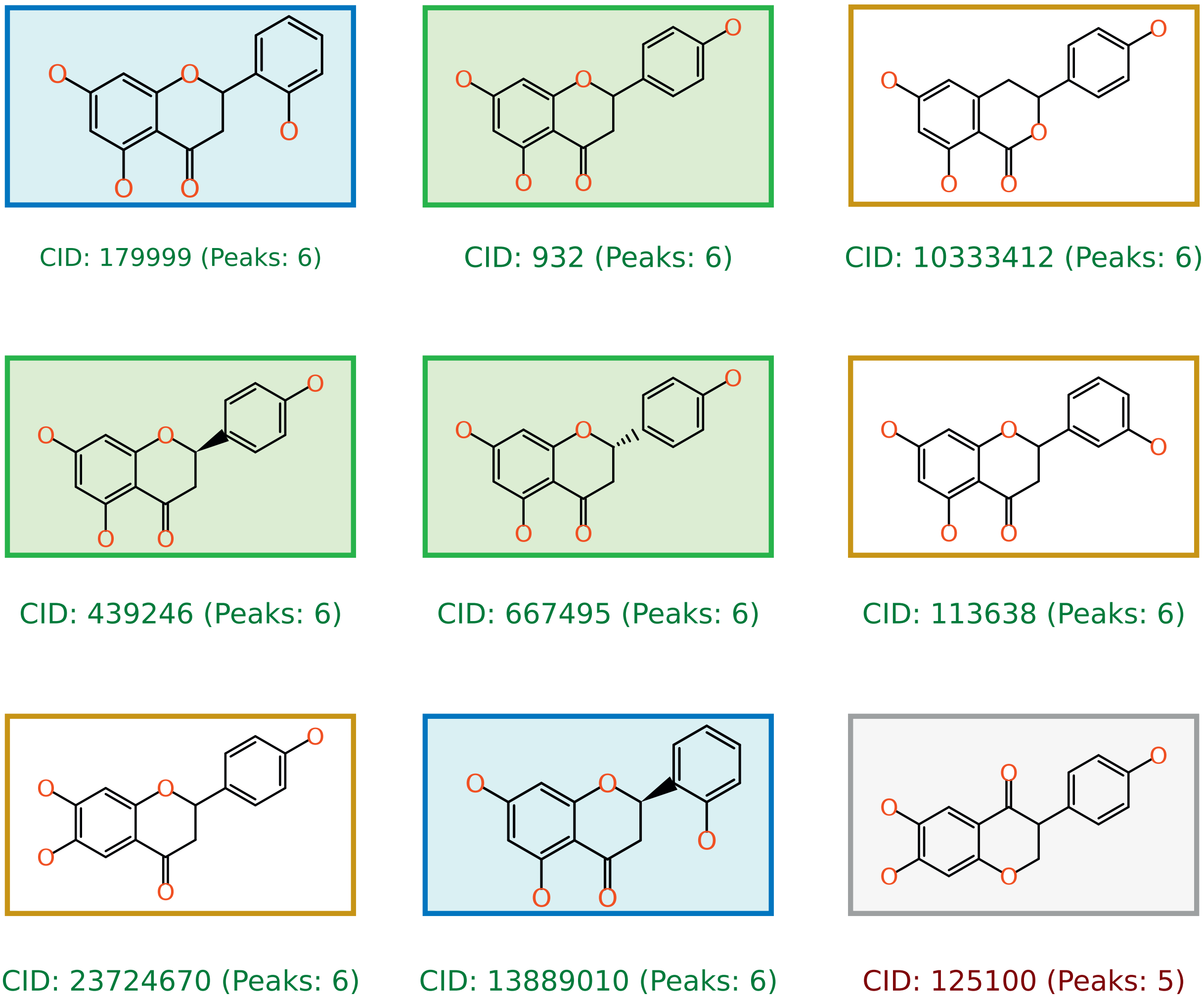
Mode: ● [M+H]+ ○ [M-H]- (Alpha)
Mzabs (e.g. 0.01): 0.01
Mzppm (e.g. 10): 10

Mass: 272.0685(Original Compound)
Mass: 153.0188 ($C_7H_4O_4$ +H)
Mass: 179.0344 ($C_9H_7O_4$)
Mass: 189.0552 ($C_{11}H_8O_3$ +H)

Score				Structure	Database ID	Actions
1.0					C00509	Show Frags
1.0	6	6,7,4'-Trihydroxyflavanone	272.0685		C16232	Show Frags
0.966	5	Naringenin chalcone 2',4,4',6'-Tetrahydroxychalcone Isosalipurpol Chalconaringenin	272.0685		C06561	Show Frags

Figure 4 MetFrag web interface. The web interface with the search parameters at the top and the result list below. The extra window can be opened for each result and shows details such as the spectrum and matching fragment structures.

DOI:10.1186/1471-2105-11-148



CID: 179999 (Peaks: 6)
CID: 932 (Peaks: 6)
CID: 10333412 (Peaks: 6)
CID: 439246 (Peaks: 6)
CID: 667495 (Peaks: 6)
CID: 113638 (Peaks: 6)
CID: 23724670 (Peaks: 6)
CID: 13889010 (Peaks: 6)
CID: 125100 (Peaks: 5)

Figure 5 Top candidates for Naringenin against PubChem. The 9 top ranked compounds where the correct solution (CID 932) is reported at (tied) rank 8. Two clusters of structures (green and blue) are identical apart from their stereochemistry, the remaining three structures (yellow) that explain all six tandem MS peaks have a Tanimoto similarity < 0.95. After clustering with a similarity ≥ 0.95 the stereoisomers are collapsed into one cluster, resulting in a cluster rank 5 for the correct solution.

Micromass Q-TOF II in positive mode and published by Hill et al. in [8]. Each compound was measured at five different collision energies: 10, 20, 30, 40 and 50 eV, for an overall of 510 spectra. All spectra are available from MassBank as well, see additional file 2 for a list of accession numbers. For the spectra from this instrument we used 10 ppm (mzabs = 0) as mass deviation and a maximum tree depth of two. Based on a PubChem snapshot (June 2009) we retrieved on average 2508 candidate compounds.

After the MetFrag scoring, the correct candidate occurred at median rank 31.5, with the structure clustering the median decreased to 14.5. The complete results are shown in additional file 2.

We were also interested in the effect of a larger tree depth: raising the tree depth to three increases the average runtime 5-fold, and worse, the prediction accuracy decreases. The median of the correct compound degraded to 39 (cluster rank 18). This behaviour can be explained with the positive predictive value (PPV):

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

where

TP = peaks explained by correct compound
FP = peaks explained by other candidates.

DOI:10.1186/1471-2105-11-148

The more (smaller) fragments are generated, the more peaks can be matched, which leads to more false positive hits. This dependency is the reason to include the exponent $mass_f^3$ in the scoring function. The higher number of false positives results in a PPV of only 0.017 (tree depth three) versus 0.028 using tree depth of two.

Similarly, we applied the neutral loss rules (Table 1) to *every* generated fragment, not just the initial candidates. Again, we obtained more matching fragments, and the PPV decreased from 0.028 to 0.017, with an even higher median of the correct compound cluster of 67.

Another aspect of the evaluation was to use individual spectra instead of the composite spectra. MetFrag showed a poor performance resulting in a median of 43 using 10 ppm. An interesting observation is that the median improved to 39.5 if the allowed mass deviation is increased from 10 ppm to 20 ppm. Because the merging (and averaging) of peaks in the composite spectra usually results in a more accurate mass, some peaks in individual spectra with a deviation beyond 10 ppm are only matched after relaxing the allowed error window to 20 ppm.

Finally, we interpreted some of the cases where MetFrag did not return good results. Table 2 shows many top 10 hits, but also several cases where MetFrag is not able to rank the correct compound even among the top 100. Some of the problematic compounds are Ormetoprim, Strychnine N-oxide and Tetramisole. One reason is the high number of very similar candidate structures, and the difficulty to distinguish them based on the predicted spectra. Another example where many similar structures occur is Tetracycline, but here the rather high rank decreased from 92 to cluster rank 10. Even these large result lists with many similar entries will still give a very good estimation of the possible compound class, which simplifies the subsequent (manual) interpretation and identification.

We also evaluated data set I (measured on the API QSTAR Pulsar I) against PubChem 2009. Because this older mass spectrometer has a much lower mass accuracy than the Micromass Q-TOF II, both the candidate search and the scoring found more false positive matches. Within the 3896 (average) candidates, the median of the correct solution is only 91. This leads to the conclusion that a good mass accuracy of ≤10 ppm is required. Almost all current QTOF instruments are specified at 5 ppm or less, and even higher accuracies are available from Orbitrap or FTICR-MS instruments.

Comparison between MetFrag and MassFrontier

In their paper [8] Hill et al. evaluate the prediction performance of MassFrontier 4.0 with an approach similar to MetFrag, using PubChem (in the version from February 2006, with $6{\cdot}10^6$ entries) as compound database. We added a constraint to our candidate search to include only compounds added in or before February 2006. Our simulated PubChem snapshot returns on average 338 candidates, the previous study only 272 structures. Nevertheless, we use following results to compare MetFrag and MassFrontier. Both MetFrag and the search procedure by Hill consider only compounds containing the elements CHNOPS and ignore molecules which consist of C, H only. The previous study reports two separate evaluation strategies: the first combines the automatic ranking with the manual a-posteriori selection of the best spectrum, obtaining the correct result on a median rank 2.5. In practice, this knowledge will not be readily available. The more realistic results are presented in the supplementary material of [8], where a heuristic was used to select one spectrum per compound. The heuristic rule chooses the spectrum with the lowest collision energy which has at most 22% of the precursor ion intensity. In this case the median drops to 4 (3^{rd} quantile at 17.5).

The median for MetFrag is 8 (3^{rd} quantile at 19), and decreases to 4 (3^{rd} quantile at 11.75) if the 95% similarity criterion is used. If the results are compared in more detail, this improvement is significant (p = 0.01), tested with a one-tailed, paired Wilcoxon signed rank test. The results for both systems are available as additional file 3.

It would be interesting to evaluate the MassFrontier approach with composite or ramp spectra, where neither automatic nor manual spectra selection would be required.

Empirical runtime evaluation

The naïve and recursive bond-disconnection approach has very high theoretical complexity. We evaluated the real-world runtime by sampling 5900 compounds (unrelated to the test sets) from PubChem with a mass between 100 and 1000 Da. In metabolomics research, only few compounds exceed a mass of 1000. Each compound was fragmented (minimum fragment mass 30 Da) to a given tree depth of two and three. Figure 6 shows the runtime of MetFrag on a PC with Intel Q9400 CPU at 2.66 Ghz and 8 Gb RAM with Ubuntu 8.04, and JVM Sun Java 1.6.0_16-b01. Each point shows the time needed to compute all fragments above 30 Da. The yellow and red lines show the non-linear runtime for tree depth two (on average 0.2 s) or three (on average 3.4s), respectively. In practice a tree depth of two has the best prediction accuracy (see above) and is fast enough to analyse compounds on demand, even with masses up to 1000 Da.

Conclusions

We have presented an algorithm which is able to identify small molecules from tandem MS measurements among a large set of candidate structures. The scoring function

DOI:10.1186/1471-2105-11-148

Table 2 Results for data set II searched against PubChem

Compound	Candidates	MassFrontier Rank	Candidates	MetFrag Rank	Cluster Rank
Thioridazine	849	1	1091	1	1
Bumetanide	619	10	768	1	1
Piperacetazine	494	1	626	1	1
Sufentanil	445	1	512	1	1
Diphenoxylate	333	4	369	1	1
Tetracaine	308	22	362	1	1
Remifentanil	246	1	286	1	1
Hydroxybutorphanol	180	2	201	1	1
Alfentanil	134	1	162	1	1
Etamiphylline	100	3	104	1	1
Ergoloid Mesylate	7	1	10	1	1
Gallamine	10	1	8	1	1
Thonzide	4	1	4	1	1
Spectinomycin	310	1	361	2	1
Methionine Enkephalin	66	1	68	2	1
Leucine Enkephalin	53	2	60	2	1
Dihydroergotamine	35	1	38	2	1
Thiothixene	726	1	909	3	1
Etodolac	420	1	580	3	1
Prednisolone Tebutate	143	4	165	3	1
Oxybutynin	114	6	156	3	1
Apramycin	54	1	60	3	1
Tenoxicam	28	1	34	3	1
Vecuronium	3	1	4	3	1
Methylergonovine	515	1	629	6	1
Rolitetracycline	105	1	151	6	1
Oxytetracycline	483	4	614	11	1
Tetracycline	529	5	673	19	1
Thiethylperazine	569	2	671	2	2
Acetophenazine	435	1	546	2	2
Mebeverine	96	2	112	2	2
Salmeterol	32	1	37	2	2
Terfenadine	34	1	35	2	2
Boldenone Undecylenate	21	2	32	2	2
Buspirone	36	1	31	2	2
Gingerol	182	2	195	3	2
Betaxolol	190	5	259	4	2
Fenoterol	370	5	521	6	2
Taurocholate	59	4	65	9	2
Aminophylline	94	21	176	3	3
Sulfadimethoxine	94	18	145	3	3
Adiphenine	623	6	796	4	3
Perindopril	102	2	119	6	3
Sulfasalazine	106	5	116	6	3
Anileridine	563	251	668	7	3
Prednisolone	269	13	363	8	3
Adenosine Diphosphate	32	3	46	9	3
⋮	⋮	⋮	⋮	⋮	⋮
Tetramisole	120	1	123	85	79
Oxaprozin	461	101	607	143	94

Table 2: Results for data set II searched against PubChem *(Continued)*

Antipyrine	306	97	341	122	104
Mefenamic Acid	579	328	633	146	124
Strychnine	664	575	882	259	171
Dimefline	644	644	876	294	175
Ormetoprim	270	124	317	233	191
Strychnine N-oxide	1185	1098	1672	1012	618
Average:	272.2 (± 24.2)	44.2 (± 14.1)	338.4 (± 31.5)	34.2 (± 10.9)	21.6 (± 6.8)
Median:	183.5	4	231.5	8	4
75% Quantile:	431.3	17.5	518.8	19	11.8
Std. Deviation:	244.1	142.4	318.1	109.8	69.1

The results on the left were reported in [8]. The corresponding MetFrag results are on the right where the candidate search was restricted to the PubChem as of February 2006 (we retrieved slightly more candidates than reported by Hill et. al.). Only the best 47 and eight worst Metfrag results are shown, the full table is given as additional file 3.

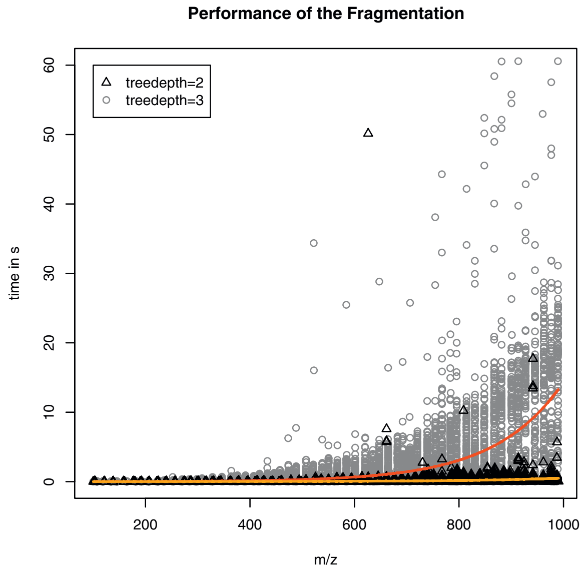


Figure 6 Empirical runtime. Runtime for the in silico fragmentation step on 5900 compounds randomly drawn from PubChem, with uniform mass distribution between 100 and 1000 Da. Limiting the tree depth of the in silico fragmentation to two (orange) results in an average runtime of 0.2 s for one compound. The exponential runtime can be seen especially when a larger tree depth (red) is used, raising the runtime to 3.4s.

does not require a set of fragmentation reactions or an actual simulation of the fragmentation process. MetFrag is able to query KEGG, PubChem and ChemSpider, and local databases can be integrated with little effort.

In comparison to the system described in [8] (which included human expertise), MetFrag achieves better results than MassFrontier.

For dedicated metabolite databases such as KEGG, the correct identification is generally among the first few candidates. Given the sheer size of generic compound libraries such as PubChem, it is no surprise that the result lists contain many structurally highly similar compounds. Hence, an unambiguous identification is generally not possible, but usually the compound class can be derived from the results. A principal limitation is the inability to distinguish stereoisomers which is not possible from MS data alone. The final identification according to MSI recommendations [24] requires the comparison against spectra of authentic standards, or even complementary analysis methods such as NMR.

Our tool MetFrag improves the identification of unknown substances from tandem MS spectra. It is fast enough to be used in the interactive web application, and has a user-friendly interface and result browser.

Availability and Requirements

- Project home page: http://metware.org/
- Operating system(s): Platform independent
- Programming language: Java
- Other requirements: Java ≥ 1.6, Tomcat ≥ 6.0
- License: GNU LGPL v3 (or later)

Additional file 1: MassBank_KEGG_results. Full list of mass spectra and compounds used in section "Data set I searched against KEGG". This includes accession numbers in the MassBank system. For each compound the number of candidates and the rank of the correct solution is given.

Additional file 2: HillData_PubChem2009. Full list of mass spectra and compounds used in section "Data set II searched against PubChem". This includes accession numbers in the MassBank system. For each compound the number of candidates and the rank of the correct solution is given.

Additional file 3: Comparison_MassFrontier_MetFrag_PubChem2006. This file includes the full results from table 2 in section "Data set II searched against PubChem". The candidate search was restricted to the PubChem as of February 2006. For convenience, we also include the results reported in [8].

Acknowledgements
We thank the CDK team (especially Egon Willighagen and Rajarshi Guha) for this cheminformatics library, and Michael Gerlich, Carsten Kuhl and Ralf

Tautenhahn for helpful discussions. SS is funded by the BMBF (GABI-ProTect 0315051C).

Author details

[1]Leibniz Institute of Plant Biochemistry- Department of Stress- and Developmental Biology, Weinberg 3, 06120 Halle(Saale), Germany. [2]Institut für Informatik, Martin-Luther-Universität, Halle-Wittenberg, Von-Seckendorffplatz 1, 06120 Halle (Saale), Germany.

Authors' contributions

SW implemented the MetFrag application, web interface and performed the evaluation. SS provided the MS expertise, MM-H and SN provided input for the requirements, the algorithmic design and architecture. All authors contributed to, read and approved the final manuscript.

Received: 24 November 2009 Accepted: 22 March 2010
Published: 22 March 2010

References

1. Dunn WB: **Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes.** *Physical Biology* 2008, **5**:011001, (24pp).
2. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D: **GMD@CSB.DB: the Golm Metabolome Database.** *Bioinformatics* 2005, **21(8)**:1635-1638.
3. Horai H, Arita M, Nishioka T: **Comparison of ESI-MS Spectra in MassBank Database.** *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on* 2008, **2**:853-857.
4. Smith CA, Maille GO, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G: **METLIN: A Metabolite Mass Spectral Database.** *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology* Louisville, Kentucky 2005, **27**:747-751.
5. **ACD/MS Fragmenter.** [http://www.acdlabs.com/products/adh/ms/ms_frag/].
6. **Mass Frontier.** [http://www.highchem.com/].
7. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, Rousu J: **FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data.** *Rapid Communications in Mass Spectrometry* 2008, **22(19)**:3043-3052.
8. Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF: **Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra.** *Analytical Chemistry* 2008, **80(14)**:5574-5582.
9. Heinonen M, Rantanen A, Mielikäinen T, Pitkanen E, Kokkonen J, Rousu J: **Ab initio prediction of molecular fragments from tandem mass spectrometry data.** *German Conference on Bioinformatics* 2006, 40-53.
10. Hill AW, Mortishire-Smith RJ: **Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach.** *Rapid Communications in Mass Spectrometry* 2005, **19(21)**:3111-3118.
11. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E: **The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics.** *Journal of Chemical Information and Computer Sciences* 2003, **43(2)**:493-500.
12. **KEGG Compound.** [http://www.genome.jp/kegg/compound/].
13. **PubChem.** [http://pubchem.ncbi.nlm.nih.gov/].
14. **Chemspider.** [http://www.chemspider.com/].
15. Stein SE, Scott DR: **Optimization and testing of mass spectral library search algorithms for compound identification.** *Journal of the American Society for Mass Spectrometry* 1994, **5(9)**:859-866.
16. Luo Y: *Handbook of bond dissociation energies in organic compounds* Boca Raton, CRC Press 2003.
17. Gross JH: *Mass Spectrometry: A Textbook* Springer, Berlin, 1 2004, corr. 2nd printing edition 2004.
18. Butina D: **Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets.** *Journal of Chemical Information and Computer Sciences* 1999, **39(4)**:747-750.
19. **Java Server Faces.** [http://java.sun.com/javaee/javaserverfaces/].
20. **Apache My Faces.** [http://myfaces.apache.org/core12/index.html].
21. **ICEfaces.** [http://www.icefaces.org/].
22. **Apache Tomcat 6.** [http://tomcat.apache.org/].
23. **Biomoby.** [http://www.biomoby.org/].
24. Sumner LW, Amberg A, Barrett D, Beale M, Beger R, Daykin C, Fan T, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane A, Lindon JC, Marriott P, Nicholls A, Reily M, Thaden J, Viant MR: **Proposed minimum reporting standards for chemical analysis.** *Metabolomics* 2007, **3(3)**:211-221.

doi:10.1186/1471-2105-11-148
Cite this article as: Wolf *et al.*: **In silico fragmentation for computer assisted identification of metabolite mass spectra.** *BMC Bioinformatics* 2010 **11**:148.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Journal of Integrative Bioinformatics, 8(2):157, 2011 http://journal.imbio.de

Database supported candidate search for Metabolite identification

Christian Hildebrandt[1,2], Sebastian Wolf[2], Steffen Neumann[2*]

[1]Anhalt University of Applied Sciences, Department of Computer Science, Lohmannstr. 23,
06366 Köthen (Anhalt), Germany, `http://www.hs-anhalt.com`

[2]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3,
06120 Halle (Saale), Germany, `http://www.ipb-halle.de`

Summary

Mass spectrometry is an important analytical technology for the identification of metabolites and small compounds by their exact mass. But dozens or hundreds of different compounds may have a similar mass or even the same molecule formula. Further elucidation requires tandem mass spectrometry, which provides the masses of compound fragments, but *in silico* fragmentation programs require substantial computational resources if applied to large numbers of candidate structures.
We present and evaluate an approach to obtain candidates from a relational database which contains 28 million compounds from PubChem.
A training phase associates tandem-MS peaks with corresponding fragment structures. For the candidate search, the peaks in a query spectrum are translated to fragment structures, and the candidates are retrieved and sorted by the number of matching fragment structures. In the cross validation the evaluation of the *relative ranking positions* (RRP) using different sizes of training sets confirms that a larger coverage of training data improves the average RRP from 0.65 to 0.72. Our approach allows downstream algorithms to process candidates in order of importance.

1 Introduction

Mass spectrometry is an important analytical technology in systems biology, and allows the detection of a large number of metabolites in biological samples. For a biological interpretation, their structures and/or accession numbers are required. Individual metabolites can be identified by their accurate mass, but dozens or hundreds of different compounds may have a similar mass or even the same molecular formula (and hence identical mass).

In tandem mass spectrometers, such as hybrid instruments like a triple quadrupole (QqQ), or quadrupole coupled to a time-of-flight analyser (QqTOF), the molecules of interest are isolated in the first quadrupole. This filter allows only the molecules within a narrow *precursor mass* window to pass through, and other molecules are discarded. These filtered molecules undergo collision induced dissociation (CID) in the second quadrupole (the so called collision cell), where they literally break apart. The masses (more correctly, the mass-over charge ratio m/z) of the resulting fragments are measured in the final mass analyser, either another quadrupole, or

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

---

[*] To whom correspondence should be addressed. Email: Steffen.Neumann@ipb-halle.de

DOI:10.2390/biecoll-jib-2011-157

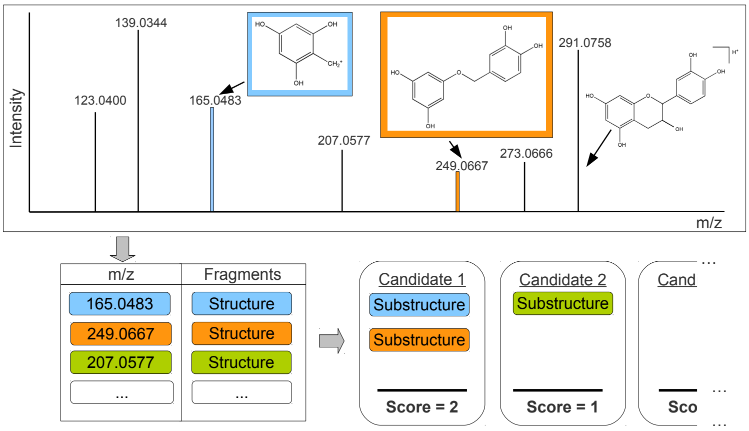


Figure 1: Top: tandem MS spectrum of Epicatechin, with some manually annotated fragment peaks. Below: database table with m/z → fragment association. Candidate compounds are scored, based on the number of substructures they contain.

a high-resolution time-of-flight (TOF) analyser. Other instruments such as Iontraps or Orbitrap perform these steps sequentially in time, rather than in different instrument compartments. A typical result of the fragmentation is shown in the tandem mass spectrum of Epicatechin in Figure 1.

For metabolite identification, a query spectrum can be compared with reference spectra from databases like MassBank [6] or commercial libraries provided by several vendors [7]. However, their chemical coverage is far from complete, especially in areas such as plant metabolomics, where most of the estimated 200 000 compounds are still uncharacterised [1].

If reference spectra are not available, the spectra can be interpreted using computational mass spectrometry methods, such as FiD [3], or the commercial ACD Fragmenter and HighChem's MassFrontier – see [7] for a review. These programs can also be used to search general purpose compound libraries, such as KEGG with about 14 215 metabolite structures or the much larger PubChem database with 28 million compounds [5, 10].

The MetFrag approach is designed to search online accessible compound databases with the accurate mass of the unfragmented metabolite. MetFrag obtains *candidates* from the compound databases, fragments these *candidates in-silico*, and scores the match between the query spectrum and the *in silico* fragments.

However, analysing thousands of candidate structures is a time-consuming process, especially for non-trivial compounds, and may take hours on a single machine. For example it takes ≈3 minutes to process about 1 672 candidates of strychnine N-oxide [10]. But for some spectra there are even more candidates. All hypotheses are processed in the (arbitrary) order determined by the candidate search. The correct compound might appear first, or towards the end of the list. If the candidate search would already pre-sort the corresponding candidates, it would be possible to process and display the correct one earlier.

The MetFrag web application will implement a dynamically updated user interface, and process all candidates in smaller batches. That way it is possible to present informative (but still preliminary) results almost from the beginning. The final result in MetFrag after completion of all candidates remains the same. Alternatively, the set of candidates can be filtered based on

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

the preliminary scores, and the subsequent MetFrag runtime would be reduced.

In this paper, we present the MassStruct approach to learn the association between the measured mass peaks and fragment structures, which allows to integrate the accurate molecule mass search with the score-based ordering. The next section describes the system architecture, the training phase and the candidate retrieval with dynamically generated SQL queries during operation. In section 3 we evaluate our approach on a dataset of 240 spectra from 218 unique compounds, and assess the runtime of the dynamically generated query.

2 Implementation

The MassStruct approach requires an offline preprocessing step to associate measured peak masses to the corresponding fragment structures in a set of training spectra. Afterwards these fragments are grouped by their mass. During a candidate search, the molecule mass and the peaks of a query spectrum are both combined into a single dynamically generated SQL query. If one or more fragment structures of a given mass exist within one candidate, one match is counted and added to the score of this candidate.

2.1 Learning the association between mass and fragment structure

The training spectra are processed with the MetFrag algorithm, to obtain a set of m/z → structure associations as shown in Figure 2. The training set of tandem MS spectra is synthetically fragmented with MetFrag and all annotated fragments are stored with their corresponding mass into a relational database.

MetFrag usually is not able to annotate every measured peak with a structure, and it is possible that one observed m/z value can be explained by different structures in different compounds, therefore all alternatives are stored.

We developed a batch import to store the fragments and their masses into a PostgreSQL 9.0 RDBMS[1] with the chemistry extension pgchem[2] 1.3-GiST [8]. All of the chemical algorithms and datatypes are handled by functions in the chemistry library OpenBabel[3] 2.3.0 [2]. The RDBMS integration allows chemical calculations, comparisons and predicates as a part of SQL statements. The ER diagram of the developed database is shown in Appendix B.

2.2 Multiple substructure database queries

The candidate retrieval query (see abbreviated query in Figure 3) selects all compounds within an error margin around the precursor mass. Then, any fragments matching the measured peak masses (within an error window) are joined with the condition `fragment.structure <= compound.structure`. This is provided by the OpenBabel chemistry algorithms and tests whether the fragment is a substructure of the candidate. The sum of the matched substructures is used as `score` for the `accession`.

---

[1] `http://www.postgresql.org`

[2] `http://pgfoundry.org/projects/pgchem`

[3] `http://www.openbabel.org/`

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

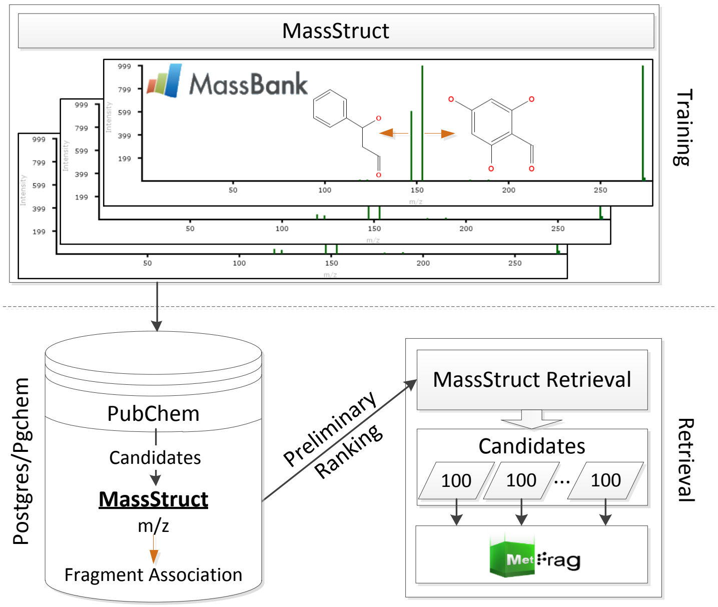


Figure 2: The training step (top) shows the MassBank spectra and the fragments predicted with MetFrag. The peaks annotated with a fragment structure are stored in the MassStruct database (lower left). In the operation phase the stored m/z → structure associations are used to retrieve the ordered candidates in batches of e.g. 100 structures.

To determine whether a fragment is a substructure of a molecule, their chemical fingerprints are compared. These 1536 bit fingerprints store characteristic chemical properties (such as bond- and atom counts or functional groups). For substructure searches, pgchem compares these fingerprints between the query molecule and the database content (primary filtering), using a Generalized Search Tree index (GiST) [4]. Afterwards, a time consuming substructure matching (secondary filtering) of the molecular structures on the previously selected records is done. All chemical operations benefit from the PostgreSQL query planning optimization.

The unabbreviated query in Appendix A also takes into account that 1) the database contains compounds from multiple compound libraries, and candidates can be restricted to a certain library 2) the compounds in the strcuture libraries might occur in multiple stereo conformations. Since mass spectrometry can hardly distinguish stereo isomers, MetFrag ignores the stereochemistry. Redundant candidates are removed by the query, such that only the first compound is considered. Because the fragments are measured with a certain error, the fragment masses are grouped into m/z cluster by hierarchical clustering analysis (HCA). The actual score counts at most one matching substructure per m/z cluster.

3 Results and Discussion

In the following we are going to present an example, and assess two separate performance aspects of the system. We evaluate the ability of the scoring to obtain the correct compound with a good rank, simulating various training set sizes. Second, we report the runtime on a snapshot (Q4 2010) of the PubChem compound database.

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

```sql
SELECT accession, count(fragment.id) AS score
FROM compound, fragment
WHERE compound.mass BETWEEN 290.2 AND 290.3
AND ( fragment.mass BETWEEN 123.0 AND 123.1
   OR fragment.mass BETWEEN 139.0 AND 139.1
   OR fragment.mass BETWEEN 165.0 AND 165.1
   OR fragment.mass BETWEEN 207.0 AND 207.1
   OR fragment.mass BETWEEN 249.0 AND 249.1
   OR fragment.mass BETWEEN 273.0 AND 273.1)
AND fragment.structure <= compound.structure
GROUP BY accession
ORDER BY score;
```

Figure 3: A SQL statement performing a combined search for the molecule's `mass`, and `fragments` which are a substructure (the <= predicate) ranked by `score`, where <= is the `chemical_substructure` operator. The peak data corresponds to the example spectrum in Figure 1.

3.1 Metabolite identification results

We used 240 metabolite spectra (see Appendix C or supplementary files as xls or csv hosted on `http://msbi.ipb-halle.de/msbi/massstruct`) with known PubChem accessions obtained from MassBank. These spectra contain data of several compounds, some of them were measured repeatedly with different instrument settings, so they covered 218 different compounds. Together, all spectra contained 2 083 peaks, and MetFrag was able to annotate 1 280 fragments with the parameters reported earlier [10]. The PubChem compound snapshot (Q4 2010) contained 28 838 421 structures. Including the indices, the database occupied ≈150 GB storage space.

To evaluate our approach, we annotated a randomly drawn sample of the 240 spectra, and used the remaining spectra as query spectra. For each query spectrum, we count the total number of candidates (TC), those with a better and those with a score worse than the correct compound (BC and WC, respectively). This allows to calculate a *relative ranking position* $RRP = 0.5\left(1 - \frac{BC-WC}{TC-1}\right)$, where the first position results in $RRP = 1$, and $RRP = 0$ in the worst case. A similar RRP was introduced in [9], where the authors used $RRP = 0$ for the best case. We modified the scoring to keep it consistent with the MetFrag scoring.

If all candidates have the same score, then $BC = WC = 0$, and hence $RRP = 0.5$. Similarly, a random score would also lead to an average $RRP = 0.5$ on a larger test set.

For evaluation we partitioned the set of spectra, again storing one subset of m/z → structure associations in the database, and used the remaining ones to evaluate the rank of the correct solution in the ordered result set. We used different ratios (1:1, 2:1, 3:1, 4:1 and 9:1) for partitioning, to simulate an increasing coverage of the training spectra in the dataset. The results are shown in Table 2. The average RRP increases from 0.65 to 0.72, and even more apparent the median RRP raises to 0.84 if the large training sets are used. An extract of an example for one evaluation run is summarized in Table 1.

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

Table 1: The best and worst examples from one of the evaluation runs.

CID	formula	mass	TC	RRP	runtime in s
13804	$C_{15}H_{16}O_9$	340.07	50 910	0.999	476
834	$C_7H_{14}N_2O_4S$	222.06	34 491	0.999	282
5319853	$C_{21}H_{22}O_{11}$	450.11	72 127	0.999	867
165627	$C_6H_{11}NO_4$	161.06	8 743	0.999	99
439155	$C_{14}H_{20}N_6O_5S$	384.12	105 691	0.998	949
442456	$C_{28}H_{34}O_{14}$	594.19	9 745	0.997	200
160556	$C_{11}H_{20}N_2O_6$	276.13	63 768	0.997	497
5318759	$C_{21}H_{18}O_{12}$	462.07	54 545	0.996	686
2901	$C_6H_{11}NO_2$	129.07	4 974	0.995	25
101781	$C_{21}H_{22}O_{11}$	450.11	72 127	0.995	1 147
5281673	$C_{21}H_{20}O_{12}$	464.09	59 828	0.995	469
5316673	$C_{21}H_{20}O_{10}$	432.10	78 553	0.995	638
...	...	...	...	...	...
637540	$C_9H_8O_3$	164.04	13 098	0.296	191
649	$C_4H_6N_2O_2$	114.04	4 175	0.261	41
70346	$C_7H_8N_4O_3$	196.05	22 378	0.250	208

Table 2: RRP **of different partition sizes.**

	RRP		
Partition	median	ø	Std. Err.
1:1	0.50	0.65	± 0.018
2:1	0.70	0.71	± 0.019
3:1	0.69	0.70	± 0.019
4:1	0.75	0.71	± 0.019
9:1	0.84	0.72	± 0.019

3.2 Runtime and PostgreSQL database tuning

The query spectra result in 31 700 candidates on average (green circle in Figure 4), 16 630 in the median and in a few cases up to 100 000. The mean runtime of a query is 330s, or roughly 10ms per candidate.

The (virtual) database server had 2 CPUs, 2 GB RAM, and was hosted on a VMWare ESX cluster with 2.6 GHz Intel Xeon CPUs. The data partition was kept on a FC-SAN storage system.

The runtime clearly depends on the number of candidates. Therefore, any increase in e.g. instrument accuracy will decrease both the number of candidates and the runtime. The performance of an RDBMS often depends on the speed of the storage subsystem, but not in this case: the majority of time is spent in the actual sub-structure search, and the CPU speed is the limiting factor. Latencies for multiple concurrent queries can best be reduced using a server with a sufficient number of CPU cores.

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

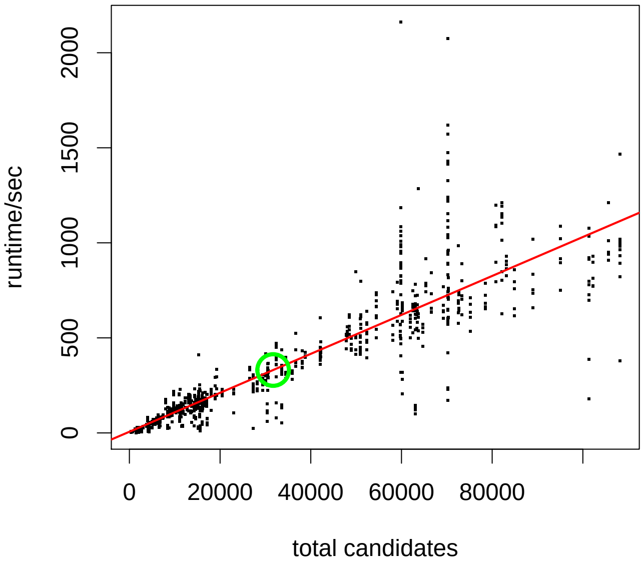


Figure 4: Runtimes of all candidate queries in the data set. The slope of the regression line is 10ms/candidate, the average (32 000 candidates in ≈ 5min) is encircled.

4 Conclusion

The process of structure elucidation with mass spectrometry data has been – and still is – the major bottleneck in metabolomics experiments. Starting from the mass of the molecule, dozens to thousands of candidates can be retrieved from compound databases like KEGG or PubChem, and subsequently analysed with computer aided structure elucidation (CASE) systems.

We introduced the MassStruct approach, improving the initial candidate query step to provide an *ordered* list of candidates. We evaluated the method with a medium sized test dataset. The benefit is that the interactive MetFrag web application can then process the candidates in batches of 100 or 1000 structures, and present intermediate results. Since the candidates are pre-sorted, the user might be satisfied after the first few iterations. The source code (including training procedure and dynamic queries) is available under the GNU General Public License from `https://github.com/childebr/MassStruct/`.

Future developments will be reducing the number of candidates to consider, e.g. by filtering the common ranges of ratios between elements for biological compounds. The growing number of spectral data in reference libraries such as MassBank will further improve the performance of the system by adding more m/z → structure associations.

References

[1] Oliver Fiehn. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2(3):155–168, 2001.

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

[2] Rajarshi Guha, Michael T Howard, Geoffrey R Hutchison, Peter Murray-Rust, Henry Rzepa, Christoph Steinbeck, Jörg Wegner, and Egon L Willighagen. The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model*, 46(3):991–998, 2006.

[3] Markus Heinonen, Ari Rantanen, Taneli Mielikäinen, Juha Kokkonen, Jari Kiuru, Raimo A. Ketola, and Juho Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom*, 22(19):3043–3052, September 2008.

[4] Joseph M. Hellerstein, Jeffrey F. Naughton, and Avi Pfeffer. Generalized Search Trees for Database Systems. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, pages 562–573. Morgan Kaufmann, 1995.

[5] Dennis W Hill, Tzipporah M Kertesz, Dan Fontaine, Robert Friedman, and David F Grant. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Anal Chem*, 80(14):5574–5582, Jul 2008.

[6] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, Jul 2010.

[7] Steffen Neumann and Sebastian Böcker. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal Bioanal Chem*, 398(7-8):2779–2788, Dec 2010.

[8] Ernst-Georg Schmid. *Database-driven procurement of substances in the researching chemical industry - An algorithmic optimization approach*. PhD thesis, Mercator School of Management - Fakultät für Betriebswirtschaftslehre - Technology and Operations Management - Wirtschaftsinformatik und Operations Research, June 2010.

[9] Emma L Schymanski, Markus Meringer, and Werner Brack. Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Anal Chem*, 81(9):3608–3617, May 2009.

[10] Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148, 2010.

Copyright 2011 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

DOI:10.2390/biecoll-jib-2011-157

Metabolomics (2013) 9:S84–S91
DOI 10.1007/s11306-012-0401-0

ORIGINAL ARTICLE

Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data

Steffen Neumann · Andrea Thum · Christoph Böttcher

Received: 9 November 2011 / Accepted: 22 January 2012 / Published online: 15 February 2012
© Springer Science+Business Media, LLC 2012

Abstract Liquid chromatography–mass spectrometry (LC–MS) is a commonly used analytical platform for non-targeted metabolite profiling experiments. Although data acquisition, processing and statistical analyses are almost routine in such experiments, further annotation and subsequent identification of chemical compounds are not. For identification, tandem mass spectra provide valuable information towards the structure of chemical compounds. These are typically acquired *online*, in data-dependent mode, or *offline*, using handcrafted acquisition methods and manually extracted from raw data. Here, we present several methods to fast-track and improve both the acquisition and processing of LC–MS/MS data. Our nearly online (*nearline*) data-dependent tandem MS strategy creates a minimal set of LC–MS/MS acquisition methods for relevant features revealed by a preceding non-targeted profiling experiment. Using different filtering criteria, such as intensity or ion type, the acquisition of irrelevant spectra is minimized. Afterwards, LC–MS/MS raw data are processed with feature detection and grouping algorithms. The extracted tandem mass spectra can be used for both library search and de-novo identification methods. The algorithms are implemented in the R package MetShot and support the export to Bruker, Agilent or Waters QTOF instruments and the vendor-independent TraML standard. We evaluate the performance of our workflow on a Bruker micrOTOF-Q by comparison of automatically acquired and extracted tandem mass spectra obtained from a mixture of natural product standards against manually extracted reference spectra. Using *Arabidopsis thaliana* wild-type and biosynthetic gene knockout plants, we characterize the metabolic products of a biosynthetic pathway and demonstrate the integration of our approach into a typical non-targeted metabolite profiling workflow.

Keywords Metabolomics · Tandem mass spectrometry · Data-dependent acquisition · Feature detection · Feature grouping · Collision-induced dissociation

1 Motivation

Today, liquid chromatography–mass spectrometry (LC–MS) is a key technology for targeted and non-targeted profiling of small molecules. In contrast to targeted approaches where a small set of known compounds is analyzed, non-targeted profiling aims at a comprehensive analysis of all detectable compounds without any prior knowledge. When applied to biological samples, non-targeted approaches have a high potential to reveal novel biomarkers predicting e.g. a disease state (Jansson et al. 2009; Wang et al. 2011) or to explore biosynthetic pathways (Böttcher et al. 2008; Okazaki et al. 2009). However, they require efficient means to assign detected molecular entities characterized by unique pairs of mass-to-charge

---

Electronic supplementary material The online version of this article (doi:10.1007/s11306-012-0401-0) contains supplementary material, which is available to authorized users.

S. Neumann (✉) · C. Böttcher
Department of Stress and Developmental Biology,
Leibniz Institute of Plant Biochemistry, Weinberg 3,
06120 Halle, Germany
e-mail: sneumann@ipb-halle.de

C. Böttcher
e-mail: cboettch@ipb-halle.de

A. Thum
Institute of Computer Science, Martin-Luther-Universität
Halle-Wittenberg, 06099 Halle, Germany
e-mail: thum@informatik.uni-halle.de

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

ratios (*m*/*z*) and retention times (features) to individual chemical compounds and to subsequently elucidate their molecular structure.

The first step of a non-targeted profiling experiment is the acquisition and analysis of a (high) number of LC–MS profiles. Besides the instrument specific vendor software, several open source software packages exist to process the resulting raw data including XCMS (Smith et al. 2006) and MZmine (Pluskal et al. 2010), or closed-source tools like MetAlign (Tikunov et al. 2005). In combination with downstream statistical analyses, the profiling data can reveal a potentially large number of “interesting” features with different intensities between sample classes or characteristic trends in time series experiments. For any further interpretation, the chemical compounds underlying these features have to be identified.

The mass and relative isotope abundance accuracy of modern high-resolution MS platforms facilitate the establishment of putative elemental compositions (Kind and Fiehn 2006; Böcker et al. 2008). In addition, tandem mass spectra provide valuable structural information for the elucidation of the underlying molecular structure. Hybrid instruments such as quadrupole time-of-flight (QTOF) or linear quadrupole trap Orbitrap mass spectrometers permit fragmentation of individual precursor ions by collision-induced dissociation (CID) and detection of the resulting fragment ions with high mass accuracy. The elemental composition of fragment ions which can be readily derived from such spectra can provide valuable hints for structural elucidation.

Most mass spectrometers allow acquisition of tandem mass spectra *online* in data-dependent acquisition (DDA) mode: the instrument performs a survey scan and selects the most abundant peak(s) for the following tandem MS scan(s). Both static and dynamic exclusion lists try to reduce the acquisition of redundant or uninteresting (solvent, plasticizer and other chemicals) tandem mass spectra. For proteomics applications Hoopmann et al. (2009) describes a method of iterative DDA, which increases the ratio of useful peptide- to unrelated spectrum acquisitions. The first problem of the existing DDA approaches is the need for survey scans. In particular when using fast chromatographic separations resulting in narrow peaks of a few seconds, the number of tandem MS scans across a chromatographic peak is drastically reduced. Closely related is the problem that the tandem MS scan should ideally be measured at the apex of a chromatographic peak to obtain a high quality spectrum. Not all vendors provide such an “apex-prediction” algorithm in their software. Low molecular weight compounds exhibit non-uniform ionization properties, and form, in contrast to peptides, different types of adduct, fragment and cluster ions upon electrospray ionization (ESI) (Brown et al. 2009, 2011; Draper et al. 2009; Kuhl et al. 2011). Usually, the quasi-molecular ions are subjected to tandem MS analysis, but the DDA approach completely ignores origin and type of detected ion species. Finally, the DDA strategy has no knowledge about “interesting” features, revealed by a preceding non-targeted profiling experiment, and instead acquires spectra irrelevant to the biological experiment at hand.

For these reasons, LC–MS/MS experiments in metabolomics are usually done in a targeted way with manually created lists of retention time and precursor *m/z* windows. The tandem mass spectra are extracted from the resulting LC–MS/MS raw data by averaging an (again manually selected) range of spectra, using the instrument vendor’s software. This is an *offline* process, sometimes days or weeks after the profiling experiments have been performed. Our nearly online (*nearline*) data-dependent tandem MS strategy aims to close the gap between instant online tandem MS acquisition which has no knowledge about features of interest, and the tedious manual approach.

The obtained tandem mass spectra can be searched against spectral libraries such as the commercial NIST library ’08 (Gaithersburg, MD) or spectral libraries from the academic community, including METLIN (Smith et al. 2005) or MassBank (Horai et al. 2010). If no reference spectra are available, tools such as MetFrag (Wolf et al. 2010) can support the identification.

In this paper we present our *nearline* data-dependent tandem MS data acquisition and processing strategies. In the next section we will explain how to schedule a minimum number of LC–MS/MS methods to cover a set of interesting features for CID experiments, and how these are translated to machine control methods. After that, we show how the resulting LC–MS/MS raw data are processed with feature detection and grouping algorithms to obtain tandem mass spectra. We apply this strategy to a mixture of natural product standards and discuss the quality of the resulting tandem mass spectra. Finally, we show integration of our workflow into a non-targeted metabolite profiling experiment.

2 Methods

The general workflow for our *nearline* data-dependent tandem MS strategy is shown in Fig. 1. It consists of the following steps: ① creation of a target list of interesting features for tandem MS analysis; ② preparation and export of LC–MS/MS methods, followed by data acquisition; ③ processing of the resulting LC–MS/MS raw data and extraction of tandem mass spectra into compound spectra; ④ querying compound spectra against reference libraries or subjecting to de-novo identification approaches.

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

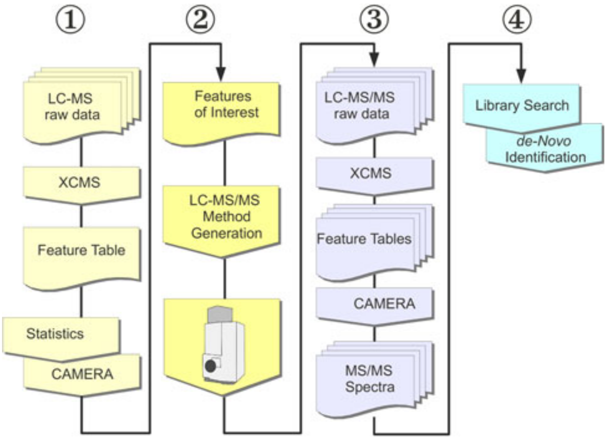


Fig. 1 The workflow for *nearline* data-dependent tandem MS data acquisition and processing. The steps ② and ③ are covered by the MetShot software package

2.1 Target selection and LC–MS/MS method generation

Target lists for *nearline* data-dependent tandem MS include m/z and retention time ranges of features to be analyzed. The retention time ranges should cover the complete chromatographic peak. In addition, specific instrument parameters are required, such as isolation width and collision energy for the Bruker micrOTOF-Q. Target lists can be generated in the simplest way manually, but here we describe a way to automate this task.

Typical LC–MS profiles of biological samples contain thousands of features. It is impossible to perform routine tandem MS analysis on each of them within a non-targeted profiling experiment. However, statistical analysis of LC–MS profiles in the context of the experimental design usually reveals tens to a few hundred "interesting" features which have to be characterized further. Such a set of features serves as a target list for our *nearline* data-dependent tandem MS approach, and a set of filters and sorting can be applied:

- Sorting and prioritizing features by *p*-value or fold-change revealed by the preceding statistical analyses.
- Selection of features related to quasi-molecular ions using a CAMERA-based annotation (Kuhl et al. 2011) of the feature list.
- Removal of features with insufficient intensity for tandem MS analysis.

After these steps, the set of interesting features is reduced to a number which can be annotated and identified in more detail. In particular for close- or co-eluting features, tandem mass spectra have to be acquired in multiple analytical runs. The creation of LC–MS/MS acquisition methods can be interpreted as an optimization problem, where as many of the features to be analyzed as possible have to be arranged in a given number of methods.

This problem is known in computer science as "Fixed Interval Scheduling" (Gertsbakh and Stern 1978; Kleinberg and Tardos 2005). One can either choose to create the smallest number of methods to cover all interesting features, or maximize the number (and their importance) of features covered using a given, limited number of methods. The maximum number of methods necessary is determined by the highest number of overlapping features. We have implemented the algorithm to either stop after a given percentage of features covered, or create a given number of LC–MS/MS methods.

After scheduling the target list, it has to be transferred into a set of acquisition methods for the instrument. To render the approach generally applicable, we support the export of scheduled target lists into the TraML format (Deutsch et al. 2011), a vendor independent XML data exchange standard for transition- and target lists which is currently under development by the proteomics standards initiative (PSI). In addition, the package allows to directly export method files for controlling a Bruker micrOTOF-Q instrument, which also records all method parameters in an XML file, but uses a different XML structure. Because the Bruker format is not documented in detail, it is difficult to create such a file from scratch. Instead, we chose an approach where a *template* method file contains a single prototype segment which describes the desired tandem MS settings. For each entry in the scheduled target list this segment is cloned and precursor m/z and retention time ranges are replaced by the values from the scheduled target list. An example is shown in Fig. 2. Other currently supported platforms are the Agilent QTOF instruments, which accepts specially formatted tabular CSV files, and the Waters QTOF instruments, which encode the acquisition segments similar to the Bruker format, but use a plain text *.EXP file instead of XML. The collision energies (either

```
<segment endtime="8.3">
 <param name="Mode_ScanMode" value="MRM" />
 <dependent polarity="positive">
  <param name="MSMSManual_IsolationMass">
   <entry value="695.3"/>
  </param>
  <param name="MSMSManual_IsolationWidth">
   <entry value="8"/>
  </param>
  <param name="MSMSManual_CollisionEnergy">
   <entry value="15"/>
  </param>
 </dependent>
</segment>
```

Fig. 2 Single segment of a Bruker micrOTOF-Q method for tandem MS acquisition (some names have been edited for brevity and readability). This segment is copied for each target feature, using the actual `values` from the precursor feature

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

one or more individual CID eV settings, or the range of the collision energy in RAMP mode) should be chosen based on experience with the instrumental setup. This approach can be generalized to other instruments which take the LC–MS/MS parameters from external files such as spreadsheets, although these formats do not usually allow to include many machine configuration parameters.

2.2 Acquisition of LC–MS and LC–MS/MS data

Chromatographic separations were performed on an Acquity UPLC system (Waters) equipped with a HSS T3 column (100 × 1.0 mm, particle size 1.8 µm, Waters) applying a binary gradient of water and acetonitrile, both acidified with 0.1% formic acid at a flow rate of 150 µL/min. Eluted compounds were detected from *m/z* 90–1,000 at a scan rate of 3 Hz in centroid mode using micrOTOF-Q hybrid quadrupole time-of-flight mass spectrometers (Bruker Daltonics) equipped with Apollo II electrospray ion sources in positive ion mode. For acquisition of tandem mass spectra precursor ions were selected using an isolation width of ±4 Da and fragmented using argon or nitrogen as collision gases and collision energies in the range of 10–30 eV. For detailed information on gradients, mass spectrometer settings and calibration see Supplemental Material S1.

2.3 Processing of LC–MS/MS raw data

LC–MS/MS raw data represent a continuous set of mass spectra and can in general be treated like LC–MS raw data. However, they contain orders of magnitudes less features, with much lower intensities compared to LC–MS raw data. Depending on the sample complexity and the precursor isolation width, individual LC–MS/MS raw spectra may comprise peaks originating from co-eluting and co-fragmenting precursor ion species. E.g. lipids with different degrees of saturation and a resulting *m/z* difference of only 2 or 4 Da between the precursor ions might elute closely together. The resulting mixed spectra will be difficult to interpret, unless very good precursor isolation is available. For this reason, extraction of pure tandem mass spectra requires a set of processing steps. For processing of LC–MS/MS raw data, we first perform a feature detection step using centWave (Tautenhahn et al. 2008). The centWave algorithm was developed for high resolution LC/MS data and is characterized by a higher sensitivity and a lower false-positive rate compared to several other feature detection algorithms. After the feature detection step, the relationship between features is lost. Therefore, features have to be subsequently assigned to compound spectra. This is done using the grouping algorithms implemented in the Bioconductor package CAMERA (Kuhl et al. 2011). The features detected within a retention time window constitute an initial compound spectrum, which can be subsequently refined using chromatographic peak shape similarity. We identify the compound spectrum best matching the retention time of the targeted precursor feature as the desired tandem mass spectrum.

2.4 Availability

We have created the MetShot package for the R statistics environment, which performs the steps ② and ③, interfaces to other R packages (XCMS and CAMERA) and external services (MassBank and MetFrag). The package is available under the Open Source GPL license together with extensive documentation and examples from our website (http://msbi.ipb-halle.de/msbi/MetShot/). The related XCMS and CAMERA packages are available from the Bioconductor web site (http://bioconductor.org/).

3 Results and discussion

3.1 Acquisition and processing of tandem MS data of standard compounds

The aim of the first experiment was to acquire and extract tandem mass spectra of protonated molecular ions of a set of standard compounds using the *nearline* data-dependent tandem MS approach and to evaluate their quality. For that purpose, we prepared a mixture of 27 natural product standards (Supplemental Material S2), each at a concentration of 20 µmol/L and analyzed it by UPLC/ESI-QTOFMS in positive ion mode. We manually collected retention times and *m/z* of corresponding protonated molecular ions for preparation of the target precursor list. The MetShot package was used to create a minimal number of LC–MS/MS methods which were subsequently used for acquisition of raw data at a collision energy of 15 eV and a precursor isolation width of ± 4 Da.

To estimate the difficulty of matching a precursor ion from the target list to the detected fragment ions, we determined the number of features being detectable at the retention time and precursor isolation windows for each acquisition segment. For an isolation width of ±0.5 Da, each acquisition segment contains 1.5 precursor features on average. This number raises to 6.2 for an ±4 Da isolation window used for acquisition of tandem mass spectra on the micrOTOF-Q in this study. This demonstrates the need for a careful assignment between precursor and tandem MS features.

We used the XCMS centWave algorithm to detect features in the LC–MS/MS raw data, and obtained between 49 and 380 features per analytical run, on average 18 per

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

acquisition segment. The feature detection also helps to improve the *m/z* accuracy, because the centroids of the raw spectra are averaged and weighted by their intensity. In the raw spectra, the maximum deviation of the centroids across a chromatographic peak is on average 16.8 ppm (std. deviation 9.7 ppm) from the averaged *m/z*, especially in the low-intensity flanks of the chromatographic peaks (Supplementary Material S3). As the last step, we used CAMERA and grouped the detected features into compound spectra to obtain the final tandem mass spectra.

As a reference, we used tandem mass spectra from our spectral library, which were previously extracted manually using the Bruker Data Analysis software. These curated spectra have been deposited in the MassBank database, their accession numbers are shown in the Supplemental Material S2.

To evaluate the quality of the spectra, we calculated the MassBank similarity score of an automatically extracted spectrum against the reference spectrum. In 25 cases, the MassBank similarity score was above 0.95, in only two cases it was below 0.95 (Supplemental Material S4).

For one of these two cases (emetine), our algorithm created spectra with several additional peaks compared to the reference spectrum. On closer examination, we found that those extra peaks can be easily explained as true fragments, which were erroneously missing in the library spectrum. In the other case (indole-3-acetonitrile), the extracted spectrum was missing several peaks, all of which had a very low intensity (below 150 cps) in the raw data.

3.2 Product analysis of a biosynthetic pathway in *Arabidopsis thaliana* by non-targeted profiling of wild-type and biosynthetic gene knockout plants

To demonstrate the usability of our approach in a non-targeted metabolomics workflow, we set up an profiling experiment using *Arabidopsis thaliana* wild-type and *cyp79B2 cyp79B3* double knockout plants (Zhao et al. 2002). The latter are completely impaired in the conversion of tryptophan to indole-3-acetaldoxime (IAOx) and do not accumulate IAOx-derived metabolites. Consequently, comparative analysis of wild-type and *cyp79B2 cyp79B3* plants has the potential to reveal the complete range of IAOx-derived indolic secondary metabolites.

Wild-type and *cyp79B2 cyp79B3* plants were grown in parallel in two independent experiments and sprayed with silver nitrate to induce accumulation of indolic secondary metabolites (Böttcher et al. 2009). For each independent experiment, leaves of six plants per genotype were harvested, pooled, extracted with aqueous methanol in quadruplicate and analyzed by UPLC/ESI-QTOFMS in positive ion mode. The resulting 16 raw data files were processed with XCMS using the centWave feature detection algorithm. Subsequent alignment gave 15,272 features with signal-to-noise ratios >3 which could be reliably detected in 75% of the technical replicates of a sample class. To identify features associated with IAOx-derived metabolites we filtered features whose median intensity within an independent experiment was 4-fold increased in wild-type samples in comparison to *cyp79B2 cyp79B3* samples at a significance level of $p < 0.01$ (Student's t test, two-sided, uncorrected). A total of 327 features met these criteria in both independent experiments (Supplemental Material S5). Using the CAMERA package, retention time grouping and chromatographic peak shape analysis allowed reconstruction of a set of 729 non-trivial compound spectra from the raw data whose automatic annotation revealed 1,085 and 1,140 putative protonated and sodiated molecular ions, respectively. The intersection of these annotated features and the differential features gave 72 features, which were passed without any intensity threshold into *nearline* data-dependent tandem MS analysis. As a result of the experimental design, these features are expected to be related to the IAOx-derived indolic secondary metabolites. The scheduling was able to distribute these features into 11 LC–MS/MS acquisition methods (see Fig. 3), where the first three already covered 39 (54%) of the features to be analyzed. Raw data were acquired at three different collision energies (10, 20 and 30 eV) using an isolation width of ±4 Da and processed as

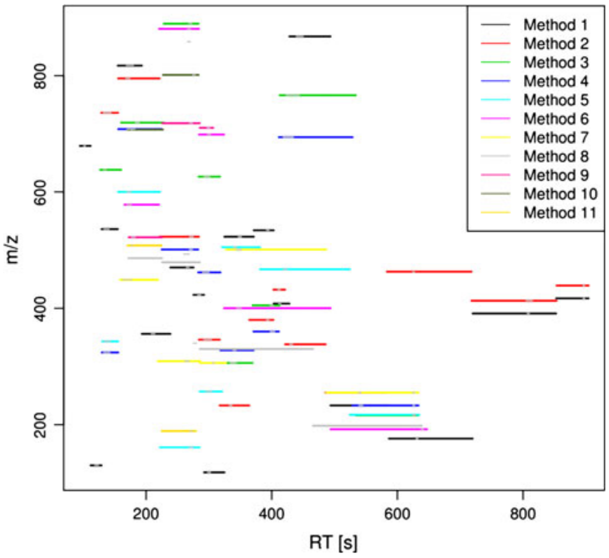


Fig. 3 Overview of the acquisition segments of the 11 LC–MS/MS methods generated by MetShot based on the target list of 72 precursor ions. The target list comprises putative protonated and sodiated molecular ions which were identified from a set of 327 differential features revealed by comparative analysis of *Arabidopsis* Col-0 and *cyp79B2 cyp79B3* leaf extracts. Target windows of the same color (*black*, *red*, *green*, *blue*, …) belong to an individual acquisition method

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

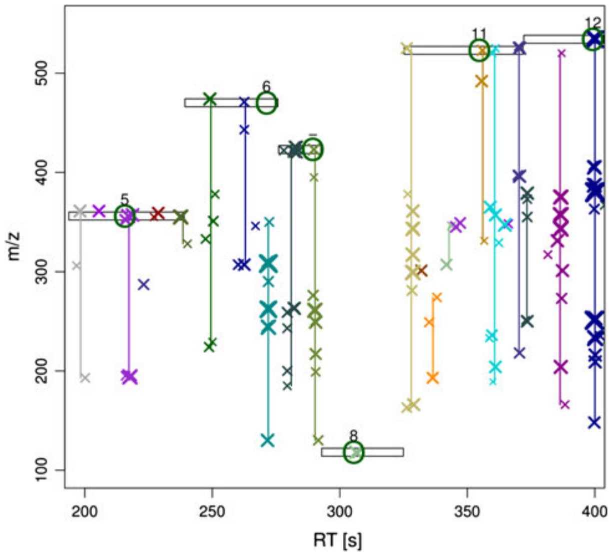


Fig. 4 Zoom into processed LC–MS/MS raw data. The target windows for the tandem MS precursor selection are shown as *black rectangles*. The precursor features (see Supplemental Metarial S6 for annotation and numbering) are *encircled in green*. The extracted LC–MS/MS features are shown as (×), scaled logarithmically by their intensity. The colors indicate the grouping into compound spectra, and non-trivial compound spectra are connected by vertical lines

described above. A zoom into the visualization of the features of interest, their isolation windows and the extracted tandem mass spectra is shown in Fig. 4.

To evaluate the obtained results (Supplemental Material S6) we first checked the quality of automatic annotation of protonated and sodiated molecular ions within the identified differential features. Based on the published indolic secondary metabolites known to be derived from IAOx (Böttcher et al. 2009) and manual analyses, 22 out of 72 putative annotations could be confirmed, representing 15 unique compounds, 9 of which were already described. However, features of numerous known IAOx-derived metabolites were not properly annotated and therefore not included in the subsequent tandem MS analysis. Reasons for this are to be found in the lack of characteristic pairs of quasi-molecular ions (e.g. $[M + H]^+$ and $[M + Na]^+$) or neutral losses in the ESI mass spectra of these metabolites, whose presence is imperative for successful spectra annotation on basis of the CAMERA package. The quality of tandem mass spectra was mainly dictated by the intensity of the corresponding precursor ions. Only 18 of the 72 analyzed features had sufficient intensity to result in tandem mass spectra with fair signal-to-noise ratios. Consequently, high quality tandem mass spectra were obtained for 7 out of 15 properly annotated compounds (Supplemental Material S6). For the sake of a simpler evaluation we had limited *nearline* data-dependent tandem MS analysis to differential features which were putatively annotated as $[M + H]^+$ and $[M + Na]^+$ ions. An alternative strategy would be to acquire tandem mass spectra of all differential features of sufficient intensity, only excluding isotopologues (which can be annotated more reliably than the ion species). In that case, all relevant tandem mass spectra would be available, at the expense of more irrelevant ones. Nevertheless, the acquired tandem mass spectra would have a high quality and relevance to the biological experiment.

The raw data are available as mzData files, together with extensive metadata annotation in the ISAtab format (Rocca-Serra et al. 2010) from our website (http://msbi.ipb-halle.de/msbi/MetShot/) and from the upcoming Metabolights repository (accession number MTBLS2, http://www.ebi.ac.uk/metabolights/MTBLS2).

4 Conclusion

Current acquisition strategies for tandem MS data are either designed manually, or performed using the data-dependent acquisition mode of the mass spectrometer. The first is very time consuming, the latter does not take into account which compounds are of (biological) interest, and the data processing is limited to individual spectra of possibly co-eluting and co-fragmenting compounds.

We have presented our *nearline* data-dependent tandem MS approach, which creates a minimal set of LC–MS/MS methods to obtain tandem mass spectra for a ranked list of interesting precursor features. We have also shown how to create these ranked lists, including a real metabolomics experiment on an *Arabidopsis thaliana* biosynthetic pathway mutant. The automated selection of quasi-molecular ions for fragmentation remains a challenge, because not all metabolites produce ion series to allow a reliable annotation, and the quasi-molecular ions might not be present in first place.

Compared to a manual analysis, the time between LC–MS profiling and the LC–MS/MS identification measurements can be drastically reduced. We envision that in a typical metabolomics study, the *nearline* data-dependent tandem MS steps are performed right after the metabolite profiling measurements, within hours after the profiling data was acquired, converted and processed with XCMS. The tandem MS acquisition thus benefits from a much better retention time stability, which in turn allows to reduce the acquisition windows to a minimum. The resulting smaller retention time uncertainty allows to cover more peaks of interest with a single tandem MS method, because the LC–MS/MS segments can focus on the chromatographic peaks without the need for a "safety-margin". The LC–MS/MS data processing performs a feature

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

detection and separates features from co-eluting and co-fragmenting compounds.

The *nearline* data-dependent tandem MS approach is suitable for any kind of LC-coupled tandem MS platform and experiments where a subsequent acquisition of tandem MS data is a requirement for further identification and biological interpretation. The R package MetShot includes the functions to schedule and write the instrument method files in several formats, and to process the resulting LC–MS/MS data.

Acknowledgements Ralf Tautenhahn (The Scripps Research Institute, La Jolla, CA) tested the *nearline* data-dependent tandem MS approach and contributed the export for the Agilent QTOF instruments.

References

Böcker, S., Letzel, M., Lipták, Z., & Pervukhin, A. (2008). SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics, 25*(2), 218–224.

Böttcher, C., von Roepenack-Lahaye, E., Schmidt, J., Schmotz, C., Neumann, S., Scheel, D., Clemens, S. (2008). Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and allows identification of a large number of new compounds in *Arabidopsis*. *Plant Physiology, 147*(4), 2107–2120.

Böttcher, C., Westphal, L., Schmotz, C., Prade, E., Scheel, D., & Glawischnig, E. (2009). The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of *Arabidopsis thaliana*. *The Plant Cell, 21*(6), 1830–1845.

Brown, M., Dunn, W. B. Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., Swainston, N., Spasic, I., Goodacre, R., & Kell, D. B. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst, 134*(7), 1322–1332.

Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., Mamas, M. A., Neyses, L., & Dunn, W. B. (2011). Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics, 27*(8), 1108–1112.

Deutsch, E. W., Chambers, M., Neumann, S., Levander, F., Binz, P.-A., Shofstahl, J., Campbell, D. S., Mendoza, L., Ovelleiro, D., Helsens, K., Martens, L., Aebersold, R., Moritz, R. L., & Brusniak, M.-Y. (Dec 2011). TraML: A standard format for exchange of selected reaction monitoring transition lists. *Molecular & Cellular Proteomics* (in press).

Draper, J., Enot, D. P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., & Zubair, H. (2009). Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive *m*/*z* annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics, 10*, 227.

Gertsbakh, I., & Stern, H. I. (1978). Minimal resources for fixed and variable job schedules. *Operations Research, 26*(1), 68–85.

Hoopmann, M. R., Merrihew, G. E., von Haller, P. D., MacCoss, M. J. (2009). Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. *Journal of Proteome Research, 8*(4), 1870–1875.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., & Nishioka, T. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45*(7), 703–714.

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., & Schmitt-Kopplin, P. (2009). Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One, 4*(7), e6386.

Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics, 7*(1), 234.

Kleinberg, J., & Tardos, E. (2005). *Algorithm Design*. Boston, MA: Addison-Wesley Longman Publishing Co Inc.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., & Neumann, S. (2011). CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Anal Chem, 84*(1), 283–289.

Okazaki, Y., Shimojima, M., Sawada, Y., Toyooka, K., Narisawa, T., Mochida, K., Tanaka, H., Matsuda, F., Hirai, A., Hirai, M. Y., Ohta, H.,& Saito, K. (2009). A chloroplastic UDP-glucose pyrophosphorylase from *Arabidopsis* is the committed enzyme for the first step of sulfolipid biosynthesis. *Plant Cell, 21*(3), 892–909.

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics, 11*(1), 395. ISSN 1471-2105.

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., & Sansone, S.-A. (2010). ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics, 26*(18), 2354–2356.

Smith, C., Want, E., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Analitical Chemistry, 78*(3), 779–787.

Smith, C. A., Maille, G. O., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., & Siuzdak, G. (2005). METLIN: A metabolite mass spectral database. In: *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology*, vol 27, pp. 747–751. Louisville, Kentucky.

Tautenhahn, R., Böttcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics, 9*(1), 504. ISSN 1471-2105.

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

Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., Lewis, G. D., Fox, C. S., Jacques, P. F., Fernandez, C., O'Donnell, C. J., Carr, S. A., Mootha, V. K., Florez, J. C., Souza, A., Melander, O., Clish, C. B., & Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine, 17*(4), 448–453.

Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

identification of metabolite mass spectra. *BMC Bioinformatics, 11*(1), 148. ISSN 1471-2105.

Zhao, Y., Hull, A. K., Gupta, N. R., Goss, K. A., Alonso, J., Ecker, J. R., Normanly, J., Chory, J., & Celenza, J. L. (2002). Trp-dependent auxin biosynthesis in *Arabidopsis*: Involvement of cytochrome P450s CYP79B2 and CYP79B3. *Genes Dev, 16*(23), 3100–3112.

DOI:10.1007/s11306-012-0401-0 Reprinted with permission. © 2012 Springer Science+Business Media

Research article

Journal of MASS SPECTROMETRY

Received: 20 April 2012 Revised: 1 October 2012 Accepted: 11 October 2012

Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI 10.1002/jms.3123

MetFusion: integration of compound identification strategies

Michael Gerlich* and Steffen Neumann

Mass spectrometry (MS) is an important analytical technique for the detection and identification of small compounds. The main bottleneck in the interpretation of metabolite profiling or screening experiments is the identification of unknown compounds from tandem mass spectra.

Spectral libraries for tandem MS, such as MassBank or NIST, contain reference spectra for many compounds, but their limited chemical coverage reduces the chance for a correct and reliable identification of unknown spectra outside the database domain.

On the other hand, compound databases like PubChem or ChemSpider have a much larger coverage of the chemical space, but they cannot be queried with spectral information directly. Recently, computational mass spectrometry methods and *in silico* fragmentation prediction allow users to search such databases of chemical structures.

We present a new strategy called MetFusion to combine identification results from several resources, in particular, from the *in silico* fragmenter MetFrag with the spectral library MassBank to improve compound identification. We evaluate the performance on a set of 1062 spectra and achieve an improved ranking of the correct compound from rank 28 using MetFrag alone, to rank 7 with MetFusion, even if the correct compound and similar compounds are absent from the spectral library. On the basis of the evaluation, we extrapolate the performance of MetFusion to the KEGG compound database. Copyright © 2013 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this article.

Keywords: metabolomics; integrated identification; MassBank; MetFrag; *in silico* fragmentation

Introduction

Soft ionization mass spectrometry, often coupled to liquid chromatography (LC-ESI-MS), has been established as an important analytical technology in several applications, such as metabolomics or screening of unknowns in the environmental sciences.[1,2] In untargeted approaches, complex samples are analyzed by LC-ESI-MS and can lead to elucidation of metabolites in biosynthetic pathways,[3,4] discovery of biomarkers,[5,6] prediction of disease states or detection of emerging pollutants in water samples.[7,8] However, these compounds are only characterized by their mass-to-charge ratio (*m/z*) and retention time, and subsequent identification requires substantial effort.

Tandem mass spectra provide valuable structural hints for the identification and structure elucidation of compounds and can be obtained from ion-trap or hybrid instruments, such as triple-quadrupole (QqQ) or quadrupole-time-of-flight (QqTOF). Collision-induced dissociation (CID) is a common fragmentation method for small compounds, resulting in a detailed fragmentation spectrum.[9]

These characteristic fragmentation patterns are available from spectral libraries such as MassBank,[10] HMDB,[11] where version v2.5 contains 2654 compounds with three MS/MS reference spectra on average, and NIST'11[12] that provides an MS/MS library with a total of 95 409 spectra representing 5843 compounds, including dipeptides and tripeptides, and METLIN,[13] which contains 48 596 high-resolution spectra for 10 076 metabolites as of February 2012.

MassBank is the first open community repository for mass spectral data (including spectral information, as well as analytical conditions) and provides both a web interface for human interaction and an application programming interface for programmatic access to the data and search functions. MassBank contains spectra from different instruments, including QqTOF, QqQ and ion-trap from different vendors. Most compounds are measured under various analytical conditions, for example in both positive and negative mode, or at several collision energies. The federated architecture of MassBank provides access to distributed data contributed by various institutes. There are approximately about 13 623 spectra high-resolution ESI-spectra representing about 2000 compounds in MassBank as of February 2012 (including redundancies, where the same compound was measured with different analytical settings on various instruments). A sample query result is shown in Figure 1.

Compound databases like PubChem,[14] KEGG[15] or ChemSpider[16] provide information on a huge number of both natural products and synthetic compounds. Although these databases excel in terms of chemical information (measured and predicted chemical properties, structure information and for some also assay results), they do not support queries using mass spectral measurements. The acquisition of reference spectra of all known compounds is unfeasible because of the enormous efforts required and the limited availability of commercial standards.

To alleviate the problem of limited availability of reference spectra, computational mass spectrometry tools with *in silico* spectra prediction have been developed.[17] *MetFrag* is a free and open-source program for compound identification based

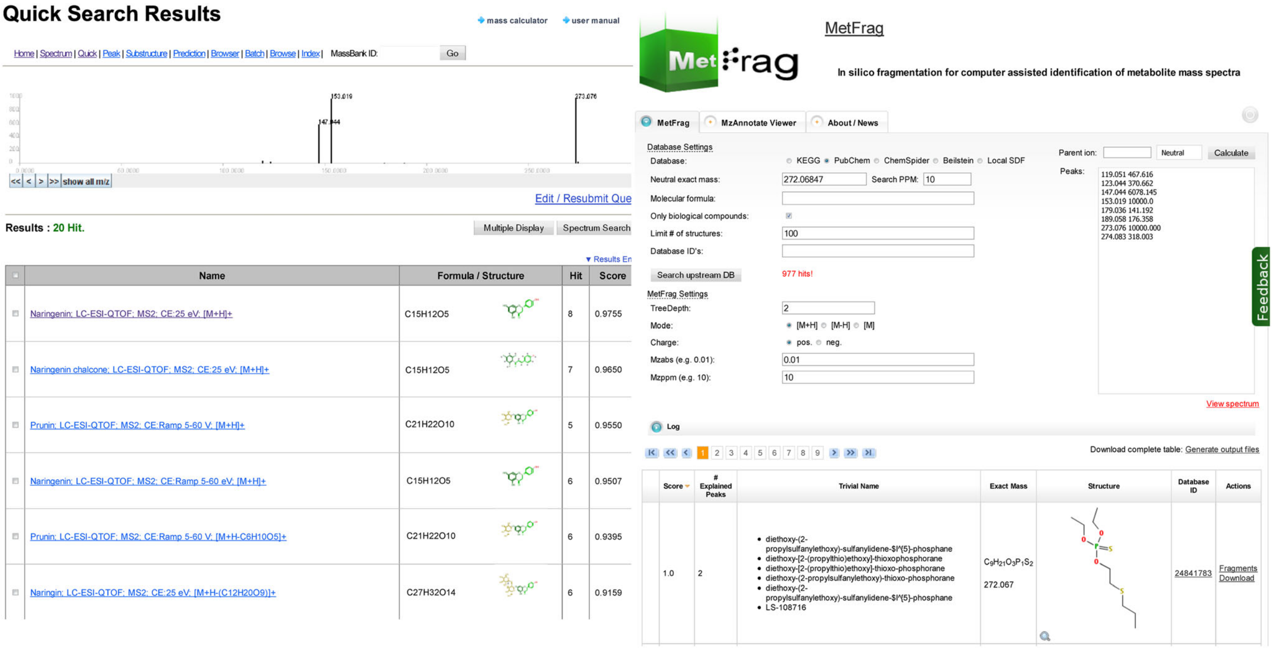
---

* *Correspondence to: Michael Gerlich, Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Germany. E-mail: mgerlich@ipb-halle.de*

Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Germany

Copyright © 2013 John Wiley & Sons, Ltd.

DOI:10.1002/jms.3123 Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.



Quick Search Results

Home | Spectrum | Quick | Peak | Substructure | Prediction | Browser | Batch | Browse | Index | MassBank ID: Go

Edit / Resubmit Que

Results : 20 Hit.

☐	Name	Formula / Structure	Hit	Score
☐	Naringenin; LC-ESI-QTOF; MS2; CE:25 eV; [M+H]+	C15H12O5	8	0.9755
☐	Naringenin chalcone; LC-ESI-QTOF; MS2; CE:25 eV; [M+H]+	C15H12O5	7	0.9650
☐	Prunin; LC-ESI-QTOF; MS2; CE:Ramp 5-60 V; [M+H]+	C21H22O10	5	0.9550
☐	Naringenin; LC-ESI-QTOF; MS2; CE:Ramp 5-60 eV; [M+H]+	C15H12O5	6	0.9507
☐	Prunin; LC-ESI-QTOF; MS2; CE:Ramp 5-60 V; [M+H-C6H10O5]+	C21H22O10	6	0.9395
☐	Naringin; LC-ESI-QTOF; MS2; CE:25 eV; [M+H-(C12H20O9)]+	C27H32O14	6	0.9159

MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

Database Settings

Database: ○ KEGG ● PubChem ○ ChemSpider ○ Beilstein ○ Local SDF

Neutral exact mass: 272.06847 Search PPM: 10

Molecular formula:

Only biological compounds: ☑

Limit # of structures: 100

Database ID's:

977 hits!

MetFrag Settings

TreeDepth: 2

Mode: ● [M+H] ○ [M-H] ○ [M]

Charge: ● pos. ○ neg.

Mzabs (e.g. 0.01): 0.01

Mzppm (e.g. 10): 10

Parent ion: Neutral

Peaks:

119.051 467.616
123.044 370.662
147.044 6078.145
153.019 10000.0
179.036 141.192
189.058 176.358
273.076 10000.000
274.083 318.003

Download complete table: Generate output files

Score	# Explained Peaks	Trivial Name	Exact Mass	Structure	Database ID	Actions
1.0	2	diethoxy-(2-propylsulfanylethoxy)-sulfanylidene-$l^{5}-phosphane diethoxy-[2-(propylthio)ethoxy]-thioxophosphorane diethoxy-[2-(propylthio)ethoxy]-thioxo-phosphorane diethoxy-(2-propylsulfanylethoxy)-thioxo-phosphorane diethoxy-(2-propylsulfanylethoxy)-sulfanylidene-$l^{5}-phosphane LS-108716	$C_9H_{21}O_3P_1S_2$ 272.067		24841783	Fragments Download

Figure 1. Individual results from MassBank and MetFrag. Left: The upper part of the MassBank screenshot shows the query spectrum, below are the resulting matches, sorted by spectrum similarity. Right: The top part of MetFrag contains the query input, the result list is presented below.

on compound databases.[18] The MetFrag web interface is shown in Figure 1. MetFrag obtains candidate structures from a compound database and matches *in silico* predicted fragments to the query spectrum. Each candidate is ranked according to a fragmentation score.

But despite the advent of *in silico* tools, reference spectra obtained under comparable analytical conditions are still the preferred way to achieve a reliable compound identification. To the best of our knowledge, there is currently no approach to integrate both strategies, where the most reliable answers of the two are returned.

In this paper, we present *MetFusion*, a strategy and system to *combine* the compound hypotheses obtained by complementary identification approaches. Here, we integrate the results from MassBank and MetFrag. This strategy combines the best of both worlds: the identification using spectral libraries if similar spectra are available and the huge chemical coverage of the compound databases queried by MetFrag.

Methods

In the following, we describe how the individual compound identification sources are queried, show the mathematical background for the integrated score and depict the web application. Subsequently, we explain the evaluation dataset and our evaluation approach to make the results more realistic and finally extrapolate the generalization to any compound in KEGG.

System architecture

The underlying assumption in MetFusion is that the correct compound is present in the compound database and consequently among the structure candidates in the MetFrag result. The idea of MetFusion is to confirm the *in silico* predicted results with spectral reference data and calculate a new integrated score for each candidate processed by MetFrag. This is depicted in the workflow shown in Figure 2.

The MassBank scores are calculated on the basis of a modified cosine distance to compute the similarity between the query spectrum and the reference spectra.[10] Results are ranked according to this spectral similarity. MassBank is accessed using a Java library application programming interface, which queries the individual servers, and passes the relevant parameters (intensity cutoff, ionization mode and instrument filter) directly to the servers.

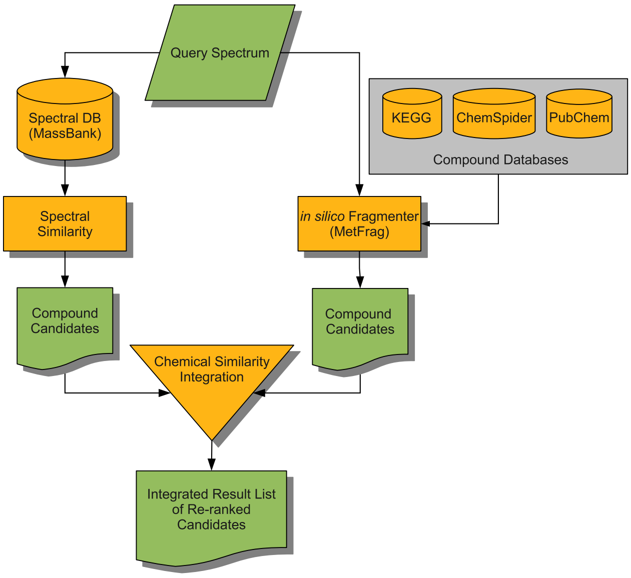


Figure 2. The MetFusion workflow: the query spectrum is passed to both the MassBank and the MetFrag query tools. Both return ranked lists, providing spectra matches and compound candidates, respectively. These lists are combined by calculating the chemical similarity between all structures. The integrated score is used to re-rank the list of MetFrag candidates from the compound database.

DOI:10.1002/jms.3123

Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

The *in silico* fragmentation is performed with an embedded MetFrag module, which queries KEGG, PubChem or ChemSpider as a compound database. In addition, local compound libraries can be used, which allows the use of in-house compound databases or mirrors of, e.g. PubChem. This is performed either by SD file upload or direct database access. Likewise, this upload allows users to submit their own generated structures as candidates for *in silico* fragmentation.

MetFusion requires a spectrum with *m/z* and intensity values as query input and passes the provided settings to the corresponding databases. MetFrag settings can also be adjusted, most importantly the allowed *mass deviation* for the generated fragments.

The core of MetFusion is implemented as a Java library, which is used both by the command line and web interface. Both the MassBank and MetFrag queries are performed in separate threads and run in parallel.

Integration of spectral matches, in silico scores and chemical similarity

The identification strategies return two individual lists of spectra matches and candidate compounds, both with associated scores. The spectra scores are combined into a *spectral summary*. This is an aggregation of similar spectra and their respective chemical similarity to a candidate compound.

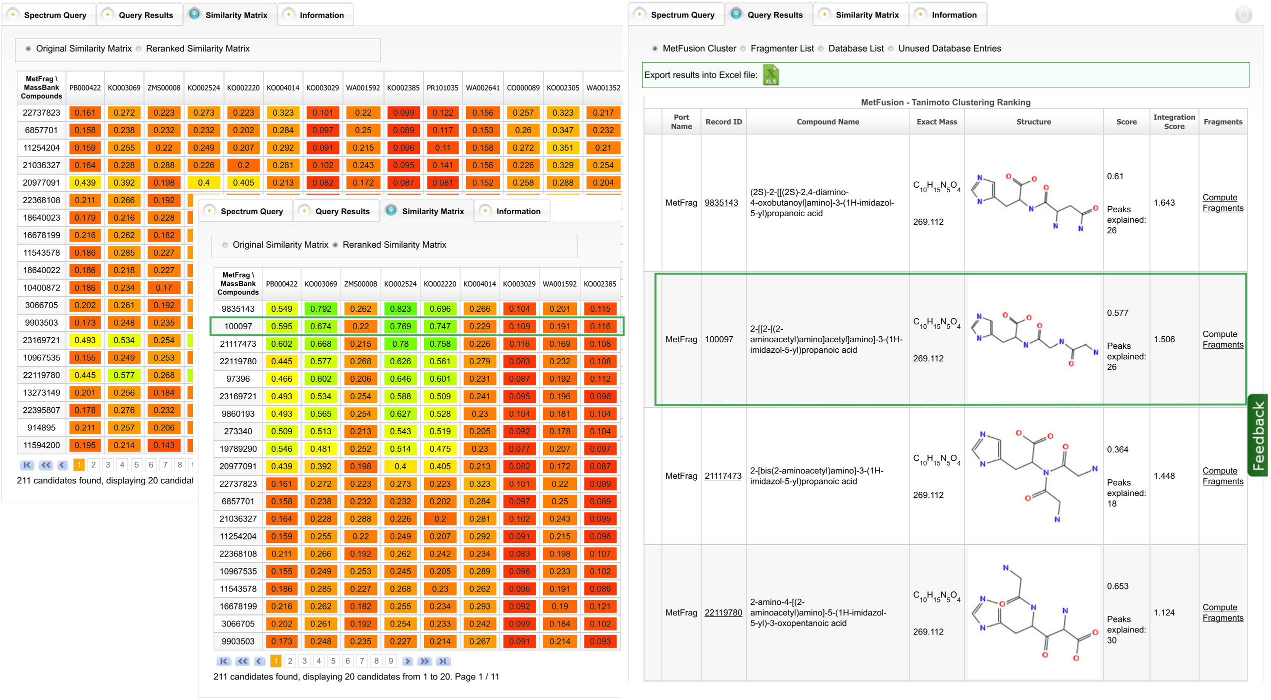
The spectral library can contain multiple measurements of a single compound or its isomeric variants, so we use an InChIKey-based filtering of the original MassBank result list, which only retains the highest-scoring record for each compound constitution. This is also justified because distinguishing between stereoisomers is hardly possible with mass spectrometry alone. For this filter step, we rely on the connectivity information stored in the first block of the InChIKey.[19]

Equation (1) describes the integrated MetFusion score s_c: for each MetFrag candidate c, we calculate s_c as a sum of the MetFrag score f_c and the ‘spectral summary’ on the basis of the scores m_j for all MassBank results j, and the chemical similarity t_{cj} between MetFrag candidate c and each MassBank result j. The number of results from MetFrag is denoted by N, and the number of MassBank results is denoted by M. This leads to an $N \times M$ matrix of chemical similarities. An excerpt of such a matrix can be seen in Figure 3.

The chemical similarity t_{cj} between a MetFrag candidate c and each MassBank result j allows us to determine how similar each pair of compounds is. This provides a validation of *in silico* generated spectra with measured spectra, based not on spectral similarity but rather on the chemical similarity between the corresponding compounds represented by their spectra. This approach results in the integrated score, allowing us to rank the MetFrag candidates with an additional level of information.

We use the sigmoid function *sig(x)* shown in Equation (2) to introduce a non-linear behavior, which reduces the influence of mediocre spectral matches and chemical similarities. Further information about the sigmoid function is available in Supplemental Material S-1, describing the impact of the parameters β and γ.

$$s_c = \underbrace{\alpha * f_c}_{MetFrag} + \overbrace{(1-\alpha) * \sum_{j=1}^{M} sig\left(m_j * t_{cj}\right)}^{\text{“spectral summary”}} \qquad (1)$$



MetFrag \ MassBank Compounds	PB000422	KO003069	ZMS00008	KO002524	KO002220	KO004014	KO003029	WA001592	KO002385	PR101035	WA002641	CO000089	KO002305	WA001352
22737823	0.161	0.272	0.223	0.273	0.223	0.323	0.101	0.22	0.099	0.122	0.156	0.257	0.323	0.217
6857701	0.158	0.238	0.232	0.232	0.202	0.284	0.097	0.25	0.089	0.117	0.153	0.26	0.347	0.232
11254204	0.159	0.255	0.22	0.249	0.207	0.292	0.091	0.215	0.096	0.11	0.158	0.272	0.351	0.21
21036327	0.164	0.228	0.288	0.226	0.2	0.281	0.102	0.243	0.095	0.141	0.156	0.226	0.329	0.254
20977091	0.439	0.392	0.198	0.4	0.405	0.213	0.082	0.172	0.087	0.081	0.152	0.258	0.288	0.204
22368108	0.211	0.266	0.192											
18640023	0.179	0.216	0.228											
16678199	0.216	0.262	0.182											
11543578	0.186	0.285	0.227											
18640022	0.186	0.218	0.227											
10400872	0.186	0.234	0.17											
3066705	0.202	0.261	0.192											
9903503	0.173	0.248	0.235											
23169721	0.493	0.534	0.254											
10967535	0.155	0.249	0.253											
22119780	0.445	0.577	0.268											
13273149	0.201	0.256	0.184											
22395807	0.178	0.276	0.232											
914895	0.211	0.257	0.206											
11594200	0.195	0.214	0.143											

211 candidates found, displaying 20 candidat

MetFrag \ MassBank Compounds	PB000422	KO003069	ZMS00008	KO002524	KO002220	KO004014	KO003029	WA001592	KO002385
9835143	0.549	0.792	0.262	0.823	0.696	0.266	0.104	0.201	0.115
100097	0.595	0.674	0.22	0.769	0.747	0.229	0.109	0.191	0.116
21117473	0.602	0.668	0.215	0.78	0.758	0.226	0.116	0.169	0.108
22119780	0.445	0.577	0.268	0.626	0.561	0.279	0.083	0.232	0.108
97396	0.466	0.602	0.206	0.646	0.601	0.231	0.087	0.192	0.112
23169721	0.493	0.534	0.254	0.588	0.509	0.241	0.095	0.196	0.096
9860193	0.493	0.565	0.254	0.627	0.528	0.23	0.104	0.181	0.104
273340	0.509	0.513	0.213	0.543	0.519	0.205	0.092	0.178	0.104
19789290	0.546	0.481	0.252	0.514	0.475	0.23	0.077	0.207	0.097
20977091	0.439	0.392	0.198	0.4	0.405	0.213	0.082	0.172	0.087
22737823	0.161	0.272	0.223	0.273	0.223	0.323	0.101	0.22	0.099
6857701	0.158	0.238	0.232	0.232	0.202	0.284	0.097	0.25	0.089
21036327	0.164	0.228	0.288	0.226	0.2	0.281	0.102	0.243	0.095
11254204	0.159	0.255	0.22	0.249	0.207	0.292	0.091	0.215	0.096
22368108	0.211	0.266	0.192	0.262	0.242	0.234	0.083	0.198	0.107
10967535	0.155	0.249	0.253	0.245	0.205	0.289	0.096	0.233	0.102
11543578	0.186	0.285	0.227	0.268	0.23	0.262	0.096	0.191	0.086
16678199	0.216	0.262	0.182	0.255	0.234	0.293	0.092	0.19	0.121
3066705	0.202	0.261	0.192	0.254	0.233	0.242	0.099	0.184	0.102
9903503	0.173	0.248	0.235	0.227	0.214	0.267	0.091	0.214	0.093

211 candidates found, displaying 20 candidates from 1 to 20. Page 1 / 11

Spectrum Query
Query Results
Similarity Matrix
Information
○ MetFusion Cluster ○ Fragmenter List ○ Database List ○ Unused Database Entries
Export results into Excel file:
XLS

MetFusion - Tanimoto Clustering Ranking

Port Name	Record ID	Compound Name	Exact Mass	Structure	Score	Integration Score	Fragments
MetFrag	9835143	(2S)-2-[[(2S)-2,4-diamino-4-oxobutanoyl]amino]-3-(1H-imidazol-5-yl)propanoic acid	$C_{10}H_{15}N_5O_4$ 269.112		0.61 Peaks explained: 26	1.643	Compute Fragments
MetFrag	100097	2-[[2-[(2-aminoacetyl)amino]acetyl]amino]-3-(1H-imidazol-5-yl)propanoic acid	$C_{10}H_{15}N_5O_4$ 269.112		0.577 Peaks explained: 26	1.506	Compute Fragments
MetFrag	21117473	2-[bis(2-aminoacetyl)amino]-3-(1H-imidazol-5-yl)propanoic acid	$C_{10}H_{15}N_5O_4$ 269.112		0.364 Peaks explained: 18	1.448	Compute Fragments
MetFrag	22119780	2-amino-4-[(2-aminoacetyl)amino]-5-(1H-imidazol-5-yl)-3-oxopentanoic acid	$C_{10}H_{15}N_5O_4$ 269.112		0.653 Peaks explained: 30	1.124	Compute Fragments

Figure 3. The MetFusion web application. Left: The background shows an excerpt of the similarity matrix for a query with the Gly-Gly-His spectrum (NIST# 1012075, CID: 100097) prior to re-ranking. Columns contain MassBank results, and rows correspond to MetFrag results. Each cell shows the chemical similarity between 0 (red) to 1 (green). The correct structure appears at tied rank 23 (not visible). None of the top MetFrag candidates show high chemical similarity. Overlaid is the similarity matrix after re-ranking. Here, several MetFrag candidates have a reasonable similarity to MassBank results, and the correct candidate is circled in green. The combination via chemical similarity improves the rank of the correct structure to two. Right: The head of the re-ranked MetFusion output, showing the results with structure formula, database link and scores.

DOI:10.1002/jms.3123

Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

$$sig(x) = \frac{1}{1 + e^{(\beta*(x-\gamma))}} \tag{2}$$

The ‘spectral summary’ for the candidate c is then the sum of all MassBank scores m_j, weighted by their chemical similarity t_{cj} to the candidate c. The MassBank spectral scores m_j use a modified cosine distance in the range from 0 to 1, where values ≥ 0.65 indicate reasonable spectral similarity.[10]

For the chemical similarity calculation, we use the Chemistry Development Kit (CDK, version 1.4.7).[20] The chemical similarity t_{cj} between the molecular fingerprints (CDK standard fingerprint with 1 024 bit length) of the compounds c and j is calculated using the Tanimoto (also known as Jaccard) coefficient.[21,22]

The balance between the individual identification approaches is determined by the weight α, where $\alpha = 1$ uses exclusively the MetFrag scores and $\alpha = 0$ results in a compound library search for those compounds that have the most similar high-scoring MassBank hits. Although both individual MetFrag and MassBank scores fall in the range of 0–1, the MetFusion result score has no upper bound and depends on the original MetFrag score f_c, the number of spectral database hits and the corresponding chemical similarity. The lower limit of the MetFusion score is 0.

Evaluation method and dataset

MetFusion was evaluated on a dataset of 1099 spectra, containing compounds ranging in molecular weight from 89 to 837 Da. A wide range of compound classes is covered, including flavonoids, steroids, amino acids, carboxylic acids, glucosides, drugs and toxins. Nine hundred and eighteen spectra were measured with a single collision energy (such as 10, 20, 30, 40 and 50 eV). The remaining 181 spectra were created by merging spectra at several collision energies for a single compound. The corresponding spectra were measured on the same instrument type, and only the collision energies differed. In this way, more informative peaks are present in a merged spectrum. The use of merged spectra for similarity search is also recommended by the MassBank consortium.[10]

The reference spectra used to evaluate MetFrag[18,23] are a subset of this evaluation dataset. All spectra are available from MassBank; for details, see Supplemental Material S-7.

The dataset contains 37 spectra, which contain only the precursor ion information, resulting from soft ionization with 10 eV. The results presented in the main article exclude these spectra, but the complete results can be found in Supplemental Materials S-5 to S-8.

For each test spectrum, we determine the rank of the correct candidate obtained with MetFusion. For the evaluation, we consider all different configurations of a candidate compound as a single constitution because the compound databases often include several stereoisomers and unspecified stereo configurations. We again use an InChIKey-based filtering of the candidate list.

The *relative ranking position* (RRP) describes the position of the correct compound in relation to the whole result set.[24,25]

$$RRP = \frac{1}{2}\left(1 - \frac{BC - WC}{TC - 1}\right) \tag{3}$$

In Equation (3), BC denotes the number of candidates that have a higher MetFusion score than the correct compound. WC denotes the number of candidates that have a lower score. TC denotes the total number of candidates, i.e. the number of MetFrag results N.

We have defined Equation (3) such that an RRP of 1.0 is equivalent to the correct compound at the first position, this value also implies that no other compound is ranked first. If the compound is ranked last, this results in an RRP of 0.0. If all compounds share the same score, this results in an RRP of 0.5.

We also report the median rank of the correct solution, which indicates how many candidates have to be considered before the correct solution appears in the web application. If several compounds (including the correct solution) have an identical score, we use the most conservative approach and report the maximum (worst case) rank of equally scored candidates.

Simulation of real world queries for training and evaluation

We cannot use the ‘normal’ operation mode of MetFusion to evaluate the identification performance, as all test spectra are also present in MassBank. If we did so, MetFusion would be simply ‘too good’. This is because of the fact that querying MassBank with spectra from our dataset is guaranteed to find matches at the top positions, as these spectra are present in MassBank. The correct candidate would also have a Tanimoto similarity of 1.0, which would favor its scoring even more because the parameter optimization would result in a scoring function strongly biased towards MassBank.

To avoid this, we simulated the identification of *unknown* spectra: we removed not only the query spectrum from the MassBank results in our evaluation but also any spectra whose compounds were above a certain chemical similarity to the query compound. This filtering approach provides controlled conditions for the evaluation because for an independent set of evaluation spectra, we also would need to specify to what degree the compounds are present in the reference library. We used 0.7 as the most stringent Tanimoto similarity threshold, which removed on average 2.4 MassBank records for each test spectrum. Less stringent filters are 0.8 and 0.9, which removed on average 1.8 and 1.3 results, respectively. A threshold of 1.0 removed only the correct compound. So with just one set of test spectra, we can evaluate our approach against several levels of completeness of the spectral library used and find to what degree the identification depends on the reference spectra.

MetFusion web application and availability

The MetFusion application is available at http://msbi.ipb-halle.de/MetFusion/ and features a user-friendly interface.

The query spectrum should contain at least *m/z* and *intensity* values. The three column MassBank peaklist format is supported as well. Additional search parameters allow users to modify the search behavior for compound database and spectral database, respectively. The results are presented in a table with 20 entries per page, ordered by decreasing MetFusion score. It is also possible to download all results as a spreadsheet, which contains the result lists of MetFusion, MetFrag and MassBank, with corresponding scores and images of the molecular structures and the computed similarity matrix. Users can then add this report to the supplemental information of publications to support their findings.

It is possible that several candidate compounds obtain identical scores and thus have tied ranks. To improve the overview of the MetFusion result list in the web application, we perform a

Copyright © 2013 John Wiley & Sons, Ltd.

DOI:10.1002/jms.3123

Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

structural clustering of all compounds that have the same MetFusion score and a Tanimoto fingerprint similarity ≥ 0.95 and join them into a cluster. This applies in particular to stereo isomers, which can in general not be distinguished with mass spectrometry. In contrast to our evaluation, the web application does not perform any filtering of the candidates because for a downstream analysis (such as citation counts), the full candidate list could be relevant. The clustered results can be expanded and viewed in detail by the user.

The web interface is based on Java Server Faces 2, ICEfaces 2 (component library with AJAX capabilities) and an Apache Tomcat 7 server.

The application is suitable for browsers with JavaScript enabled.

The MetFusion implementation is available as a Java library from the project repository at https://github.com/mgerlich/MetFusion, which can be used to perform batch searches on a local computer or cluster. The code is available under the open-source GPL license.

Results and discussion

MassBank contains spectra from various MS instruments and chromatography types. For this paper, we focused on ESI tandem MS spectra. This includes 13 instrument types with a total of 13 623 spectra as of February 2012. The result size for a MassBank query was limited to the best 100 records.

MetFrag was used with two different values for the *mzabs* parameter. This parameter defines the allowed absolute mass deviation between the *in silico* generated fragments and the measured peaks. Spectra that were measured on high-resolution devices with good accuracy used *mzabs* = 0.0, so only the relative *mzppm* error threshold is used. For less accurate spectra, we increased this value to *mzabs* = 0.01, allowing a broader range for the exact mass of generated fragments to match. The additional parameter *mzppm* was set to 10 ppm in all cases. After filtering for unique InChIKeys, we found that the result list contained on average 1247 candidates per query spectrum from the PubChem database.

Optimization of scoring function parameters

The integrated scoring function has three internal parameters: α balances the *in silico* prediction and the spectral summary, and β and γ determine the shape of the sigmoid function in the spectral summary. For an optimal choice of these parameters, we performed a parameter scan using the complete set of 1 099 spectra for each of the filter thresholds. The resulting parameter sets are shown in Supplemental Material S-2. To assess the stability and generalization of such a parameter optimization, we performed a tenfold cross-validation. Across all ten partitions, we obtained very similar optimal parameter combinations when optimizing the mean rank of the correct compound, which suggests that the scoring function is robust to parameter and data variations. The detailed results of the cross-validation are shown in Supplemental Material S-2.

For the remainder of the paper and for the web application, we have chosen the parameter set obtained with the similarity filtering threshold of 0.9. The corresponding optimal parameters are thus $\alpha = 0.3$, $\beta = -9$ and $\gamma = 0.6$. Additional information on the performance of MetFusion is available in Supplemental Material S-3 to S-8.

Examples: Gly-Gly-His and naringenin

First, we have selected two example query spectra to demonstrate MetFusion and discuss the results.

We selected a spectrum for the tripeptide Glycine-Glycine-Histidine (Gly-Gly-His, Pubchem CID: 100097) with 42 peaks, measured on a Micromass Quattro Micro QqQ device with nominal mass resolution from the NIST 2008 database. MassBank has very little spectral information on dipeptides and almost none for tripeptides or polypeptides. However, the basic amino acids are present in MassBank. So the challenges for MetFusion are to deal with the low mass resolution and the lack of a reference spectrum for this compound.

Gly-Gly-His has an exact mass of 269.112 Da. We modified the MetFrag parameters and increased *mzabs* to 0.1 Da and *mzppm* to 30 ppm to account for the low resolution spectrum.

With MetFrag alone, the top ranked candidates explain up to 35 fragments, and many have purine or furan substructures. The correct structure explains 26 fragments and is returned at tied rank 23. The first MassBank hits contain spectra for Histidine (155.069 Da, best score 0.856), Carnosine (a dipeptide of β-Alanine and Histidine, 226.106 Da) and L-Homocarnosine (240.122 Da, score of 0.798). Figure 3 shows the similarity matrix dominated by chemical similarities < 0.3.

Although none of the MassBank hits fully resemble the tripeptide of interest, the basic building blocks and their corresponding characteristic fragment peaks provide enough information to obtain the higher rank of the correct compound. The rank of Gly-Gly-His is improved from rank 23 in MetFrag to rank 2 in MetFusion. After MetFusion was run, the visual inspection of the similarity matrix helps to interpret the result and avoid some pitfalls. In another example, for the naringenin chalcone (CID: 155802) spectrum, both MetFrag and MassBank results also contain the related naringenin (CID: 932). The spectrum PB000129 of naringenin chalcone has a MassBank score of 0.98 compared with the spectrum PB000125 of naringenin. The ring break of naringenin chalcone leads to very low chemical similarity scores, thus promoting the rank of naringenin with intact rings and its higher spectral and chemical similarity towards naringin (spectrum PB000804, CID: 442428), favoring naringenin over naringenin chalcone. Additional information is available in Supplemental Material S-3.

Evaluation with benchmark dataset

We performed the evaluation on both the reduced dataset of 1062 spectra, which excluded spectra that contain only precursor ion information, as well as the complete dataset with 1099 spectra. The latter are available in Supplemental Material S-5. We first queried MetFrag separately and use these results as baseline. As stated before, we applied the InChIKey-based filter step to the candidate lists prior to the MetFusion combination to remove duplicated constitutions. Here, the correct compound had a median tied rank position of 28 and a mean rank position of 164. The corresponding median RRP is 0.959, and the mean RRP is 0.886. The discrepancy between mean and median shows that the distribution is skewed and the low performance for several compounds increases the mean considerably.

For the evaluation of our MetFusion strategy, we used the simulated real world queries for the evaluation dataset. We chose the similarity filter of 0.9 as the basis for the optimization and evaluation. With this setting, we obtain the correct solution among the

Copyright © 2013 John Wiley & Sons, Ltd.

DOI:10.1002/jms.3123

Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

top 2% (median RRP 0.991) in the result list or at an absolute rank 7 (median). Without filtering the correct compound from the MassBank results, we obtain a median RRP of 1.

With the most restrictive filtering we used (similarity threshold 0.7), the median RRP drops to 0.986, with a median rank of 10. This filter setting removed on average 2.4 spectra from the MassBank results, and in one case up to 23. With these results, one can expect to find the correct solution on the first page of the result list of the web application. The other (less pessimistic) filter settings are shown in Table 1.

The main advantage of this approach is the combination of two separate identification approaches: (1) Instead of dealing with multiple interfaces, all results are available in a single application. More importantly, (2) MetFusion does neither depend solely on *in silico* prediction nor on the possibly poor coverage of reference spectra. A distinct advantage is that the spectrum search from MassBank will not only retrieve spectra from compounds with the actual precursor mass but also includes related compounds with different masses that share similar fragment peaks. These peaks can be attributed to similar structural features of a compound. MetFrag usually retrieves the candidates on the basis of the precursor mass or elemental composition, so all candidates of a query will have the same mass. Hence, if the correct compound is contained in the compound database, it is also included in the MetFusion result list.

Please also note that in this evaluation, we used PubChem to demonstrate the ability to process large compound databases, although for metabolomics applications, many of the candidates will be irrelevant. Generally, an experimentalist will have additional prior information, which can be used to ignore candidates that could not occur in the sample under investigation.

These results show that combining an *in silico* approach with curated reference measurements can directly improve compound identification and give the best of both worlds.

Extrapolation to KEGG

The results presented earlier show the performance of MetFusion on the benchmark dataset from MassBank. But what performance can we expect for arbitrary compounds in, e.g. KEGG? This depends on the number of 'similar' reference spectra available in MassBank for each KEGG compound so that the results in Table 1 can be extrapolated to the KEGG compound database.

We calculated the pairwise chemical similarity between compounds in MassBank and KEGG. Using the last publically available KEGG COMPOUND snapshot (15 499 entries as of 24 June 2011) and a local MassBank database of 5 063 compound structures for which ESI reference spectra are available, we found that for 2690 KEGG entries, there is a MassBank record with a Tanimoto similarity of 0.9 or better. Additional information is available in Supplemental Material S-4.

Table 1. Results of MetFusion for the 1062 spectra dataset with the (artificially) filtered MassBank

		Similarity filter				
		0.7	0.8	**0.9**	1	none
MetFusion	Rank	10	8	7	4	1
	RRP	0.986	0.990	0.991	0.993	1
KEGG only	Rank	8	6	6	4	1
	RRP	0.976	0.984	0.987	0.989	1

Both the median rank and the median relative ranking position (RRP) are shown for a given filter stringency. In addition, the results for 180 unique KEGG compounds are presented.

Under the assumption that our compound selection in the test data is unbiased and that Table 1 can be generalized to all KEGG compounds, we would expect that half of these 2690 compounds can be ranked among the first seven MetFusion results, even if they are searched against the whole of PubChem. If we relax the restrictions, we find that for 5513 entries in KEGG, there is a MassBank spectrum with a Tanimoto similarity of at least 0.7, so the extrapolation from Table 1 suggests a median rank of 10 for the correct compound.

We were able to validate this extrapolation on the subset of 180 unique compounds from our dataset that also provide a KEGG identifier. We used the identical settings as for the full benchmark dataset and also retrieved the candidates from PubChem. The results are shown in Table 1. Although the RRPs are slightly lower, the absolute ranks are even slightly better. One reason is that PubChem returned fewer candidates for the compounds also present in KEGG.

These calculations are just an extrapolation under several assumptions, which cannot be taken for granted. If a compound is not amendable to mass spectrometry, e.g. because of low ionization efficiency, the identification is impossible with MetFusion, and other analytical methods have to be used. Secondly, the extrapolation assumes the same performance on compounds not contained in the benchmark dataset. Although there are several diverse compound classes in the evaluation dataset we used, the benchmark data could be biased and MetFusion could have a different performance (lower, but also higher) for classes not taken into account here. On the other hand, because of the distributed nature of MassBank, we only considered those structures that were available in our local database mirror, so the number of KEGG structures for which similar reference spectra are available is definitely higher than 2690.

Conclusions

The MetFusion approach was developed to combine the knowledge from reference spectra with the *in silico* prediction tool MetFrag for structure identification in metabolomics studies and small molecules in general. We showed that merging this information via chemical similarity improves the position of the correct compound from rank 28 to rank 7 compared with *in silico* prediction alone. This improvement is even more remarkable, given that in the evaluation, we made sure that reference spectra of the correct (and similar) compounds were excluded. Solely relying on the spectral library as identification strategy would result in no – or a wrong – identification for these cases.

As this paper used existing benchmark data from spectral libraries, we have described a metabolomics workflow elsewhere.[26] The MetShot approach first obtains a list of peaks of interest from metabolite profiling and statistical analysis and acquires high-quality tandem mass spectra, which can be converted to MetFusion batch query files.

The metabolomics standards initiative has defined four levels of confidence in metabolite identification.[27] The most confident level 1 identification requires the comparison of an unknown compound with authentic standards under the same analytical conditions, and level 2 can be achieved by a comparison of spectral data with literature or database information of the same compound. MetFrag alone can be considered to achieve the

Copyright © 2013 John Wiley & Sons, Ltd.

DOI:10.1002/jms.3123

Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

annotation of compound classes, resulting in a level 3 identification. With MetFusion, we can often achieve a level between two and three, even if a reference spectrum of the actual unknown is absent from MassBank. Beyond MetFusion, additional steps such as the comparison of retention time and predicted logP ranges, UV spectra or filtering for known substructures can further improve the confidence of the identification. Some of these aspects have been evaluated elsewhere.[28]

The approach is generally applicable to any identification strategies that return compound structures and can be modified for other spectral libraries (such as Metlin, HMDB or GMD[29]) as well as other identification strategies, such as the recently published analysis of fragmentation trees.[30–32] Furthermore, it is not only restricted to tandem MS spectra but can readily be applied to MS1 spectra with informative in-source fragments, MS*n* data and GC-MS spectra.

Because the number of known metabolites will grow faster than the coverage of spectral libraries, our approach to integrate multiple identification strategies will remain of high importance in the future. Even if all KEGG compounds (as of today) were available in MassBank, the huge number of 200 000 metabolites estimated in the plant kingdom[33] will not be part of reference libraries any time soon.

Of course, a major improvement in general would be an increase in high-resolution spectra contributions to reference databases. The analysis of the chemical similarities between MassBank and KEGG allows to prioritize future efforts and select substances for which spectra should be added to MassBank to improve the coverage of biologically relevant reference spectra.

Acknowledgements

The authors thank all contributors for providing mass spectra to MassBank. We also like to thank the CDK developers, Sebastian Wolf who developed MetFrag and Emma Schymanski for helpful discussions.

Supporting information

Supporting information may be found in the online version of this article.

References

[1] W. Dunn, A. Erban, R. Weber, D. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, M. Viant. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2012**, 1–23. 10.1007/s11306-012-0434-4.

[2] M. Krauss, H. Singer, J. Hollender. LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.* **2010**, *397*, 943–951. 10.1007/s00216-010-3608-9.

[3] C. Böttcher, E. von Roepenack-Lahaye, J. Schmidt, C. Schmotz, S. Neumann, D. Scheel, S. Clemens. Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and allows identification of a large number of new compounds in arabidopsis. *Plant Physiol.* **2008**, *147*(4), 2107–2120.

[4] Y. Okazaki, M. Shimojima, Y. Sawada, K. Toyooka, T. Narisawa, K. Mochida, H. Tanaka, F. Matsuda, A. Hirai, M. Y. Hirai, H. Ohta, K. Saito. A chloroplastic UDP-glucose pyrophosphorylase from Arabidopsis is the committed enzyme for the first step of sulfolipid biosynthesis. *Plant Cell* **2009**, *21*(3), 892–909.

[5] G. Glauser, D. Guillarme, E. Grata, J. Boccard, A. Thiocone, P.-A. Carrupt, J.-L. Veuthey, S. Rudaz, J.-L. Wolfender. Optimized liquid chromatography-mass spectrometry approach for the isolation of minor stress biomarkers in plant extracts and their identification by capillary nuclear magnetic resonance. *J. Chromatogr. A* **2008**, *1180* (1-2), 90–98.

[6] R. Mohamed, E. Varesio, G. Ivosev, L. Burton, R. Bonner, G. Hopfgartner. Comprehensive analytical strategy for biomarker identification based on liquid chromatography coupled to mass spectrometry and new candidate confirmation tools. *Anal. Chem.* **2009**, *81*(18), 7677–7694.

[7] S. Kern, K. Fenner, H. P. Singer, R. P. Schwarzenbach, J. Hollender. Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. *Environ. Sci. Technol.* **2009**, *43*(18), 7039–7046.

[8] S. D. Richardson. Environmental mass spectrometry: emerging contaminants and current issues. *Anal. Chem.* **2012**, *84*(2), 747–778.

[9] J. M. Wells, S. A. McLuckey. Collision-induced dissociation (CID) of peptides and proteins. In Biological Mass Spectrometry, volume 402 of Methods in Enzymology **2005**, 148–185. Academic Press.

[10] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*(7), 703–714.

[11] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, L. Querengesser. HMDB: the human metabolome database. *Nucleic Acids Res.* **2007**, *35*(suppl1), D521–526.

[12] S. Stein. Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644–655.

[13] C. A. Smith, G. O. Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak. METLIN: a metabolite mass spectral database. In *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology* **2005**, *27*, 747–751. Louisville, Kentucky.

[14] E. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant, R. A. Wheeler, D. C. Spellmeyer. Chapter 12 PubChem: integrated platform of small molecules and biological activities. In *Annual Reports in Computational Chemistry* **2008**, *4*, 217–241. Elsevier.

[15] M. Kanehisa, S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*(1), 27–30.

[16] H. E. Pence, A. Williams. ChemSpider: an online chemical information resource. *J. Chem. Educ.* **2010**, *87*(11), 1123–1124.

[17] S. Neumann, S. Böcker. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem.* **2010**, *398*(7-8), 2779–2788.

[18] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **2010**, *11*(1), 148.

[19] H. Collier. Proceedings of the 2003 International Chemical Information Conference: Nîmes, France, 19–22 October 2003. Infonortics, **2003**.

[20] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*(2), 493–500. PMID: 12653513.

[21] P. Willett, J. M. Barnard, G. M. Downs. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*(6), 983–996.

[22] D. Butina. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (4), 747–750.

[23] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman, D. F. Grant. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.* **2008**, *80*(14), 5574–5582.

[24] A. Kerber, M. Meringer, C. Rücker. CASE via MS: ranking structure candidates by mass spectra. *Croatica chemica acta* **2006**, *79*(3), 449–464.

Copyright © 2013 John Wiley & Sons, Ltd.

DOI:10.1002/jms.3123

Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

[25] E. L. Schymanski, M. Meringer, W. Brack. Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Anal. Chem.* **2009**, *81*(9), 3608–3617.
[26] S. Neumann, A. Thum, C. Böttcher. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics* **2012**, 1–8. 10.1007/s11306-012-0401-0.
[27] L. W. Sumner, A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. Lane, J. C. Lindon, P. Marriott, A. Nicholls, M. Reily, J. Thaden, M. R. Viant. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*(3), 211–221.
[28] E. L. Schymanski, C. M. J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Anal. Chem.* **2012**, *84*(7), 3287–3295.
[29] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **2005**, *21*(8), 1635–1638.
[30] F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, S. Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.* **2012**, *84*(7), 3417–3426.
[31] F. Hufsky, K. Dührkop, F. Rasche, M. Chimani, S. Böcker. Fast alignment of fragmentation trees. *Bioinformatics* **2012**, *28*(12), i265–i273.
[32] M. Rojas-Cherto, J. E. Peironcely, P. T. Kasper, J. J. J. van der Hooft, R. C. H. de Vos, R. Vreeken, T. Hankemeier, T. Reijmers. Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. *Anal. Chem.* **2012**, *84*(13), 5524–5534.
[33] R. A. Dixon, D. Strack. Phytochemistry meets genome analysis, and beyond. *Phytochemistry* **2003**, *62*(6), 815–816.

Copyright © 2013 John Wiley & Sons, Ltd.

DOI:10.1002/jms.3123
Reprinted with permission. © 2013 John Wiley & Sons, Inc. All Rights Reserved.

analytical chemistry

Article

pubs.acs.org/ac

Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies

Hendrik Treutler,[‡] Hiroshi Tsugawa,[∥] Andrea Porzel,[§] Karin Gorzolka,[‡] Alain Tissier,[†] Steffen Neumann,[‡] and Gerd Ulrich Balcke[*,†]

[†]Leibniz Institute of Plant Biochemistry, Department of Cell and Metabolic Biology, Weinberg 3, D-06120 Halle/Saale, Germany

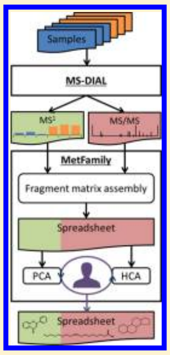
[‡]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, D-06120 Halle/Saale, Germany

[§]Leibniz Institute of Plant Biochemistry, Department of Bioorganic Chemistry, Weinberg 3, D-06120 Halle/Saale, Germany

[∥]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045, Japan

Supporting Information

ABSTRACT: The identification of metabolites by mass spectrometry constitutes a major bottleneck which considerably limits the throughput of metabolomics studies in biomedical or plant research. Here, we present a novel approach to analyze metabolomics data from untargeted, data-independent LC-MS/MS measurements. By integrated analysis of MS^1 abundances and MS/MS spectra, the identification of regulated metabolite families is achieved. This approach offers a global view on metabolic regulation in comparative metabolomics. We implemented our approach in the web application "MetFamily", which is freely available at http://msbi.ipb-halle.de/MetFamily/. MetFamily provides a dynamic link between the patterns based on MS^1-signal intensity and the corresponding structural similarity at the MS/MS level. Structurally related metabolites are annotated as metabolite families based on a hierarchical cluster analysis of measured MS/MS spectra. Joint examination with principal component analysis of MS^1 patterns, where this annotation is preserved in the loadings, facilitates the interpretation of comparative metabolomics data at the level of metabolite families. As a proof of concept, we identified two trichome-specific metabolite families from wild-type tomato *Solanum habrochaites* LA1777 in a fully unsupervised manner and validated our findings based on earlier publications and with NMR.



INTRODUCTION

Metabolomics experiments provide small molecule measurements from biological samples in a broad range of applications including cancer research, drug development, and plant science.[1−5] Mass spectrometry (MS) coupled to liquid chromatography (LC) is an essential analytical technology to acquire a snapshot of the metabolic state of a sample. On the basis of untargeted MS measurements, it is possible to measure thousands of detectable signals as MS^1 features per chromatographic run and to acquire signal profiles of small molecules based on retention time (RT), accurate mass-to-charge ratio (m/z), and abundance.[6] Univariate or multivariate statistical analysis is then applied to signal profiles of different sample groups to detect MS^1 features that are group-discriminating or of interest based on the experimental design.

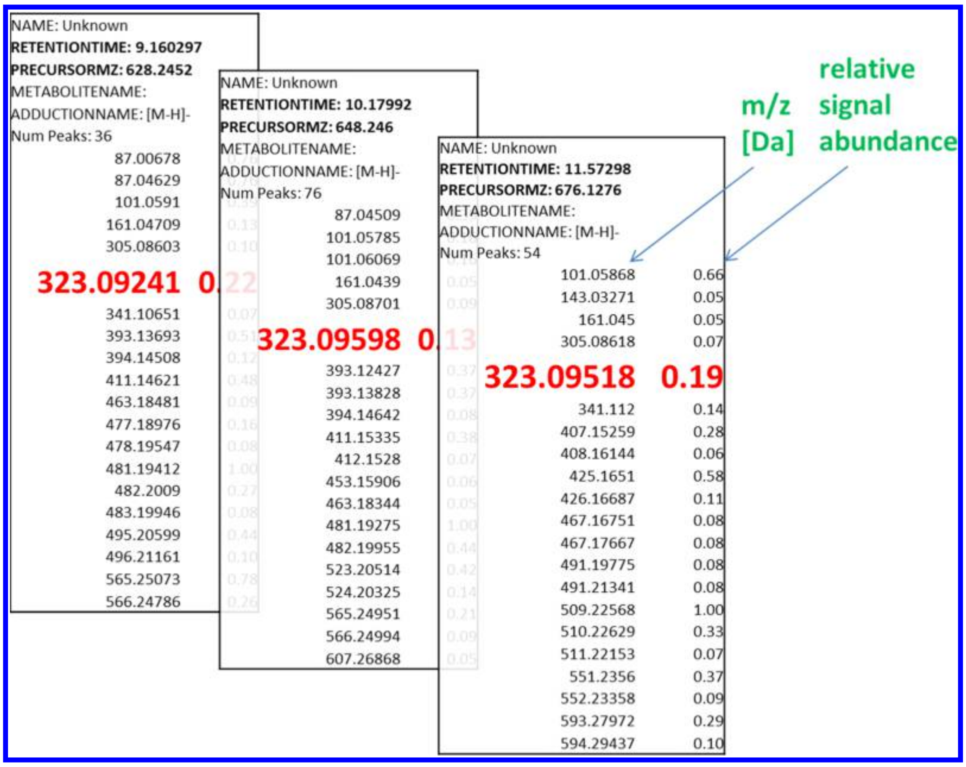
Hints for the structural characterization or even identification of MS^1 features are obtained from tandem MS measurements (MS/MS), where the metabolites undergo fragmentation resulting in MS/MS spectra. MS/MS spectra can be collected by data-dependent acquisition (DDA) or in data-independent acquisition (DIA) mode, requiring a trade-off between dwell time and spectral purity.[7,8] Using DIA, it is possible to collect thousands of MS^1 features from a single LC run as well as the associated MS/MS spectra.[9] However, in most studies, the identity of the vast majority of MS^1 features is unknown. Structure elucidation of each individual MS^1 feature from complex biological samples, e.g., by NMR and interpretation of MS/MS spectra, is currently out of reach. Thus, the biochemical relation between MS^1 features remains largely unexplained.

Group-discriminating MS^1 features are often structurally related, e.g., if particular metabolic pathways are differentially regulated as a consequence of disease,[10] stress,[11] genetic manipulation,[12] or in the case of organ-specific accumulation of structurally related metabolites.[13] Structurally related metabolites often exhibit latent similarity in their MS/MS spectra in which characteristic fragmentation patterns arise from common functional groups or structural features. For instance, upon negative mode ionization and collision-induced dissociation (CID), adenylated metabolites such as adenyl nucleotides, CoA esters, and NAD cofactors form a fragment ion of m/z 134.0472 Da ($C_5N_5H_4^-$), which corresponds to the mass of the purine core element. Under the same conditions, glucosides often form a fragment ion of m/z 161.0455 Da ($C_6H_9O_5^-$), characteristic of the hexose side-chain. Thus, on the basis of existing information, precursor ions showing these character-

Received: April 21, 2016
Accepted: July 24, 2016
Published: July 24, 2016

© 2016 American Chemical Society

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.



```
NAME: Unknown
RETENTIONTIME: 9.160297
PRECURSORMZ: 628.2452
METABOLITENAME:
ADDUCTIONNAME: [M-H]-
Num Peaks: 36
    87.00678
    87.04629
    101.0591
   161.04709   0.13
   305.08603   0.10
   323.09241   0.22
   341.10651   0.07
   393.13693   0.51
   394.14508   0.12
   411.14621   0.48
   463.18481   0.09
   477.18976   0.16
   478.19547   0.08
   481.19412   1.00
    482.2009   0.27
   483.19946   0.08
   495.20599   0.44
   496.21161   0.10
   565.25073   0.78
   566.24786   0.26
```

```
NAME: Unknown
RETENTIONTIME: 10.17992
PRECURSORMZ: 648.246
METABOLITENAME:
ADDUCTIONNAME: [M-H]-
Num Peaks: 76
    87.04509
   101.05785
   101.06069
    161.0439   0.05
   305.08701   0.09
   323.09598   0.13
   393.12427   0.37
   393.13828   0.37
   394.14642   0.08
   411.15335   0.38
    412.1528   0.07
   453.15906   0.06
   463.18344   0.05
   481.19275   1.00
   482.19955   0.44
   523.20514   0.42
   524.20325   0.14
   565.24951   0.21
   566.24994   0.09
   607.26868   0.05
```

```
NAME: Unknown
RETENTIONTIME: 11.57298
PRECURSORMZ: 676.1276
METABOLITENAME:
ADDUCTIONNAME: [M-H]-
Num Peaks: 54
   101.05868   0.66
   143.03271   0.05
     161.045   0.05
   305.08618   0.07
   323.09518   0.19
     341.112   0.14
   407.15259   0.28
   408.16144   0.06
    425.1651   0.58
   426.16687   0.11
   467.16751   0.08
   467.17667   0.08
   491.19775   0.08
   491.21341   0.08
   509.22568   1.00
   510.22629   0.33
   511.22153   0.07
    551.2356   0.37
   552.23358   0.09
   593.27972   0.29
   594.29437   0.10
```

Figure 1. MS/MS library format before upload into MetFamily.

istic fragments could be grouped together as metabolites sharing common structural features, or *metabolite families*. However, even pre-existing MS/MS information characteristic of certain metabolite families is sparse. Hence, novel approaches that turn MS^1- and MS/MS-features into interpretable information within a reasonable amount of time are urgently needed. These approaches should be able to relate MS^1 abundances to latent similarity at the MS/MS spectral level.

Recently, several studies reported on the organization of hundreds of MS^1 features by molecular networking depicting relationships between structurally related molecules based on their spectral similarity.[14−17] However, an explicit assignment of MS^1 features to similarity clusters and the source of structural similarity between up- or downregulated MS^1 features was not apparent. Previously, Wagner et al. used GC-MS data for hierarchical cluster analysis (HCA) to arrange known and structurally related metabolites.[18] Using HCA, it was possible to identify structural classes among 59 metabolites. Rasche et al. described FT-BLAST[19] to compare spectra and computationally derived fragmentation trees, revealing clusters of structurally closely related compounds. However, neither Wagner et al. nor Rasche et al. considered the abundance of MS^1 features in different samples.

Inspired by the idea to comprehensively analyze molecular networks and to explicitly group MS^1 features, we performed HCA across hundreds of MS/MS spectra obtained from glandular trichomes of wild-type tomato *Solanum habrochaites* LA1777. Glandular trichomes of vascular plants such as tomato are metabolic factories producing a plethora of secondary metabolites involved in plant defense and the communication with its environment.[13,20] We considered characteristic fragments prevalent in MS/MS similarity clusters to assign MS^1 features to certain trichome-specific metabolite families. In addition, we applied principal component analysis (PCA) to metabolite profiles for the discovery of group-discriminating MS^1 features and combined the information on metabolite families obtained from HCA (MS/MS feature similarity) with the PCA loadings (sample-specific MS^1 abundance). This combination of statistical analyses of MS^1 feature abundances and MS/MS structural annotations can not only speed-up the individual analysis steps, but allows us to address new questions, such as the discovery of group-discriminating metabolite families with biochemical relevance. Here, we exemplarily selected two metabolite families being produced by tomato glandular trichomes which play important roles in the plant defense against herbivores, namely the branched chain acyl sugars[21−24] and the sesquiterpene glucosides which are potentially poisonous to plant herbivores.[25,26] We implemented the proposed methodology in the Open Source web application "MetFamily" and made our approach freely available (accessible via http://msbi.ipb-halle.de/MetFamily/).

■ MATERIALS AND METHODS

Fragment Matrix Assembly. MetFamily processes a metabolite profile of a set of MS^1 features together with an MS/MS library comprising MS/MS spectra for these MS^1 features. We obtain both data sets as output of MS-DIAL,[9] where the metabolite profile contains extracted m/z/retention time features from MS^1 scans with the corresponding feature abundances (Data S-1 of the Supporting Information, SI) and the MS/MS library contains deconvoluted MS/MS spectra of the MS^1 features with relative intensities of the fragment ions (Figure 1, Data S-2). Instead of MS-DIAL, other tools can produce similar input data as described in Note S-3. Upon data

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

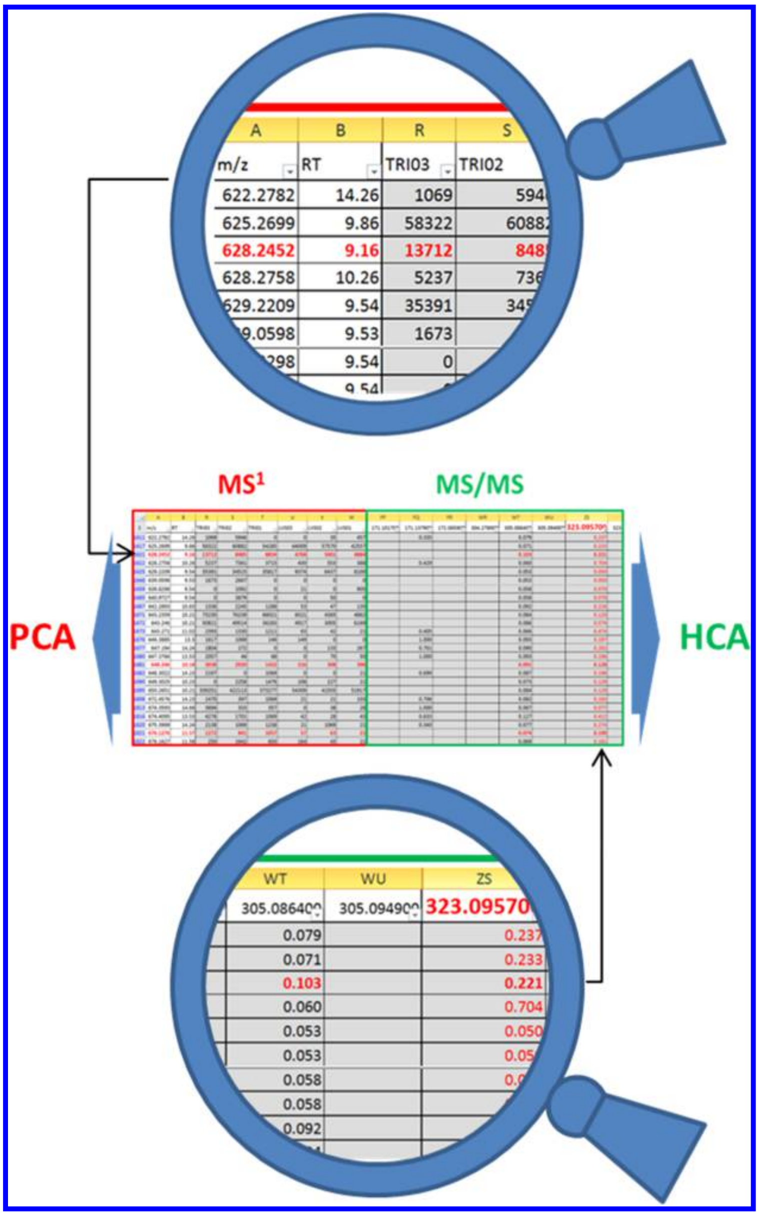


Figure 2. Combined data matrix after data preprocessing by MetFamily. The quantification part (red, left) contains the MS^1 features (rows; precursor ions) and the MS^1 abundances in individual samples. In the fragment part (green, right), the column headers are the mean of binned MS/MS features (m/z or neutral loss) from the MS/MS library. Upper zoom: m/z; retention time of feature (628.2452; 9.16) and its respective peak heights in two trichome samples. Lower zoom: relative MS/MS intensities of fragment ion m/z 323.09570 Da. Arrows to the left and to the right: MS^1 abundances are analyzed using PCA and MS/MS spectra are analyzed using HCA.

import, MetFamily aligns all MS/MS spectra with a user-defined m/z error to create the *fragment matrix* as shown in Figure 2, where the relative intensity of unique MS/MS fragments is associated with the corresponding MS^1 feature (i.e., precursor ion) and its MS^1 abundance in individual samples (Data S-3). For our showcase, this preprocessing step takes one or 2 min. The fragment matrix is assembled as follows.

First, we process the set of all fragments. Here, we remove fragments with an intensity below a user-defined noise threshold. We normalize fragment intensities within each MS/MS spectrum to a maximum of 1 (base peak). In addition, we add one neutral loss (NL) for each fragment by calculating the mass of the neutral loss as the difference of fragment m/z and precursor m/z in MS^1 (intentionally a negative m/z value). The intensity of the NLs is chosen equal to the intensity of the corresponding fragment. In this manuscript, we treat fragments and NLs equally by denoting both as fragments.

Second, we align the individual MS/MS spectra (Figures 1 and 2). Here, we match fragments from different MS/MS spectra with similar m/z and merge these to *fragment groups* of unique m/z. We call the mean of all fragment m/z's of one fragment group the *fragment group mean*. For the alignment of the individual MS/MS spectra, we use an efficient algorithm

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

implemented in the R package *xcms*[31] (version 1.44.0). This algorithm avoids the usage of fixed m/z bins with a heuristic approach that groups fragments with similar m/z and decomposes contiguous fragment groups using hierarchical clustering. Here, a fragment m/z matches a fragment group, if the following:

$$|m - m_{\text{group}}| \leq mz\text{Abs}_{\text{MS/MS}} + m \times mz\text{PPM}_{\text{MS/MS}}/1E6$$

where m is the fragment m/z, m_{group} is the fragment group mean, $mz\text{Abs}_{\text{MS/MS}}$ is a parameter representing the absolute m/z error, and $mz\text{PPM}_{\text{MS/MS}}$ is representing the relative m/z error in ppm (parts per million). See Table S-3 for a summary of user-customizable parameters. After fragment group assembly, we remove fragment groups which correspond to isotopic ions. Specifically, we detect fragment groups with a m/z difference of 1.0033 Da (regarding the fragment group means ± m/z error) which correspond to ^{13}C isotopes. Third, we create the fragment matrix with one row for each unique MS^1 precursor and columns of fragment groups (Figure 2). We register the intensity of each fragment in the row and column of the corresponding MS^1 feature and fragment group, respectively. For each MS^1 feature, we generate an ID given by "m/z/ retention time" in MS^1.

Finally, we add the set of MS^1 abundances in all samples and other annotations to each row resulting in a combined data matrix. The combined data matrix represents the data basis for subsequent analyses and can be examined in a spreadsheet program for complementing analyses (Figure 2 and Data S-3).

MS^1/MS/MS Combined Data Analysis. A principal component analysis (PCA) for the set of m MS^1 features in n samples is performed as follows. Given the m by n matrix of scaled MS^1 abundances, we calculate the scores and the loadings. Here, MetFamily supports the scaling functions *log*$_2$ *transformation*, *Pareto scaling*, *Centering*, and *Autoscaling*.[27] The scores comprise one data point per sample and reflect differences between samples. The loadings comprise one data point per MS^1 feature and emphasize MS^1 features with differential abundance between samples.

We perform a hierarchical cluster analysis (HCA) on MS/MS spectra of a set of MS^1 precursor features as follows. We calculate the distance matrix of pairwise dissimilarities between the MS/MS spectra of all MS^1 features. Here, we provide different distance functions to score common and distinct fragments. Specifically, we recommend the distance function 'Jaccard (intensity-weighted)', which sums the intensities of common and disjoint fragments:

$$f(s_i, s_j) = 1 - \frac{sum(map(s_i \cap s_j))}{sum(map(s_i \cup s_j))}$$

where s_i and s_j are the fragments in the MS/MS spectrum of MS^1 feature i and j. To suppress noise and emphasize the importance of intense fragments, *map* discretizes the intensities of the fragments as follows. Intensities smaller than 0.2 are mapped to 0.01, intensities greater or equal than 0.2 and smaller than 0.4 are mapped to 0.2, and intensities greater or equal than 0.4 are mapped to 1. Given the distance matrix, we calculate a hierarchical cluster dendrogram where each cluster of MS^1 features represents a putative metabolite family.

For each cluster of MS/MS spectra, we calculate the *cluster-discriminating power* for prevalent fragments as follows. For each fragment present in more than 50% of the MS/MS spectra in a cluster, we measure the ability of this fragment to discriminate spectra in the cluster from spectra outside the cluster as

$$\text{cdp}(f_{k,l}) = \frac{p_{\text{in}} - p_{\text{out}}}{n}$$

where $f_{k,l}$ is the l-th fragment of the k-th cluster, p_{in} is the number of MS/MS spectra in the k-th cluster containing the fragment $f_{k,l}$, p_{out} is the number of MS/MS spectra outside the k-th cluster containing the fragment $f_{k,l}$, and n is the total number of MS/MS spectra in the k-th cluster. If $p_{\text{out}} > p_{\text{in}}$, then we define $\text{cdp}(f_{k,l}) = 0$. The cluster-discriminating power of a fragment is in the range from zero to one, and a fragment with a cluster-discriminating power close to one indicates a very specific fragment.

Clusters containing fragments with a cluster-discriminating power close to one indicate metabolite families. Currently, the annotation of metabolite families based on characteristic MS/MS fragments is performed by a mass spectrometry expert who manually evaluates the hierarchy of putative metabolite families and labels a set of clusters with functional and/or structural annotations based on characteristic fragment patterns. Each MS^1 feature can be labeled with one annotation, i.e., membership in a metabolite family.

Plant Growth and Harvest. *Solanum habrochaites* LA1777 was grown on soil in a greenhouse (65% humidity, light intensity: 165 μmol s^{-1} mm^2, 21−24 °C, 16 h light period) and watered with tap water every 2 days. The plant material was harvested 12 weeks after germination during the light phase in the early afternoon. For trichome harvest, tomato leaves were put on the hand palm (using gloves) and trichomes were quickly brushed off the leaves by a 2 cm broad paint brush which was dipped in liquid nitrogen. The frozen trichomes were collected in a mortar filled with liquid nitrogen. Trichomes from 15 plant leaves were pooled under cryogenic conditions and further purified by sieving through steel sieves of 150 μm mesh width (Retsch, Hahn, Germany). After removal of trichomes, the plant leaves were immediately quenched in liquid nitrogen. Pooled leaves were ground in a mortar under liquid nitrogen conditions. After evaporation of all liquid nitrogen during storage at −80 °C leaves and trichomes were lyophilized overnight and stored in a deep freezer until extraction.

Metabolite Extraction. Using wall-reinforced cryo-tubes of 1.6 mL volume (Precellys Steel Kit 2.8 mm, Peqlab Biotechnologie GmbH, Erlangen, Germany) filled with 5 steel beads (3 mm), 25 mg aliquots of dry leaf or trichome powder was suspended in 900 μL dichloromethane/ethanol (−80 °C). Then, 200 μL of 50 mM aqueous ammonium formate/formic acid buffer (0 °C, pH 3) was added to each vial, and two rounds of cell rupture/metabolite extraction were conducted by FastPrep bead beating (60 s, speed 5.5 m/s, first round −80 °C, second round room temperature, FastPrep24 instrument with cryo adapter, MP Biomedicals LLC, Santa Ana, CA, U.S.A.). After phase separation by centrifugation at 20 000g (2 min, 0 °C) the aqueous phase was removed, and 600 μL of the organic phase was collected. Following, 500 uL tetrahydrofuran (THF) was added to exhaustively extract hydrophobic metabolites and the Fastprep and centrifugation were repeated accordingly. The THF supernatant was combined with the first organic phase extract and dried in a stream of nitrogen gas. The dried extract was resuspended in 150 μL 75% methanol (aqueous) and filtered over 0.2 μm PVDF.

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

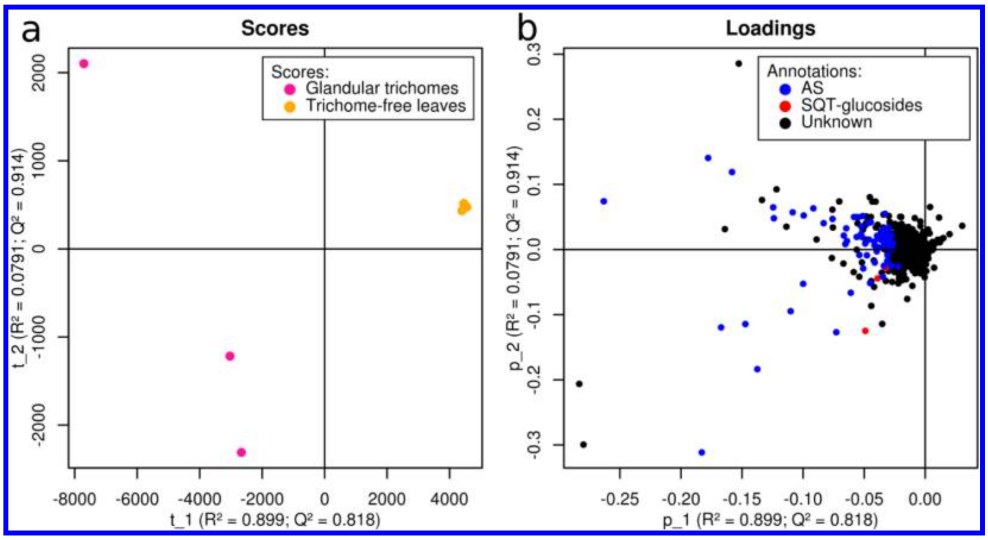


Figure 3. Principal component analysis of metabolite extracts of glandular trichomes and leaves of *Solanum habrochaites* LA1777. Comparison of 2585 MS^1 features from TOF-MS measurements ($n = 6$). (a) scores and (b) loadings with annotations. The PCA loadings with annotations indicate a predominant enrichment of acyl sugars in glandular trichomes. AS: acyl sugars, SQT-glucosides: sesquiterpene glucosides, and Unknown: Not characterized here.

Analytical Conditions for Liquid Chromatography and Mass Spectrometry. 0.5 μL methanolic extract was injected into an Acquity-UPLC (Waters Inc.) and separated on a Nucleoshell RP18 (150 mm × 2 mm × 2.7 μm; Macherey & Nagel, Düren, Germany) at 40 °C. The mobile phase A was 0.33 mM ammonium formate with 0.66 mM formic acid in water; mobile phase B was acetonitrile. The gradient was 0 min, 5% B; 2 min, 5% B; 19 min, 95% B; 21 min, 95% B; 21.1 min, 5% B; and 24 min, 5% B. The column flow rate was 0.4 mL/min, the autosampler temperature was 4 °C.

ESI-(−)-Mass Spectrometry was performed on an AB Sciex TripleTOF 5600 system (Q-TOF) equipped with a DuoSpray ion source. All analyses were performed at the high sensitivity mode for both TOF MS^1 and product ion scan. The mass calibration was automatically performed every 20 injections using an APCI calibrant solution via a calibration delivery system (CDS). The instrument (TripleTOF 5600, Sciex, Toronto, Canada) was configured to simultaneously acquire high resolution MS/MS spectra for all MS^1 features (sequential window acquisition of all theoretical fragment-ion spectra, SWATH)[28] (Figure S-1). The SWATH parameters were MS^1 accumulation time, 150 ms; MS^2 accumulation time, 20 ms; collision energy, −45 V; collision energy spread, 35 V; cycle time, 1160 ms; Q1 window, 25 Da; mass range, m/z 65−1250. The other parameters were curtain gas, 35; ion source gas 1, 60; ion source gas 2, 70; temperature, 600 °C; ion spray voltage floating, −4.5 kV; declustering potential, 35 V.

Raw Data Processing. After measurement, raw data of triplicate trichome and trichome-free leaf material was converted from the vendor file format (in our case *.wiff) into the common file format of Reifycs Inc. (Analysis Base File format *.abf) using the freely available Reifycs ABF converter (http://www.reifycs.com/AbfConverter/index.html).This process took about 1 min per sample. After conversion, the freely available MS-Dial software was used for feature detection, ion species annotation, compound spectra extraction, and peak alignment between samples.[9] Data processing by MS-Dial using the parameters in Table S-1 took about 30 min. Data processing by MetFamily using the parameters in Table S-2 took 1 min.

Notably, neither the use of SWATH-triggered CID fragmentation nor the use of MS-Dial are prerequisite to run MetFamily. Any data independent or data dependent acquisition to collect MS/MS spectra and other peak picking and deconvolution software can alternatively be used.[29−32] In that case, their output has to be provided as a text file containing the peak intensities and a msp-type spectral library which are formatted as exemplified in Data S-1 and Figure 1, and described in Note S-3. However, as unique feature, MS-Dial jointly deconvolutes MS^1 and MS/MS features and automatically predicts the precursor ion when DIA was applied. Via the Reifycs ABF converter, MS-DIAL accepts all of major MS vendor-formats as well as the common mzML data and is applicable to either DIA or DDA MS/MS fragmentation methods.

Substance Purification. Since NMR requires purified analytes in the upper μm range, 1 kg of LA1777 leaf material was surface-extracted with methanol for 2 h. After evaporation, a methanolic concentrate of this extract was produced and injected into a LC system in 100 μL increments. For peak separation using semipreparative HPLC and an analysis by mass spectrometry (1260 Infinity system, Agilent), a full scan between 200 and 800 m/z was performed after negative electrospray ionization (ion source: API-ES, gas temperature: 350 °C, drying gas 10 mL/min, nebulizer pressure 35 psig, capillary voltage 4500 V). For HPLC, a XTerra prep MS C18 column (5 μm × 7.8 mm × 150 mm; Waters) was used and run at a flow rate of 6 mL/min at 25 °C. Solvent A was 0.3 mM ammonium formate acidified with formic acid to pH 6.2. Solvent B was acetonitrile. Gradient conditions were: 0−5 min 5% B; 5−87 min linear gradient to 95% B; 87−88 min 95% B; and 88−90 min 5% B. For fractionation, m/z 605.5, 737.5, and 751.5 triggered the selective collection. A makeup pump that transferred an aliquot of the eluate to the mass analyzer was set to 0.5 mL/min 50% A - 50% B. Subsequently, all collected fractions were dried by lyophilization prior to NMR analysis.

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

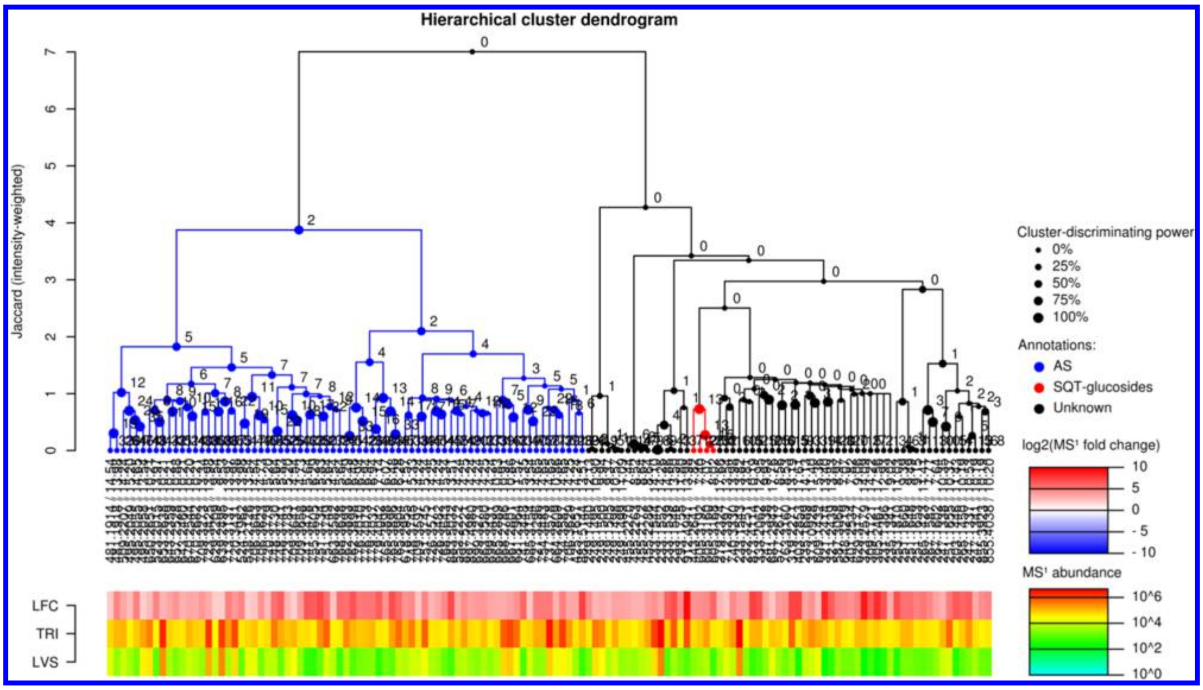


Figure 4. Hierarchical cluster analysis of 135 trichome-specific MS^1 features using the corresponding MS/MS spectra obtained from organic extracts of *S. habrochaites* LA1777. For comparison of the groups trichomes versus leaf focusing on trichome-specific features, the set of 2585 MS^1 features was filtered using an MS^1 abundance threshold of 20 000 counts and a $\log_2$-fold change (LFC) of two. The heatmap below depicts the LFC and the absolute MS^1 abundance in glandular trichomes (TRI) and trichome-free leaves (LVS), respectively. The 135 filtered MS^1 features clearly segregated into two main signal-clusters which in turn further segregated into signal-clusters with different levels of similarity between MS/MS spectra. Specifically, we identified a cluster of 73 short branched chain acyl sugars (AS, in blue) and a cluster of four sesquiterpene glucosides (SQT-glucosides, in red) on the basis of a set of characteristic fragments which were prevalent in both clusters (see legend "Annotations" on the right). Both signal-clusters show characteristic fragments with a cluster-discriminating power of 80% and more (size of the branch nodes, see legend "Cluster-discriminating power" on the right). 58 trichome-specific MS^1 features partially showed further clusters, but remained uncharacterized in this study (Unknown in black).

Analytical Conditions for NMR. NMR spectra were recorded on an Agilent/Varian VNMRS 600 NMR spectrometer operating at a proton NMR frequency of 599.83 MHz using a 5 mm inverse detection cryoprobe. 2D NMR spectra were recorded using standard pulse sequences (gDQCOSY, zTOCSY, gHSQCAD, gHMBCAD) implemented in Agilent (Varian) VNMRJ 4.2A (CHEMPACK 7.1) spectrometer software. A TOCSY mixing time of 80 ms was used. HSQC experiments were run with multiplicity editing and optimized for $^1J_{CH}$ = 146 Hz. HMBC experiments were optimized for a long-range coupling constant of 8 Hz; a 2-step $^1J_{CH}$ filter was used (130−165 Hz). Proton and carbon chemical shifts are referenced to internal TMS (0 ppm).

■ RESULTS AND DISCUSSION

As a proof of concept, we applied MS signal profiles to compare the metabolism of a special plant organ in tomato, the glandular trichomes, to tomato leaves. Plant glandular trichomes are secretory cells that protrude from the epidermis of many vascular plants. As "metabolic factories", they produce important drugs such as the antimalaria artemisinin or compounds known to be involved in plant defense.[20,33] Here, we used *Solanum habrochaites* LA1777, a wild type tomato accession with a rich profile of secondary metabolites produced in the glandular trichomes.[34] We used six UPLC-(−)ESI-SWATH-MS/MS runs of triplicate trichome and trichome-free leaf extracts (cf. Materials and Methods). However, MetFamily is applicable to a larger number of samples and sample groups. We used MS-DIAL[9] for data preprocessing and exported (i) a signal profile with MS^1 features and (ii) a spectral library with deconvoluted MS/MS spectra extracted from the raw data (Data S-1 and Data S-2). Using the software MetFamily, we aligned the MS/MS spectra of the spectral library resulting in a novel fragment matrix structure, and we fused this fragment matrix with the matching set of MS^1 features from the six individual samples to a single matrix (cf. Materials and Methods, Figures 1 and 2, Data S-3, Table S-3).

MetFamily provides options to perform principal component analyses (PCA). Here, we performed a PCA on 2585 MS^1 features detected in glandular trichomes or leaves of LA1777 using Pareto-scaled data. In our example, PC1 shows a clear separation between trichomes and leaves with $R^2 = 0.90$, $Q^2 = 0.82$ and a large number of MS^1 features more abundant in glandular trichomes (Figure 3A,B). A Scree plot on additional principal components is provided in Figure S-2. Up to this point, all data have been acquired in a fully untargeted manner and traditionally this is where group-discriminating MS^1 features would be subjected to tedious manual structure elucidation. In our approach, we amended the loadings plot of the PCA (Figure 3B) with a set of structural annotations based on characteristic MS/MS fragments which we identified in different signal-clusters using HCA (Figure 4). Using MetFamily, we performed a hierarchical cluster analysis (HCA) on MS/MS spectra of the fragment matrix (Data S-

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

3). For PCA as well as for HCA, MetFamily allows the usage of thresholds for the MS^1 abundance of individual MS^1 features (average of all samples) and for the $\log_2$-fold change between the average MS^1 abundance of two sample groups. Since we were interested in abundant trichome-specific metabolites, we retained 135 MS^1 features in the HCA with MS^1 abundances ≥20 000 counts and a $\log_2$-fold change ≥2 comparing trichomes versus leaf. After hierarchical cluster analysis, the resulting dendrogram indicated a clear segregation into two main clades with internal spectral similarity (Figure 4).

The first signal-cluster contained 73 MS^1 features which correspond to short branched chain acyl sugars[21] (AS, blue in Figure 4). The structural similarities among members of this clade was supported by prevalent fragment ions 87.0451 Da (theoretical mass for $C_4H_7O_2^-$ is 87.0452) and 101.0603 Da (theoretical mass for $C_5H_9O_2^-$ is 101.0608), which are indicative for short branched acyl groups. These acyl moieties were esterified to sucrose as reflected by the fragments 323.0957 Da (theoretical mass for $C_{12}H_{19}O_{10}^-$ is 323.0984; sucrose-H_2O—H^-) and 305.0864 Da (theoretical mass for $C_{12}H_{17}O_9^-$ is 305.0878; sucrose-$2H_2O$—$H]^-$). MS/MS fragmentation patterns and NMR analysis of two selected MS^1 features of this clade ([m/z; RT]: [737.3578; 14.65] and [751.3749; 15.64]) confirmed the membership to the metabolite family of short branched chain acyl sugars (Figures S-3, S-4, S-7, S-8, S-11−S-14, S-15−S-19 and Tables S-5, S-6). Our NMR analysis revealed that the feature [737.3578; 14.65] comprised an isomeric mixture of isobutyl, isopentyl, and anteisobutyl acyl moieties, which were not resolvable using our chromatography. MS/MS fragmentation and NMR of various AS have been thoroughly studied earlier by Ghosh et al., where compounds selected here for analysis were annotated as acylsucrose S4:21[2] (theoretical m/z:737.36012 Da (formate adduct-H)) and acylsucrose S4:22[6] (theoretical m/z: 751.37577 Da (formate adduct-H)), respectively.[21]

The second signal-cluster contained a group of four MS^1 features which correspond to sesquiterpene glycosides (SQT-glucosides, red in Figure 4). The structural similarities among members of this clade was supported by three prevalent fragment ions: m/z 401.2548 Da (theoretical mass for $C_{21}H_{37}O_7^-$ is 401.2545), 563.3051 Da (theoretical mass for $C_{27}H_{47}O_{12}^-$ is 563.3073), and 605.3176 Da (theoretical mass for $C_{29}H_{49}O_{13}^-$ is 605.3179) (Figure S-5). Recently, Ekanayaka et al. identified a novel class of trichome-specific sesquiterpene glucosides from *S. habrochaites* using these fragment ions and elucidated the structures of purified representatives by NMR.[26] In our study, CID fragmentation and preparative isolation of MS^1 feature [605.3160; 7.07, an abundant in-source fragment] with subsequent NMR confirmed the structure of 12-*O*-(6″-*O*-malonyl-*β*-D-glucopyranosyl-(1 → 2)-*β*-D-glucopyranosyl)-campherenane-2-endo,12-diol, a member of the novel sesquiterpene glucoside metabolite family (Figures S-3−S-6, S-9, S-10 and Tables S-3, S-4).

After annotation of both metabolite families, the corresponding MS^1 features are highlighted by their color-code in the PCA loadings (Figure 3B). In our case, it was evident that the representatives of both metabolite families were enriched in glandular trichomes, indicating a trichome-specific upregulation of short branched chain acyl sugars and sesquiterpene glucosides. Please note that the hierarchical cluster dendrogram comprised more clades with internal spectral similarity, but we concentrated on the short branched chain acyl sugars and sesquiterpene glucosides whose structures were confirmed by NMR. A detailed workflow exemplified here is given in Figure S-1, and the full showcase protocol is given in Note S-1. A general user guide for MetFamily is given in Note S-2.

Additional Features of MetFamily. MetFamily also supports semitargeted analyses. In this case, sets of MS^1 features can be selected by certain fragment masses, neutral losses, or combinations thereof within a user-defined mass error in ppm as filter criteria. Using this option, only selected MS^1 features are considered in subsequent PCA or HCA calculations and the data analysis is consequently constrained to selected metabolite families. For example, to isolate only glycosylated MS^1 features from all data the user can specify a fragment ion of m/z 161.0455 Da ($C_6H_9O_5^-$) from MS/MS spectra in negative mode and can then focus on the regulation of enzymatic glycosylations in a biological context (for details, see the MetFamily user guide in Note S-2). When we applied this filter with a mass error of 25 ppm, we obtained 568 MS^1 features from our example data, presumably containing a hexose as a structural moiety. In addition, it is possible to search MS^1 features with certain fragments or neutral losses postanalysis. The corresponding MS^1 features can then be jointly visualized in the PCA loadings and the hierarchical cluster dendrogram.

It is possible to export different kinds of results from MetFamily. Selected sets of precursor ions can be exported and, e.g., reloaded into the original MS data acquisition software. Further, it is possible to export both the hierarchical cluster dendrogram and the PCA plots as publication-ready high quality images. The set of parameters used for the initial data import can be exported and imported. Finally, it is possible to export the whole project (including all annotations and color codes) to enable the user to share the project or to continue the data analysis at a later time (Data S-4).

■ CONCLUSIONS

The web application "MetFamily" presented here constitutes a novel approach to analyze metabolomics data from untargeted, data-independent LC-MS/MS measurements. Rather than relying on the time-consuming structure identification of individual metabolites, MetFamily assists in the interpretation of complex metabolomics data by identifying metabolite families through patterns in MS/MS. These are generated by similarity clustering of associated MS/MS spectra and can be annotated with names and colors. After preprocessing of LC-MS/MS raw data, MetFamily performs a joint data analysis of MS^1 abundances and MS/MS spectra in which the annotation of metabolite families facilitates the interpretation of comparative data sets. Structure elucidation at the metabolite level can be performed afterward in a much more focused way. As a proof of concept, we identified two trichome-specific metabolite families from wild type *Solanum habrochaites* LA1777 in a fully unsupervised manner and validated our findings based on earlier publications and with NMR. The plethora of identified trichome-specific acyl sucroses correlates with upregulation of acyltransferases of the BAHD family in tomato glandular trichomes (Schilmiller 2012). In addition, the size of the clade "acyl sugar" is related to a low substrate specificity of BAHD acyltransferases, illustrating that MetFamily can uncover links between enzymatic promiscuity and organ-specific regulation of enzymes.

Using the proposed approach, it is now possible to obtain a comprehensive overview of data sets containing thousands of mass features within a reasonable amount of time. Thus, by

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

providing a dynamic link between structural similarity at the MS/MS level (HCA) and the corresponding MS^1-signal intensity-based patterns (PCA) we bridge the gap between raw data and structural information. Moreover, using MetFamily, precursor ions can now be filtered via combinations of fragment ions and neutral losses, permitting the selection of metabolite families based on characteristic fragmentation patterns.

While traditional compound identification is based on the comparison of MS/MS spectra (or electron impact MS spectra) with reference spectra from known compounds, future developments should exploit spectral patterns of MS/MS features being characteristic of certain metabolite families. Public knowledge on such characteristic fragment ions or neutral losses, e.g., based on metabolite families, can assist mass spectrometry specialists in the elucidation of unknown features and will open new perspectives in life science.

■ AVAILABILITY

Project name: MetFamily
Source code: https://github.com/Treutler/MetFamily
Availability: http://msbi.ipb-halle.de/MetFamily/
Operating system(s): Platform independent
Programming language: R
Other requirements: Installation of R 3.2.2 or higher; License: GPL 3
Any restrictions to use by nonacademics: None

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b01569.

General work flow as flowchart (Figure S-1) Scree plot of the first five principle components of Figure 3 (Figure S-2). Exported parameter file of MS-DIAL (Table S-1) and exported parameter file from MetFamily and parameter explanation (Tables S-2, S-3). Structure elucidation of three selected MS^1 features (Figures S-3–S-19, Tables S-4–S-6). Metabolite profile of the showcase (Data S-1), MS/MS library of the showcase (Data S-2), matrix of the showcase (Data S-3), and annotated MetFamily project file (Data S-4) (PDF)
Protocol for the presented showcase (PDF)
user guide for MetFamily (PDF)
Detailed specification of MetFamily input files (PDF)
(TXT)
(ZIP)
(ZIP)
(ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gerd.balcke@ipb-halle.de (G.U.B.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Anja Ehrlich for the preparative isolation of metabolites, Anja Henning and Nick Bergau for trichome harvest and data acquisition. We thank Prof. Masanori Arita and K.G. for fruitful discussions and review of the manuscript.

■ REFERENCES

(1) Jorge, T. F.; Rodrigues, J. A.; Caldana, C.; Schmidt, R.; van Dongen, J. T.; Thomas-Oates, J.; Antonio, C. *Mass Spectrom. Rev.* **2016**, *35*, 620.

(2) Tonoli, D.; Varesio, E.; Hopfgartner, G. *Chimia* **2012**, *66*, 218–222.

(3) Wishart, D. S.; Mandal, R.; Stanislaus, A.; Ramirez-Gaona, M. *Metabolites* **2016**, *6*.

(4) Suhre, K.; Shin, S. Y.; Petersen, A. K.; Mohney, R. P.; Meredith, D.; Wagele, B.; Altmaier, E.; CardioGram; Deloukas, P.; Erdmann, J.; Grundberg, E.; Hammond, C. J.; de Angelis, M. H.; Kastenmuller, G.; Kottgen, A.; Kronenberg, F.; Mangino, M.; Meisinger, C.; Meitinger, T.; Mewes, H. W.; Milburn, M. V.; Prehn, C.; Raffler, J.; Ried, J. S.; Romisch-Margl, W.; Samani, N. J.; Small, K. S.; Wichmann, H. E.; Zhai, G.; Illig, T.; Spector, T. D.; Adamski, J.; Soranzo, N.; Gieger, C. *Nature* **2011**, *477*, 54–60.

(5) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–171.

(6) Fernie, A. R.; Trethewey, R. N.; Krotzky, A. J.; Willmitzer, L. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 763–769.

(7) Roemmelt, A. T.; Steuer, A. E.; Poetzsch, M.; Kraemer, T. *Anal. Chem.* **2014**, *86*, 11742–11749.

(8) Zhu, X.; Chen, Y.; Subramanian, R. *Anal. Chem.* **2014**, *86*, 1202–1209.

(9) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12*, 523–526.

(10) Ferslew, B. C.; Xie, G.; Johnston, C. K.; Su, M.; Stewart, P. W.; Jia, W.; Brouwer, K. L.; Sidney Barritt, A. t. *Dig. Dis. Sci.* **2015**, *60*, 3318–3328.

(11) Kaling, M.; Kanawati, B.; Ghirardo, A.; Albert, A.; Winkler, J. B.; Heller, W.; Barta, C.; Loreto, F.; Schmitt-Kopplin, P.; Schnitzler, J. P. *Plant, Cell Environ.* **2015**, *38*, 892–904.

(12) Qu, G.; Quan, S.; Mondol, P.; Xu, J.; Zhang, D.; Shi, J. *J. Integr. Plant Biol.* **2014**, *56*, 849–863.

(13) Glas, J. J.; Schimmel, B. C. J.; Alba, J. M.; Escobar-Bravo, R.; Schuurink, R. C.; Kant, M. R. *Int. J. Mol. Sci.* **2012**, *13*, 17077–17103.

(14) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743–1752.

(15) Garg, N.; Kapono, C. A.; Lim, Y. W.; Koyama, N.; Vermeij, M. J. A.; Conrad, D.; Rohwer, F.; Dorrestein, P. C. *Int. J. Mass Spectrom.* **2015**, *377*, 719–727.

(16) Nguyen, D. D.; Wu, C. H.; Moree, W. J.; Lamsa, A.; Medema, M. H.; Zhao, X. L.; Gavilan, R. G.; Aparicio, M.; Atencio, L.; Jackson, C.; Ballesteros, J.; Sanchez, J.; Watrous, J. D.; Phelan, V. V.; van de Wiel, C.; Kersten, R. D.; Mehnaz, S.; De Mot, R.; Shank, E. A.; Charusanti, P.; Nagarajan, H.; Duggan, B. M.; Moore, B. S.; Bandeira, N.; Palsson, B. O.; Pogliano, K.; Gutierrez, M.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E2611–E2620.

(17) Li, D.; Baldwin, I. T.; Gaquerel, E. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E4147–4155.

(18) Wagner, C.; Sefkow, M.; Kopka, J. *Phytochemistry* **2003**, *62*, 887–900.

(19) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatos, A.; Bocker, S. *Anal. Chem.* **2012**, *84*, 3417–3426.

(20) Tissier, A. *Plant J.* **2012**, *70*, 51–68.

(21) Ghosh, B.; Westbrook, T. C.; Jones, A. D. *Metabolomics* **2014**, *10*, 496–507.

(22) Kim, J.; Kang, K.; Gonzales-Vigil, E.; Shi, F.; Jones, A. D.; Barry, C. S.; Last, R. L. *Plant Physiol.* **2012**, *160*, 1854–1870.

(23) Schilmiller, A.; Shi, F.; Kim, J.; Charbonneau, A. L.; Holmes, D.; Jones, A. D.; Last, R. L. *Plant J.* **2010**, *62*, 391–403.

(24) Schilmiller, A. L.; Charbonneau, A. L.; Last, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 16377–16382.

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

(25) Ekanayaka, E. A.; Celiz, M. D.; Jones, A. D. *Plant Physiol.* **2015**, *167*, 1221.

(26) Ekanayaka, E. A. P.; Li, C.; Jones, A. D. *Phytochemistry* **2014**, *98*, 223−231.

(27) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.

(28) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, 016717.

(29) Lommen, A. *Anal. Chem.* **2009**, *81*, 3079−3086.

(30) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8*, 719−726.

(31) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(32) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.

(33) Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell, M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Jiang, H.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D. *Nature* **2013**, *496*, 528−532.

(34) McDowell, E. T.; Kapteyn, J.; Schmidt, A.; Li, C.; Kang, J. H.; Descour, A.; Shi, F.; Larson, M.; Schilmiller, A.; An, L. L.; Jones, A. D.; Pichersky, E.; Soderlund, C. A.; Gang, D. R. *Plant Physiol.* **2011**, *155*, 524−539.

DOI:10.1021/acs.analchem.6b01569 Reprinted with permission. © 2016 American Chemical Society.

Ruttkies *et al. J Cheminform (2016) 8:3*
DOI 10.1186/s13321-016-0115-9

Journal of Cheminformatics
a SpringerOpen Journal

SOFTWARE
Open Access

MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation

Christoph Ruttkies[1*†], Emma L. Schymanski[2†], Sebastian Wolf[1,4], Juliane Hollender[2,3] and Steffen Neumann[1]

Abstract

Background: The *in silico* fragmenter MetFrag, launched in 2010, was one of the first approaches combining compound database searching and fragmentation prediction for small molecule identification from tandem mass spectrometry data. Since then many new approaches have evolved, as has MetFrag itself. This article details the latest developments to MetFrag and its use in small molecule identification since the original publication.

Results: MetFrag has gone through algorithmic and scoring refinements. New features include the retrieval of reference, data source and patent information via ChemSpider and PubChem web services, as well as InChIKey filtering to reduce candidate redundancy due to stereoisomerism. Candidates can be filtered or scored differently based on criteria like occurence of certain elements and/or substructures prior to fragmentation, or presence in so-called "suspect lists". Retention time information can now be calculated either within MetFrag with a sufficient amount of user-provided retention times, or incorporated separately as "user-defined scores" to be included in candidate ranking. The changes to MetFrag were evaluated on the original dataset as well as a dataset of 473 merged high resolution tandem mass spectra (HR-MS/MS) and compared with another open source *in silico* fragmenter, CFM-ID. Using HR-MS/MS information only, MetFrag2.2 and CFM-ID had 30 and 43 Top 1 ranks, respectively, using PubChem as a database. Including reference and retention information in MetFrag2.2 improved this to 420 and 336 Top 1 ranks with ChemSpider and PubChem (89 and 71 %), respectively, and even up to 343 Top 1 ranks (PubChem) when combining with CFM-ID. The optimal parameters and weights were verified using three additional datasets of 824 merged HR-MS/MS spectra in total. Further examples are given to demonstrate flexibility of the enhanced features.

Conclusions: In many cases additional information is available from the experimental context to add to small molecule identification, which is especially useful where the mass spectrum alone is not sufficient for candidate selection from a large number of candidates. The results achieved with MetFrag2.2 clearly show the benefit of considering this additional information. The new functions greatly enhance the chance of identification success and have been incorporated into a command line interface in a flexible way designed to be integrated into high throughput workflows. Feedback on the command line version of MetFrag2.2 available at http://c-ruttkies.github.io/MetFrag/ is welcome.

Keywords: Compound identification, *In silico* fragmentation, High resolution mass spectrometry, Metabolomics, Structure elucidation

Background

The identification of unknown small molecules from mass spectral data is one of the most commonly-mentioned bottlenecks in several scientific fields, including metabolomic, forensic, environmental, pharmaceutical and medical sciences. Recent developments to high resolution, accurate mass spectrometry coupled with chromatographic separation has revolutionized high-throughput analysis and opened up whole new ranges of substances that can be detected at ever decreasing detection limits. However, where "peak inventories" are reported, the vast majority of the substances or peaks detected in samples typically remain unidentified [1–3]. Although targeted analysis, where a reference standard is available, remains

---

*Correspondence: cruttkie@ipb-halle.de
†Christoph Ruttkies, Emma L. Schymanski contributed equally to this work
[1] Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany
Full list of author information is available at the end of the article

© 2016 Ruttkies et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

DOI:10.1186/s13321-016-0115-9

the best way to confirm the identification of a compound, it is no longer possible to have access to reference standards for the 100s–1000s of substances of interest in complex samples. While mass spectral libraries are growing for high accuracy tandem and MS^n spectra, the coverage is still relatively small compared with the number of compounds that could potentially be present in typical samples [4, 5]. Thus, for substances without reference standards or not present in the spectral libraries, the challenge of identification still remains. This has spurred activities in computational mass spectrometry, aimed at proposing tentative identifications for the cases where the mass spectrum is not (yet) in a mass spectral library.

The *in silico* fragmenter MetFrag, launched in 2010, was one of the first approaches to address this niche for accurate tandem mass spectra in a fast, combinatorial manner [6]. The MetFrag workflow starts by retrieving candidate structures from the compound databases PubChem [7], ChemSpider [8] or KEGG [9, 10], or accepting the upload of a structure data file (SDF) containing candidates. Candidates are then fragmented using a bond dissociation approach and these fragments are compared with the product ions in the measured mass spectrum to determine which candidates best explain the measured data. The candidate scoring is a function of the mass to charge ratio (m/z), intensity and bond dissociation energy (BDE) of the matched peaks, while a limited number of neutral loss rules (5 in total) account for rearrangements [6]. Searching PubChem, the original MetFrag (hereafter termed “MetFrag2010” for readability) achieved a median rank of 8 (with an average of 338 candidates per compound) when restricted to a Feb. 2006 version of PubChem, and 31.5 querying PubChem in 2009 (average of 2508 candidates per compound) on a 102 compound dataset from Hill et al. [11]. As PubChem is now double the size of the 2009 version, the candidate ranking becomes more challenging over time due to the increase in numbers of candidates. Thus, innovations are required to improve performance and efficiency.

Other methods for *in silico* fragmentation are also available. The commercial software Mass Frontier [12] uses rule–based fragmentation prediction based on standard reactions, a comprehensive library of over 100,000 fragmentation rules, or both. The approaches of MetFrag and Mass Frontier are complementary and have been used in combination to support structure elucidation [13, 14], but Mass Frontier does not perform candidate retrieval or scoring by itself. With increasing amounts of data available, machine learning approaches have been used to train models of the fragmentation process. Heinonen et al. [15] introduced FingerID, which uses a support vector machine to learn the mapping between the mass spectra and molecular fingerprints of the candidates. Allen et al. [16] use a stochastic, generative Markov model for the fragmentation. Implemented in CFM-ID (competitive fragment modelling), this can be used to assign fragments to spectra to rank the candidates, but also to predict spectra from structures alone. The MAGMa algorithm [17] includes information from MS^n fragmentation data, but also uses the number of references as an additional scoring term. The latest fragmenter, CSI:FingerID combines fragmentation trees and molecular fingerprinting to achieve up to 39 % Top 1 ranks, outperforming all other fragmenters [18]. The MetFusion [19] approach takes advantage of the availability of spectral data for some compounds and performs a combined query of both MetFrag and MassBank [20], such that the scores of candidates with high chemical similarity to high-scoring reference spectra are increased.

Lessons from recent critical assessment of small molecule identification contests (CASMI) [21, 22], which included many of the above-mentioned algorithms, show that the use of smaller, specific databases greatly improves the chance of obtaining the correct answer ranked highly and that the winners gathered information from many different sources, rather than relying on the *in silico* fragmentation alone. Furthermore, performing candidate selection by molecular formula can risk losing the correct candidate if the formula prediction is not certain, such that an exact mass search can be more appropriate in cases where more than one formula is possible. Despite the progress achieved for *in silico* fragmentation approaches, there are still some fundamental limitations to mass spectrometry that mean that candidate ranking cannot be solved by fragment prediction alone. For example, mass spectra that are dominated by one or only a few fragments (e.g. a water loss) that can be explained by most of the candidates simply do not contain enough information to distinguish candidates. Further examples and limitations are discussed extensively in [4].

The aim of MetFrag2.2 was to incorporate many additional features into the original MetFrag *in silico* fragmenter, considering all the information presented above. Features to explicitly include or exclude combinations of elements and substructures by either filtering or scoring were added. Suspect screening approaches, growing in popularity in environmental analysis [1], were also incorporated to allow users to screen large databases (i.e. PubChem and ChemSpider) while being able to check for candidates present in smaller, more specific databases (e.g. KEGG [9], HMDB [23], STOFF-IDENT [24], MassBank [20] or NORMAN suspects [25]), enabling users to “flag” potential structures of interest. The number of references, data sources and/or patents for a substance are now accessible via PubChem and/or ChemSpider web services, and a PubChem reference score has already

DOI:10.1186/s13321-016-0115-9

been included in the MAGMa web interface [26]. A high number of literature references or patent listings may indicate that the substance is of high use and thus more likely to be found in the environment. Similarly, a higher number of scientific articles for a metabolite could indicate that this has been observed in biological samples before. Reference information has been shown to increase identification "success" in many cases, for example [17, 27, 28], by providing additional information completely independent of the analytical evidence. However, as this information can introduce a bias towards known compounds, this information should be incorporated with caution, depending on the experimental context.

Retention time information is often used for candidate selection in LC/MS. Unlike the retention index (RI) in GC, where the Kovats RI [29] is quite widely applied, there is not yet an established RI per se for LC/MS despite a high interest. Instead, where a reverse phase column is used for the LC method, the octanol–water partitioning coefficient (log *P*) and retention times (RT) of substances can be correlated due to the column properties [30]. The log *P* of the measured standards can be predicted with various software approaches and correlated with the retention times (see e.g. [31] for an overview on different methods). This has already been used in candidate selection (e.g. [13, 32–34]), with various log *P* predictions. The orthogonal information proved useful despite the large errors associated with the predictions (e.g. over 1 log unit or up to several minutes retention time window depending on the LC run length). These are due to uncertainties in log *P* prediction that are common among different prediction implementations when considering a broad range of substances with different (and many) functional groups and ionization behaviour. As the Chemical Development Kit (CDK [35, 36]) offers log *P* calculations, this can be incorporated within MetFrag2.2. Alternative approaches with log *D*, accounting for ionization, or those requiring more extensive calculations (e.g. [37–39]) can be included via a user-defined score, described further below.

This article details the developments and improvements that have been made to MetFrag since the original publication, including a detailed evaluation on several datasets and specific examples to demonstrate the use of MetFrag2.2 in small molecule identification.

Implementation

MetFrag architecture

MetFrag2.2 is written in Java and uses the CDK [35] to read, write and process chemical structures. To start, candidates are selected from a compound database based on the neutral monoisotopic precursor mass and a given relative mass deviation (e.g. 229.1089 ± 5 ppm), the neutral molecular formula of the precursor or a set of database-dependent compound accession numbers. Currently, the online databases KEGG [9, 10], PubChem [7] or ChemSpider [8] can be used with MetFrag2.2, as well as offline databases in the form of a structure data file (SDF) or, new to MetFrag2.2, a CSV file that contains structures in the form of InChIs [40] together with their identifiers and other properties. Furthermore, MetFrag2.2 is able to query local compound database systems in MySQL or PostgreSQL, as performed in [41].

MetFrag2010 considered the ion species $[M + H]^+$, $[M]^+$, $[M]^-$ and $[M - H]^-$ during candidate retrieval and fragment generation. While the web interface contained an adduct mass adjustment feature, the presence of adducts was not considered in the fragments. MetFrag2.2 can also handle adducts also appearing in the product ions associated with $[M + Na]^+$, $[M + K]^+$, $[M + NH_4]^+$ for positive ionization and $[M + Cl]^-$, $[M + HCOO]^-$ and $[M + CH_3COO]^-$ for negative ionization. As the candidate retrieval is performed on neutral molecules, the precursor adduct type must still be known beforehand; for high-throughput workflows this information is intended to come from the workflow output.

Additive relative and absolute mass deviation values are used to perform the MS/MS peak matching and can be adjusted according to the instrument type used for MS/MS spectra acquisition. The number of fragmentation steps performed by MetFrag2.2 can be limited by setting the tree depth (default is 2).

The overall score of a given candidate is calculated as shown in Eq. 1.

$$\begin{aligned} S_{C_{\text{Final}}} = {} & \omega_{\text{Frag}} \cdot S_{C_{\text{Frag}}} + \omega_{\text{RT}} \cdot S_{C_{\text{RT}}} + \omega_{\text{Refs}} \cdot S_{C_{\text{Refs}}} \\ & + \omega_{\text{Incl}} \cdot S_{C_{\text{Incl}}} \\ & + \omega_{\text{Excl}} \cdot S_{C_{\text{Excl}}} + \omega_{\text{Suspects}} \cdot S_{C_{\text{Suspects}}} \\ & + \cdots + \omega_n \cdot S_{C_{\text{n}}} \end{aligned} \tag{1}$$

The final candidate score $S_{C_{\text{Final}}}$ is the weighted sum of all single scoring terms used, where the weights given by ω_i specify the contribution of each term. All S_C scoring terms used to calculate $S_{C_{\text{Final}}}$ are normalized to the maximum value within the candidate result list for a given MS/MS input. The calculation of individual scoring terms are detailed in the subsections below; all terms besides $S_{C_{\text{Frag}}}$ are new to MetFrag2.2.

A variety of output options are available. Output SDFs contain all compounds with a structure connection table and all additional information stored in property fields. For the CSV and XLS format, the structures are encoded by SMILES [42] and InChI codes, while an extended XLS option is available that includes images of the compounds and/or fragments. In all cases the compounds are sorted by the calculated score by default.

In silico fragmentation refinements

The *in silico* fragmentation part of MetFrag2.2 has undergone extensive algorithmic and scoring refinements. The fragmentation algorithm still uses a top-down approach, starting with an entire molecular graph and removing each bond successively. However, the generated fragments are now stored more efficiently by using only the indexes of removed bonds and atoms, similar to the MAGMa approach [43]. This not only increases processing speed and decreases memory usage, but still allows the fast calculation of the masses and molecular formulas of each fragment. This makes it possible to process MS/MS spectra with higher tree depths to generate reliable fragments for molecules with complex ring structures with lower CPU and memory requirements. As a result, fragment filters such as the molecular formula duplicate filter used in MetFrag2010 to decrease the number of generated structures were no longer required, their removal reduces the risk of missing a potentially correct fragment. The calculation of the fragmentation score, $S_{C_{\mathrm{Frag}}}$, modified from the score given in [6], is shown in Eq. 2 for a given candidate C:

$$S_{C_{\mathrm{Frag}}} = \sum_{p \in P} \frac{\mathrm{RelMass}_p{}^{\alpha} \cdot \mathrm{RelInt}_p{}^{\beta}}{\left(\sum_{b \in B_f} \mathrm{BDE_b}\right)^{\gamma}} \tag{2}$$

For each peak p matching a generated fragment, the relative mass $\mathrm{RelMass}_p$ and intensity RelInt_p as well as the sum of all cleaved bonds b of the fragment f assigned to p are considered. Where more than one fragment could be assigned to p, the fragment with the lowest denominator value is considered. In contrast to Eq. 2, the MetFrag2010 scoring used the difference between $1/max(w_c)$ and $1/max(e) \cdot e_c$, which could lead to negative scores if the BDE penalty was large. The weights α, β and γ were optimized on a smaller subset of spectra from Gerlich and Neumann [19] that was not used further in this work including merged MassBank IPB (PB) and RIKEN (PR) MS/MS spectra and were set to $\alpha = 1.84$, $\beta = 0.59$ and $\gamma = 0.47$. Once $S_{C_{\mathrm{Frag}}}$ has been calculated for all candidates within a candidate list, it is normalised so that the highest score is one.

Compound filters, element and substructure options

The *unconnected compound filter* was already implemented in MetFrag2010 to remove salts and other unconnected substances that could not possibly have the correct neutral mass from the candidate list. InChIKey filtering has now been added to reduce candidate redundancy due to stereoisomerism, as stereoisomers inflate candidate numbers but cannot (usually) be distinguished with MS/MS. The InChIKey filtering is performed using the first block, which encodes the molecular skeleton (or connectivity), but not the stereochemistry. While this is generally reasonable, some tautomers may have differing InChIKey first blocks (see e.g. [40]), such that not all tautomers will be filtered out. The highest-scoring stereoisomers overall with a matching first block are retained.

Element restrictions have been added to enhance the specificity of the exact mass search. Three options are available to restrict the elements considered: (a) include *only* the given elements, (b) the given elements have to be present, but other elements can also be present (as long as they are not explicitly excluded) and (c) exclude certain elements. Options (b) and (c) can be used in combination. These filters can be used for example to incorporate isotope information (e.g. Cl, S) that has been detected in the full scan (MS1) data.

Substructure restrictions allow the inclusion and exclusion of certain molecular substructures, encoded in SMARTS [44]. Each substructure is searched independently, thus overlapping substructures can also be considered. This option is particularly useful for cases where detailed information about a parent substance is known (e.g. transformation product, metabolite elucidation), or complementary substructure information is available from elsewhere (e.g. MS2Analyzer [45] or other MS classifiers [13]). Candidates containing certain substructures can either be included and/or excluded prior to fragmentation, or scored differently. To calculate a score, the number of matches in the inclusion or exclusion list containing n substructures are added per candidate as given in Eq. 3 (where $M_i = 1$, if substructure i matches candidate C from the given candidate list L or 0 otherwise):

$$N_{C_{\mathrm{Match}}} = \sum M_1 + M_2 + \cdots + M_n; \quad M_i \in \{0, 1\} \tag{3}$$

The inclusion ($S_{C_{\mathrm{Incl}}}$) and/or exclusion ($S_{C_{\mathrm{Excl}}}$) score(s) per candidate are then calcualted as shown in Eq. 4:

$$\begin{aligned} S_{C_{\mathrm{Incl}}} &= \frac{N_{C_{\mathrm{Match}}}}{max_{C' \in L}\left(N_{C'_{\mathrm{Match}}}\right)}; \\ S_{C_{\mathrm{Excl}}} &= \frac{n - N_{C_{\mathrm{Match}}}}{max_{C' \in L}\left(n - N_{C'_{\mathrm{Match}}}\right)} \end{aligned} \tag{4}$$

where $max_{C' \in L}(N_{C'_{\mathrm{Match}}})$ is the maximal value of $N_{C_{\mathrm{Match}}}$ within the candidate list and the scores $S_{C_{\mathrm{Incl}}}$ or $S_{C_{\mathrm{Excl}}}$ are set to 0 when $max_{C' \in L}(N_{C'_{\mathrm{Match}}}) = 0$ or $max_{C' \in L}(n - N_{C'_{\mathrm{Match}}}) = 0$, respectively.

Additional substance information

Reference and patent information

While the reference and patent information is represented by the placeholder term $\omega_{\mathrm{Refs}} \cdot S_{C_{\mathrm{Refs}}}$ in Eq. 1, the score can either be composed of several terms or added as a combined term, as described below.

DOI:10.1186/s13321-016-0115-9

If the query databases is PubChem, the number of patents (PubChemNumberPatents, PNP) and PubMed references (PubChemPubMedCount, PPC) are retrieved for each candidate via the PubChem PUG REST API [46]. These values result in the scoring terms $S_{C_{\text{PNP}}}$ and $S_{C_{\text{PPC}}}$, which can be weighted individually, or a combined term with either or both parameters. For the latter, first, a cumulative reference term is calculated as shown in Eq. 5, before the PubChem combined reference score ($S_{C_{\text{PCR}}}$) is calculated for candidate C in candidate list L as shown in Eq. 6 for PubChem:

$$N_{C_{\text{PCR}}} = a_1 \cdot \text{PNP}_C + a_2 \cdot \text{PPC}_C, \quad a_1, a_2 \in \{0, 1\} \tag{5}$$

$$S_{C_{\text{PCR}}} = \frac{N_{C_{\text{PCR}}}}{max_{C' \in L} N_{C'_{\text{PCR}}}} \tag{6}$$

For ChemSpider, five values with reference information can be retrieved using the ChemSpider web services [47]), including the number of data sources (ChemSpiderDataSourceCount, CDC), references (ChemspiderReferenceCount, CRC), PubMed references (ChemSpiderPubMedCount, CPC), Royal Society for Chemistry (RSC) references (ChemSpiderRSCCount, CRSC) and external references (ChemSpiderExternalReferenceCount, CERC). Any combination of these reference sources can be used and weighted individually, yielding the score terms $S_{C_{\text{CDC}}}$, $S_{C_{\text{CRC}}}$, $S_{C_{\text{CPC}}}$, $S_{C_{\text{CRSC}}}$ and $S_{C_{\text{CERC}}}$. Alternatively, the ChemSpider Combined Reference Scoring term ($S_{C_{\text{CCR}}}$) can be calculated, as shown below in Eqs. 7 and 8:

$$\begin{aligned} N_{C_{\text{CCR}}} = {} & b_1 \cdot \text{CRC}_C + b_2 \cdot \text{CERC}_C + b_3 \cdot \text{CRSC}_C \\ & + b_4 \cdot \text{CPC}_C + b_5 \cdot \text{CDC}_C \\ & b_1, b_2, b_3, b_4, b_5 \in \{0, 1\} \end{aligned} \tag{7}$$

$$S_{C_{\text{CCR}}} = \frac{N_{C_{\text{CCR}}}}{max_{C' \in L}\, N_{C'_{\text{CCR}}}} \tag{8}$$

The corresponding command line terms are given in the additional information (see Additional files 1, 2, 3).

Suspect lists

Additional lists of substances (so-called “suspect lists”) can be used to screen for the presence of retrieved candidates in alternative databases. The suspect lists are input as a text file containing InChIKeys (one key per line) for fast screening. The first block of the InChIKey is used to determine matches. Example files are available from [25]. This “suspect screening” can be used as an inclusion filter (include only those substances that are in the suspect list) or as an additional scoring term for the ranking of the candidates, yielding the term $\omega_{\text{Suspects}} \cdot S_{C_{\text{Suspects}}}$ given in Eq. 1.

Retention time score via log *P*

The retention time (RT) scores offered within MetFrag2.2 are based on the correlation of log *P* and user-provided RT information. The RTs must be associated with sufficient analytical standards measured under the same conditions as the unknown spectrum (a minimum of ten data points are recommended, depending on the distribution over the chromatographic run). By default, the log *P* is calculated using the XlogP algorithm in the CDK library [36, 48, 49]. Alternatively, if PubChem is used as a candidate source, the XLOGP3 value retrieved from PubChem can also be used [50]. The user-provided RTs and their associated log *P* values comprise a training dataset to generate a linear model between RT and the log *P*, shown in Eq. 9, where *a* and *b* are determined using least squares regression:

$$\log P_{\text{Unknown}} = a \cdot \text{RT}_{\text{Unknown}} + b \tag{9}$$

This equation is then used to estimate $\log P_{\text{Unknown}}$, given the measured RT associated with the unknown spectrum, and compared with $\log P_{\text{C}}$ calculated for each candidate. It is imperative that the log *P* calculated for each candidate arises from the same source as the log *P* used to build the model in Eq. 9. Lower log *P* deviations result in a higher score for a candidate; the score is calculated using density functions assuming a normal distribution with $\sigma = 1.5$ (chosen arbitrarily), as shown in Eq. 10:

$$S_{C_{\text{RT}}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(|\log P_{\text{Unknown}} - \log P_{\text{C}}|)^2/2\sigma^2} \tag{10}$$

Alternative log *P* values that are not available within MetFrag2.2 can also be used to establish a model and calculate a different $S_{C_{\text{RT}}}$ in a two-step approach. First, MetFrag2.2 can be run either with or without one of the built-in models, so that candidates and all other scores can be obtained. The InChIs or SMILES in the output CSV, or structures in the output SDF can then be used by the user to calculate their own log *P* values. These should be included in the output CSV or SDF using the “UserLogP” tag (or a self-defined alternative) and used as input for MetFrag2.2 with the Local Database option and a RT training file containing retention times and the user log *P*s with the column header matching the tag in the results file. The values *a* and *b* in Eq. 9 are then determined and used to calculate $S_{C_{\text{RT}}}$ for the final scoring. Alternative RT models that do not use log *P* should be included as a “user-defined score”, as described below.

User-defined scoring functions

The final term in Eq. 1, $\omega_n \cdot S_{C_n}$, represents the “user-defined scoring function”, which allows users to incorporate any additional information into the final candidate scoring. The MetFrag2.2 output (InChIs, SMILES, SDF) can be used to calculate additional "scores" for the candidates using external methods and these scores can be reimported with the candidates and all other MetFrag2.2 scores in the pipe-separated (|) format for final scoring. The scores and weights are matched from the column headers in the input file and the parameter names added to the score list. The commands are given in a additional table (see Additional files 1, 2, 3), with an example ("terbutylazine and isomers") below.

Results and discussion

The changes to MetFrag2.2 were evaluated on several datasets, described in the following. Further examples are given to demonstrate the use of different new features. Unless mentioned otherwise, candidate structures were retrieved from the compound databases PubChem and ChemSpider in June, 2015. If not stated explicitly, the datasets were processed with a relative and absolute fragment mass deviation of 5 ppm and 0.001 Da, respectively. The resulting ranks, if not specified explicitly, correspond to pessimistic ranks, where the worst rank is reported in the case where the correct candidate has the same score as other candidates. Stereoisomers were filtered to keep only the best scored candidate based on the comparison of the first part of the candidates' InChIKeys. The expected top ranks calculated as in Allen et al. [16], which handles ties of equally scored candidates in a uniformly random manner, are also given when comparing the two *in silico* fragmenters. This demonstrates the effect of equally scored candidates on ranking results.

The datasets from Eawag and UFZ used in this publication arose from the measurement of reference standard collections at Eawag and UFZ, which comprise small molecules of environmental relevance such as pharmaceuticals and pesticides with a wide range of physicochemical properties and functional groups, and also include several transformation products which typically have lower reference counts. All spectra are publicly available in MassBank.

In Silico fragmentation performance

Comparison with MetFrag2010

The merged spectra from 102 compounds published in Hill et al. [11], also used in [6, 19], formed the first evaluation set. The candidate sets from Gerlich and Neumann [19] were used as input for MetFrag2.2 and processed with consistent settings: relative mass deviation of 10 ppm and absolute mass deviation of 0 Da, i.e. no absolute error, for a direct comparison with MetFrag2010. With MetFrag2.2, the median rank improved from 18.5 to 14.5, while the number of correct ranked candidates in the top 1, 3 and 5 improved from 8 to 9, 20 to 24 and 28 to 34, respectively.

Baseline performance on Orbitrap XL Dataset

A set of 473 LTQ Orbitrap XL spectra resulting from 359 reference standards formed the second dataset. The spectra were measured at several collision energies with both collision-induced ionization (CID) 35 and higher-energy CID (HCD) 15, 30, 45, 60, 75 and 90 normalized units (see [51] for more details) coupled with liquid chromatography (LC) with a 25 min program on an Xbridge C18 column. The raw files were processed with RMassBank [51, 52], yielding the "EA" records in MassBank. These spectra were merged using the mzClust_hclust function in xcms [53] (parameters eppm $= 5 \times 10^{-6}$ and eabs $= 0.001$ Da) to create peaks with the mean m/z value and highest (relative) intensity and retained where they contained at least one fragment peak other than the precursor. In total 473 spectra (319 $[M + H]^+$and 154 $[M - H]^-$) were evaluated with MetFrag2010 using ChemSpider, as well as MetFrag2.2 using either PubChem or ChemSpider. The correct molecular formula was used to retrieve candidates. The results, given in Table 1, show the clear improvement between MetFrag2010 (73 Top 1 ranks with ChemSpider) and MetFrag2.2 (105 top 1 ranks with ChemSpider). This is also indicated by the higher relative ranking positions (RRP) [19] retrieved by MetFrag2.2 where a value of 1 marks the best possible result and 0 the worst possible result. Note that the version used here is 1-RRP as defined in Kerber et al. [54] and Schymanski et al. [55]. The results show that the algorithmic refinements improved the baseline *in silico* fragmentation performance, although it is difficult to tell which of the changes had the greatest influence.

Comparison with CFM-ID using Orbitrap XL Dataset

The same dataset of 473 merged spectra and the corresponding PubChem candidate sets were used as input for CFM-ID [16] version 2.0 ("Jaccard", RDKit 2015.03.1, lpsolve 5.5.2.0, Boost 1.55.0), to form a baseline comparison with an alternative *in silico* fragmenter. The results, given in Table 1, show that CFM-ID generally performed better, indicated by the higher number of correct first ranked candidates (43 vs. 30), top 5 (170 vs. 145), top 10 (232 vs. 226) and a lower median and mean rank of 11 versus 12 and 127 versus 141. The expected ranks, including equal ranked candidates, also implied a better performance of CFM-ID (top 1: 43 vs. 57, top 5: 163 vs. 193, top 10: 245 vs. 261). This was not entirely unexpected as CFM-ID uses a more sophisticated fragmentation approach, but also requires a much longer computation time. For run time analysis, 84 of the 473 queries, selected at random, were processed (single-threaded) with MetFrag2.2 and CFM-ID in parallel on a computer cluster with a maximum of 28 (virtual) computer nodes with 12 CPU cores each. The total run times (system +

DOI:10.1186/s13321-016-0115-9

Table 1 Comparison of *in silico* fragmentation results for 473 Eawag Orbitrap spectra (formula search)

	MetFrag2010	MetFrag2.2		CFM-ID	MetFrag2.2 + CFM-ID
	ChemSpider	ChemSpider	PubChem	PubChem	PubChem
Pessimistic ranks					
Median rank	8	4	12	11	8
Mean rank	74	38	141	127	85
Mean RRP	0.859	0.894	0.880	0.881	0.901
Top 1 ranks	73 (15 %)	105 (22 %)	30 (6 %)	43 (9 %)	62 (13 %)
Top 5 ranks	202	267	145	170	202
Top 10 ranks	258	320	226	232	276
Expected top ranks					
Top 1 ranks	90 (19 %)	124 (26 %)	43 (9 %)	57 (12 %)	70 (15 %)
Top 5 ranks	218	280	163	193	213
Top 10 ranks	274	329	245	261	288

MetFrag2010 and MetFrag2.2 were compared with the same ChemSpider candidate sets; MetFrag2.2 and CFM-ID with the same PubChem candidate sets. Far right: Best top 1 pessimistic ranks obtained by combining MetFrag2.2 and CFM-ID 2.0 with the weights $\omega_{\text{Frag}} = 0.67$ and $\omega_{\text{CFM-ID}} = 0.33$. The expected ranks, which partially account for equally scored candidates as calculated in [16], are shown in the lower part of the table

user runtime, retrieved by linux bash command *time*) were 75 min for MetFrag2.2 and 12,570 min (209.5 h) for CFM-ID. These values represent the runtime on a single CPU core for all 84 queries in series. The average run time per query amounts to 54 s for MetFrag2.2 and 8979 s (150 min) for CFM-ID.

As CFM-ID and MetFrag2.2 use independent *in silico* fragmentation approaches, one can hypothesize that the combination of the approaches should improve the results further. To demonstrate this, the CFM-ID results were incorporated into MetFrag2.2 by introducing an additional scoring term $\omega_{\text{CFM-ID}} \cdot S_{C_{\text{CFM-ID}}}$, where $S_{C_{\text{CFM-ID}}}$ defines the normalized CFM-ID probability of candidate C. Different contributions of each fragmenter relative to another was determined by randomly drawing 100 combinations of ω_{Frag} and $\omega_{\text{CFM-ID}}$ such that $(\omega_{\text{Frag}} + \omega_{\text{CFM-ID}} = 1)$. The best results, shown in Table 1, were obtained with $\omega_{\text{Frag}} = 0.67$ and $\omega_{\text{CFM-ID}} = 0.33$, where the change in number 1 ranks with weight is shown in Additional file 4. With this best combination, the number of Top 1 ranks improved from 30 to 61, while the median rank improved to 8. This shows that the combination of independent fragmentation methods can indeed yield valuable improvements to the results, shown again in the next paragraph after including the additional information. Further validation was beyond the scope of the current article, as further improvements could be made by retraining CFM-ID on Orbitrap data, but would be of interest in the future.

Adding retention time and reference information

Parameter selection on Orbitrap XL Dataset

The next stage was to assess the effect of references and retention time information on the MetFrag results. Firstly, each score term (i.e. fragmenter, retention time and/or reference information) was either included or excluded by setting the weight $(\omega_{\text{Frag}}, \omega_{\text{RT}}, \omega_{\text{Refs}})$ to 1 or 0, to assess the impact of the various combinations on the number of correctly-ranked number 1 substances. The results are shown in Table 2. The best result was obtained when all three “score terms” (fragmenter, RT and references) were included in candidate ranking. For PubChem, both RT/log P models (CDK XlogP and XLOGP3 from PubChem directly) were assessed and thus two sets of results are reported. The reference information was included using the combined reference scores introduced in Eqs. 6 and 8, where all combinations of the reference values described above (1–2 for PubChem, 1–5 for ChemSpider, i.e. 3 and 31 combinations in total, respectively), were used to form a cumulative total reference term, shown in Eq. 5 for PubChem and Eq. 7 for ChemSpider. The best results were achieved with PubChem when using both patents and PubMed references ($S_{C_{\text{PNP+PPC}}}$; $a_1 = 1$, $a_2 = 1$), while for ChemSpider using the ReferenceCount, ExternalReferenceCount and the DataSourceCount ($S_{C_{\text{CRC+CERC+CDC}}}$) proved best, i.e. $b_1 = 1, b_2 = 1, b_3 = 0, b_4 = 0, b_5 = 1$. Table 2 contains the number of Top 1 ranks for each combination of $\omega_{\text{Frag}}, \omega_{\text{RT}}, \omega_{\text{Refs}} \in \{0, 1\}$. The results show clearly that, while references alone result in over 311 top 1 ranks (65 % for PubChem), the addition of both fragmentation and retention time information improves the results further, to 69 % of candidates ranked first (PubChem) and even 87 % of candidates ranked first (ChemSpider). For PubChem the distribution of the number of CombinedReferences (including patents and PubMed references) for the 359 queries of the (unique) correct candidates is shown in Additional file 5.

DOI:10.1186/s13321-016-0115-9

Table 2 PubChem and ChemSpider results (number of pessimistic top 1 ranks) for 473 Eawag Orbitrap spectra

Weight term	Score term	Weights						
ω_{Frag}	$S_{C_{Frag}}$	1	1	1	0	1	0	0
ω_{RT}	$S_{C_{RT}}$	1	1	0	1	0	1	0
ω_{Refs}	$S_{C_{Refs}}$	1	0	1	1	0	0	1
Database	**RT source**	**Top 1 ranks**						
PubChem	XLOGP3	325 (69 %)	53	322	315	30	10	311
PubChem	CDK XlogP	326 (69 %)	43	322	316	30	8	311
ChemSpider	CDK XlogP	411 (87 %)	113	411	376	105	41	376

The weights indicate where the score term was included (1) or excluded (0) from the candidate ranking. For PubChem $\omega_{Refs} \cdot S_{C_{Refs}} = \omega_{Refs} \cdot (S_{C_{PNP+PPC}})$; for ChemSpider $S_{C_{Refs}} = S_{C_{CRC+CERC+CDC}}$ only. See text for explanations

Following this, the combination of each scoring term was assessed by randomly drawing 1000 different weight combinations such that $(\omega_{Frag} + \omega_{RT} + \omega_{Refs} = 1)$ to determine the optimal relative contributions of each term for the best results. This was performed for all combinations of reference sources (3 for PubChem, 31 for ChemSpider). The best result was obtained again when using both patents and PubMed references for PubChem ($S_{C_{PNP+PPC}}$; $a_1 = 1$, $a_2 = 1$), but using only the ReferenceCount ($S_{C_{CRC}}$; $b_1 = 1$, $b_2 = 0$, $b_3 = 0$, $b_4 = 0$, $b_5 = 0$) for ChemSpider. The results are summarized in Table 3 (including the weight terms) and shown in Figs. 1 and 2 for PubChem and ChemSpider respectively. These triangle plots show the top 1 candidates for all ω_i combinations, colour-coded (black—0 % of the correct candidates ranked first, yellow—10 0 % of the correct candidates ranked first) with the ω_i per category increasing in the direction of the arrow. Each corner is $\omega_i = 1$. The 25th and 75th percentiles are shown to give an idea of the distribution of the ranks. The equivalent plots for the number of top 5 and top 10 ranks are given in Additional files 6, 7, 8 and 9. Although the results from (ω_{Frag}, ω_{RT}, $\omega_{Refs} \in \{0, 1\}$) above indicated that the term $S_{C_{CRC+CERC+CDC}}$ yielded the best result for ChemSpider with 411 top 1 ranks, $S_{C_{CRC}}$ yielded 410 top 1 ranks for the same calculations, indicating that there is little difference between the two combinations. Using the randomly-drawn weights, the top 1 ranks improved to 420 (ChemSpider) and 336 (PubChem). This proves without a doubt that the addition of reference and retention time information drastically improves the performance, going from 22 to 89 % top 1 ranks (ChemSpider) and 6.3 to 71 % (PubChem).

As above, it was interesting to investigate whether the addition of a complementary fragmentation technique, i.e. CFM-ID, would improve the results even further. MetFrag2.2 and CFM-ID were combined with retention time and reference information using 1000 randomly drawn combinations of ω_{Frag}, $\omega_{CFM\text{-}ID}$, ω_{RT} and $\omega_{PNP+PPC}$ such that $(\omega_{Frag} + \omega_{CFM\text{-}ID} + \omega_{RT} + \omega_{PNP+PPC} = 1)$. The results, shown in Table 3, indicate that the PubChem results can be improved further, to 343 top 1 ranks (73 %). This is a drastic improvement from the performance of both original fragmenters alone, with CFM-ID alone yielding between 10 and 12 % top 1 hits (expected rank) in their original publication [16] with an older PubChem, the combination of both fragmenters alone yielding 15 % (expected rank) here. These combined results are also drastically better than the latest *in silico* fragmentation results just published for CSI:FingerID. Dührkop et al. [18] investigated each individual fragmenter currently available and compared the results with

Table 3 PubChem and ChemSpider results for 473 Eawag orbitrap spectra with formula retrieval, including *in silico* fragmentation, RT and reference information as shown, with the given ω_i for the highest number of Top 1 ranks

	MetFrag2.2			**MetFrag2.2 + CFM-ID**
Database	ChemSpider	PubChem	PubChem	PubChem
RT/log *P* Model	CDK XlogP	CDK XlogP	XLOGP3	CDK XlogP
ω_{Frag} ($S_{C_{Frag}}$)	0.49	0.57	0.50	0.33
ω_{RT} ($S_{C_{RT}}$)	0.19	0.02	0.16	0.03
ω_{Refs} ($S_{C_{Refs}}$)	0.32	0.41	0.34	0.35
ω_{CFMID} ($S_{C_{CFMID}}$)	–	–	–	0.29
Median rank	1	1	1	1
Mean rank	6.5	35	41	18
Mean RRP	0.990	0.977	0.977	0.978
Top 1 ranks	420 (89 %)	336 (71 %)	336 (71 %)	343 (73 %)
Top 5 ranks	447	396	398	411
Top 10 ranks	454	422	414	429

For PubChem $\omega_{Refs} \cdot S_{C_{Refs}} = \omega_{Refs} \cdot (S_{C_{PNP+PPC}})$; for ChemSpider $S_{C_{Refs}} = S_{C_{CRC}}$ only. See text for explanations. Far right: combining CFM-ID results to incorporate complementary fragmentation information

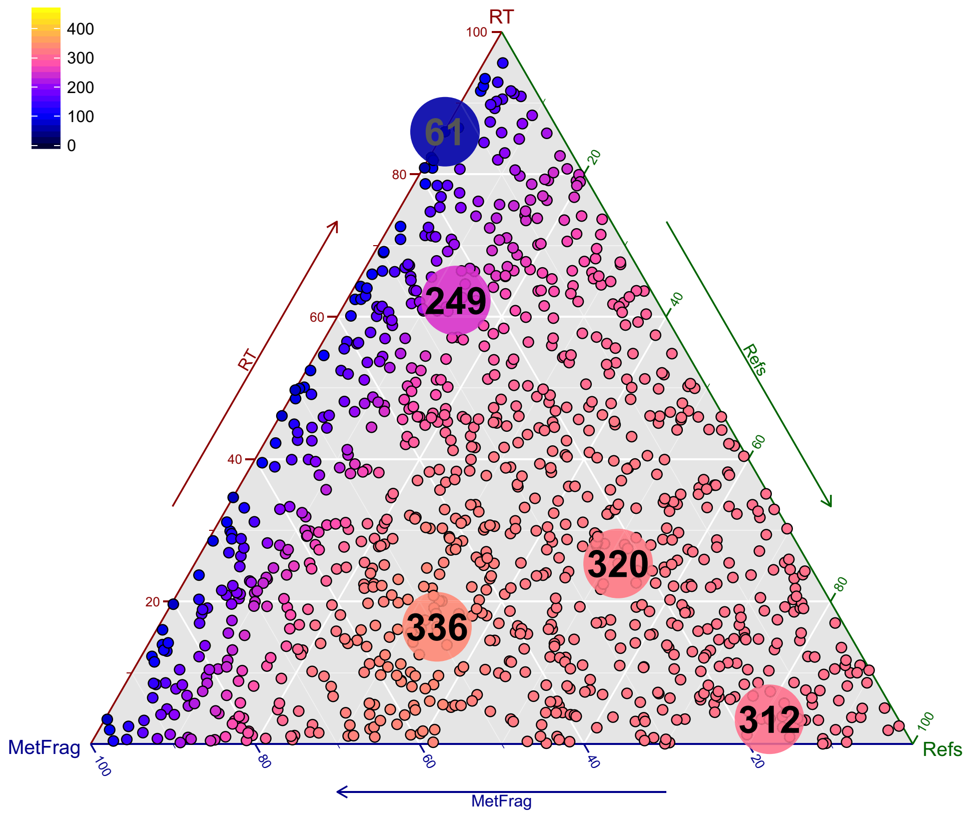


Fig. 1 Top 1 ranks with PubChem (XlogP3) on the Orbitrap XL Dataset. The results were obtained with MetFrag formula query and the inclusion of references and retention time. The reference score was calculated with the number of patents (PNP) and PubMed references (PPC). The *larger dots* show the best result (336 number 1 ranks), 75th percentile (320), median (312), 25th percentile (249) and worst result (61). For the best result, the weights were $\omega_{\mathrm{Frag}} = 0.50$, $\omega_{\mathrm{RT}} = 0.16$ and $\omega_{\mathrm{Refs}} = 0.34$

the CSI:FingerID. Despite using different data and settings to those here, their results on the Agilent dataset indicated that MetFrag2010 and CFM-ID achieved 9 and 12 % top 1 (expected) ranks, which are reasonably comparable with the results presented above. FingerID [15] achieved 19.6 %, while CSI:FingerID achieved 39 % top 1 results, which is a dramatic improvement over the other fragmenters. Since the external information boosted the top 1 ranks to 73 % for MetFrag2.2 plus CFM-ID, one could speculate that the combination of CSI:FingerID, MetFrag2.2 and CFM-ID would result in an even greater performance.

Cross-evaluation on additional datasets

As the RT and reference scores are very subjective to experimental context, MetFrag2.2 now contains so many tuneable parameters that it will be beneficial to users when a few default cases are suggested. Thus, once the optimal reference source combinations were determined as described above, alternative datasets were used to re-determine the optimal weights ω_{Frag}, ω_{RT} and ω_{Refs} to investigate the variation over different datasets. Three sufficiently large datasets available on MassBank contained good quality MS/MS and RT data, all processed with RMassBank [51].

UF dataset: A susbset of the 2758 UFZ Orbitrap XL records were acquired on an Kinetex Core-Shell C18 column from Phenomenex with a 40 min chromatographic program (all others were direct infusion experiments). These MS/MS spectra, arising from $[M + H]^+$ and $[M - H]^-$ precursors, were recorded at four collision energies: CID 35 and 55 as well as HCD 50 and 80. These spectra were merged and processed as described above for the Orbitrap XL dataset, resulting in 225 merged spectra ("UF" dataset) from 195 substances (184 $[M + H]^+$ and 41 $[M - H]^-$).

EQex and EQxPlus datasets: Two additional Eawag datasets were also available. The "EQex" dataset, measured on a Q Exactive Orbitrap, contained MS/MS spectra associated with $[M + H]^+$ and $[M - H]^-$ precursors recorded at six different collision energies (HCD 15, 30, 45, 60, 75 and 90). The "EQExPlus" dataset, measured

DOI:10.1186/s13321-016-0115-9

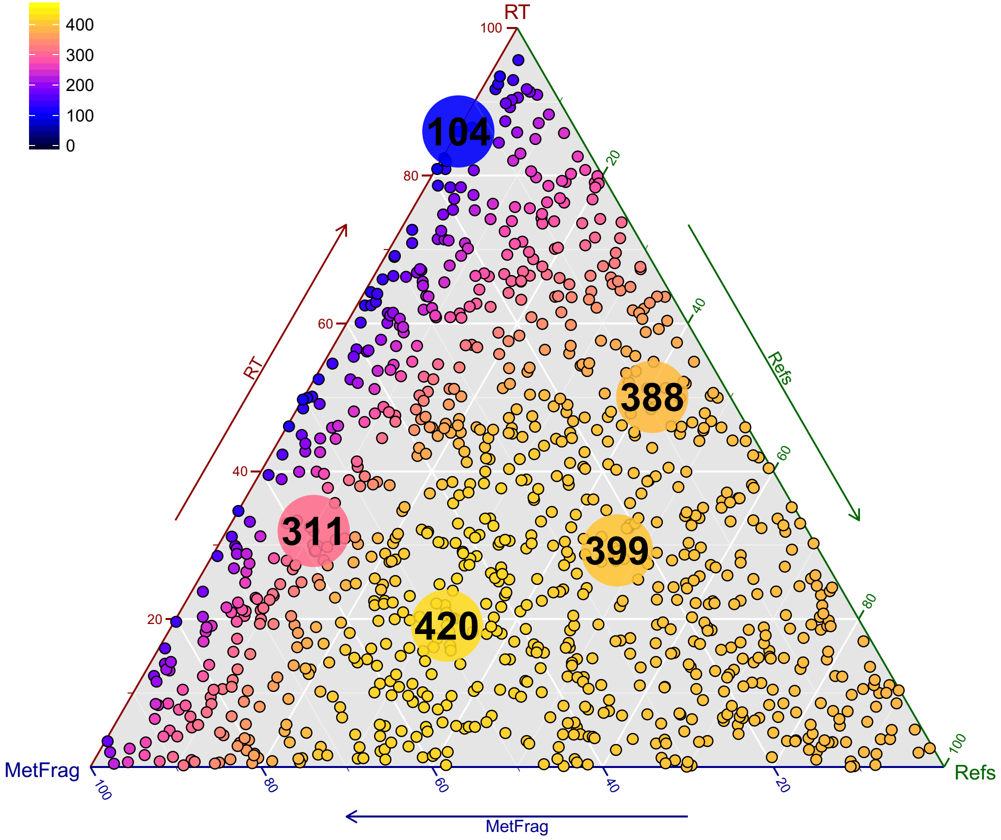


Fig. 2 Top 1 ranks with ChemSpider on the Orbitrap XL Dataset. The results were obtained with MetFrag formula query and the inclusion of references and retention time. The reference score was calculated with the ChemSpider reference count (CRC). The *larger dots* show the best result (420), 75th percentile (399), median (388), 25th percentile (311) and worst result (104). The weights for the best result were $\omega_{\text{Frag}} = 0.49, \omega_{\text{RT}} = 0.19$ and $\omega_{\text{Refs}} = 0.32$

on a Q Exactive Plus Orbitrap, contained MS/MS spectra associated with $[M + H]^+$ and $[M - H]^-$ precursors recorded at nine different collision energies (HCD 15, 30, 45, 60, 75, 90, 120, 150, 180).

Both datasets were acquired using the same LC setup as the other Eawag dataset. The MS/MS from these two datasets were merged as above to yield 294 merged spectra from 204 compounds (195 $[M + H]^+$ and 94 $[M - H]^-$) for the “EQEx” dataset and 314 merged spectra from 232 compounds (219 $[M + H]^+$ and 91 $[M - H]^-$) for the “EQExPlus” dataset. There was a very small overlap between the different Eawag datasets (5, 2 and 2 substance overlap between EA and EQEx, EA and EQExPlus and EQEx and EQExPlus, respectively).

The overlap between the UFZ and Eawag datasets was larger, with 97, 16 and 21 substances in common between the UFZ and EA, EQEx and EQExPlus datasets, respectively. The overlap was determined using the first block of the InChIKey. As the spectral and retention time data for the substances in the individual datasets were processed independently with different collision energies and ionization modes, none of the overlapping substances were removed from the datasets. The retention times extracted from the MassBank records per substance were used to establish the RT–log *P* model (see Eq. 9) for each dataset independently based on a tenfold cross-validation.

The influence of the different parameters was assessed for each dataset by setting $\omega_{\text{Frag}}, \omega_{\text{RT}}$ and ω_{Refs} to either 0 or 1 again; these results are presented in Table 4. As above, the performance improved from between 2 and 9 % of the candidates ranked first using fragmentation alone, through to 64–82 % ranked first when all ω_x were weighted equally, although the results varied quite dramatically between the datasets. The 473 spectrum dataset used above thus fell within this range.

Similarly, the optimization of $\omega_{\text{Frag}}, \omega_{\text{RT}}$ and ω_{Refs} was performed again for each dataset independently using the 1000 randomly-drawn weights. The results are presented in Table 5 and show that the percentage of top 1 ranks varies widely between the datasets, from 63 to 82 %; the

DOI:10.1186/s13321-016-0115-9

Table 4 Results (Top 1, 5 and 10 ranks) using PubChem formula queries on three additional datasets

Weight term	Score Term	Weights						
ω_{Frag}	$S_{C_{\text{Frag}}}$	1	1	1	0	1	0	0
ω_{RTs}	$S_{C_{\text{RT}}}$	1	1	0	1	0	1	0
ω_{Refs}	$S_{C_{\text{Refs}}}$	1	0	1	1	0	0	1
Dataset	**Metric**	**Ranks**						
UF (n = 225)	Top 1 ranks	164 (73 %)	9	163	159	3	2	157
UF (n = 225)	Top 5 ranks	186 (83 %)	48	189	189	36	13	199
UF (n = 225)	Top 10 ranks	191 (53 %)	77	196	192	61	25	204
EQex (n = 289)	Top 1 ranks	235 (81 %)	33	232	230	26	11	223
EQex (n = 289)	Top 5 ranks	263 (91 %)	87	260	258	88	38	276
EQex (n = 289)	Top 10 ranks	270 (93 %)	132	269	263	139	55	280
EQexPlus (n = 310)	Top 1 ranks	190 (61 %)	32	183	182	21	8	181
EQexPlus (n = 310)	Top 5 ranks	238 (77 %)	84	246	238	83	28	243
EQexPlus (n = 310)	Top 10 ranks	254 (82 %)	115	258	247	121	37	256

The weights indicate where ranking parameters were included (1) or excluded (0) from the candidate ranking. Retention time score calculation was performed using the XLOGP3 values of PubChem. $\omega_{\text{Refs}} \cdot S_{C_{\text{Refs}}} = \omega_{\text{Refs}} \cdot S_{C_{\text{PNP+PPC}}}$. See text for explanations

original dataset falls in the middle with 71 %. The results in Table 5 also show that the suggested relative weights to one another remain consistent enough to enable default parameter suggestion, with $\omega_{\text{Frag}} \approx 0.5, \omega_{\text{RT}} \approx 0.2$ and $\omega_{\text{Refs}} \approx 0.3$. All results for the number of top 1 ranks for the three additional datasets are shown in Additional files 10, 11 and 12.

Specific examples

As the additional features are more difficult to evaluate using large datasets, individual examples are presented below to demonstrate the flexibility of MetFrag2.2 command line (CL), with the corresponding commands give in a different font. Lists of the available parameters are given in Additional files 1, 2 and 3. These examples serve to show how MetFrag2.2 can also be adjusted by the user to explore individual cases in greater detail than during e.g. a high-throughput screening.

Table 5 Best Top 1 rank results on three additional datasets using PubChem formula queries including *in silico* fragmentation, RT and reference information as shown, with the given ω_i

Dataset	MetFrag2.2		
	UFZ (n = 225)	EQex (n = 289)	EQexPlus (n = 310)
ω_{Frag} ($S_{C_{\text{Frag}}}$)	0.40	0.38	0.61
ω_{RT} ($S_{C_{\text{RT}}}$)	0.23	0.27	0.11
ω_{Refs} ($S_{C_{\text{Refs}}}$)	0.37	0.35	0.28
Median rank	1	1	1
Mean rank	58.0	14.6	46.2
Mean RRP	0.972	0.981	0.976
Top 1 ranks	165 (73 %)	236 (82 %)	196 (63 %)
Top 5 ranks	188	261	233
Top 10 ranks	191	268	247

Retention time score calculation was performed using the XLOGP3 values of PubChem. $\omega_{\text{Refs}} \cdot S_{C_{\text{Refs}}} = \omega_{\text{Refs}} \cdot S_{C_{\text{PNP+PPC}}}$. See text for explanations

Gathering evidence for unknown 199.0428

During the NORMAN Collaborative Non-target Screening Trial [1], a tentatively identified non-target substance of m/z $[\text{M}-\text{H}]^-$ 199.0431 was reported by one participant as mesitylenesulfonic acid (ChemSpider ID (CSID) 69438, formula $C_9H_{12}O_3S$, neutral monoisotopic mass 200.0507) or isomer. The same unknown was detected in the same sample measured at a second institute, where the standard of mesitylenesulfonic acid was available. Although the retention time was plausible (5.96 min), comparing the MS/MS spectra clearly disproved the proposed identification, with many fragments from the unknown absent in the standard spectrum. Thus, MetFrag2.2 was used to investigate other possibilities.

Firstly, the following parameter combination was used, taking the unknown MS/MS peak list from the second participant: ChemSpider exact mass search, fragment error 0.001 Da + 5 ppm, tree depth 2, unconnected compound and InChIKey filter, filter included elements = C, S (as isotope signals were detected in the full scan), experimental RT = 6.20 min, an RT training set of 355 InChIs and RTs measured on the same system and score weights of 1 (fragmenter and RT score) and 0.25 each for four ChemSpider reference sources. This yielded 134 candidates with four different formulas ($C_9H_{12}O_3S, C_8H_{16}SSi_2, C_7H_{13}BO_2SSi, C_7H_{10}N_3O_2S$), all fulfiling the element filter (C, S). $S_{C_{\text{Final}}}$ ranged from 0.70 to 2.12, where several candidates had high numbers of references and similar number of peaks explained. Three candidates are shown in Table 6, along with a summary of the information retrieved. The clear top match, ethyl *p*-toluenesulfonate (CSID 6386, shown to the left) was unlikely to be correct, as the MS/MS contained no evidence of an ethyl loss and also had a clear fragment peak at m/z 79.9556, corresponding with an SO_3H group (thus eliminating alkyl sulfonates from consideration).

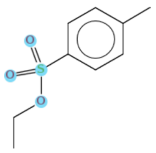
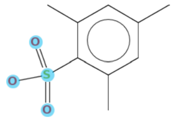
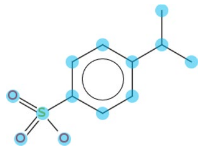
MetFrag2.2 was run again with the SMARTS substructure inclusion filter, which resulted in 31 candidates but with the same top matching structure. However, adding the SMARTS S(=O)(=O)OC to the exclusion list eliminates the alkyl sulfonate species and resulted in 18 candidates, where the top candidate was now the originally proposed (and rejected) identification mesitylenesulfonic acid, shown in the middle of Table 6. The next matches were substitution isomers. Referring to the MS/MS again, another large peak was present at m/z 183.0115, which is often observed in surfactant spectra corresponding with a *p*-ethyl benzenesulfonic acid moiety. Running MetFrag2.2 again with a substructure inclusion of CCc1ccc(cc1)S(=O)(=O)O yielded only two candidates, 4-isopropylbenzenesulfonic acid ($S_{C_{\text{Final}}} = 2.5$, CSID 6388), shown to the right in Table 6 and 4-propylbenzenesulfonic acid ($S_{C_{\text{Final}}} = 2.0$, CSID 5506213).

To check the relevance of the proposed candidates in an environmental sample, a "suspect screening" was performed. The STOFF-IDENT database [24] contains over 8000 substances including those in high volume production and use in Europe registered under the European REACH (Registration, Evaluation, Authorisation and Restriction of CHemicals) Legislation. The STOFF-IDENT contents were downloaded (February 2015) and the SMILES were converted to InChIKeys using OpenBabel and given as input to MetFrag as a suspect list. Of the 134 original candidates, only one, 4-isopropylbenzenesulfonic acid, was tagged as being present in the STOFF-IDENT database. This gives additional evidence that indeed 4-isopropylbenzenesulfonic acid is the substance behind the unknown spectrum, however it has not been possible to confirm this identification at this stage due to the lack of a sufficiently pure reference standard.

Terbutylazine and isobars

The example of terbutylazine (CSID 20848, see Table 7) shows how MetFrag2.2 can help in gathering the evidence supporting the identification of isobaric substances. Terbutylazine and secbutylazine (CSID 22172) often co-elute in generic non-target chromatographic methods and have very similar fragmentation patterns, but can usually be distinguished from the other common triazine isobars propazine (CSID 4768) and triethazine (CSID 15157) via MS/MS information. However, during the NORMAN non-target screening collaborative trial [1], all four substances were reported as potential matches for the same mass, showing clearly the danger of suspect screening based only on exact mass. For this example, the merged $[M + H]^+$ MS/MS spectrum of terbutylazine from the EA dataset above (EA02840X) was used as a peak list to run MetFrag2.2, as the correct answer is clear with a reference

Table 6 Top MetFrag2.2 candidates for unknown at *m/z* 199.0428 with different settings

CSID	6386	69438	6388
			
Original results (134 candidates)			
Rank (n = 134)	1	6	90
#Peaks explained	5	5	5
CDK log $P/S_{C_{RT}}$	1.44/0.167	1.50/0.161	2.02/0.107
$\sum S_{C_{Refs}}$	94 + 15 + 7 + 70 = 186	179 + 1 + 0 + 40 = 220	32 + 0 + 0 + 21 = 53
Substructure interpretation			
Included	S(=O)(=O)O	S(=O)(=O)O	CCc1ccc(cc1)S(=O)(=O)O
Excluded	–	S(=O)(=O)OC	–
Comment	No ethyl loss in MS/MS	Disproven via standard	Present in suspect list

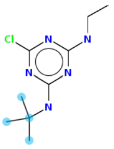
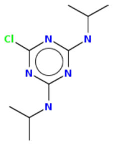
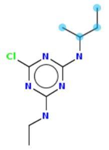
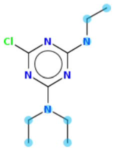
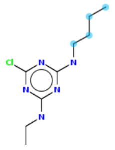
Structures overlaid with the included substructure were generated with AMBIT [57]. See text for details spectrum. Table 7 shows the data for the four substances mentioned above plus the top match based on fragmentation data alone, *N*-butyl-6-chloro-*N*′-ethyl-1,3,5-triazine-2,4-diamine (CSID 4954587, given the synonym "*n*Butylazine" hereafter to save space). ChemSpider was used to perform an exact mass search, resulting in a total of 112 structures (data from only five are shown). Five scores were used, all with weight 1: FragmenterScore, ChemSpiderReferenceCount, RetentionTimeScore, SuspectListsScore and SmartsSubstructureInclusionScore. To show the inclusion of external log *P* calculations, ChemAxon JChem for Excel [56] was used to predict log *P* and log *D* at pH 6.8 (the pH of the chromatographic program used) for a training dataset of the 810 substances in the Eawag database on MassBank. The log *P* and log *D* predictions were then performed externally for all MetFrag candidates on the dominant tautomeric species and added to the MetFrag CSV file for final scoring. The scores, shown in Table 7, showed that different candidates were the best match for different categories, indicated in italics. The candidates are ordered by the number of references. As above, STOFF-IDENT was used as a suspect list and all four of the substances reported by trial participants were indeed in STOFF-IDENT. However, Table 7 clearly shows that two can be eliminated using SC_{Frag} and substructure matches (as the MS/MS clearly displays the loss of a C_2H_5 and C_4H_9 group, indicating these are likey attached to a heteroatom, in this case N). Although secbutylazine is scored lower than terbutylazine, the reference count is the main influence here and both substances could be present in an environmental sample—depending on the context.

The large dataset evaluations show that MetFrag2.2 is suitable for high-throughput workflows, with a relatively quick runtime. On the other hand, the detailed examples shows how the various features of MetFrag2.2 can be used to investigate the top candidates in more detail and enhance the interpretation of the results, including the inclusion of external RT/log *P* and/or log *D* information that cannot be calculated within MetFrag2.2 (e.g. due to license restrictions, as in the case of ChemAxon).

Conclusions

In many cases additional information is available and needed from the experimental context to complement small molecule identification, especially where the mass spectrum alone is not sufficient for candidate

Table 7 Summary of MetFrag2.2 results for terbutylazine and four isobars

Name CSID	Terbutylazine 20848	Propazine 4768	Secbutylazine 22172	Triethazine 15157	*n*Butylazine[a] 4954587
					
SC_{Frag}	0.958	0.765	0.997	0.653	*1.0*
#Peaks explained	11/15	10/15	12/15	8/15	*12/15*
SC_{CSRefs}	*286*	204	56	45	4
ChemAxon log *P*	1.65	2.75	2.28	1.11	2.31
SC_{RT} log *P*	0.159	*0.256*	0.223	0.103	0.225
ChemAxon log *D*	1.63	2.75	2.19	0.97	2.23
SC_{RT} log *D*	0.249	0.247	*0.266*	0.192	0.266
Suspect hit	*1*	*1*	*1*	*1*	0
Substructure hits	*2*	0	*2*	1	*2*
Matches	NC(C)(C)C N[CH$_2$][CH$_3$]	–	NC(C)CC N[CH$_2$][CH$_3$]	N[CH$_2$][CH$_3$]	NCCCC N[CH$_2$][CH$_3$]
SC_{Final} (log *P*)	*4.22*	3.43	3.69	2.53	2.52
SC_{Final} (log *D*)	*4.56*	3.41	3.85	2.87	2.68
Comment	Correct substance	No longer in use	Can co-elute with 20848		

The predicted log *P* and log *D* from the retention time was 3.17 and 2.18 using a training set of 810 substances calculated externally with ChemAxon and added to MetFrag2.2 via the UserLogP option. Included substructure SMARTS were N[CH$_2$][CH$_3$], NCCCC, NC(C)CC, NC(C)(C)C

[a] Name synonym assigned for space reasons. The values in italics indicates the best result per category. Structures overlaid with the included substructure were generated with AMBIT [57]. See text for details and weights

selection from a large number of candidates. The results for MetFrag2.2 clearly show the benefit of considering this additional information, with a tenfold improvement compared with MetFrag2.2 fragmentation information alone. The flexibility of the new features in addition to the ability to add user-defined scores means that MetFrag2.2 is ideally suited to high-throughput workflows, but can also be used to perform individual elucidation efforts in greater detail. The ability to incorporate CFM-ID as an additional scoring function shows the potential to improve these results further using complementary *in silico* fragmentation approaches. The parameter files including the spectral data, the candidate, result and ranking files of the used EA, UF, EQEx, EQExPlus and HILL datasets are available at http://msbi.ipb-halle.de/download/CHIN-D-15-00088/ and can be downloaded as ZIP archives. Feedback on the command line version available at http://c-ruttkies.github.io/MetFrag/ is welcome. The new functions greatly reduce the burden on users to collect and merge ever increasing amounts of information available for substances present in different compound databases, thus enabling them to consider much more evidence during their screening efforts.

Availability and requirements

- Project name: MetFrag2.2;
- Project home page: http://c-ruttkies.github.io/MetFrag/;
- Operating system(s): Platform independent;
- Programming language: Java;
- Other requirements: Java ≥1.6, Apache Maven ≥3.0.4 (for developers);
- License: GNU LGPL version 2.1 or later;
- Any restrictions to use by non-academics: none.
-

Additional files

Additional file 1. MetFrag2.2 Command Line (CL) general parameters.

Additional file 2. MetFrag2.2 CL local database parameters (*MySQL*, *PostgresSQL*)

Additional file 3. MetFrag2.2 CL - Different Scoring terms (MetFragScoreTypes) available for online databases used by MetFrag All or a subset of these values can also be used as a total with CombinedReferenceScore (Table in Additional file 1).

Additional file 4. Top 1 ranks of MetFrag2.2. combined with CFM--ID This figure shows the distribution of the number of top 1 ranks with different weights (100 drawn randomly between 0 and 1) for MetFrag2.2 and CFM--ID. Lightestyellow dot marks the maximum, 62 top 1 ranks at $_{\text{MetFrag}} = 0.67$ and $_{\text{CFM-ID}} = 0.33$. The red dot at the right marks the minimum, 36 top 1 ranks at $_{\text{MetFrag}} = 0.997$ and $_{\text{CFM-ID}} = 0.003$. The most left dot marks 49 top 1 ranks at $_{\text{MetFrag}} = 0.02$ and $_{\text{CFM-ID}} = 0.98$.

Additional file 5. Number of patents and PubMed references shown as CombinedReferences retrieved from PubChem for the Orbitrap XL dataset This figure shows the distribution of the number of references and patents for all candidates (marked by black dots) retrieved from PubChem for the 359 (unqiue) correct candidates (marked with green line) and the additional (wrong) candidates retrieved for each query. The queries are sorted by the number of CombinedReferences for the correct candidate, respectively. The intensity of the black dots indicate the number of candidates which overlap at that position.

Additional file 6. Top 5 ranks with PubChem (XlogP3) on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (402 in the top 5), 90th percentile (386), median (375), 10th percentile (325) and worst result (145).

Additional file 7. Top 5 ranks with ChemSpider on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (463 in the top 5), 90th percentile (452), median (440), 10th percentile (385) and worst result (195).

Additional file 8. Top 10 ranks with PubChem (XlogP3) on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (422 in the top 10), 90th percentile (406), median (391), 10th percentile (351) and worst result (187).

Additional file 9. Top 10 ranks with ChemSpider on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (471 in the top 10), 90th percentile (460), median (450), 10th percentile (404) and worst result (223).

Additional file 10. Top 1 ranks with PubChem (XlogP3) on the UFZ dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (165 in the top 1), 90th percentile (159), median (156), 10th percentile (112) and worst result (11).

Additional file 11. Top 1 ranks with PubChem (XlogP3) on the EQex dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (236 in the top 1), 90th percentile (230), median (225), 10th percentile (162) and worst result (29).

Additional file 12. Top 1 ranks with PubChem (XlogP3) on the EQexPlus dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (196 in the top 1), 90th percentile (184), median (181), 10th percentile (142) and worst result (28).

Authors' contributions

SW developed the original MetFrag code under supervision of SN. CR rewrote MetFrag and performed all programming. CR and ES developed the additional features for MetFrag together and performed the data evaluation and interpretation. ES generated the test datasets and examples, CR and ES both prepared the manuscript. JH provided analytical advice on the concepts and integration of additional strategies; SN conceptualized MetFrag and provided advice on the informatics optimization. All authors read and approved the final manuscript.

Author details

[1] Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany. [2] Eawag: Swiss

DOI:10.1186/s13321-016-0115-9

Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland. [3] Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland. [4] Present Address: R&D NMR Software, Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany.

Acknowledgements
The authors thank Heinz Singer, Michael Stravs, Birgit Beck and other members of the Environmental Chemistry Department at Eawag; Tobias Schulze, Martin Krauss and others at the Helmholtz Centre for Environmental Research (UFZ) involved in the acquisition of the Eawag and UFZ datasets, as well as Jennifer Schollée for performing the ChemAxon calculations, Martin Krauss for the unknown at *m/z* 199 data and Felicity Allen for her assistance with CFM-ID. CR acknowledges funding from DFG grant NE/1396/5-1, CR and ES acknowledge funding from EU FP7 project SOLUTIONS under Grant Agreement No. 603437.

Competing interests
The authors declare that they have no competing interests. The contributions of SW were independent of his employment at Bruker.

Received: 3 October 2015 Accepted: 8 January 2016
Published online: 29 January 2016

References

1. Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M, Schulze T, Haglund P, Letzel T, Grosse S et al (2015) Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Anal Bioanal Chem 407(21):6237–6255
2. Hug C, Ulrich N, Schulze T, Brack W, Krauss M (2014) Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. Environ Pollut 184:25–32
3. Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, Ripollés Vidal C, Hollender J (2014) Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. Environ Sci Technol 48(3):1811–1818
4. Stein S (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. Anal Chem 84(17):7274–7282
5. Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O (2015) Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. Trends Anal Chem (TrAC). doi:10.1016/j.trac.2015.09.005
6. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinform 11:148
7. National Center for Biotechnology Information (2016) PubChem Database. https://pubchem.ncbi.nlm.nih.gov/search/search.cgi. Accessed 14 Jan 2016
8. Royal Society of Chemistry (2016) ChemSpider. http://www.chemspider.com/
9. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
10. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34(suppl 1):354–357
11. Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF (2008) Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. Anal Chem 80(14):5574–5582
12. HighChem Ltd. (2015) Mass Frontier v. 7. HighChem Ltd., Bratislava
13. Schymanski EL, Gallampois CMJ, Krauss M, Meringer M, Neumann S, Schulze T, Wolf S, Brack W (2012) Consensus structure elucidation combining GC/EI–MS, structure generation, and calculated properties. Anal Chem 84:3287–3295
14. Chiaia-Hernandez AC, Schymanski EL, Kumar P, Singer HP, Hollender J (2014) Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments. Anal Bioanal Chem 406(28):7323–7335
15. Heinonen M, Shen H, Zamboni N, Rousu J (2012) Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics 28(18):2333–2341
16. Allen F, Greiner R, Wishart D (2015) Competitive fragmentation modeling of ESI–MS/MS spectra for putative metabolite identification. Metabolomics 11(1):98–110. doi:10.1007/s11306-014-0676-4
17. Ridder L, van der Hooft JJJ, Verhoeven S (2014) Automatic compound annotation from mass spectrometry data using MAGMa. Mass Spectrom 3(Special Issue 2):0033. doi:10.5702/massspectrometry.S0033
18. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci 112(41):12580–12585. doi:10.1073/pnas.1509788112
19. Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. J Mass Spectrom 48(3):291–298. doi:10.1002/jms.3123
20. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45:703–714
21. Kasama T, Kinumi T, Makabe H, Matsuda F, Miura D, Miyashita M, Nakamura T, Tanaka K, Yamamoto A, Nishioka T (2014) Winners of CASMI2013: automated tools and challenge data. Mass Spectrom 3(Special_Issue_2):S0039. doi:10.5702/massspectrometry.S0039
22. Schymanski EL, Neumann S (2013) CASMI: and the winner is . . . Metabolites 3(2):412–439
23. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E et al (2013) HMDB 3.0—the human metabolome database in 2013. Nucleic Acids Res 41(Database issue):D801–D807. doi:10.1093/nar/gks1065
24. LfU: Bayerisches Landesamt für Umwelt (2016) STOFF-IDENT (login required). http://bb-x-stoffident.hswt.de/. Accessed 14 Jan 2016
25. NORMAN Association (2016) NORMAN Suspect List Exchange. http://www.norman-network.com/?q=node/236. Accessed 14 Jan 2016
26. Netherlands eScience Center (2016) MAGMa Web Interface. http://www.emetabolomics.org/magma. Accessed 14 Jan 2016
27. Little J, Cleven C, Brown S (2011) Identification of known unknown utilizing accurate mass data and chemical abstracts service databases. J Am Soc Mass Spectrom 22:348–359. doi:10.1007/s13361-010-0034-3
28. Little J, Williams A, Pshenichnov A, Tkachenko V (2012) Identification of known unknowns utilizing accurate mass data and ChemSpider. J Am Soc Mass Spectrom 23:179–185. doi:10.1007/s13361-011-0265-y
29. Kováts E (1958) Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. Helv Chim Acta 41(7):1915–1932. doi:10.1002/hlca.19580410703
30. Dunn WJ, Block JH, PR S (1986) Partition coefficient, determination and estimation. Pergamon Press, Oxford
31. Mannhold R, Poda GI, Ostermann C, Tetko IV (2009) Calculation of molecular lipophilicity: state-of-the-art and comparison of log P methods on more than 96,000 compounds. J Pharm Sci 98(3):861–893. doi:10.1002/jps.21494
32. Kern S, Fenner K, Singer HP, Schwarzenbach RP, Hollender J (2009) Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. Environmental Sci Technol 43(18):7039–7046
33. Bade R, Bijlsma L, Sancho JV, Hernández F (2015) Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. Talanta 139:143–149
34. Hogenboom A, Van Leerdam J, de Voogt P (2009) Accurate mass screening and identification of emerging contaminants in environmental samples by liquid chromatography–hybrid linear ion trap Orbitrap mass spectrometry. J Chromatogr A 1216(3):510–519
35. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bio-informatics. J Chem Inf Comput Sci 43(2):493–500

DOI:10.1186/s13321-016-0115-9

36. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bio-informatics. Curr Pharm Des 12(17):2111–2120
37. Ulrich N, Schüürmann G, Brack W (2011) Linear solvation energy relationships as classifiers in non-target analysis—a capillary liquid chromatography approach. J Chromatogr A 1218(45):8192–8196. doi:10.1016/j.chroma.2011.09.031
38. Miller TH, Musenga A, Cowan DA, Barron LP (2013) Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. Anal Chem 85(21):10330–10337. doi:10.1021/ac4024878
39. Cao M, Fraser K, Huege J, Featonby T, Rasmussen S, Jones C (2015) Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. Metabolomics 11(3):696–706. doi:10.1007/s11306-014-0727-x
40. Heller SR, McNaught A, Stein S, Tchekhovskoi D, Pletnev IV (2013) InChI—the worldwide chemical structure identifier standard. J Cheminform 5(7). doi:10.1186/1758-2946-5-7
41. Ruttkies C, Strehmel N, Scheel D, Neumann S (2015) Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an in silico generated compound database and MetFrag. Rapid Commun Mass Spectrom 29(16):1521–1529
42. Daylight Chemical Information Systems, Inc. (2016) SMILES—a simplified chemical language. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. Accessed 14 Jan 2016
43. Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RCH, van Schaik R, Vervoort J (2012) Substructure-based annotation of high-resolution multistage MSn spectral trees. Rapid Commun Mass Spectrom 26(20):2461–2471. doi:10.1002/rcm.6364
44. Daylight Chemical Information Systems, Inc. (2016) SMARTS—a language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 14 Jan 2016
45. Ma Y, Kind T, Yang D, Leon C, Fiehn O (2014) MS2Analyzer: a software for small molecule substructure annotations from accurate tandem mass spectra. Anal Chem 86(21):10724–10731
46. National Center for Biotechnology Information (2016) PubChem REST Services. https://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST_Tutorial.html. Accessed 14 Jan 2016
47. Royal Society of Chemistry (2016) ChemSpider MassSpec API. http://www.chemspider.com/MassSpecAPI.asmx. Accessed 14 Jan 2016
48. Leo AJ (1993) Calculating log Poct from structures. Chem Rev 93(4):1281–1306
49. Wang R, LL Fu Y (1997) A new atom-additive method for calculating partition coefficients. J Chem Inf Comput Sci 37(3):615–621
50. Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, Li Y, Wang R, Lai L (2007) Computation of octanol-water partition coefficients by guiding an additive model with knowledge. J Chem Inf Model 47(6):2140–2148
51. Stravs MA, Schymanski EL, Singer HP, Hollender J (2013) Automatic recalibration and processing of tandem mass spectra using formula annotation. J Mass Spectrom 48(1):89–99
52. Stravs MA, Schymanski EL (2016) RMassBank Package. http://www.bioconductor.org/packages/devel/bioc/html/RMassBank.html. Accessed 14 Jan 2016
53. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78(3):779–787. doi:10.1021/ac051437y
54. Kerber A, Meringer M, Rücker C (2006) CASE via MS: ranking structure candidates by mass spectra. Croat Chem Acta 79(3):449–464
55. Schymanski EL, Meringer M, Brack W (2009) Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? Anal Chem 81(9):3608–3617. doi:10.1021/ac802715e
56. ChemAxon (2016) JChem for Excel 15.7.2700.2799. http://www.chemaxon.com. Accessed 14 Jan 2016
57. AMBIT (2016) AMBIT Web. https://apps.ideaconsult.net/ambit2/depict. Accessed 14 Jan 2016

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com

DOI:10.1186/s13321-016-0115-9

Ruttkies *et al. BMC Bioinformatics* (2019) 20:376
https://doi.org/10.1186/s12859-019-2954-7

BMC Bioinformatics

RESEARCH ARTICLE **Open Access**

Improving MetFrag with statistical learning of fragment annotations

Christoph Ruttkies[1*], Steffen Neumann[1,2] and Stefan Posch[3]

Abstract

Background: Molecule identification is a crucial step in metabolomics and environmental sciences. Besides in silico fragmentation, as performed by MetFrag, also machine learning and statistical methods evolved, showing an improvement in molecule annotation based on MS/MS data. In this work we present a new statistical scoring method where annotations of *m/z* fragment peaks to fragment-structures are learned in a training step. Based on a Bayesian model, two additional scoring terms are integrated into the new MetFrag2.4.5 and evaluated on the test data set of the CASMI 2016 contest.

Results: The results on the 87 MS/MS spectra from positive and negative mode show a substantial improvement of the results compared to submissions made by the former MetFrag approach. Top1 rankings increased from 5 to 21 and Top10 rankings from 39 to 55 both showing higher values than for CSI:IOKR, the winner of the CASMI 2016 contest. For the negative mode spectra, MetFrag's statistical scoring outperforms all other participants which submitted results for this type of spectra.

Conclusions: This study shows how statistical learning can improve molecular structure identification based on MS/MS data compared on the same method using combinatorial in silico fragmentation only. MetFrag2.4.5 shows especially in negative mode a better performance compared to the other participating approaches.

Keywords: Mass spectrometry, Statistical modeling, Identification

Background

The identification of small molecules such as metabolites is a crucial step in metabolomics and environmental sciences. The analytical tool of choice to achieve this goal is mass spectrometry (MS) where ionized molecules can be differentiated by their mass-to-charge (*m/z*) ratio. As a single *m/z* value is not sufficient for the unequivocal determination of the molecular structure, tandem mass spectrometry (MS/MS) is applied, which results in the formation of fragment ions of the entire molecule. These fragments result in fragment peaks that are characterized by their *m/z* and intensity value. The intensity correlates with the amount of ions detected with that particular *m/z* value. These *m/z* fragment peaks can be used to infer additional hints about the underlying molecular structure.

The interpretation of the generated data is complex and usually requires expert knowledge. Over the past years, several software tools have been developed to overcome the time-consuming manual analysis of the growing amount of MS/MS spectra in an automated way. The first approaches tried to reconstruct observed fragment spectra by performing in silico fragmentation in either a rule based (e.g. MassFrontier [1]) or combinatorial manner such as MetFrag [2, 3], MIDAS [4], MS-Finder [5] and MAGMa [6].

MetFrag was one of the first combinatorial approaches developed and performs in silico fragmentation of molecular structures. Given a single MS/MS spectrum of an unknown molecule, MetFrag first selects molecular candidates from databases given the neutral mass of the parent ion. In the next step, each of the retrieved candidates is treated individually and fragmented in silico using a bond-disconnection approach. The generated fragment-structures are assigned to the *m/z* fragment peaks of the

---

*Correspondence: christoph.ruttkies@ipb-halle.de
[1] Department Biochemistry of Plant Interactions, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

DOI:10.1186/s12859-019-2954-7

MS/MS spectrum, based on the comparison of the theoretical mass of the generated structure and the *m/z* value of the acquired fragment peak. Given a set of assignments of *m/z* fragment peaks to fragment-structures for one candidate, MetFrag calculates a score that indicates how well the candidate matches the given MS/MS spectrum. These scores are used to rank all retrieved candidates. Ideally, the correct one is ranked in first place.

Statistical approaches have evolved, which are learning fragmentation processes on the basis of annotated experimental MS/MS data. CFM-ID [7] is using Markov-chains to model transitions of fragment-structures for the prediction of MS/MS spectra. Generated spectra can be aligned with the spectrum of interest and report the candidates with the best matching spectral prediction. FingerID [8] uses MS/MS spectra to predict molecular fingerprints. These Fingerprints are bit-wise representations of molecular structures where each position in the fingerprint encodes a structural property of the underlying molecule. FingerID uses support vector machines (SVM) and is enhanced by CSI:FingerID (CSI:FID) [9], integrating fragmentation trees which are calculated by SIRIUS [10]. CSI:IOKR [11] replaces the SVM prediction by an input-output kernel regression approach. Recent analysis in one of the latest CASMI (Critical Assessment of Small Molecule Identification) contests (2016) [12] reveal that techniques supported by statistical learning (i.e. CSI:FID and CSI:IOKR) are the most promising and powerful methods used to perform structure elucidation if only the MS/MS data is considered.

In this work we introduce a new statistical approach to evaluate candidates for MS/MS spectra. Using training data, probabilities of the predicted fragment-structures given the observed *m/z* peaks are estimated with a Bayesian approach. These probabilities are integrated as new scoring terms for MetFrag to rank candidates. The new scoring schema is tested on the challenge data sets of the CASMI contest 2016. The method shown here complements the different machine learning and statistical approches that perform MS/MS spectra prediction (CFM-ID), prediction of molecular fingerprints (CSI:FID, CSI:IOKR) and now combining in silico fragmentation and statistical scoring for the evaluation of retrieved molecular candidates. The new scoring functions are available with the new MetFrag version 2.4.5.

Methods

This section introduces the notation and the Bayesian model approach used to evaluate how likely a fragment-structure is in the presence of an *m/z* fragment peak. The resulting probabilities are defined across the domain of all possible fragment-structures and all *m/z* fragment peaks, but can be reduced to become tractable. The resulting probability distribution will be used in the candidate score $S^{c}_{RawPeak}$ indicating whether a candidate can explain the *m/z* fragment peaks with fragment-structures seen in the training spectra. In analogy, neutral losses will also be considered. The parameter estimation to model the probability distribution is at the heart of our approach. We describe how they are estimated from training data, taking care to clearly separate training data from evaluation data. Finally we describe the evaluation using the CASMI 2016 challenge data and comparison to the results obtained by other approaches and state-of-the art small molecule identification programs.

First, we introduce notations required for our approach. A summary of the notation used in the following and their description can be found in Additional files 4 and 5: Tables S1 and S2. Consider a set of N centroided MS/MS spectra $\underline{m} = \{\underline{m}_n | n = 1, \ldots N\}$ where $\underline{m}_n = (m_{n1}, \ldots m_{nK_n})$ consists of K_n *m/z* fragment peaks m_{nk}. Furthermore, for each spectrum $\underline{m}_n$ a set of candidates $\underline{c}_n$ of length C_n is given, typically retrieved from a database. For a given candidate $c_{nc} \in \underline{c}_n$, MetFrag performs an in silico fragmentation and assigns each observed *m/z* fragment peak m_{nk} to one of the generated fragment-structures, denoted f_{nck} in the following. This can be interpreted as explaining the *m/z* fragment peak m_{nk} with the fragment-structure f_{nkc}. On the basis of the in silico fragmentation, assignments of *m/z* fragment peaks to fragment-structures $(\underline{m}_n, \underline{f}_{nc}), c = 1, \ldots C_n$, are determined. As there is not necessarily a matching fragment-structure for every *m/z* fragment peak m_{nk}, we introduce $\perp$ in case an *m/z* fragment peak m_{nk} cannot be annotated, and denote $f_{nck} = \perp$ in this case.

As stated in the introduction, we want to evaluate candidates for an MS/MS spectrum by a statistical scoring approach to be integrated into MetFrag. Therefore, we apply a scoring term based on the probability $P(\underline{f}_{nc} | \underline{m}_n)$. The distribution $P(\underline{f} | \underline{m})$ models the occurence of fragment-structures in $\underline{f}$ in the correct candidate for a given list $\underline{m}$ of *m/z* fragment peaks in an observed spectrum. In the following we assume the independence of the assignments of *m/z* fragment peaks to fragment-structures yielding

$$P(\underline{f} | \underline{m}) = \prod_{k=1}^{K} P(f_k | m_k),$$

with $\underline{m} = (m_1,, \ldots, m_K)$ and $\underline{f} = (f_1, \ldots f_K)$. From a chemical point of view, we know that certain *m/z* fragment peaks occur concurrently with other *m/z* fragment peaks (or at least with a higher certainty) due to multi-stage fragmentation pathways that lead to a further fragmentation of a generated fragment-structure. However, for the sake of model simplification we do not consider this information when assuming independence of assignments of *m/z* fragment peaks to fragment-structures.

A fragment-structure can be regarded as a connected charged molecular structure consisting of atoms connected via bonds. A graph can be used as data structure to represent a fragment-structure, as atoms and bonds can be represented by graph nodes and edges, respectively. However, to reduce the computational costs for comparing graphs by determining graph isomorphisms, especially when working with thousands or even hundreds of thousands of fragment-structures, we use molecular fingerprints as a bit-string representation of a molecular structure. Each bit of the fingerprint describes the presence or absence of a molecular feature within the structure. As different fragment-structures may share the same fingerprint, this approach reduces the the domain size and also generalizes very similar fragment-structures that would explain the same *m/z* fragment peak. There are different molecular fingerprint functions available, e.g., the MACCSFingerPrint [13] and the LingoFingerprint [14]. A fragment-structure fingerprint is defined as $\widetilde{f}_k = MolFing(f_k)$, calculated by the fingerprint function *MolFing*.

We regard two fragment-structures f and f' to be equal, if $\widetilde{f}$ and $\widetilde{f'}$ are equal, although f and f' might be structurally different. This reduces the comparison to constant time as the fingerprint length is independent of the size of the fragment-structure. The distribution can now be re-defined as

$$P(\underline{\widetilde{f}}|\underline{m}) = \prod_{k=1}^{K} P(\widetilde{f}_k|m_k).$$

The comparison of two *m/z* fragment peaks m and m' can not be performed as a simple test for equality by $m = m'$. This is impractical for MS measurements as they show a certain degree of deviation depending on the mass accuracy of the instrument. For this reason, the m/z range covered by training and test spectra is discretized into non-equidistant bins $[\,b_i, b_{i+1}]$. The boundaries are calculated as $b_{i+1} = b_i + 2 \cdot (mzppm(b_i) + mzabs)$ with b_0 set to the minimum mass value of this range. The values $mzabs$ and $mzppm(b_i)$ represent the absolute (in *m/z*) and relative mass (in ppm) deviation given by the MS setup.

Two *m/z* fragment peaks m and m' are considered to be equal if they fall into the same bin. In the following each *m/z* fragment peak m is discretized to the central value of its bin.

Domains and Parameters

As a next step, the two domains M of *m/z* values m and F of all fragment-structure fingerprints $\widetilde{f}$ need to be defined. For M one could consider all bins resulting from discretization. However, this is impractical as the major part of this domain is not observed for a given data set. Likewise, the domain F can be defined to contain all possible fragment-structure fingerprints. Using the MACCSFingerprint with 166 bits would result in $2^{166} \approx 9.35 \cdot 10^{49}$ different fingerprints. In practice this space needs to be reduced to be tractable, and again only a fraction will be observed for a given problem. For a spectral training data set of N MS/MS spectra and C_n candidates each, we define a reduced peak domain $\widetilde{M}_{tr}$ and a reduced fingerprint domain $\widetilde{F}_{tr}$ as

$$\begin{aligned}\widetilde{M}_{tr} &= \{m_{nk}|n \in 1, \ldots N, k = 1, \ldots K_n\} \subseteq M\\ \widetilde{F}_{tr} &= \left\{\widetilde{f}_{nck}|n \in 1, \ldots N, c = 1, \ldots C_n, k = 1, \ldots K_n\right\} \subseteq F,\end{aligned}$$

which are the *m/z* fragment peaks and fragment-structure fingerprints observed in this data set.

Furthermore, we define $\mathcal{D}_{train}$ as a list of all assignments of *m/z* fragment peaks to fragment-structures in the training data, i.e.

$$\mathcal{D}_{train} = \left((m_{nk}, f_{nck})|n = 1, \ldots N, c = 1, \ldots C_n, k = 1, \ldots K_n\right).$$

Besides the MS/MS spectra given in this training data set we also need to address observations of an additional centroided MS/MS query spectrum $\underline{m}_q$ that is not part of the training data set. The processing of $\underline{m}_q$ is illustrated in Fig. 1. The domains are extended by the observations retrieved from this single query spectrum with C_q candidates and K_q *m/z* fragment peaks, i.e.

$$\begin{aligned}\widetilde{M} &= \widetilde{M}_{tr} \cup \{m_{qk}|k = 1, \ldots K_q\}\\ \widetilde{F} &= \widetilde{F}_{tr} \cup \{\widetilde{f}_{qck}|c = 1, \ldots C_q, k = 1, \ldots K_q\}.\end{aligned}$$

To define the distribution $P(\underline{\widetilde{f}}|\underline{m})$ with $m \in \widetilde{M}$ and $\widetilde{f} \in \widetilde{M}$, we introduce the notation $\theta_{m\widetilde{f}} := P(\widetilde{f}|m)$, which is the probability of fragment-structure fingerprint $\widetilde{f}$ given an observed mass m. The complete set of parameters is given as

$$\underline{\theta} = (\theta_{m\widetilde{f}}), \quad \text{for} \quad m \in \widetilde{M}, \widetilde{f} \in \widetilde{F}.$$

Parameter estimation

The parameters are initially not known and need to be estimated from the training data. In the process of parameter estimation $\underline{c}_n$ is set to only contain the known correct candidate ($C_n = 1$) for the generation of $\mathcal{D}_{train}$ as this results in mainly correct predicted fragment-structure assignments as ground truth. The generation

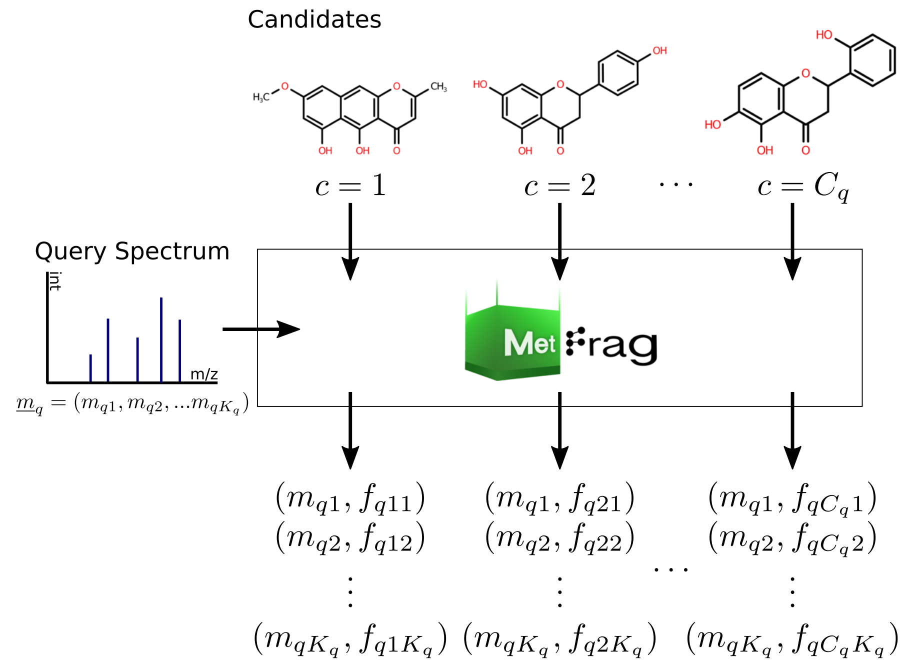


Fig. 1 MetFrag processing of a single query spectrum ($\underline{m}_q$). The input for a MetFrag processing run is a query MS/MS spectrum and the candidate list. Fragments are generated in silico for each candidate and mapped to *m/z* fragment peaks in the given spectrum. The output is a list of assignments of *m/z* fragment peaks to fragment-structures for each candidate

of $\mathcal{D}_{train}$ is illustrated in Fig. 2 where only the correct candidate for each spectrum is processed. One paradigm for parameter estimation is the maximum likelihood principle

$$\underline{\hat{\theta}}^{ML} = \underset{\underline{\theta}}{\operatorname{argmax}}\ P(\mathcal{D}_{train}|\underline{\theta}),$$

which results in

$$\hat{\theta}^{ML}_{m\tilde{f}} = \frac{N_{m\tilde{f}}}{\sum_{\tilde{f}'\in\tilde{F}} N_{m\tilde{f}'}},$$
$$\text{with}\quad N_{m\tilde{f}} = \sum_{(m_t,\tilde{f}_t)\in\mathcal{D}_{train}} \delta(\tilde{f}_t,\tilde{f})\delta(m_t,m)$$

$N_{m\tilde{f}}$ is the absolute frequency of the assignments of *m/z* fragment peaks to fragment-structures $(m,\tilde{f})$ in the training data set $\mathcal{D}_{train}$.

If such an assignment $(m,\tilde{f})$ resulting from the query spectrum is not contained in the training data, a probability $\hat{\theta}^{ML}_{m\tilde{f}}=0$ is estimated. As a consequence the probability $P(\underline{\tilde{f}}|\underline{m})$ for the query will be zero.

Due to the limitation of the available training data, this situation will arise quite often. To avoid this problem, we use the Bayes paradigm including a priori distribution for the parameters to be estimated. In addition, as we only consider the correct candidate for each spectrum in $\mathcal{D}_{train}$ it is not possible to reliably estimate parameters in case $\tilde{f}\ =\perp$, which is the probability for an *m/z* fragment peak without an assigned fragment-structure. Within the Bayesian approach we model this probability with the prior distribution and set $N_{m\perp}=0$.

In the following we will use the mean posterior (MP) principle

$$\hat{\theta}^{MP}_{m\tilde{f}} = E_{P(\underline{\theta}|\mathcal{D}_{train},\pi)}[\,\underline{\theta}]$$

where

$$P(\underline{\theta}|\mathcal{D}_{train},\pi) = \frac{P(\underline{\theta}|\underline{\pi})P(\mathcal{D}_{train}|\underline{\theta})}{P(\mathcal{D}_{train}|\pi)}$$

is the a posteriori distribution of parameters $\underline{\theta}$. As a prior distribution $P(\underline{\theta}|\underline{\pi})$ on the parameters we use a product Dirichlet distribution with hyper parameters $\pi_{m\tilde{f}}$, $m\in\tilde{M},\tilde{f}\in\tilde{F}$ defined as

$$\pi_{m\tilde{f}} = \begin{cases} \alpha, & \tilde{f}\neq\perp \\ \beta, & \tilde{f}=\perp \end{cases}$$

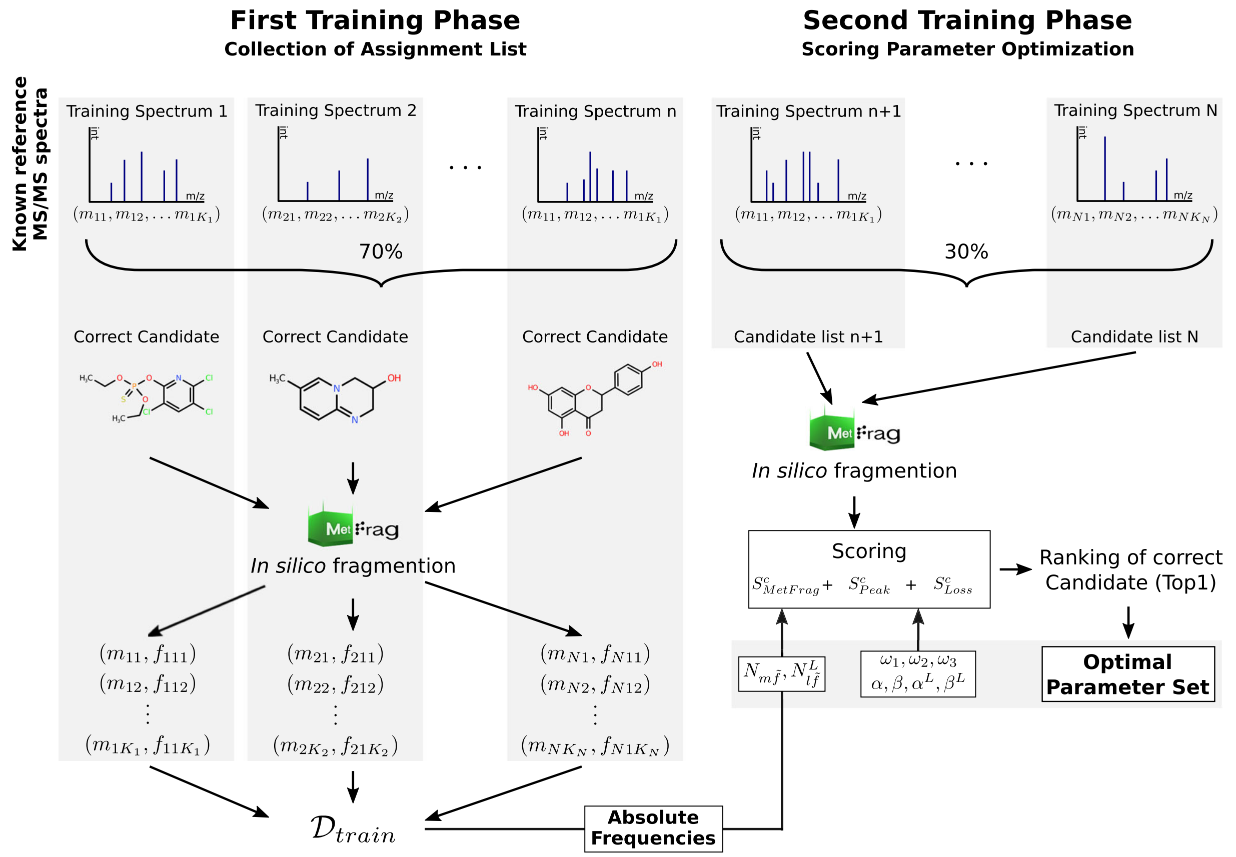


Fig. 2 The training phase. The training consists of two major phases. For each phase a subset of the known reference MS/MS spectra is used. In the first phase MetFrag generates a list of assignments of *m/z* fragment peaks to fragment-structures for the given MS/MS spectra and their correct candidates. These assignments are generated by the in silico fragmentation of the correct candidate and the mapping of the generated fragment-structures to the *m/z* fragment peaks in the training spectrum. This assignments list ($\mathcal{D}_{train}$) is used in the second training phase along with the second subset of the reference spectra. Here, for each MS/MS spectrum the correct candidate is ranked with a candidate list using the consensus candidate score integrating besides the fragmenter ($S^c_{MetFrag}$) the two new statistical scoring terms (S^c_{Peak}, S^c_{Loss}). The number of correct Top1 rankings is used to optimize pseudo count and scoring weight parameters. The first training phase is used in analogy for the generation of the list containing assignments of *m/z* fragment losses to fragment-structures ($\mathcal{D}^L_{train}$)

where α and β are also called pseudo counts.

The parameter estimation is given by

$$\hat{\theta}^{MP}_{m\tilde{f}} = \frac{N_{m\tilde{f}} + \pi_{m\tilde{f}}}{\sum_{\tilde{f}'\in\tilde{F}} \left(N_{m\tilde{f}'} + \pi_{m\tilde{f}'}\right)}.$$

Fragment losses

Fragment losses can provide additional evidence for a molecular structure as the difference between two *m/z* fragment peaks provides hints about a substructure that was lost but not observed directly by an *m/z* fragment peak (neutral loss). However, we want to include this information in the evaluation of candidates for a given MS/MS spectrum. We define l_{nkh} to be the *m/z* fragment loss between two different *m/z* fragment peaks m_{nk} and m_{nh} from the spectrum $\underline{m}_n$, where

$$l_{nkh} = m_{nk} - m_{nh}, \qquad m_{nk} > m_{nh}.$$

For each pair of assignments of *m/z* fragment peaks to fragment-structures (m_{nk}, f_{nck}) and (m_{nh}, f_{nch}) with f_{nch} being a genuine substructure of f_{nck} $(f_{nck} \neq f_{nch})$, we introduce f_{nckh} as a loss fragment-structure. This fragment-structure is a substructure of f_{nck}, that is generated if all bonds and atoms present in f_{nch} are removed $(f_{nckh} = f_{nck} \setminus f_{nch})$. If f_{nckh} is connected, we define (l_{nkh}, f_{nckh}) to be an assignment of an *m/z* fragment loss to a fragment-structure.

In analogy to the pairs of *m/z* fragment peaks and fragment-structures (m_{nk}, f_{nck}), we define the domains for

DOI:10.1186/s12859-019-2954-7

the *m/z* fragment losses and loss fragment-structures for the N MS/MS training spectra as

$$\widetilde{L}_{tr} = \{l_{nkh} | n \in 1, \ldots N, k = 1, \ldots K_n, h = 1, \ldots K_n\}$$
$$\widetilde{F}^L_{tr} = \Big\{\widetilde{f}_{nckh} | n \in 1, \ldots N, c = 1, \ldots C_n, \\ k = 1, \ldots K_n, h = 1, \ldots K_n\Big\}$$

for a given training data set

$$\mathcal{D}^L_{train} = \big((l_{nkh}, f_{nckh}) | n = 1, \ldots N, c = 1, \ldots C_n, \\ k = 1, \ \ldots K_n, h = 1, \ldots K_n\big)$$

of assignments of *m/z* fragment losses to fragment-structures.

In addition, both domains need to be extended for the additional query MS/MS spectrum $\underline{m}_q$

$$\widetilde{L} = \widetilde{L}_{tr} \cup \{l_{qkh} | k = 1, \ldots K_q, h = 1, \ldots K_q\},$$
$$\widetilde{F}^L = \widetilde{F}^L_{tr} \cup \Big\{\widetilde{f}_{qckh} | c = 1, \ldots C_q, k = 1, \ldots K_q, h = 1, \ldots K_q\Big\}.$$

We consider the distribution $P(\underline{\widetilde{f}} | \underline{l})$ for assignments of fragment-structures to *m/z* fragment losses with $l \in \widetilde{L}$ and $\widetilde{f} \in \widetilde{F}^L$, and denote $\phi^L_{l\widetilde{f}} := P(\underline{\widetilde{f}} | \underline{l})$. In analogy to the estimation of the parameters $\theta_{m\widetilde{f}}$, we can now formulate the estimation of $\phi^L_{l\widetilde{f}}$ including a Dirichlet a priori distribution with the additional hyper parameters $\psi_{l\widetilde{f}}$:

$$\psi_{l\widetilde{f}} = \begin{cases} \alpha^L, \widetilde{f} \neq \perp \\ \beta^L, \widetilde{f} = \perp \end{cases}$$

This yields the mean posterior estimates

$$\hat{\phi}^{MP}_{l\widetilde{f}} = \frac{N^L_{l\widetilde{f}} + \psi_{l\widetilde{f}}}{\sum_{f' \in \widetilde{F}^L} \left(N^L_{l\widetilde{f}'} + \psi_{l\widetilde{f}'}\right)},$$
$$\text{with} \quad N^L_{l\widetilde{f}} = \sum_{(l_t, \widetilde{f}_t) \in \mathcal{D}^L_{train}} \delta(\widetilde{f}_t, \widetilde{f}) \delta(l_t, l)$$

analogous to the parameter estimation for the assignments of *m/z* fragment peaks to fragment-structures, where $N^L_{l\widetilde{f}}$ is the absolute frequency of the *m/z* fragment loss and fragment-structure pair $(l, \widetilde{f})$ observed in the training data set $\mathcal{D}^L_{train}$.

Evaluation of the assignments of fragment-structures to *m/z* fragment peaks and losses in MetFrag candidate scoring

To evaluate a given candidate c retrieved from a compound database for an MS/MS query spectrum $\underline{m}_q$ based on the statistical models, we define a score for both the models of the assignments of *m/z* fragment peaks/losses to fragment-structures. In addition, the MetFrag fragmenter score $S^c_{MetFrag}$ as defined in [3] is also integrated in this candidate evaluation. We define the score S^c_{Fin} as the final or consensus score for a candidate c to be the weighted sum of these three scoring terms

$$S^c_{Fin} = \omega_1 \cdot S^c_{MetFrag} + \omega_2 \cdot S^c_{Peak} + \omega_3 \cdot S^c_{Loss}$$
$$\omega_i \geq 0, \sum_{i=1,2,3} \omega_i = 1.$$

To define S^c_{Peak} and S^c_{Loss}, we first introduce the raw score of a candidate as

$$S^c_{RawPeak} = \frac{1}{-\log P\left(\underline{\widetilde{f}}_{nc} | \underline{m}_n, \underline{\hat{\theta}}^{MP}\right)}$$

using the log likelihood based on the estimated parameters $\underline{\theta}^{MP}$ for the assignment of an *m/z* fragment peak to a fragment-structure $(\underline{m}_n, \underline{\widetilde{f}}_{nc})$ for candidate c. With $\underline{\widetilde{f}}_{nc} = (\widetilde{f}_{nc1}, \ldots, \widetilde{f}_{ncK_n})$ and $\underline{m}_n = (m_{n1}, \ldots, m_{nK_n})$ the log likelihood decomposes as

$$\log P\left(\underline{\widetilde{f}}_{nc} | \underline{m}_n, \underline{\hat{\theta}}^{MP}\right) = \sum_{k=1}^{K_n} \log P\left(\widetilde{f}_{nck} | m_{nk}, \underline{\hat{\theta}}^{MP}\right).$$

Furthermore, the raw score is normalized to the interval $[0, 1]$ by

$$S^c_{Peak} = \frac{S^c_{RawPeak}}{\max_{c' \in C_q} S^{c'}_{RawPeak}}.$$

Using identical ranges for the different scoring terms as for the MetFrag fragmenter score simplifies their integration into the weighted sum of the final score. The score for including the assignments of *m/z* fragment losses to fragment-structures S^c_{Loss} is defined in analogy.

Method evaluation

For the evaluation of the presented approach we used the challenge data set and evaluation procedures of the CASMI 2016 contest. In this contest candidate lists were provided by the organizers along with the spectra to be used by all participants. After the contest, several participants which used statistical learning (e.g. CSI:FID, CSI:IOKR, CFM-ID) coordinated which compounds were used in the training steps to improve the comparability between methods. They exchanged the InChIKeys (InChI: International Chemical Identifier) [15] of the spectra used in training their approaches, although it was not guaranteed that two participants used exactly the same MS/MS spectrum for a compound identified by a common InChIKey if they used different spectral databases. This evaluation is based on 87 of the 208 spectra provided originally in the challenge, as the remaining 121 spectra were removed as they were included in the training data of at least one participant. The results for this subset of the challenge spectra were published in [12] and used here in Table 2 for comparison against MetFrag2.4.5. We used the same set of InChIKeys to obtain the training spectra for this paper. The training data is available from the github repository accompanying the paper.

Preparation of the training data set

The training data set includes MS/MS spectra provided by the contest organizers consisting of 312 CASMI training spectra. Participants were allowed to use additional training spectra retrieved from spectral databases e.g. the MassBank of North America (MoNA) [16] and the Global Natural Products Social Molecular Networking (GNPS) [17] spectral library. The InChIKeys of the molecules of these additional spectra were provided by the participants.

We used the provided InChIKeys to retrieve the additional training spectra by querying the MoNA and GNPS spectral databases. For MoNA, retrieved MS/MS spectra from one institution were merged in case more than one spectrum was present for a molecule based on the first block the InChIKey. Thus for one InChIKey several merged spectra can be present in case they originate from different sources. Spectra originating from GNPS spectral database were merged independently of their source. The spectra merging was performed by averaging *m/z* fragment peaks within a specified mass range (given by MS setup of the MS/MS spectra) and retaining the peak of maximum intensity. This resulted in 5 622 spectra (4728 positive and 884 negative) which were used for training. To reduce the spectral complexity only the 40 most abundant (based on intensity) *m/z* peaks in each spectrum were used. The same applies to test spectra used for evaluation.

Training of parameters

In the training phase the optimal parameters used to calculate the candidates' consensus score need to be determined. This parameter set consists of the absolute frequencies $N_{m\tilde{f}}$ and $N^L_{l\tilde{f}}$ of the assignments of *m/z* fragment peaks and losses to fragment-structures, the hyper parameters α, β, α^L and β^L, and the score weights ω_1, ω_2 and ω_3. The whole training phase described in this paragraph is illustrated in Fig. 2.

Training was separated into two phases where in the first phase the $N_{m\tilde{f}}$ and $N^L_{l\tilde{f}}$ parameters were determined using only the correct candidate for each training spectrum. Based on these absolute frequencies the optimal hyper parameters and weight scores are determined in the second phase.

If we had used the same data set for the estimation of all parameters, $\mathcal{D}_{train}$ and $\mathcal{D}^L_{train}$ would have contained the same pairs of *m/z* fragment peaks/losses and fragment-structures for the correct candidate to be ranked in the second phase. The correct candidate would then be favoured during candidate ranking. This is not representing a realistic case when a query spectrum of an unobserved molecule is processed where we expect also *m/z* fragment peak and loss assignments not previously observed in the optimization phase.

For this reason the complete training data set was split randomly into two disjunct groups of spectra. The splitting was performed by dividing the unique list of InChIKeys (first block) with a ratio of 70:30 and collecting each corresponding spectrum to a group based on the InChIKey of the underlying molecule. The larger group is used in the first phase to calculate the $N_{m\tilde{f}}$ and $N^L_{l\tilde{f}}$.

In the first phase the correct candidate of each spectrum was processed by MetFrag's in silico fragmentation. The *m/z* fragment peaks explained by a fragment-structure were corrected to the mass of the molecular formula of the assigned fragment-structure. This is required to be independent of the different mass accuracies of MS/MS spectra acquired under different instrument conditions. Thus the list of assignments of *m/z* fragment peaks/losses to fragment-structures $\mathcal{D}_{train}$ and $\mathcal{D}^L_{train}$ contained assignments with the corrected *m/z* values used for the calculation of $N_{m\tilde{f}}$ and $N^L_{l\tilde{f}}$.

In the second training phase candidates were retrieved from a local PubChem [18] mirror (June 2016) using the monoisotopic mass of the correct candidate of each spectrum and a relative mass deviation dependent on the experimental conditions of the underlying MS measurement. To reduce runtime the correct and at most 500 randomly sampled candidates were processed from the retrieved list of candidates. The rank of the correct candidate was determined and the overall number of Top1 ranks was used as optimization criterion.

For the hyper parameters the optimization was performed by a grid search over an initial domain including a set of all combinations of the values 0.0025, 0.0005 and 0.0001 resulting in a total of $3^4 = 81$ sets of hyper parameters. If the optimal number of Top1 ranks was located at the border of this hyper parameter domain the search space was extended by increasing or decreasing the parameter by a factor of 5 or 1/5 respectively. This procedure was continued until an optimum was found with an improvement of less than 1% compared to the previous optimum of Top1 ranks. For the score weights a set of 1000 parameter combinations was sampled equally distributed on the simplex. Consensus scores and the rankings of the correct candidates were calculated for all combinations of hyper parameters and weights resulting in initially 81.000 combinations.

Subsequent to this training procedure, the absolute frequencies $N_{m\tilde{f}}$ and $N^L_{l\tilde{f}}$ were recalculated using the entire training data set to increase the observation domain of assignments of *m/z* fragment peaks/losses to fragment-structures used for the processing of the challenge data set.

Fingerprint function

To investigate the effect of the fingerprint function *MolFing* on the results, the complete training phase was performed four times with different fingerprint functions for the same training spectra. For comparison the Lingo- [14], the MACCS- [13], the Circular- [19], and the GraphOnlyFingerprint were used. For calculation of the different fingerprints CDK (version 2.1) [20] implementations were used. The fingerprint with the best training result was selected for the processing of the challenge data set.

Processing of the CASMI challenge data set

After the training phase and the selection of the fingerprint function, the in silico fragmentation and scoring was performed for the 87 challenge spectra using the provided candidate lists. Candidates that included non-connected substructures or non-natural isotopes (like deuterium) were discarded from the candidate lists. The candidate ranking was performed after the removal of multiple stereoisomers in compliance with the contest rules and evaluation. Stereoisomers were detected based on the first block of the candidates' InChIKey representing the molecular skeleton and only the best scoring stereoisomer was regarded for candidate ranking. The results were evaluated and compared on the basis of the average Top1, Top3, and Top10 rankings and the median and mean average rankings of the correct candidate as in [12].

Stability of parameter optima and ranking results

Splitting of the training data set for the two phases was performed randomly. As the resulting parameters depend on the splitting, we performed ten independent trials with different splits of the training data. The resulting parameters and their performance on the challenge data set were reported to investigate the effect of randomization.

Results

Comparison of different fingerprint functions

The ranking results obtained in the training phase on the basis of the different fingerprint functions (*MolFing*) are shown in Fig. 3. The fingerprints used are the Lingo-, MACCS-, Circular-, and GraphOnlyFingerprint. The training results are based on the spectra processed in the second phase during training consisting of 1389 to 1471 spectra in positive and 255 to 279 spectra in negative mode depending on the run and the spectra splitting.

Comparable results are obtained with the Circular- and LingoFingerprint across both ion modes and across the different rankings as shown in Fig. 3 by the similar curve for the Top1, Top3 and Top10 rankings. Similar means of the rankings across the ten runs confirm this observation with 402.3, 639.8, and 881.2 for the mean Top1, Top3 and Top10 rankings using the Circular- and 398.4, 640.0 and 881.9 using the LingoFingerprint. These two fingerprint functions show superior results for the Top1 rankings compared to MACCS with 371.0 and GraphOnly 328.6. For Top3 and Top10 rankings and positive mode the MACCSFingerprint gives comparable results. Top3 and Top10 rankings in negative mode are comparable for all fingerprint functions.

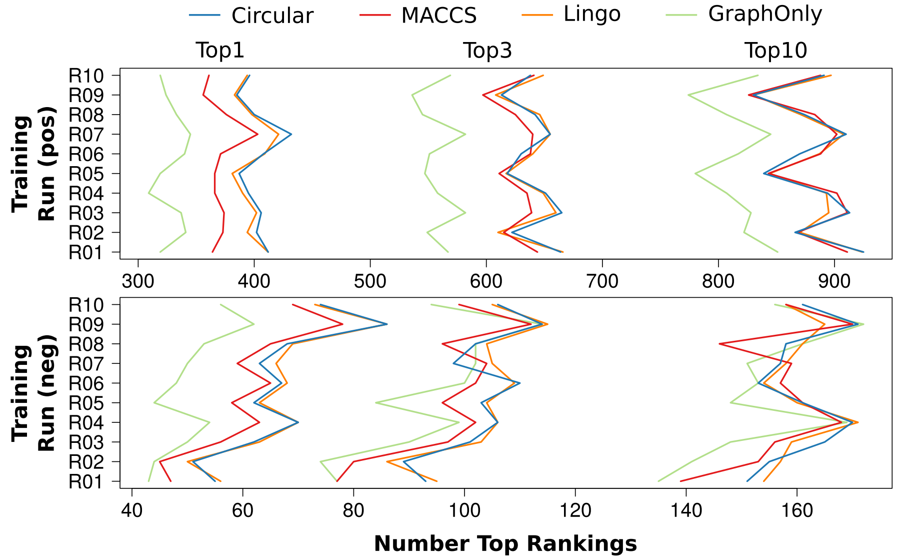


Fig. 3 Top rankings of training results. The Top rankings (Top1, Top3, Top10) of the ten training runs are shown for the different fingerprint function. The results are based on the rankings of the correct candidates of the training data used in the second training phase consisting of 1389 to 1471 spectra in positive mode (top) and 255 to 279 spectra in negative mode (bottom)

The CircularFingerprint shows with the runs R07 in positive and R09 in negative mode the overall highest number of Top1 rankings with 518 of the 1686 training spectra. Due to this performance the CircularFingerprint is used for subsequent investigations and the evaluation of the challenge data set.

Randomization of training data sets

In this section we evaluate the impact of the randomization of the training data on parameter optimization. Table 1 shows the optimal parameter sets and the performance achieved on the training data using the CircularFingerprint. The overall ranking results vary across the ten runs for the Top1, Top3 and Top10 numbers in both positive and negative ion mode as expected. Boxplots of the parameter sets are shown in Fig. 4. The variation of optimal hyper parameters as well as weights shows a similar pattern for both positive and negative ion mode where a larger variation can be observed in negative mode. Particularly the pseudo counts for annotated *m/z* fragment peaks show a broader variation with 5e-04 to 2e-05 (α) and 1e-03 to 2e-05 (α^L) compared to positive mode with 1e-04 as optimum for α and an interval of 2e-03 to 1e-04 for α^L.

The largest of the weights combining the three scores is ω_2 which gives the score S^c_{Peak} the largest influence in the overall assessment. The median of ω_2 is 0.4855 in positive and 0.4935 in negative mode. The impact of the original MetFrag score $S^c_{MetFrag}$ and S^c_{Loss} are distinctively lower and comparable to each other. The weight ω_1 for the MetFrag score has a median of 0.2875 in positive and 0.2840 in negative mode. The weights for ω_3 are 0.2355 respectively 0.2045.

In the following we analyze the robustness and the homogeneity of the results on the challenge data set with regard to varying parameters across the parameter space evaluted during optimization. This also helped to obtain a better explanation on the deviation of optimized parameters. Specifically we compare the distribution of the Top1 rankings considering (i) the ten optimal parameter sets from the ten randomizations, (ii) the parameter sets within the convex hull constituted by these ten optimal

Table 1 Ranking results in the training phase based on the CircularFingerprint

Top1	Top3	Top10	Top1 (%)	α	β	α^L	β^L	ω_1	ω_2	ω_3	# Spectra
Negative Mode											
55	93	151	20.8	0.00002	0.00250	0.00050	0.00050	0.268	0.460	0.272	265
51	89	155	19.5	0.00002	0.06250	0.01250	0.00050	0.434	0.380	0.186	261
62	101	165	22.9	0.00050	0.01250	0.00010	0.01250	0.309	0.508	0.184	271
70	106	170	25.8	0.00050	0.00250	0.00002	0.01250	0.317	0.494	0.189	271
62	103	161	23.8	0.00010	0.00010	0.00010	0.00250	0.170	0.616	0.214	260
67	110	153	24.0	0.00010	0.00250	0.00250	0.00010	0.300	0.493	0.207	279
63	98	157	22.9	0.00010	0.00050	0.00010	0.00050	0.054	0.512	0.434	275
68	102	158	25.0	0.00002	0.00250	0.00250	0.00250	0.240	0.558	0.202	272
86	114	171	31.2*	0.00010	0.00250	0.00250	0.00010	0.413	0.398	0.189	276
74	106	161	29.0	0.00010	0.00010	0.00002	0.00010	0.189	0.465	0.346	255
Positive Mode											
412	664	925	28.0	0.00010	0.00250	0.00010	0.00250	0.333	0.438	0.229	1471
402	622	866	28.2	0.00010	0.00050	0.00010	0.00250	0.208	0.483	0.309	1426
406	665	913	29.0	0.00010	0.01250	0.00250	0.00250	0.333	0.438	0.229	1399
395	651	894	27.6	0.00010	0.00250	0.00250	0.00250	0.309	0.503	0.188	1432
387	618	839	27.4	0.00010	0.00250	0.00050	0.00050	0.413	0.398	0.189	1413
408	630	870	28.6	0.00010	0.00050	0.00050	0.00050	0.165	0.584	0.251	1428
432	655	910	30.6*	0.00010	0.01250	0.00250	0.00050	0.378	0.488	0.134	1410
400	642	874	28.2	0.00010	0.00250	0.00250	0.00050	0.210	0.488	0.302	1420
385	613	830	27.7	0.00010	0.00250	0.00010	0.00010	0.266	0.388	0.346	1389
396	638	891	27.7	0.00010	0.00050	0.00050	0.00010	0.165	0.593	0.242	1428

The optimization of the parameters was performed on the training data set with ten different random splits of the MS/MS training spectra to be used for first and second training phase. Optimization was performed separately for positive and negative mode. *Runs with the best results based on the relative correct Top1 rankings (neg: R09, pos: R07)

DOI:10.1186/s12859-019-2954-7

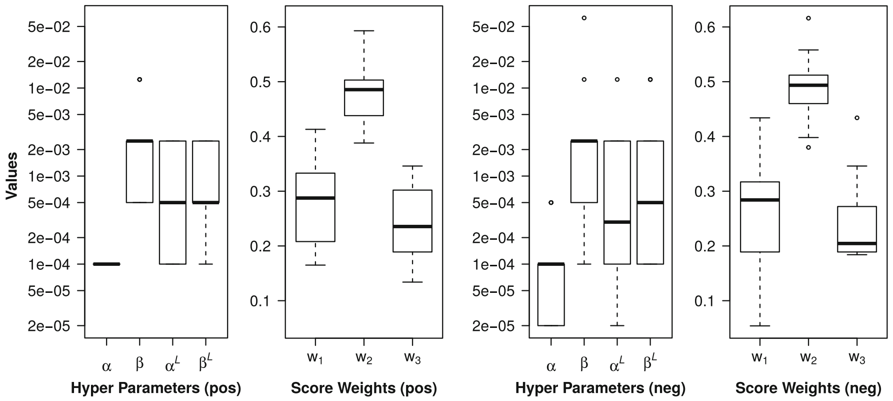


Fig. 4 Boxplots of optimal weight and hyper parameters retrieved in the training phase. The parameters were obtained from the ten training runs with randomized splits of the training set and the CircularFingerprint. The rankings results show the optimal weight and hyper parameters for positive and negative mode

parameter sets in the six dimensional parameter space, and (iii) the complete parameter space evaluated during training of the parameters. The convex hull over the ten optimal parameter sets was calculated using the six degrees of freedom (α, β, α^L, β^L, ω_1, ω_2) from the seven parameters with the Python *Numpy* package.

Figure 5 shows in yellow the distribution of the Top1 rankings of the CASMI challenge data set for the complete parameter space. Top1 ranking vary from 1 to 12 for the positive and from 4 to 14 for the negative challenge spectra, where the maximum of the distributions are six and ten for positive and negative mode, respectively. If parameter sets are restricted to the convex hull the distribution is clearly shifted to better performance, where Top1 rankings vary between 8 to 11 for positive and 10 to 13 for negative mode. This range of Top1 rankings is almost identical to the one resulting from the ten optimal parameter sets. The only exception are nine Top1 rankings for parameter sets within the convex hull in negative mode. In positive mode about 76% of the investigated parameters show worse results than achieved by the parameters contained in the convex hull. For negative mode this proportion is reduced to around 15% which can again be explained by the smaller number of available training data.

For the subsequent comparison to other methods on the challenge data set we use the parameter sets resulting in the best relative Top1 ranking performance in the training

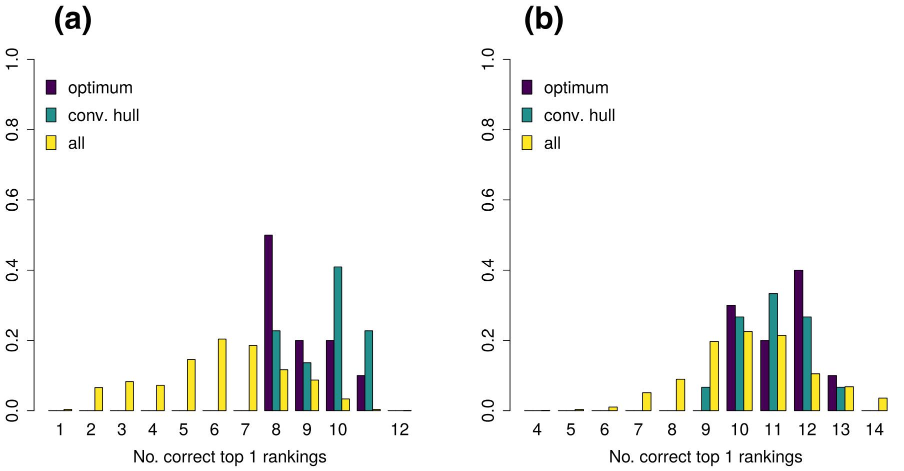


Fig. 5 Distribution of Top1 rankings on the challenge data set. The collection of barcharts show the Top1 rankings retrieved using the CircularFingerprint for selected parameter sets. Yellow bars show the normalized Top1 counts for all parameter sets used in the training phase. The green bars show the normalized rankings for all parameter sets within the convex hull spanned by the ten optimal parameter sets retrieved from the ten randomized training runs. The violet bars show the normalized counts from these optimal parameter sets. **a** Positive mode **b** Negative mode

DOI:10.1186/s12859-019-2954-7

phase. The corresponding runs are highlighted in Table 1 and are R07 for positive and R09 in negative mode.

Comparison with MetFrag2.3

The main goal of the integration of the proposed approach into MetFrag was to improve the candidate ranking augmenting the fragmenter score with statistical scores. The MetFrag versions 2.3 and 2.4.5 use exactly the same in silico fragmentation approach. MetFrag2.4.5 scoring was extended with the statistical scoring terms which make the difference in the comparison of both version. The results of MetFrag version 2.4.5 show a drastic improvement of the rankings for the CASMI challenge data compared to its older version 2.3 with regard to all performance measures as given in the first two columns of Table 2. The correct Top1 rankings show a more than four fold increase from 5 to 21 Top1 rankings. The improvement is especially distinct for positive mode with 9 Top1 rankings where MetFrag2.3 resulted in one single query correctly ranked at first position. The number of Top1 hits in negative mode is also increased three fold from 4 to 12. The improvement is also illustrated by the reduced mean and median ranks. Where the mean rank halved to 34.6 the median rank was even reduced by two third to 5. All three scores contribute substantially to these improvements and Top1 rankings vary smoothly with the weight scores (see Additional file 1: Figure S1).

Comparison with other CASMI participants

The MetFrag2.4.5 results were compared to the results obtained by all other participants of CASMI 2016, i.e., CFM_retrain, CSI_IOKR_AR, and CSI:FID_leaveout (abbreviated by CFM-ID, CSI:IOKR, and CSI:FID), MS-Finder and MAGMa. Table 2 shows the original data from Table 7 of [12] with the ranking results for the 87 Challenge MS/MS spectra. The additional MetFrag2.4.5 column summarizes the results achieved using the new MetFrag statistical scoring terms.

In positive mode, MetFrag2.4.5 obtains nine Top1 rankings and shows a similar performace as CFM-ID (9) and CSI:IOKR (10). CSI:FID (13) outperforms all other approaches with regard to Top1 rankings in positive mode, however did not submit results for negative mode spectra. Figure 6b shows the overlap of the Top1 ranked challenges in positive mode for MetFrag2.4.5 and CSI:FID. There are only five challenges ranked first by both tools and thus a large degree of divergence between the correct predictions.

For the negative mode spectra MetFrag2.4.5 considerably outperformed all participants with 12 Top1 rankings. These are five more queries than MS–Finder could rank in first position and even twice as many than the other statistical approaches CFM-ID and CSI:IOKR.

Considering the complete test data set MetFrag2.4.5 outperforms all participants with regard to Top1, Top3, and Top10 rankings including the declared winner of the contest CSI:IOKR (Top1: 21, Top3: 38, Top10: 55 vs. Top1: 16, Top3: 26, Top10: 46). The improved results are also confirmed by the smaller median and mean rankings of 5 and 34.6 compared to 10 and 97.9. We note that considering the median, CSI:FID shows a better performance than MetFrag2.4.5, however did only submit results for positive mode.

Figure 6a shows the overlap of correctly identified Top1 challenges of the participants which use statistical approaches. Interestingly, there is a relatively large number of challenges that are identified by only one of the approaches. With 10 challenges MetFrag2.4.5 shows the highest amount of unique queries ranked correctly in first place, which is predominantly caused by the eight Top1 negative mode challenges.

Discussion

The results obtained by the combination of MetFrag's in silico fragmentation approach and statistical fragment annotation learning have shown an overall improvement of the ranking results of the relevant CASMI 2016 test set. Different fingerprint functions have been tested to avoid the expensive graph isomorphism problem to find matching fragments. The training phase revealed a dependency

Table 2 Results for the 87 MS/MS test spectra from the CASMI 2016 Challenge taken from Table 7 in [12] augmented with the results of the proposed approach (MetFrag 2.4.5). For the participants of the challenge the best result is given

	MetFrag 2.4.5	MetFrag 2.3	CFM-ID	CSI:IOKR	CSI:FID	MS-Finder	MAGMa
Top 1 Pos.	9	1	9	10	13	3	2
Top 1 Neg.	12	4	6	6	–*	7	4
Top 1	21	5	15	16	13*	10	6
Top 3	38	16	24	26	23*	25	16
Top 10	55	39	40	46	32*	38	35
Mean rank	34.6	68.4	64.1	97.9	41.5*	28.7	76.8
Med. rank	5	14.5	12.5	10	3*	17.5	23.5

*CSI:FID did not submit results for negative mode spectra

DOI:10.1186/s12859-019-2954-7

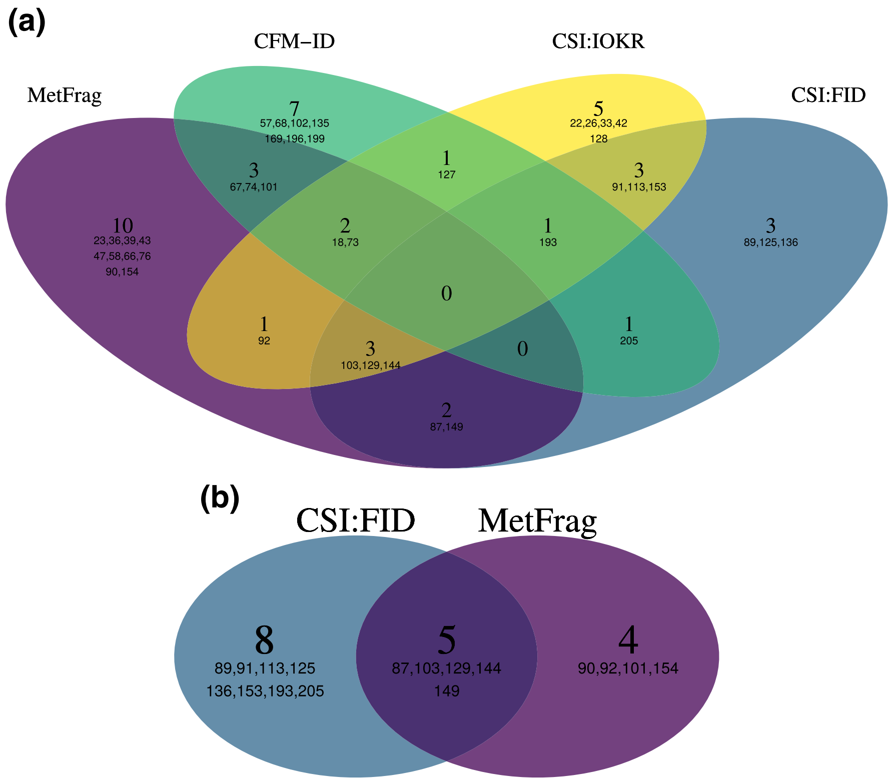


Fig. 6 Overlap of the correctly identified Top1 spectra of the challenge data set for selected participants. The Venn diagram (**a**) includes the four tools using statistical approaches (MetFrag2.4.5, CFM-ID, CSI:IOKR, CSI:FID) and shows the overlap of correcly identified challenges out of the 87 spectra (positive and negative mode). The diagram (**b**) shows the overlap of CSI:FID and MetFrag2.4.5 for the positive mode challenges. The large numbers indicate the amount of common challenges and the numbers listed underneath their challenge IDs

between the number of correct top hits and the fingerprint used. While MACCS- and especially Lingo- and the CircularFingerprint showed the best and also comparable results, the GraphOnlyFingerprint showed a significantly lower number of correct top rankings on the training set. We attribute the inferior performance of the GraphOnlyFingerprint primarily to the lack of representing bond orders and hence encoding less chemical information than all other fingerprint types evaluated. Due to the best performance in the training phase the CircularFingerprint was selected for further investigation on the test set.

Ten different hyper and weight parameter sets resulting from optimization with ten randomized splits of the training data were used to investigate the robustness and the distribution of these parameters accross the different training sets. While the optima of the seven parameters varied slightly between the different splits, the parameter sets still showed a clear trend across all ten runs. Especially the effect of the S^{c}_{Peak} score weight ω_2 was predominantly higher compared to ω_1 and ω_3 for both positive and negative ion mode. The assumption that the observed parameter variation is an indication for a relatively broad and homogenious parameter optimum was confirmed by the investigation of the ranking results retrieved using parameters located in the convex hull spanned by the ten optima. These distributions also indicate a high robustness of the performance with varying parameter sets across these parameter optima.

An important outcome of this study is the significant improvement of the ranking results retrieved adding the presented Bayesian approach to MetFrag's native in silico fragment annotation. While the improvement gain for the Top3 and Top10 rankings are less pronounced, this comparison impressively demonstrates the benefit including statistical approaches for MS based compound identification. This corresponds to the outcome of CASMI 2016 where a comparison of different statistical and non-statistical approaches was made [12].

The proposed Bayesian approach follows a different mechanism than the existing statistical compound identification methods predicting molecular fingerprints

DOI:10.1186/s12859-019-2954-7

(CSI:FingerID, CSI:IOKR) or MS/MS spectra (CFM-ID). The comparison of the different approaches on the CASMI 2016 test set used in this study shows on the one hand that the presented approach compares well to the existing ones and on the other hand that a relatively large number of challenges are identified by only one of the approaches (Fig. 6a). From the latter finding it may be concluded that there are different preferences for certain types of spectra of the CASMI 2016 contest. The comparison also revealed that for MetFrag2.4.5 the performance is comparable between positive and negative mode (9 vs. 12). CSI:IOKR shows lower performance ranking result for the negative mode spectra compared to positive mode (6 vs. 10). We assume the combination of in silico fragmentation and statistical scoring has a positive effect in case only limited training data is available. Only a small fraction of negative mode training data was available for this contest and resulted in generally worse results of the statistical approaches in negative mode.

Conclusions

In this work new statistical scoring terms are introduced to MetFrag. This model assesses the assignments of *m/z* fragment peaks/losses to fragment-structures derived from in silico fragmentation of a candidate and assumes independence of the individual assignments. The model parameters are estimated using the mean posterior approach. Hyper parameters of the statistical model as well as score weights are optimized by a grid search. The performance is evalutated on a subset of the CASMI 2016 contest challenge spectra for which the spectrum was not among the training data set of any participant. The results show that with the integration of the two new statistical scoring terms MetFrag could be improved four fold regarding the number of Top1 rankings. In addition it showed a better performance than the declared winner of the contest CSI:IOKR regarding the number of correctly ranked Top1, Top3 and Top10 candidates. The new scoring terms are now available in the command line tool (version 2.4.5) as AutomatedPeakFingerprintAnnotationScore and AutomatedLossFingerprintAnnotationScore and also in the web interface (https://msbi.ipb-halle.de/MetFrag) as "Statistical Scoring" trained on extended data set than used in this work. The additional scoring terms complement current scoring terms based on experimental data and can also be combined with additional meta information if available as described in [3].

We also want to stress that once the method is trained on spectra in the training phase, it can be applied and used for annotation on any data set. The data set can vary whereas the training data set is fixed once the method was trained, which is similar to all other machine learning and statistical methods mentioned in this work.

Additional files

Additional file 1: Figure S1 - Weight Parameter Scan for the test dataset. (PDF 767 kb)

Additional file 2: Figure S2 - Maximum spectral similarities. (PDF 196 kb)

Additional file 3: Figure S3 - Rankings of the correct candidates (test) vs. max. spectral similarity. (PDF 204 kb)

Additional file 4: Table S1 - Notation summary. (PDF 109 kb)

Additional file 5: Table S2 - Notation summary (Scores). (PDF 70.4 kb)

Abbreviations

CASMI: Critical assessment of small molecule identification; CSI:FID: CSI:fingerID; InChI: International chemical identifier; MP: Mean posterior; MS/MS: Tandem mass spectrometry; m/z: Mass-to-charge ratio; mzabs: Absolute mass deviation; mzppm: Relative mass deviation; SVM: Support vector machines

Acknowledgements

We thank all CASMI 2016 participants for generating and providing all result sets of their used software and methods. We acknowledge Emma Schymanski (Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg) for valuable discussions and proof-reading the manuscript. CR and SN acknowledge support from the Leibniz Association's Open Access Publishing Fund.

Authors' contributions

SP, SN, CR contributed to method development, manuscript preparation and revision, discussion. CR implemented all neccessary changes to MetFrag and performed data analysis to generate presented results. All authors read and approved the final version of the manuscript.

Funding

CR acknowledge funding from the European Commission for the FP7 project SOLUTIONS under Grant Agreement No. 603437 and for the H2020 project PhenoMeNal under Grant Agreement No. 654241. Funding bodies played no role in study design, data analysis and interpretation, nor manuscript development.

Availability of data and materials

The *m/z* peak and candidate lists used in this study is available on the official CASMI website, http://www.casmi-contest.org/2016/index.shtml. A complete list of the used MassBank and GNPS training spectra and the ranking data sets generated during the current study are available on GitHub, https://github.com/c-ruttkies/metfrag_statistical_annotation. Further information on how to use the new scoring terms with the commandline version of MetFrag can be found on the project website http://ipb-halle.github.io/MetFrag/projects/metfragcl. The source code is published on GitHub (https://github.com/ipb-halle/MetFragRelaunched (branch: feature/statistical_scoring)).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

SN is Associate Editor for BMC Bioinformatics.

Author details

[1]Department Biochemistry of Plant Interactions, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany. [2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. [3]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06099 Halle (Saale), Germany.

Received: 16 November 2018 Accepted: 17 June 2019
Published online: 05 July 2019

References

1. MassFrontier. http://www.highchem.com/. Accessed 19 June 2018.

DOI:10.1186/s12859-019-2954-7

2. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics. 2010;11:148.
3. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. J Cheminformatics. 2016;8(1):1.
4. Wang Y, Kora G, Bowen BP, Pan C. Midas: A database-searching algorithm for metabolite identification in metabolomics. Anal Chem. 2014;86(19):9496–503.
5. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M. Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS–FINDER software. Anal Chem. 2016;88(16):7946–58.
6. Ridder L, van der Hooft JJJ, Verhoeven S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. Mass Spectrom. 2014;3(Special Issue 2):0033.
7. Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics. 2015;11:98.
8. Heinonen M, Shen H, Zamboni N, Rousu J. Metabolite identification and molecular fingerprint prediction through machine learning. Bioinformatics. 2012;28(18):2333–41.
9. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci. 2015.
10. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A. 2015;112(41):12580-85.
11. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J. Fast metabolite identification with input output kernel regression. Bioinformatics. 2016;32(12):28–36.
12. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, Allen F, Vaniya A, Verdegem D, Böcker S, Rousu J, Shen H, Tsugawa H, Sajed T, Fiehn O, Ghesquière B, Neumann S. Critical assessment of small molecule identification 2016: automated methods. J Cheminformatics. 2017;9(1):22.
13. McGregor MJ, Pallai PV. Clustering of large databases of compounds: Using the mdl "keys" as structural descriptors. J Chem Inform Comput Sci. 1997;37(3):443–8.
14. Vidal D, Thormann M, Pons M. Lingo, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J Chem Inf Model. 2005;45(2):386–93.
15. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. Inchi, the iupac international chemical identifier. J Cheminformatics. 2015;7(1):23.
16. MassBank of North America. http://mona.fiehnlab.ucdavis.edu/. Accessed 8 Dec 2016.
17. Wang MX, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Criisemann M, Boudreau PD, Esquenazi E, Sandoval-Calderon M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai JQ, Neupane R, Gurr J, Rodriguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrovr T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Muller R, Waters KM, Shi WY, Liu XT, Zhang LX, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutierrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat Biotechnol. 2016;34(8):828–37. n/a.
18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, et al. Pubchem substance and compound databases. Nucleic Acids Res. 2015;44(D1):1202–13.
19. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.
20. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G, Evelo CT, Guha R, Steinbeck C. The chemistry development kit (cdk) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminformatics. 2017;9(1):33.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

DOI:10.1186/s12859-019-2954-7

Schymanski *et al. J Cheminform (2017) 9:22*
DOI 10.1186/s13321-017-0207-1

Journal of Cheminformatics

RESEARCH ARTICLE **Open Access**

Critical Assessment of Small Molecule Identification 2016: automated methods

Emma L. Schymanski[1*], Christoph Ruttkies[2], Martin Krauss[3], Céline Brouard[4,5], Tobias Kind[6], Kai Dührkop[7], Felicity Allen[8], Arpana Vaniya[6,9], Dries Verdegem[10], Sebastian Böcker[7], Juho Rousu[4,5], Huibin Shen[4,5], Hiroshi Tsugawa[11], Tanvir Sajed[8], Oliver Fiehn[6,12], Bart Ghesquière[10] and Steffen Neumann[2]

Abstract

Background: The fourth round of the Critical Assessment of Small Molecule Identification (CASMI) Contest (www.casmi-contest.org) was held in 2016, with two new categories for automated methods. This article covers the 208 challenges in Categories 2 and 3, without and with metadata, from organization, participation, results and post-contest evaluation of CASMI 2016 through to perspectives for future contests and small molecule annotation/identification.

Results: The Input Output Kernel Regression (`CSI:IOKR`) machine learning approach performed best in "Category 2: Best Automatic Structural Identification—*In Silico* Fragmentation Only", won by Team Brouard with 41% challenge wins. The winner of "Category 3: Best Automatic Structural Identification—Full Information" was Team Kind (`MS-FINDER`), with 76% challenge wins. The best methods were able to achieve over 30% Top 1 ranks in Category 2, with all methods ranking the correct candidate in the Top 10 in around 50% of challenges. This success rate rose to 70% Top 1 ranks in Category 3, with candidates in the Top 10 in over 80% of the challenges. The machine learning and chemistry-based approaches are shown to perform in complementary ways.

Conclusions: The improvement in (semi-)automated fragmentation methods for small molecule identification has been substantial. The achieved high rates of correct candidates in the Top 1 and Top 10, despite large candidate numbers, open up great possibilities for high-throughput annotation of untargeted analysis for "known unknowns". As more high quality training data becomes available, the improvements in machine learning methods will likely continue, but the alternative approaches still provide valuable complementary information. Improved integration of experimental context will also improve identification success further for "real life" annotations. The true "unknown unknowns" remain to be evaluated in future CASMI contests.

Keywords: Compound identification, *In silico* fragmentation, High resolution mass spectrometry, Metabolomics, Structure elucidation

Background

The Critical Assessment of Small Molecule Identification (CASMI) Contest [1] was founded in 2012 as an open contest for the experimental and computational mass spectrometry communities [2, 3]. Since then, CASMI contests have been held in 2013 [4], 2014 [5] and now in 2016, which is summarized in this article. The focus of CASMI has changed slightly with each contest, reflecting differences in focus of the organizers as well as the perceived interest and challenges in structure elucidation with mass spectrometry. CASMI is purely a research activity—there is no fee for participation but likewise also no prize money for the winners.

In 2016, Category 1 was "Best Structural Identification on Natural Products", with 18 challenges available, a number achievable for both manual and automatic methods. Any methods could be used to submit entries and seven groups participated in this category. The outcomes

---

*Correspondence: emma.schymanski@eawag.ch
[1] Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland
Full list of author information is available at the end of the article

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

DOI:10.1186/s13321-017-0207-1

of this category are presented separately [6] and reported here briefly for comparison purposes.

In contrast, Categories 2 and 3 were defined with 208 challenges in total. Candidate lists containing the correct solution were provided, along with training data for parameter optimization. These categories were specifically designed for automated methods, as no participant with a manual approach could be expected to invest so much time in solving all challenges. Category 2 was defined as "Best Automatic Structural Identification—*In Silico* Fragmentation Only". The aim was to compare the different fragmentation approaches, ranging from combinatorial, to rule-based, to simulations; the use of mass spectral library searching or additional information was not allowed. In contrast, Category 3 was "Best Automatic Structural Identification—Full Information". The same data files and candidate lists were provided as for Category 2, but any form of additional information could be used (retention time information, mass spectral libraries, patents, reference count, etc.). This was to assess the influence of additional information (hereafter termed metadata) on the results of the contest. Participants were required to detail their submissions in an abstract submitted with the results. The rules and submission formats were communicated on the CASMI rules website [7] prior to the release of the challenge data; the evaluation was automated provided the submission format passes all checks. In contrast to previous years, participants were allowed to submit up to three entries each, to evaluate the performance of different approaches. More details are given below.

This article summarizes Categories 2 and 3 of CASMI 2016, including organization, participation and additional post-contest analysis. Six external groups participated in these categories (see Graphical Abstract); 10 in total combined with the Category 1 participants, which is more than ever before.

Methods

Contest data for CASMI 2016

Mass spectra

All MS/MS spectra were obtained on a Q Exactive Plus Orbitrap (Thermo Scientific), with <5 ppm mass accuracy and nominal MS/MS resolving power of 35,000 at $m/z = 200$ using electrospray ionization (ESI) and stepped 20/35/50 nominal higher-energy collisional dissociation (HCD) energies. The spectra were obtained by measuring 22 mixes of authentic standards with the same liquid chromatography–mass spectrometry (LC–MS) method, in data-dependent acquisition mode using inclusion lists containing the $[M + H]^+$ (positive) and $[M - H]^-$ ion masses. Positive and negative mode data were acquired separately. Each mix contained between 10 and 94 compounds. A reversed phase column was used (Kinetex C_{18} EVO, 2.6 µm, 2.1 × 50 mm with a 2.1 × 5 mm precolumn from Phenomenex). The gradient was (A/B): 95/5 at 0 min, 95/5 at 1 min, 0/100 at 13 min, 0/100 at 24 min (A = water, B = methanol, both with 0.1% formic acid) at a flow rate of 300 µL/min.

The MS/MS peak lists were extracted with RMassBank [8] using the ion mass and a retention time window of 0.4 min around the expected retention time and reported as absolute ion intensities. To obtain high-quality spectra, the data was cleaned and recalibrated to within 5 ppm using known subformula annotation [8], all other peaks without a valid subformula within 5 ppm of the recalibrated data were removed. All substances with double chromatographic peaks, different substances with identical spectra (detected via the SPectraL hASH (SPLASH) [9, 10]), MS/MS containing only one peak or with a maximum intensity below 1×10^5 were excluded from the datasets. Substances that were measured multiple times (because they were present in more than one mix) in the same ionization mode were only included once, selected by higher intensity. MS/MS from positive and negative mode were included if the substance ionized in both modes. The final peak lists were saved in plain text format and Mascot Generic Format (MGF). All MS/MS spectra are now available on MassBank [11].

Candidates

The candidates were retrieved from ChemSpider via MetFrag2.3 [12] using the monoisotopic exact mass ±5 ppm of the correct candidate on February 14th, 2016. The SMILES from the MetFrag output were converted to standard InChIs and InChIKeys with OpenBabel (version 2.3.2) [13]. Candidates were removed if the SMILES to InChI conversion failed, all other candidates were retained without any additional filtering. The presence of the correct solution in the candidate list was verified and the lists were saved as CSV files.

Training and challenge datasets

The MS/MS spectra and corresponding candidates were split into training and challenge datasets, according to the spectral similarity to MassBank spectra (as many substances were already in MassBank). Challenge spectra were those where no MassBank spectrum was above 0.85 similarity (calculated with MetFusion [14]); all spectra where there was a match in MassBank above 0.85 were included in the CASMI training set. There were two exceptions: Alizarin, similarity 0.88 to laxapur (FIO00294), and anthrone, similarity 0.86 to phosphocreatine (KO003849), to ensure a sufficient number of natural products remained as challenges for Category 1 (see below). Many of the natural products in the mixes did not ionize well with the experimental setup used.

DOI:10.1186/s13321-017-0207-1

The challenge dataset consisted of 208 peak lists from 188 substances, 127 obtained in positive mode (all $[M+H]^+$) and 81 in negative mode (all $[M-H]^-$). The retention times for each substance was provided in a summary CSV file. The training dataset consisted of 312 MS/MS peak lists (from 285 substances), of which 254 were obtained in positive mode (all $[M+H]^+$) and 58 negative mode (all $[M-H]^-$). The identities and retention times of the substances in the training dataset were provided in a summary CSV file. All files were uploaded to the CASMI website [15]. Participants were asked to contact the organizers if they required additional formats.

To allow a comparison with manual approaches, Challenges 10–19 in Category 1 were a (re-named) subset of the dataset in Categories 2 and 3. The corresponding challenge numbers are given in Table 1.

Information about the full scan (MS1) data was not originally provided for CASMI 2016, but was provided retrospectively for Challenges 10–19 in Category 1 upon request and post-contest for Categories 2 and 3 for another publication [16]. All data is now available on the CASMI website [15].

Rules and evaluation

The goal of the CASMI contest was for participants to determine the correct molecular structure for each challenge spectrum amongst the corresponding candidate set, based on the data provided by the contest organizers. A set of rules were fixed in advance to clarify how the submissions were to be evaluated and ranked, to ensure that the evaluation criteria were transparent and objective. All participants were encouraged to follow the principles of reproducible research and accurately describe how their results were achieved in an abstract submitted with the results. Submission formats were defined in advance (described below) to satisfy the R scripts used to perform the automatic evaluation, results and web page generation. Test submissions could be submitted predeadline to check for issues; any post-deadline problems were resolved prior to the release of the solutions.

Table 1 Overlapping challenges between Category 1 and Categories 2 and 3

Name	Category 1	Categories 2 and 3	Mode
Creatinine	Challenge-010	Challenge-084	Positive
Anthrone	Challenge-011	Challenge-162	Positive
Flavone	Challenge-012	Challenge-166	Positive
Medroxyprogesterone	Challenge-013	Challenge-184	Positive
Abietic acid	Challenge-014	Challenge-207	Positive
Estrone-3-(β-D-glucuronide)	Challenge-015	Challenge-034	Negative
Alizarin	Challenge-016	Challenge-045	Negative
Thyroxine	Challenge-017	Challenge-048	Negative
Purpurin	Challenge-018	Challenge-054	Negative
Monensin	Challenge-019	Challenge-079	Negative

Participants could enter a maximum of three submissions per approach and category, provided they used these submissions to assess the influence of different strategies on the outcomes. The rationale and differences had to be detailed in the abstract. The *best overall performing submission* per participant was considered in declaring the winner(s). The submission requirements were an abstract file (per submission, see website for details) plus results files for each challenge to be considered in the contest. There was no explicit requirement to submit entries for all challenges. Valid challenge submissions were plain text, tab separated files with two columns containing the representation of the structure as the standard InChI or the SMILES code (column 1) and the score (column 2). To be evaluated properly, the score was to be non-negative with a higher score representing a better candidate.

For each challenge, the absolute rank of the correct solution (ordered by score) was determined. The average rank over all equal candidates was taken where two or more candidates had the same score. Due to inconsistencies with how participants dealt with multiple stereoisomers (and since stereoisomers amongst the candidates could not be separated with the analytical methods used), submissions were filtered post-submission to remove duplicate stereoisomers using the first block of the InChIKey. The *highest scoring* isomer was retained. The ranks were then compared across all eligible entries to declare the gold (winner), silver and bronze positions for each challenge. *Gold was awarded to the contestant(s) with the lowest rank among all contestants for that challenge*. This way, a winner could be declared even if no method ranked the correct candidate in the Top 1. Joint positions were possible in case of ties. The overall winner was determined using an Olympic medal tally scheme, i.e. the participants with the most gold medals per category won. The winners were declared on the basis of this automatic evaluation.

Additional scores

Further scores that were used to interpret the results included the mean and median ranks, Top X rank counts, relative ranking positions (RRPs, defined in [2]) and quantiles. The *Formula 1 Score*, based on the method used in Formula 1 racing [17] since 2010, is the sum of the Top 1 to 10 ranks of the correct candidates weighted by the scores 25, 18, 15, 12, 10, 8, 6, 4, 2 and 1. The *Medal Score* (as opposed to the per-challenge Gold Medal count used in CASMI to declare the winner) is the sum of weighted Top 1 ranks with 5 points (gold medal), Top 2 ranks with 3 points (silver) and Top 3 ranks (bronze) with 1. Non-integer ranks (due to equally-scoring candidates) were rounded up to the higher rank for calculating Top X, Formula 1 and medal scores (e.g. rank 1.5 was counted as 2).

Participant methods

Team Allen (Felicity Allen, Tanvir Sajed, Russ Greiner and David Wishart) processed the provided candidates for Category 2 using CFM-ID [18]. CFM-ID uses a probabilistic generative model to produce an *in silico* predicted spectrum for each candidate compound. It then uses standard spectral similarity measures to rank those candidates according to how well their predicted spectrum matches the challenge spectrum. The original Competitive Fragmentation Model (CFM) positive and negative models were used, which were trained on data from the METLIN database [19]. Mass tolerances of 10 ppm were used, the Jaccard score was applied for spectral comparisons and the input spectrum was repeated for low, medium and high energies to form the `CFM_orig` entry. The `CFM_retrain` entry consisted of a CFM model trained on data from METLIN and the NIST MS/MS library [20] for the positive mode spectra. This new model also incorporated altered chemical features and a neural network within the transition function. Mass tolerances of 10 ppm were used, and the DotProduct score was applied for spectral comparisons. This model combined the spectra across energies before training, so only one energy exists in the output. The negative mode entries were the same as for `CFM_orig`.

CFM-ID was also used to submit entries for Category 3, by combining the above CFM-based score with a database score (DB_SCORE). For each hit in the databases HMDB [21], ChEBI [22], FooDB [23], DrugBank [24] and a local database of plant-derived compounds, 10 was added to DB_SCORE. The `CFM_retrain+DB` and `CFM_orig+DB` submissions were formed by adding the DB_SCORE for each candidate to the `CFM_retrain` and `CFM_orig` entries from Category 2, respectively.

Team Brouard (Céline Brouard, Huibin Shen, Kai Dührkop, Sebastian Böcker and Juho Rousu) participated in Category 2 using CSI:FingerID [25] with an Input Output Kernel Regression (IOKR) machine learning approach to predict the candidate scores [26]. Fragmentation trees were computed with SIRIUS version 3.1.4 [27] for all the molecular formulas present in the candidate set. Only the tree associated with the best score was considered. SIRIUS uses fragment intensities to distinguish noise and signal peaks, while the intensities were weighted lowly during learning (see [25, 26]). Different kernel functions were computed for measuring the similarities between either MS/MS spectra or fragmentation trees. Multiple kernel learning (MKL, see [28]) was used to combine the kernels as input for IOKR. In the `CSI:IOKR_U` submission, the same weight was associated with each kernel (uniform multiple kernel learning or "Uni-MKL"). In the `CSI:IOKR_A` submission the kernel weights were learned with the Alignf algorithm [29] so that the combined input kernel was maximally aligned to an ideal target kernel between molecules. In both submissions, IOKR was then used for learning a kernel function measuring the similarity between pairs of molecules. The values of this kernel on the training set were defined based on molecular fingerprints, using approximately 6000 molecular fingerprints from CDK [30, 31]. Separate models were trained for the MS/MS spectra in positive and negative mode. The method was trained using the CASMI training spectra, along with additional merged spectra from GNPS [32] and MassBank [33]. For the negative ion mode spectra, 102 spectra from GNPS and 714 spectra from MassBank were used. For the positive ion mode spectra, 3868 training spectra from GNPS were used. These training sets were prepared following a procedure similar to that described in [25].

The additional post-competition submission `CSI:IOKR_AR` used the same approach as `CSI:IOKR_A`, but the positive model was learned using a larger training set containing 7352 positive mode spectra from GNPS and MassBank. This training set was effectively the same as that used by Team Dührkop, with minor differences due to the pre-selection criteria of the spectra. The negative mode training set was not modified.

Team Dührkop (Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu and Sebastian Böcker) entered Category 2 with a command line version of CSI:FingerID version 1.0.1 [25], based on the original support vector machine (SVM) machine learning method. The peaklists were processed in MGF format and fragmentation trees were computed with SIRIUS version 3.1.4 [27] using the Q-TOF instrument settings. Trees were computed for all candidate formulas in the given structure candidate list; trees with a score <80% of the optimal tree score were discarded. The remaining trees were processed with CSI:FingerID. SIRIUS uses fragment intensities to distinguish noise and signal peaks, while the intensities are weighted lowly in CSI:FingerID (see [25]). Molecular fingerprints were predicted for each tree (with Platt probability estimates [34]) and compared against the fingerprints of all structure candidates (computed with CDK [30, 31]) with the same molecular formula. The resulting hits were merged together in one list and were sorted by score. A constant value of 10,000 was added to all scores to make them positive (as required in the CASMI rules). Ties of compounds with same score (and sometimes also with same 2D structure) were ordered randomly. The machine learning method was trained on 7352 spectra (4564 compounds) downloaded from GNPS [32] and MassBank [33]. All negative ion mode challenges were omitted due to a lack of training data; i.e. entries were only submitted for positive challenges. This formed the `CSI:FID` entry.

Team Dührkop submitted a second "leave out" entry, `CSI:FID_leaveout`, during the contest. Before the correct answer was known, the team observed that the top-scoring candidate matched a compound from the `CSI:FID` training set in 67 challenges, which could indicate that the method had memorized the training spectra. To assess the generalization of their method, the classifiers were retrained on the same training set, plus CASMI training spectra, but with these top scoring candidates removed. As this entry was "guesswork" and did not affect the contest outcomes, upon request Team Dührkop resubmitted a true "leave out" entry post-contest where all CASMI challenge compounds were removed from their training set (not just their "guess" based on top scoring candidates) prior to retraining and calculating the `CSI:FID_leaveout` results. For the sake of interpretation, only these updated "leave out" results are presented in this manuscript.

Team Kind (Tobias Kind, Hiroshi Tsugawa, Masanori Arita and Oliver Fiehn) submitted entries to Category 3 using a developer version (1.60) of the freely available MS-FINDER software [35, 36] combined with MS/MS searching and structure database lookup for confirmation (entry `MS-FINDER+MD`). MS-FINDER was originally developed to theoretically assign fragment substructures to MS/MS spectra using hydrogen rearrangement (HR) rules, and was subsequently developed into a structure elucidation program consisting of formula prediction, structure searching and structure ranking methods. For CASMI, an internal database was used to prioritize existing formulas from large chemical databases over less common formulas and the top 5 molecular formulas were regarded for structure queries. Each formula was then queried in the CASMI candidate lists as well as an internal MS-FINDER structure database. A tree-depth of 2 and relative abundance cutoff of 1% as well as up to 100 possible structures were reported with MS-FINDER. The final score was calculated by the integration of mass accuracy, isotopic ratio, product ion assignment, neutral loss assignment, bond dissociation energy, penalty of fragment linkage, penalty of hydrogen rearrangement rules, and existence of the compound in the internal MS-FINDER structure databases (see Additional file 1 for full details). MS-FINDER uses ion intensities in the relative abundance cutoff and isotopic ratio calculations, but not in candidate scoring.

Secondly, MS/MS search was used for further confirmation via the NIST MS Search GUI [37] together with major MS/MS databases such as NIST [20], MassBank of North America (MoNA) [38], ReSpect [39] and MassBank [33]. The precursor was set to 5 ppm and product ion search tolerance to 200 ppm. Around 100 out of the 208 candidates had no MS/MS information. For these searches, a simple similarity search without precursor information was also used, or the precursor window was extended to 100 ppm. Finally, those results that gave overall low hit scores were also cross-referenced with the STOFF-IDENT database of environmentally-relevant substances [40, 41] to obtain information on potential hit candidates. This step was taken because the training set consisted of mostly environmentally relevant compounds.

Team Vaniya (Arpana Vaniya, Stephanie N. Samra, Sajjan S. Mehta, Diego Pedrosa, Hiroshi Tsugawa and Oliver Fiehn) participated in Category 2 using MS-FINDER [35, 36] version 1.62 (entry `MS-FINDER`). MS-FINDER uses hydrogen rearrangement rules for structure elucidation using MS and MS/MS spectra of unknown compounds. The default settings were used; precursor *m/z*, ion mode, mass accuracy of instrument, and precursor type (given in CASMI) were used to populate the respective fields in MS-FINDER. Further parameter settings were: tree depth of 2, relative abundance cutoff of 1, and maximum report number of 100. Although relative abundance cutoffs were used to filter out noisy data, ion abundances were not used by MS-FINDER for calculation of either the score or rank of candidate structures. The default formula finder settings were used, except the mass tolerance, which was set to ±5 ppm mass accuracy as given by the CASMI organizers.

MS-FINDER typically retrieves candidates from an Existing Structure Database (ESD) file compiled from 13 databases, but this was disabled as candidates were provided. Instead, one ESD was created for each of the 208 challenges, containing the information from the candidate lists provided by the CASMI organizers. A batch search of the challenge MS/MS against the challenge candidate list (in the ESD) was performed on the top 500 candidates, to avoid long computational run times. Up to 500 top candidates structures were exported as a text file from MS-FINDER. Scores for automatically matching experimental to virtual spectra were ranked based on mass error, bond dissociation energy, penalties for linkage discrepancies, or violating hydrogen rearrangement rules. Final scores and multiple candidate SMILES were reported for 199 challenges for submission to CASMI 2016. Nine challenges could not be processed due to time constraints (Challenges 13, 61, 72, 78, 80, 106, 120, 133, 203). Full details on this entry, MS-FINDER and file modifications required are given in Additional files 1 and 2.

Team Verdegem (Dries Verdegem and Bart Ghesquière) participated in Category 2 with MAGMa+ [42], which is a wrapper script for the identification engine MAGMa [43]. For any given challenge, MAGMa+ runs MAGMa twice with two different parameter sets. A total of four optimized parameter sets exist (two for positive and two for negative ionization mode), which all differ from the original MAGMa parameters. Within one ionization mode, both corresponding parameter sets were each optimized for a unique latent molecular class. Following the outcome of both MAGMa runs, MAGMa+ determines the molecular class of the top ranked candidates returned by each run using a trained two-class random forest classifier. Depending on the most prevalent molecular class, one outcome (the one from the run with the parameters corresponding to the most prevalent class) is returned to the user. The candidate lists provided were used as a structure database without any prefiltering. MAGMa determines the score by adding an intensity-weighted term for each experimental peak. If a peak is explained by the *in silico* fragmentation process, the added term reflects the difficulty with which the corresponding fragment was generated. Otherwise, an "unexplained peak penalty" is added. Consequently, MAGMa returns smaller scores for better matches, and therefore the reciprocal of the scoring values was submitted to the contest. MAGMa was run with a relative m / z precision of 10 ppm and an absolute m / z precision of 0.002 Da. Default values were taken for all other options. MAGMa+ is available from [44].

To enable a comparison between MAGMa+ (entry `MAGMa+`) and MAGMa, entries based on MAGMa were submitted post-contest (entry `MAGMa`). MAGMa was run as is, without customization of its working parameters (bond break or missing substructure penalties). Identical mass window values as for MAGMa+ were applied (see above). Default values were used for all other settings. Again, the reciprocal of the scoring values was submitted to obtain higher scores for better matches.

Additional results

Additional results were calculated using MetFrag2.3 [12] to compare these results with the other methods outside the actual contest and to investigate the influence of metadata on the competition results. MetFrag command line version 2.3 (available from [45]) was used to process the challenges, using the MS/MS peak lists and the ChemSpider IDs (CSIDs) of the candidates provided. MetFrag assigns fragment structures generated *in silico* to experimental MS/MS spectra using a defined mass difference. The candidate score considers the mass and intensity of the explained peaks, as well as the energy required to break the bond(s) to generate the fragment. Higher masses and intensities will increase the score, while higher bond energies will decrease the score. The `MetFrag` submission consisted of the MetFrag fragmentation approach only. In the `MetFrag+CFM` entry the MetFrag and CFM-ID (version 2) [18] scores were combined. The `CFM` scores were calculated independently from Team Allen. Additionally, a `Combined_MS/MS` entry was prepared, combining six different fragmenters with equal weighting: `CFM_orig`, `CSI:FID`, `CSI:IOKR_A`, `MAGMa+`, `MetFrag` and `MS-FINDER`.

Several individual metadata scores were also prepared. A retention time prediction score was based on a correlation formed from the CASMI training set (submission `Retention_time`; `+RT`, see Additional file 1: Figure S1. The reference score (submission `Refs`) was the ChemSpiderReferenceCount, retrieved from ChemSpider [46] using the CSIDs given in the CASMI data. The `MoNA` submission ranked the candidates with the MetFusion-like [14] score built into MetFrag2.3, using the MoNA LC–MS/MS spectral library downloaded January 2016 [38]. The `Lowest_CSID` entry had candidates scored according to their identifier, where the lowest ChemSpider ID was considered the best entry.

The combined submissions to test the influence of different metadata on the results were as follows: `MetFrag+RT+Refs`, `MetFrag+CFM+RT+Refs`, `MetFrag+CFM+RT+Refs +MoNA`, `Combined_MS/MS+RT+Refs` and finally `Combined_MS/MS+RT+Refs+MoNA`. Full details of how all these submission were prepared are given in Additional file 1.

Results

CASMI 2016 overall results

The sections below are broken up into the official results of the two categories during the contest, shown in Table 2, followed by the post-contest evaluation and a comparison with all approaches from Category 1.

Category 2: In silico fragmentation only

The results from Category 2 are summarized in Table 2. The participant with the highest number of wins over all challenges (i.e. gold medals) was **Team Brouard** with 86 wins over 208 challenges (41%) for `CSI:IOKR_A`. **Team Dührkop** with `CSI:FID` (82 gold, 39%) and **Team Vaniya** with `MS-FINDER` (70 gold, 34%) were in second and third place, respectively. This clearly shows that the recent machine-learning developments have greatly improved the performance relative to the bond-breaking approaches and even `CFM`. The third place for `MS-FINDER` shows that it performs in quite a complementary way to the `CSI` methods. The performance of

Table 2 Results summary for Categories 2 and 3: medal tally and other statistics

	Category 2					Category 3	
	Allen CFM orig	**Brouard** CSI: IOKR_A	**Dührkop** CSI:FID	**Vaniya** MS–FINDER	**Verdegem** MAGMa+	**Allen** CFM retrain +DB	**Kind** MS–FINDER +MD
Gold	63	**86**	82	70	44	156	**159**
Silver	**71**	50	21	26	53	**52**	38
Bronze	40	31	11	35	**65**	0	0
Gold (neg)	26	20	0	**33**	24	61	**64**
Gold (pos)	37	66	**82**	37	20	**95**	**95**
Top 1 (neg)	12	9	0	**14**	8	47	**59**
Top 1 (pos)	27	53	**70**	32	16	73	47
Top 1	39	62	**70**	46	24	120	**146**
Top 3	77	**93**	90	79	59	160	**162**
Top 10	**123**	118	100	101	105	**182**	174
Mean rank	47.98	127.34	25.17	**19.75**	70.79	13.72	**6.4**
Median rank	6	5.2	**1**	3	9.8	**1**	**1**
Mean RRP	0.906	0.874	**0.945**	0.804	0.88	**0.971**	0.904
Median RRP	0.987	0.988	**1**	0.922	0.972	**1**	**1**
Formula 1	1957	**2276**	2156	1867	1524	3861	**4011**
Medal Score	275	375	**396**	305	195	700	**766**

The first, second and third place by "Gold medals" (used to declare CASMI winners) are highlighted in red, orange and yellow, respectively. The best value per statistic is marked in bold

Team Dührkop is especially surprising considering that they did not submit any challenges in negative mode (due to a lack of training data).

Table 2 also includes the Top 1 (correct candidate ranked in first place), Top 3 (correct candidate amongst the top 3 scoring entries) and Top 10 entries per participant as well as the Formula 1 and Medal scores. The `CSI:FID` entry from Team Dührkop had the best Top 1 result (70, or 34%), followed by Team Brouard and Team Vaniya with 62 and 46 Top 1 candidates. This is an amazing improvement on previous contests and consistent with recent results [25], despite their use of larger candidate sets (PubChem instead of ChemSpider) and a slightly different ranking system. Very interesting to note is that all methods have the correct candidate in the Top 10 in ≥49% of cases, which is likewise a dramatic improvement for automatic annotation. `CFM_orig` had the most the correct candidates in the Top 10 (123 or 59%) and this is reflected in the Formula 1 Score, which weighted the `CFM_orig` performance ahead of `MS-FINDER`, despite their lower Top 1 ranks.

Separating the challenges into positive and negative modes revealed that Team Dührkop clearly led the positive mode predictions (82 wins/gold medals and 70 Top 1 candidates, versus 66 wins and 53 Top 1 candidates for Team Brouard). Both `MS-FINDER` (14 Top 1) and `CFM_orig` (12 Top 1) outperformed Team Brouard for negative mode (9 Top 1), showing that a greater amount of training data for negative spectra would likely improve the `CSI` methods in the future. The training set used by Team Brouard contained 7300 spectra for positive mode and only 816 negative mode spectra. The difference between positive and negative mode was less dramatic for the other approaches.

The results of Category 2 were dominated by the methods that use machine learning on large spectral databases (GNPS [32], MassBank [33], METLIN [19] and NIST [20]), namely Teams Brouard and Dührkop (`CSI`) and Allen (`CFM`). The great increase in data available for training these methods has led to the dramatic improvements in *in silico* methods seen in this contest—increasing the availability of open data will only improve this situation further! The performance of `MS-FINDER`, which does not use machine learning but instead chemical interpretation, is also particularly encouraging and below is shown to perform quite complementary to the machine learning methods. The influence of the training data was investigated during the contest by Teams Dührkop (`CSI:FID_leaveout`) and Allen (`CFM_retrain`); see Table 3. This was investigated for all approaches post-contest, discussed in "Machine learning approaches and training data" section.

Category 3: Full information

The results of Category 3, also summarized in Table 2, were extremely close considering the freedom given to the use of metadata in this Category. **Team Kind** was the winner with 159 gold (64 positive, 95 negative), closely followed by **Team Allen** on 156 gold (61 positive, 95 negative). Interestingly, the number of Top 1 ranks were

Table 3 Results summary for additional Category 2 entries

	Allen		Brouard			Dührkop		Ruttkies		Vaniya	Verdegem	
	CFM_orig	CFM_retrain	*CSI:IOKR_A*	CSI:IOKR_AR*	CSI:IOKR_U	*CSI:FID*	CSI:FID_leaveout*	MetFrag*	MetFrag+CFM*	*MS−FINDER*	*MAGMa+*	MAGMa*
Top 1 Neg.	12	12	9	9	8	0	0	9	**20**	14	8	7
Top 1 Pos.	27	28	53	69	50	**70**	36	15	21	32	16	14
Top 1	39	40	62	**78**	58	70	36	24	41	46	24	21
Top 3	77	73	93	**102**	95	90	70	60	84	79	59	51
Top 10	123	116	118	**131**	118	100	88	108	127	101	105	106
Mean rank	47.98	44.53	127.3	95.09	123.3	25.17	52.02	51.92	33.97	**19.75**	70.79	70.24
Med. rank	6	7	5.25	4	5	**1**	3	8.75	6	3	9.8	9.8
Mean RRP	0.906	0.917	0.874	0.887	0.857	**0.945**	0.931	0.905	0.915	0.804	0.88	0.88
Med. RRP	0.987	0.985	0.988	0.993	0.98	**1**	0.995	0.98	0.991	0.922	0.972	0.969
Gold	53	52	73	**91**	70	74	41	32	51	61	35	31
Formula 1	1957	1900	2276	**2500**	2237	2156	1596	1593	2058	1867	1524	1463
Medal Sc.	275	269	375	**442**	371	396	252	198	292	305	195	175
Q_10	1	1	1	1	1	1	1	1	1	1	1	1.4
Q_25	2	2	1	1	1	1	1	3	2	1	3	3.5
Q_50	6	7	5.25	4	5	1	3	8.75	6	3	9.8	9.8
Q_75	36.25	27.63	55.5	36	78.75	6	17	37.88	25	17	66.1	64.5
Q_90	121.8	104.6	192.9	134.9	288.9	37.5	72.4	120.9	87.65	68.75	187.1	148.5

The column header of entries used in Table 2 are given in italics. The best value per statistic is marked in bold. * indicates internal and post-competition submissions. Med. = median. Q_X indicates Xth quantile

DOI:10.1186/s13321-017-0207-1

very different, 146 (Team Kind) versus 120 (Team Allen); consistent with Category 2 `CFM_orig` had more Top 10 entries but fewer Top 1 and 3 entries than `MS-FINDER`. In this category the `CFM_retrained` model from Team Allen outperformed `CFM_orig`, which performed better in Category 2.

While very different approaches were used to obtain the "metadata", the results of Category 3 clearly demonstrate the value of using metadata when identifying "known unknowns" as was the case in this contest where candidates were provided. This decision to provide candidates was taken deliberately to remove the influence of the candidate source on the CASMI results. The role of this "metadata" is discussed further below (Category 3: Additional Results). For true unknown identification the benefit of this style of metadata could be considerably reduced depending on the context, however this would have to be the subject of an alternative category in a future contest.

Post-contest evaluation

While the best overall results per participant were used to declare the winners, each participant was able to submit up to three entries to the contest if they chose to assess the influence of different strategies on their outcome. This has revealed many interesting aspects that would otherwise have gone undetected with only one entry per participant, as in previous contests. To explore these further and take advantage of the automatic evaluation procedure offered in CASMI, several internal and post-contest entries were also evaluated, as described in the Methods section. The results of all these entries, including those run in the contest, are given in Table 3 for Category 2 and in Table 4 for Category 3.

Category 2: Additional results

The additional results for Category 2 (see Table 3) show that the retrained `CSI:IOKR_AR` entry from Team Brouard (using the more extensive `CSI:FID` training data plus negative mode results) would have outperformed their winning `CSI:IOKR_A` entry as well as the `CSI:FID` entry from Team Dührkop. The improvement with additional training data was dramatic for some challenges, e.g. Challenge 178 went from Rank 3101 with `CSI:IOKR_A` to rank 1 with `CSI:IOKR_AR`. Separating the Top 1 ranks into positive and negative mode (see Table 3) shows indeed that the performance for `CSI:IOKR_AR` and `CSI:FID` in positive mode was quite similar (69 vs. 70 wins, respectively), whereas all `CSI` methods are outperformed by `MS-FINDER` and `CFM_orig` in negative mode.

The `MetFrag` entry performed quite similarly to Team Verdegem (`MAGMa+`); as both are combinatorial fragmentation approaches this is not surprising. While the `MetFrag+CFM` entry improved these results dramatically, it was only slightly improved compared with the individual `CFM` entries of Team Allen. However, the improvement by combining the two fragmenters in negative mode was marked, increasing the Top 1 ranks from 9 (`MetFrag`) and 12 (`CFM`) to 20 (`MetFrag+CFM`).

Table 4 Results summary for additional Category 3 entries

	Allen		Kind	Ruttkies		
	CFM orig +DB	*CFMretrain+DB*	*MS-FINDER+MD*	MetFrag+RT+Refs*	MetFrag+CFM+RT+Refs*	MetFrag+CFM+RT+Refs+MoNA*
Top 1	117	120	146	162	**163**	155
Top 3	159	160	162	**183**	180	182
Top 10	182	182	174	191	**199**	194
Mean rank	14	13.62	6.4	7.04	5.39	**4.25**
Median rank	**1**	**1**	**1**	**1**	**1**	**1**
Mean RRP	0.969	0.971	0.904	0.987	0.989	**0.990**
Median RRP	**1**	**1**	**1**	**1**	**1**	**1**
Gold	124	128	148	168	**174**	167
Formula 1	3798	3861	4011	4469	**4509**	4437
Medal score	687	700	766	855	**856**	840
Q_10	1	1	1	1	1	1
Q_25	1	1	1	1	1	1
Q_50	1	1	1	1	1	1
Q_75	3	3	2	1	1	2
Q_90	13.7	14.0	15.0	5.0	5.0	4.3

The column header of entries used in Table 2 are given in italics. The best value per statistic is marked in bold. * Indicates internal and post-competition submissions. Q_X indicates Xth quantile

DOI:10.1186/s13321-017-0207-1

`MS-FINDER` still performed the best in negative mode of all the individual entries. `MAGMa+` outperformed `MAGMa` in Top 1 and Top 3 entries.

Category 3: Additional results

The additional results for Category 3 (see Table 4) show that `MetFrag+CFM+RT+Refs` outperformed the other approaches both in terms of wins and the number of Top 1 ranks. Although adding MoNA to the mix resulted in a poorer performance, this was because spectral similarity was used to separate the training and challenge sets and the resulting MoNA weight was too optimistic for the challenges.

As these results are driven more by the metadata used than the fragmenter behind, a variety of entries were created to assess the contribution of the individual metadata aspects, as well as a "Combined Fragmenter" entry (`Combined MS/MS`) to remove the influence of the fragmentation method (see "Methods" for details). These results are given in Table 5. The `Combined MS/MS` entry outperformed all of the individual Category 2 entries, showing the complementarity of the different approaches. These also outperformed the MS library (MoNA) entry. The retention time prediction alone performed poorly, because this does not contain sufficient structural information to distinguish candidates, as demonstrated in Additional file 1: Figure S2. The lowest identifier strategy, which was used as a "gut feeling" decision criteria commonly in environmental studies before retrieval of reference information could be automated, takes advantage of the fact that well known substances were added to ChemSpider earlier and thus have lower identifiers. Surprisingly this still outperformed the combined fragmenters—but again this is highly dependent on the dataset. The references outperformed all individual metadata categories and even the combined fragmenters clearly. The influence of the metadata is discussed further in "Metadata and consensus identification" section.

Comparison with results from Category 1

Challenges 10–19 in Category 1 were also present among the Category 2 and 3 challenges, as given in Table 1. The results for these challenges, separated by category, are summarized in Table 6 and visualized in Figure S3 and S4 in Additional file 1. Interestingly, this shows that the results of Categories 1 and 3 were remarkably comparable, while the ranks of Category 2, using only MS/MS data, were generally worse. Again, this shows that the incorporation of metadata in automated methods is essential to guide users to the identification for known substances—but misleading when assessing the performance of computational methods. As metadata cannot assist in the identification of true unknowns for which no data exists, more work is still needed to bring the performance of the *in silico* MS/MS identification methods (Category 2) closer to that of Categories 1 and 3. However, it is clear from this 2016 contest that much progress has been made with the new machine learning methods and—as observed above—continuing to improve the availability of training data will improve these further.

Interestingly, Challenge 14 (Abietic acid) was challenging for all participants in all categories; this was the only challenge in Category 1 where no participant had the correct answer in first place despite the fact that the challenge spectrum was very informative and the candidate numbers were relatively low (see Additional file 1: Figure S7).

Discussion

Visualization of CASMI results: clustering

To visualize the CASMI 2016 results together, a hierarchical clustering was performed. The heat map of the negative mode challenges (1–81, excluding Team Dührkop) can be seen in Fig. 1, while the heat map of the positive mode challenges (82–208) is given in Fig. 2. These are discussed below; in addition interactive plots are provided

Table 5 Contribution of Metadata to the results

	RT	MoNA	Lowest CSID	Refs	Combined MS/MS	Combined MS/MS+RT+Refs	Combined MS/MS+RT+Refs+MoNA
Top 1	1	70	113	143	82	**164**	**164**
Top 3	5	87	158	177	126	183	*187*
Top 10	20	104	177	**196**	166	194	195
Mean rank	504.5	238.3	37.7	**3.0**	13.4	3.9	3.7
Median rank	135	10.25	**1**	**1**	2	**1**	**1**
Mean RRP	0.576	0.780	0.959	**0.995**	0.955	0.990	0.991
Median RRP	0.630	0.977	**1**	*1*	0.998	**1**	**1**

The first four columns contain submissions formed using just one type of metadata, the "Combined MS/MS" column was formed by equally weighting all Category 2 entries from Table 2, while the last two columns combined this with retention time and references without and with MoNA, respectively

The best value per statistic is marked in bold

Table 6 Comparison of Categories 1, 2 and 3 results for the overlapping challenges in Category 1

Chal.	Median rank of correct candidate per Category				Number of valid entries per category			Minimum and maximum rank of correct candidate per category (min, max)		
	All	1	2	3	1	2	3	1	2	3
10	1	1	19.5	1	14	12	6	(1, 15)	(11, 63)	(1, 1)
11	9	2	21	2	11	12	6	(1, 175)	(2, 208)	(1, 9)
12	1.5	1	16	1.5	15	11	6	(1, 88)	(1, 299.5)	(1, 8)
13	3	2	20	3.5	8	12	6	(1, 146)	(1, 270)	(1, 87)
14	25	23	26.5	20	11	12	6	(2, 292)	(17, 164.5)	(12, 144)
15	1	1	1.25	1	12	10	6	(1, 4)	(1, 6)	(1, 3)
16	2.5	2	25	2	12	9	6	(1, 25)	(14, 288)	(1, 14)
17	1	1	2.5	1	10	10	6	(1, 3)	(2, 5)	(1, 1)
18	11	4	19.5	2	9	10	6	(1, 34.5)	(3, 50)	(1, 11)
19	1	1	4.5	1	12	10	6	(1, 3)	(1, 7.5)	(1, 1)

The median ranks of Categories 1 and 3 (highlighted) are remarkably similar

(see reference links provided in the captions) for readers to investigate these clusters in more detail. Corresponding clusters excluding challenges in the training sets are available in Additional file 1: Figures S5 and S6.

The dark blue areas in Fig. 1 indicate very good ranking results. It is clear for the negative spectra that the metadata (Category 3) really improved performance, with very few yellow or red entries for the Category 3 participants, which all grouped together in the cyan cluster (middle left), indicated by the dark blue participant names (middle right). What is also clear is that all methods were very good for most of the compounds in the red challenge cluster (shown at the top, right-most cluster). The combinatorial fragmenters and `CFM` also performed well on the dark blue challenge cluster (second cluster from right)—in contrast both `MS-FINDER` and the `CSI:IOKR` methods struggled for these challenges, shown with the yellow to red coloring in the heat map.

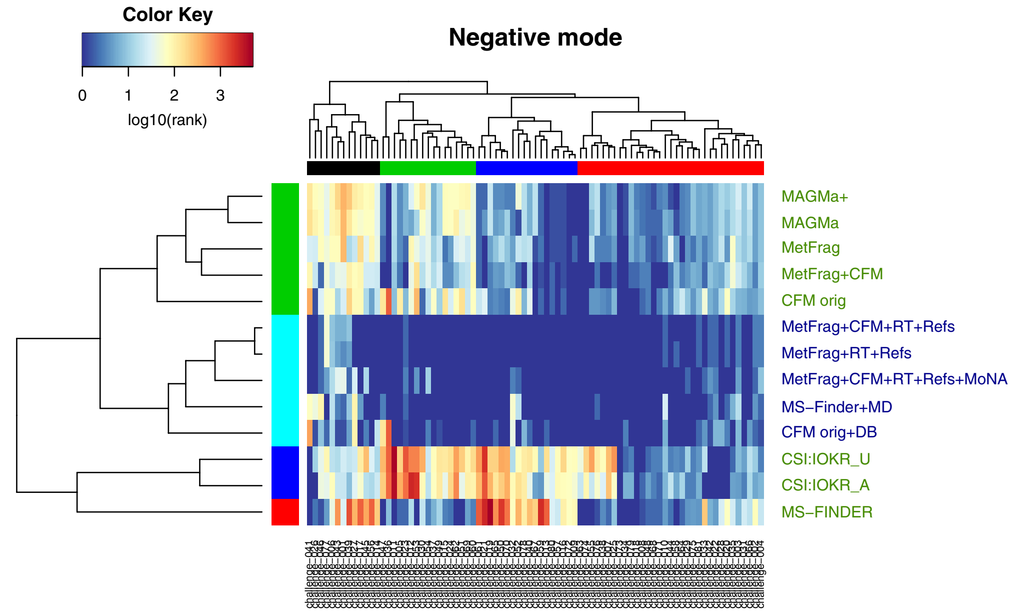


Fig. 1 Heat Map of CASMI Challenges 1–81 (negative mode). Both Category 2 (*green labels* on the *right*) and 3 (*blue labels*) participants are included. Missing values (correct solution missed, or no submission for a challenge) were replaced with the number of candidates for that challenge. Ranks are log-scaled from good (*blue*) to poor (*red*). Team Dührkop was omitted as they did not submit for any challenge, while `CSI:IOKR_AR` and `CFM_retrain` were omitted as these were identical with their original submissions. An interactive version of this plot with legible challenge numbers is available from [47]

`MS-FINDER` outperformed other Category 2 approaches in the green challenge cluster (second from left)—showing the complementarity of the different approaches. This is reinforced by the fact that `MS-FINDER` was split into a participant cluster on its own and also explains partially why the `Combined MS/MS` entry performed better than all individual participant entries. For the clusters of challenges (top), the mean candidate numbers per cluster were (left to right): black (611), green (1603), blue (1019) and red (380), compared with a mean overall of 816. Both the red ("good" overall performance) and black ("poor") clusters have mean candidates below the overall mean, whereas the poorly performing green cluster had mean candidates well above the overall mean. Thus, candidate numbers are not the only driver of performance.

Looking at individual challenges, all machine learning approaches performed poorly for Challenge 36, which was a 3 peak spectrum of a substance typically measured in positive mode (see Additional file 1: Figure S8). The combinatorial approaches performed poorly for Challenge 41 (see Additional file 1: Figure S9), monobenzyl phthalate, where the main peak is a well-known rearrangement that is not covered by these approaches. For this challenge, both `CSI:IOKR` and `MS-FINDER` performed well, indicating that this substance is in the training data domain (many phthalate spectra are in the open domain) and that `MS-FINDER` interprets the spectrum beyond combinatorial methods. The compounds in the dark blue and green challenge clusters are likely not to be covered too well in the training data for `CSI:IOKR`. While it appears that `MS-FINDER` performs very poorly for some challenges, this is in fact an artifact of their submissions; for all the red entries in the heatmap, either the correct answer was absent from their submission (as they took only the top 500 candidates—this applied for 15 challenges) or no answer was submitted (5 challenges). In these cases the total number of candidates was used for the clustering. Removing the challenges where no submission was made from the clustering did not drastically alter any of the outcomes discussed above.

The positive mode cluster (Fig. 2) revealed an even darker blue picture (and thus generally very good results) than the negative mode cluster. The large dark blue patch in the middle of the heat map indicates that for the majority of challenges, largely those in the black challenge cluster (top, middle), both the metadata but also the more extensive training data in positive mode for the machine learning approaches ensured that many Top 1 ranks were achieved. This is also shown well in the green challenge cluster, where the improvements that the metadata and machine learning add beyond the combinatorial approaches can be seen moving down and getting darker from the generally yellow top right corner. As for negative mode, the mean candidate numbers per challenge cluster were calculated (left to right): magenta (5297), cyan (1029), red (886), black (1534), blue (978), green

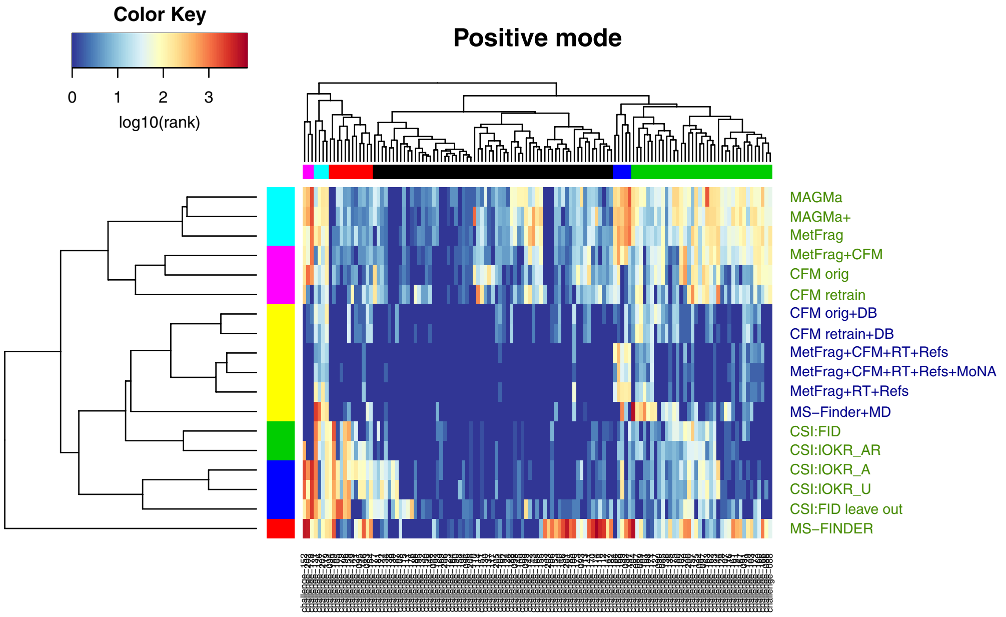


Fig. 2 Heat Map of CASMI Challenges 82–208 (positive mode) both Category 2 (*green labels* on the *right*) and 3 (*blue labels*) participants are included. Missing values (correct solution missed, or no submission for a challenge) were replaced with the number of candidates for that challenge. Ranks are log-scaled from good (*blue*) to poor (*red*). Interactive version with legible challenge numbers available from [48]

(722), with an overall mean of 1281. The performance for the magenta, cyan and blue clusters were all relatively "poor", yet only the magenta cluster contained mean candidate numbers far above the overall mean. The combinatorial fragmenters performed poorly for the green cluster, which had mean candidate numbers below the overall mean. As mentioned above, candidate numbers are again not the only driver of performance. Investigations into other parameters that may influence the challenge clusters, such as number of peaks in the spectra, revealed similarly inconclusive results.

In contrast to negative mode, several participant clusters were formed in positive mode. The top two clusters contained the combinatorial fragmenters `MAGMa`, `MAGMa+` and `MetFrag`, which clustered apart from the `CFM-ID` entries, either alone or in combination with `MetFrag`. Below this was one very large cluster with all Category 3 entries (metadata, yellow). This is followed by three smaller clusters, one in green with the two best `CSI` entries (`CSI:FID` and `CSI:IOKR_AR`), one blue cluster with the remaining `CSI` entries, followed by `MS-FINDER` by itself. Note that `MS-FINDER` still clustered by itself in both positive and negative mode, even when compensating for the challenges with no submission, as mentioned above. This is due in part to their strategy to only select the top 500—again for the vast majority of the red `MS-FINDER` entries in the heat map either the correct candidate was missing in the submission (29 challenges in positive mode), or no submission was made (4 challenges). However, their location in a separate cluster is also possibly due to the fact that `MS-FINDER` does indeed use a different approach to fragmentation than either the combinatorial fragmenters or the machine learning approaches.

The challenge clusters revealed some interesting patterns: four small clusters contained challenges that were problematic for different approaches. Most metadata-free methods performed poorly for the pink cluster (challenges 152, 202, 178); all approaches performed relatively poorly for the cyan cluster adjacent (challenges 131, 126, 207 and 119). The challenges in the red cluster were likely reasonably dissimilar to the other substances in the machine learning training sets, as the combinatorial fragmenters outperformed the `CSI` approaches clearly in this cluster. The machine learners performed well on the dark blue cluster (challenges 184, 168, 199, 92, 197), where surprisingly the metadata even failed the combinatorial fragmenters. Three of these (92, 168, 199) involve breaking an amide bond, which may be something for these approaches to investigate further. Challenge 197 is a fused N heterocycle with one fragment. Spectra of these challenges, with additional comments, are available in Additional file 1: Figures S7–S20.

Visualization of CASMI results: candidate numbers and raw scores

Additional plots have been included in Additional file 1 to provide further visualization of the results. Additional file 1: Figure S21 shows the number of candidates for each challenge, ordered by the number of candidates versus the results for all CASMI entries (during and post-contest). Interestingly, fewer Top 1 entries and higher median/mean ranks were observed for the challenges with moderate candidate numbers (200–1000 candidates); lower median ranks and more Top 1 entries were observed for lower and higher candidate numbers. Additional file 1: Figures S22–S30 show the raw scores for selected submissions per participant and category, in order: `MAGMa+`, `CSI:IOKR_A`, `CSI:FID`, `CFM_orig`, `CFM_retrain+DB`, `MS-FINDER`, `MS-FINDER+MD`, `MetFrag` and `MetFrag+CFM+RT+Refs+MoNA`. These reveal interesting differences in the raw data behind each submission, including for instance the influence of training data availability on the positive and negative challenge results for `CSI:IOKR_A`, the metadata step function in `CFM_retrain+DB` as well as the effect of score scaling on `MetFrag`.

Machine learning approaches and training data

The CASMI2016 results show very clearly how the training data influences the performance of different approaches. The difference in Top 1 positive mode ranks between `CSI:IOKR_A`, 62 and `CSI:FID`, 70 (see Table 2) were due to the different training sets used, the `CSI:IOKR_AR` results (retrained on the same data as `CSI:FID`) had 69 Top 1 ranks. The results for `CSI:IOKR` in negative mode were also generally worse than all other approaches, which shows that the decision of Team Dührkop not to submit entries due to a lack of training data was quite well justified (even though it likely cost them the overall contest "win" for Category 2).

Team Dührkop noted that there was a large overlap between the challenges and their training set and investigated this with the `CSI:FID_leaveout` entry (described in the methods). For the sake of interpretation in this manuscript, this entry was updated post-contest once the exact solutions were known to make it a true "leave out" analysis. Although the performance was reduced compared with `CSI:FID` (36 vs. 70 Top 1 ranks in positive mode), the `CSI:FID_leaveout` entry still had more Top 1 ranks than any other non-`CSI` method in the contest (for positive mode only).

Following the idea of Team Dührkop, the CASMI results were evaluated for all participants on only those challenges where no contestant had the correct candidate in their training sets. Teams Dührkop, Allen and Brouard provided comprehensive lists of their training sets. These

DOI:10.1186/s13321-017-0207-1

were used to determine the overlap between all training sets and the CASMI challenges. The results over those challenges that were not in *any* training set (44 positive and 43 negative challenges) are given in Table 7.

The general observations made on the full contest data are supported by this reduced dataset as well, despite the unsurprising fact that the results on this reduced dataset were generally worse than the official contest results (see Table 2). This demonstrates that, as expected, machine learning methods do better on compounds from within their training sets (for example, the percentage of maximum Top 1 ranks dropped from 34 to 18%). Although the median ranks were worse, the Top 10 ranks still remained around 40–50% for most methods. Cluster plots on this reduced dataset for negative and positive mode, given in the supporting information (Additional file 1: Figures S5, S6), show similar patterns to the cluster plots on the full dataset.

Interestingly, these results show that the `CSI:FID_leaveout` entry outperformed `CSI:FID`, while `CSI:IOKR_A` also outperformed `CSI:IOKR_AR`, the retrained dataset, also for some different scores—similar observations could be made for `CFM_orig` versus `CFM_retrain`. While this could be a potential sign for overfitting, this is a small dataset and some or all of these observations could be due to fluctuations in the data. Overfitting is a potential problem that developers, especially of non-standard machine learning methods should test for, *e.g.* by checking if their performance decreases significantly for compounds which are structural dissimilar to compounds in the training data. These results highlight just one means by which the choice of training set can influence the performance of automated methods. The training set can also impact challenge results in a range of other ways that are harder to disambiguate. One training set may be more or less compatible with the challenge set, even after common compounds are removed. This suggests the importance of assessing automated methods using the same training set, where at all possible.

Metadata and consensus identification

The dataset for CASMI 2016 was predominantly well-known anthropogenic substances and as a result there are many distinct and highly referenced substances in the candidate lists. This is shown in the huge improvement that the metadata made to the ranking performance (Tables 4, 5). Figure 3 shows clearly that the vast majority of substances were either ranked first or second based purely on the reference count, with most other candidates having much lower counts. Figure 4 gives an overview of the contribution the metadata made to each approach based on the CASMI 2016 entries, merging team results in the case of `MS-FINDER`. In the environmental context, it is quite common to search an exact mass or formula in databases such as ChemSpider, where e.g. the highest reference count as well as the substance with the "lowest CSID" are often picked as the most promising hit in many cases, discussed e.g. in [49]. The success with these strategies would have been quite considerable with this dataset. However, for new (emerging) anthropogenic substances and transformation products of known chemicals, these strategies would not work so well as they would have neither a high reference count nor a low database identifier. This situation is also likely to be drastically different for natural products and metabolites, where many more closely-related substances or even isomers could be expected.

The metadata results in Category 3 show that the importance of the sample context cannot be ignored during identification, especially for studies looking to find well-known substances. This is also highlighted by the comparison with the approaches used in Category 1, where also manual and semi-automatic approaches were considered. The current reality is that most automated approaches still depend on retrieving candidates from compound databases containing known structures—i.e. the situation replicated in this CASMI contest. Compound databases such as the Metabolic *In Silico* Network Expansion Databases (MINEs) [50] could be used as alternative sources of candidates for predicted metabolites in the metabolomics context, but would have had limited relevance in this contest.

While metadata, the way it was used here, will not help in the case of true unknowns, there are two cases to consider for automated approaches at this stage. For "unknowns" that happen to be in a database almost accidentally (e.g. a to-date unknown transformation product), the automated fragmentation approaches are very useful, because these structures can be retrieved from substance databases. However for true "unknown unknowns" that are not in any database, fragmenters could only be used in combination with structure generation, which is still impractical with the quality of data and methods at this stage unless candidate numbers can be restrained sufficiently. These cases are often extremely difficult to elucidate using MS^n alone and the information from additional analysis such as NMR will usually be necessary.

Stereoisomerism is another aspect of identification that was not covered in this contest. None of the current approaches are able to distinguish stereoisomers (even cis/trans isomers) using only MS/MS information for known unknowns. The evaluation of this contest addressed this by taking the best scoring stereoisomer and eliminating others (see "Methods") to reduce the

DOI:10.1186/s13321-017-0207-1

Table 7 Global leaveout analysis for additional Category 2 entries—including only challenges where the correct answer was not in any training set

	Allen		Brouard			Dührkop		Ruttkies		Vaniya	Verdegem	
	CFM_orig	CFM_re-train	*CSI_IOKR_A*	CSI_IOKR_AR*	CSI_IOKR_U	*CSI:FID_orig*	CSI:FID_leaveout*	MetFrag*	MetFrag+CFM*	*MS−FINDER*	*MAGMa+*	MAGMa*
Top 1 Neg.	6	6	6	6	4	0	0	4	10	7	4	3
Top 1 Pos.	4	9	9	10	7	9	13	1	3	3	2	2
Top 1	10	15	15	**16**	11	9	13	5	13	10	6	5
Top 3	23	24	29	26	27	17	23	16	27	25	16	14
Top 10	46	40	45	46	40	25	32	39	47	38	35	35
Mean rank	52.57	64.05	106.5	97.84	99.92	52.81	41.48	68.38	37.16	28.7	76.75	100.4
Med. rank	10	12.5	8	10	12	7	3	14.5	8	7.5	23.5	20.5
Mean RRP	0.863	0.872	0.849	0.856	0.837	0.891	0.91	0.863	0.878	0.738	0.832	0.811
Med. RRP	0.966	0.961	0.963	0.967	0.956	0.981	0.993	0.942	0.972	0.806	0.924	0.902
Gold	18	21	23	**26**	19	11	17	7	18	18	10	9
F1 score	628	654	**735**	691	632	403	557	484	707	594	462	434
Medal Sc.	79	94	**105**	98	91	59	87	50	95	85	46	46

n = 43 (negative) and n = 44 (positive)

The best value for selected statistics is marked in bold

DOI:10.1186/s13321-017-0207-1

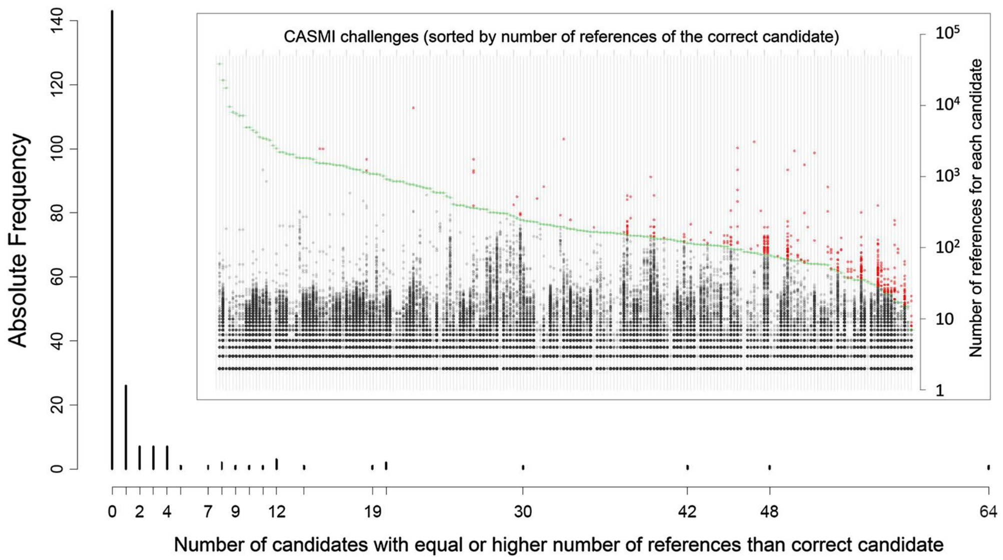


Fig. 3 The distribution of references for CASMI 2016 candidates

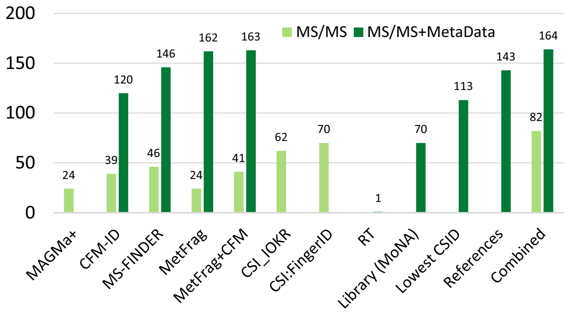


Fig. 4 The influence of Metadata on CASMI 2016 first seven groups—*light green* MS/MS information only, i.e. Category 2. *Dark green* with metadata, i.e. Category 3 participants. Note these are plotted according to the number 1 ranks, not wins. Next 4 groups: *dark green* metadata only; Last group: *light green* is the equally-weighted combination of the six individual Category 2 entries and *dark green* is this plus metadata as shown in Table 5

influence of stereoisomers on the ranking results. However, for electron ionization (EI) MS it is already possible to distinguish stereoisomers in some cases using ion abundances. This is an aspect that should be developed in the future for MS/MS once the spectrum generation is sufficiently reproducible to allow this. Coupling with suitable chromatography will potentially enhance the ability to distinguish between stereoisomers further.

Evaluating methods and winner declaration

Contests such as CASMI always generate much discussion about how the winner was evaluated and declared; this years contest was no exception. A "contest" setting is different to the way individual methods compare their performance with others and this is the role of CASMI—to look at the approaches in different ways, relative to one another. One change in CASMI 2016 was to use the "average rank" instead of the "worst-case" rank to account for equal candidate scores, as participants pointed out that for previous contests one could add small random values to break tied scores and improve results in the contest. There will be several cases where candidates are indistinguishable according to the MS and it is important to capture this aspect in CASMI. While equal scores may make most chemical sense in these cases, computational methods deal with this differently; some report equal scores, others generate slightly different scores for effectively equal candidates. The average rank deals with this better than the "worst-case" rank, but can now disadvantage methods that report equal scores compared with others, as the chances are that at least one other method will beat it each time.

The criteria for declaring the winner in this contest was that the best performing participant(s), i.e. the winner, was defined per challenge and then the wins were added to determine the overall winner. This allows the declaration of a winner per challenge, irrespective of the actual performance (i.e. the winner could have rank 100, if all other participants were worse). The drawback of this approach is that it creates cross-dependencies between participants, i.e. the removal (or addition) of one participant completely changed the rank of the other participants. `CFM` likely suffered from

DOI:10.1186/s13321-017-0207-1

this, as a machine-learning approach with similar training set coverage to `CSI`, which allowed the complementary approach of `MS-FINDER` to claim third place ahead of `CFM`. An alternative approach could be to look at this in terms of overall success and say that if a team had the correct structure as the 20th hit and other teams were even worse, none of the approaches were really sufficient to the task and nobody should then earn a 'win'. This may reflect real structure elucidation cases better, where investigators would likely also consider the Top 3, Top 5, or maybe even Top 10 structures, but is perhaps not so good to declare a winner in a contest as some (difficult) challenges would have no "winner" and the performance of methods on difficult challenges is also an important aspect of the contest. This idea was investigated in this publication by also providing the Top 1, Top 3, Top 10 ranks per participant, as well as the Formula 1 Score (scaled Top 1–10 results) and Medal Score, where the medal count is based on Top 1, 2 and 3 ranks. The results of these metrics confirm the overall pattern observed in the contest: the two `CSI` teams outperformed all others in Category 2, followed by either `MS-FINDER` or `CFM` depending on exactly which score was used. In other words, the approaches have made fantastic progress, are complementary to one another but actually quite difficult to tell apart. Although 208 challenges is an order of magnitude in terms of challenge numbers above previous CASMIs, these numbers are still quite small and almost random differences between the methods resulted sometimes in large changes in the various scores, as shown with the different `CSI` entries.

Participant perspectives

Team Allen submitted two alternative versions of `CFM`, the main difference being that for `CFM_retrain` version, additional training data was added from the 2014 NIST MS/MS database. While the addition of extra training data may have been expected to improve the results, this appears not to have been the case for this competition. One possible reason for this is that the additional data were generally of poorer (often integer) mass accuracy as compared to that used to train the original `CFM` model. This required a wider mass tolerance (0.5 Da) to be used during the retraining (compared to 0.01 Da previously), which may have hindered the training algorithm from accurately assigning explanations to peaks, and so modeling their likelihoods. This highlights that while the production of larger, more comprehensive data sets is likely crucial for better training of automated methods, the quality of these data sets is also very important. Most automated methods would likely benefit from training on cleaner data with better mass accuracies.

Team Dührkop investigated how `CSI:FingerId` compared with a direct spectral library search. A spectral library containing all structures and spectra used to train `CSI:FingerId` was created and searched with a 10 ppm precursor mass deviation. The resulting spectra were sorted via cosine similarity (normalized dot product), again with 10 ppm mass accuracy. Candidates were returned for 91 of the 127 (positive mode) challenges; the correct answer was contained in the library for 69 of these. The spectral library search correctly identified 63 of the 69 structures in total, 40 of these were "trivial" (the correct answer was the only candidate). On average, candidate lists for the spectral library search contained only 2.4 candidates, which was almost three orders of magnitude below the average CASMI candidate list of 1114 candidates. The cosine product between the challenge spectrum and the corresponding training spectrum of the same compound was only 0.76 on average; for one challenge it was below 0.01. For example, the cosine similarity between the spectrum for Challenge 202 (Pendimethalin) and the training spectrum was only 0.137, but it was still "correctly identified" as it was the only candidate with this precursor mass. This compound was correctly identified in the original `CSI:FID` submission, and ranked 569 for the `CSI:FID_leaveout` submission. This indicates that `CSI:FingerId` and other machine-learning approaches are capable of learning inherent properties from the mass spectra, beyond simple spectral similarity.

Team Vaniya The CASMI Category 2 contest was a reshuffling contest: potential structures were given to all participants, listing one to over 8000 potential structures for each challenge. These structures were within 5 ppm mass accuracy and often included different elemental formulas. Therefore, Category 2 was a 'structure dereplication' contest, finding the best structure within a pre-defined list of structures, not a completely open *in silico* test on all exhaustive structures in the chemosphere. In practical terms, it is important to note that an *in silico* software does not eliminate the time consuming aspects of data preparation, formatting, and interpretation. Counting the computing power and manual effort between two people, it took about 24 h to complete the 208 challenges for the `MS-FINDER` submission.

From Table 2, one could say that `MS-FINDER` was best based on the mean rank (19.75), but ranks lower than 10 are less relevant in reality. While MS-FINDER had almost 50% of the challenges within the top 10 ranks, so did every other software (or team). In reality, no chemist would use a software without any database or mass spectral library behind it. The importance of using *a priori* knowledge is seen by Team Allen's submission that improved the Top 1 correct structure hits from 39 to 120 challenges in Category 3, a bit more than 50% of the challenges. Hence, we conclude that the glass is half full: if only *in silico* methods are used, some 50% of the challenges are within the top 10 hits within the structures given by the CASMI organizers. However, many challenges would score much higher if other metadata are used, e.g. constraining the search database to particular classes of compounds that can be expected for a specific study. Which parameters need to be optimized, and which *a priori* metadata should be used? Those questions may be answered in a more tailored future CASMI contest.

Team Verdegem participated in Category 2 of the CASMI 2016 contest with `MAGMa+`, which is a fast, plug-and-play method relying on combinatorial fragmentation without requiring a preliminary training phase for improved performance. The entire submission, including scripting for automation and single core calculations, took less than 1 day. `MAGMa+` outperformed `MAGMa`, showing the use of the parameter optimization performed to improve several second and third ranked candidates to first place. `MAGMa+` shared the best ranking for 44 of 208 challenges (see Table 2) and performed considerably better than other contestants for nine of those challenges (21, 32, 36, 40, 52, 61, 121, 157 and 189), indicating the relevance of the underlying algorithm.

Since MAGMa+ outperformed MAGMa according to some (e.g. number of gold medals, Top 1 and 3 ranks) but not all metrics, further more advanced parameter optimizations are planned to achieve a more global performance improvement. However, further improvements to the performance of `MAGMa/MAGMa+` will require interventions of a different kind. The performance of `MAGMa+` decreases with increasing candidate numbers (in this contest 1116 on average after the removal of duplicate stereoisomers), however, in case of smaller numbers, it starts to outperform some of the other methods [25, 42]. For untargeted metabolite identification in biological/biomedical setups, it is arguably more suitable to restrict the candidate structure database to those metabolites known to exist in the organism under study, e.g. using only the ≈42,000 metabolites currently present in the HMDB [21] for samples of human origin. This was noted also in previous CASMI contests [2]. Many candidate structures had identical scores with `MAGMa+`, resulting in the correct matches being given lower ranks according to the evaluation rules. Whereas on average 1098 structures were retained from the structure database based on the parent mass match, only 616 different score values were observed (on average). Team Verdegem will investigate more discriminative scoring options for `MAGMa+` in the future.

Conclusions

This was the first CASMI contest to use a large set of challenges, targeted especially at the automated methods. This decision was taken on the basis of feedback from several representatives at the 2015 Dagstuhl seminar in Computational Metabolomics [51], to allow a statistically more robust comparison of the methods. The decision to provide candidates this year was also on the basis of Dagstuhl discussions, to eliminate the data source as an influence on the contest outcomes and thus focus more on the role of the *in silico* fragmentation approaches themselves.

From the perspective of the organizers, it was a great success to have participants contribute from each of the major different approaches; MetFrag was added internally for the sake of completion as this was not otherwise represented and allows this paper to complement the work in [25] on a different dataset. Very interesting and constructive discussions have resulted from choosing to prepare this article with "all on board" and the post-contest analysis has been instrumental in teasing apart some of the differences between the actual contest results.

The contest winners, **Team Brouard** with `CSI:IOKR_A` in Category 2 and **Team Kind** with `MS-FINDER+MD` in Category 3 prove that the latest developments in this field have indeed resulted in great progress in automated structure annotation. Despite the very large candidate sets, the majority of methods achieved around 50% in the Top 10, which is very positive for real-life annotation, especially with an outlook to higher-throughput untargeted analysis. The combination of the Category 2 submissions resulted in even better overall performance than each individual method, indicating the complementarity of the approaches and supporting the potential use of *consensus* fragmentation results as has been shown earlier for fragmenters [12, 52] and also recently for toxicity modeling using a more sophisticated weighting than that attempted here [53]. The role of the metadata and comparison with Category 1 shows that sample context cannot be ignored during identification.

In this contest, few participants used the CASMI training set provided, which was also a suggestion from Dagstuhl. In the end this was too "big" for pure parameter optimization (where a few spectra may suffice), but too small for serious method training. Team Brouard added it to their other training data in their original submissions, while it was used to determine the score weights in the `MetFrag` entries. Team Vaniya did not use this for `MS-FINDER` to avoid over-training; Team Allen due to a lack of time. One conclusion from the post-contest evaluation is that future CASMIs could consider providing an extensive, open training dataset (e.g. the GNPS/MassBank collection used by `CSI:FID`) and ensure all CASMI challenges are absent from this set. This would, however, force all machine-learning approaches to retrain their methods prior to submission. Another option is that the organizers would have to ensure that all challenges are outside all available datasets—which is possible but also difficult with the number of private and closed collections available. A compromise could be to ensure that a sufficient majority of the candidates are outside the "major" mass spectral resources, with some overlap to ensure sufficient challenges are available (finding data sources for CASMI is a challenging task!) and require participants to submit InChIKey lists of their training sets with their submissions; as done with Teams Allen, Brouard and Dührkop post-contest here.

Challenges for future contests remain true unknowns, i.e. substances that are not present in compound databases. This would currently be feasible for manual approaches and was attempted already once in CASMI 2014, Challenges 43–48 [54], albeit with limited success. Automated approaches would need either a metabolite database such as MINEs [50] or structure generation [55], but finding sufficient appropriate data for an automated category will also be a challenge for the contest organizers, let alone the participants! The ability to distinguish stereoisomers using MS/MS alone also remains a challenge for the future that is not yet ripe enough for a CASMI contest; distinguishing (positional) isomers is likely sufficient challenge for the next few years.

The huge improvements in machine learning approaches will continue as more training data becomes available—the more *high quality* data with likewise *high quality* annotations that becomes available in the open data domain will ensure that the best computational people can work on the best identification methods. The complementarity of the chemistry behind `MS-FINDER` and the machine learning behind `CSI` shows that developments in both directions will carry the field forward.

The "take home" messages of CASMI 2016 are:

- The latest developments in the field, `CSI:IOKR` and `MS-FINDER` were well-deserved winners of Categories 2 and 3, respectively.
- The complementarity of different approaches is clear; combining several *in silico* fragmentation approaches will improve annotation results further.
- The best methods are able to achieve over 30% Top 1 ranks and most methods have the correct candidate in the Top 10 for around 50% of cases using fragmentation information alone, such that the outlook for higher-throughput untargeted annotation for "known unknowns" is very positive.
- This success rate rises to 70% Top 1 ranks (`MS-FINDER`) and 87% Top 10 ranks (`CFM`) when including metadata.
- The machine learning approaches clearly improve with larger training data sets—the more high quality annotated, open data that is available, the better they will get.
- Developments that focus on the chemistry such as `MS-FINDER` are also essential, especially to cover the cases where no training data is available.
- Despite the above, several challenges remain where the simple combinatorial approach of `MetFrag` and `MAGMa` still performs best.
- Improved incorporation of experimental "metadata" will increase annotation successes further, especially for large candidate sets.
- Challenges for future contests remain true unknowns, assessing the ability of methods to distinguish positional isomers and eventually also stereoisomers.

Finally, a big thank you to all those who participated in CASMI 2016 in any way, shape or form and keep an eye on the CASMI website [1] for future editions.

Availability and requirements

- Project name: CASMI
- Project home page: http://www.casmi-contest.org/
- Operating system(s): Platform independent
- Programming language: Various
- License: N/A
- Any restrictions to use by non-academics: none.

Additional files

Additional file 1. This file contains additional content (methods, results and selected spectra) to complement the manuscript. See PDF for details.

Additional file 2. *Table A1* ESD file used in MS-FINDER version 1.62 for a total of 220,212 compounds. Additional columns for InChIKey, short InChIKey, exact mass, formula, SMILES are not shown here. The use of N/A and a database identifier represents the presence or absence of a compound in a given database. For example, 1,3-butadiyne is only present in ChEBI database (CHEBI:37820). This ESD file was replaced by a dummy file where all HMDB identifiers were modified to dummy identifiers AV001... AV00n and all other identifiers replaced by -1 or N/A. *Table A2*: Formatted ESD file for CASMI 2016 Category 2 Challenge-001. The first 10 compounds from the candidates list for Challenge-001 are listed above. Columns for InChIKey, short InChIKey, PubChem CID, exact mass, formula, SMILES are shown in this table. Databases from BMDB through PubChem are replaced by dummy information.

Abbreviations

CASMI: Critical Assessment of Small Molecule Identification; CSI:IOKR: Compound Structure Identification:Input Output Kernel Regression; MS/MS: tandem mass spectrum; ESI: electrospray ionization; HCD: higher-energy collisional dissociation; LC–MS: liquid chromatography coupled to mass spectrometry; $[M+H]^+$, $[M-H]^-$: protonated and deprotonated molecular ions; SPLASH: SPectraL hASH; MGF: Mascot Generic Format; SMILES: Simplified Molecular Input Line Entry System; InChI, InChIKey: IUPAC International Chemical Identifier and (hash) key; CSV: comma-separated values; MS1: full scan mass spectrum; RRP: relative ranking position; CFM-ID: Competitive

DOI:10.1186/s13321-017-0207-1

Fragmentation Modeling for Metabolite Identification; NIST: National Institute of Science and Technology (USA); HMDB: human metabolome database; ChEBI: Chemical Entity of Biological Interest; CSI:FID: Compound Structure Identification:FingerID; IOKR: Input Output Kernel Regression; (Uni-)MKL: (Uniform) Multiple Kernel Learning; CDK: Chemistry Development Kit; GNPS: Global Natural Products Social Networking; SVM: support vector machine; Q-TOF: Quadrupole Time of Flight; HR: hydrogen rearrangement; GUI: graphical user interface; MoNA: MassBank of North America; ESD: Existing Structure Database; CSIDs: ChemSpider Identifiers; RT: retention time; MINEs: Metabolic *In Silico* Network Expansion Databases; EI-MS: electron ionization mass spectrometry.

Authors' contributions

ES and SN jointly organized Categories 2 and 3 of CASMI 2016; MK selected the challenge compounds and recorded the spectra; ES wrote the majority of the manuscript, SN performed the majority of the automatic evaluation. CR prepared the additional post-contest results. All participants (CB, TK, KD, FA, AV, DV, SB, JR, HS, HT, TS, OF, BG) contributed via their submissions and comments/contributions to the manuscript. All authors read and approved the final manuscript.

Author details

[1] Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland. [2] Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany. [3] Department of Effect-Directed Analysis, UFZ: Helmholtz Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany. [4] Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland. [5] Helsinki Institute for Information Technology, Tekniikantie 14, 02150 Espoo, Finland. [6] West Coast Metabolomics Center and Genome Center, University of California Davis, 451 Health Sciences Drive, Davis, CA 95616, USA. [7] Chair of Bioinformatics, Friedrich-Schiller-University, Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany. [8] Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E9, Canada. [9] Department of Chemistry, University of California Davis, One Shields Avenue, Davis, CA 95616, USA. [10] Metabolomics Expertise Center, Vesalius Research Center (VRC), VIB, KU Leuven – University of Leuven, 3000 Louvain, Belgium. [11] RIKEN Center for Sustainable Resource Science (CSRS), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [12] Department of Biochemistry, Faculty of Sciences, King Abdulaziz University, Jeddah, Saudi Arabia.

Acknowledgements

CASMI is a joint effort and we wish to thank all organizers and participants of previous CASMIs as well as other interested parties for valuable discussions. For this manuscript, AV wishes to acknowledge the contributions of Stephanie N. Samra, Sajjan S. Mehta, Diego Pedrosa and Hiroshi Tsugawa. KD and team acknowledge the contributions of Marvin Meusel. FA and TS acknowledge the contributions of Russ Greiner and David Wishart. MK acknowledges the help of Janek P. Dann with spectral acquisition. We thank the reviewers for their comments and suggestions.

Competing interests

The authors declare that they have no competing interests.

Funding

CR, ES and MK acknowledge funding from the European Commission for the FP7 project SOLUTIONS under Grant Agreement No. 603437. CB, JR and HS acknowledge funding from the Academy of Finland (Grant 268874/MIDAS) and computational resources provided by the Aalto Science–IT project. SB acknowledges funding from the Deutsche Forschungsgemeinschaft (BO 1910/16). SN acknowledges basic institutional funding by the Leibniz Association. FA and team were funded by NSERC, AICML, AIHS, Genome Alberta, CIHR, The Metabolomics Innovation Centre (TMIC) and their work was carried out using the Compute Canada Westgrid facility.

Received: 14 December 2016 Accepted: 13 March 2017
Published online: 27 March 2017

References

1. Neumann S, Schymanski EL (2016) CASMI contest webpage. http://www.casmi-contest.org. Accessed 8 Dec 2016
2. Schymanski EL, Neumann S (2013) CASMI: and the winner is. Metabolites 3(2):412–439
3. Schymanski EL, Neumann S (2013) The Critical Assessment of Small Molecule Identification (CASMI): challenges and solutions. Metabolites 3(3):517–538
4. Nishioka T, Kasama T, Kinumi T, Makabe H, Matsuda F, Miura D, Miyashita M, Nakamura T, Tanaka K, Yamamoto A (2014) The winner of CASMI 2013 is... Mass Spectrom 3(Special Issue 2), 0039
5. Nikolic D, Jones M, Sumner L, Dunn W (2017) CASMI2014: challenges, solutions and results. Current Metab. doi:10.2174/2213235X04666160617113437
6. Genta-Jouve G, Thomas OP, Touboul D, Schymanski EL, Neumann S (2016) CASMI 2016: Category 1: Natural products. http://www.casmi-contest.org/2016/results-cat1.shtml. Accessed 20 Mar 2017
7. Neumann S, Schymanski EL (2016) CASMI contest rules and evaluation. http://www.casmi-contest.org/2016/rules.shtml. Accessed 8 Dec 2016
8. Stravs MA, Schymanski EL, Singer HP, Hollender J (2013) Automatic recalibration and processing of tandem mass spectra using formula annotation. J Mass Spectrom 48(1):89–99
9. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, Schymanski EL, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, Fiehn O (2016) SPLASH: The SPectraL HaSH Identifier. http://splash.fiehnlab.ucdavis.edu/. Accessed 8 Dec 2016
10. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, EL Tomáš Schymanski, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, Fiehn O (2016) SPLASH, a hashed identifier for mass spectra. Nat Biotechnol 34(11):1099–1101
11. CASMI2016 Mass Spectra. http://massbank.eu/MassBank/jsp/Result.jsp?type=rcdidx&idxtype=site&srchkey=36. Accessed 12 Dec 2016
12. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminf 8(1):1
13. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminf 3:33
14. Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. J Mass Spectrom 48(3):291–298. doi:10.1002/jms.3123
15. Neumann S, Schymanski EL (2016) CASMI contest challenges. http://www.casmi-contest.org/2016/challenges-cat2+3.shtml. Accessed 8 Dec 2016
16. Meusel M, Hufsky F, Panter F, Krug D, Möller R, Böcker S (2016) Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. Anal Chem 88(15):7556–7566. doi:10.1021/acs.analchem.6b01015
17. Formula One Scoring Systems. https://en.wikipedia.org/wiki/List_of_Formula_One_World_Championship_points_scoring_systems. Accessed 8 Dec 2016
18. Allen F, Greiner R, Wishart D (2014) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics. doi:10.1007/s11306-014-0676-4
19. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27:747–751
20. NIST, EPA, NIH: NIST Mass Spectral Library 2014 Edition. U.S. Secretary of Commerce, USA
21. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorndahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A (2013) HMDB 3.0—the Human Metabolome Database in 2013. Nucleic Acids Res 41(D1):D801–D807
22. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36(Suppl 1):344–350

DOI:10.1186/s13321-017-0207-1

23. Wishart DS (2016) FooDB. http://foodb.ca/. Accessed 8 Dec 2016
24. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(Suppl 1):901–906
25. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci. doi:10.1073/pnas.1509788112. http://www.pnas.org/content/early/2015/09/16/1509788112.full.pdf
26. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J (2016) Fast metabolite identification with input output kernel regression. Bioinformatics 32(12):28–36. doi:10.1093/bioinformatics/btw246. http://bioinformatics.oxfordjournals.org/content/32/12/i28.full.pdf+html
27. Böcker S, Dührkop K (2016) Fragmentation trees reloaded. J Cheminform 8:5. doi:10.1186/s13321-016-0116-8
28. Shen H, Dührkop K, Böcker S, Rousu J (2014) Metabolite identification through multiple kernel learning on fragmentation trees. Bioinformatics 30(12):157–164
29. Cortes C, Mehryar M, Rostamizadeh A (2012) Algorithms for learning kernels based on centered alignments. J Mach Learn Res 13:795–828
30. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the Chemistry Development Kit (CDK)—an open-source java library for chemo- and bioinformatics. Curr Pharmaceut Des 12(17):2111–2120
31. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): an open-source java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43(2):493–500
32. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T et al (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat Biotechnol 34(8):828–837
33. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45:703–714
34. Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Schölkopf B (eds) Advances in large margin classifiers, vol 5. MIT Press, Cambridge
35. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M (2016) Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Anal Chem 88(16):7946–7958
36. Tsugawa H et al (2016) MS–FINDER. http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/index.html. Accessed 8 Dec 2016
37. NIST MS Search GUI. http://chemdata.nist.gov/. Accessed 8 Dec 2016
38. MassBank of North America. http://mona.fiehnlab.ucdavis.edu/. Accessed 8 Dec 2016
39. Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, Akiyama K, Sakurai T, Matsuda F, Aoki T et al (2012) RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. Phytochemistry 82:38–45
40. LfU: Bayerisches Landesamt für Umwelt: STOFF-IDENT (login Required). http://bb-x-stoffident.hswt.de/stoffidentjpa/app. Accessed 13 June 2016
41. NORMAN Association: NORMAN Suspect List Exchange. http://www.norman-network.com/?q=node/236. Accessed 8 Dec 2016
42. Verdegem D, Lambrechts D, Carmeliet P, Ghesquière B (2016) Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization. Metabolomics 12(6):1–16. doi:10.1007/s11306-016-1036-3
43. Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RCH, van Schaik R, Vervoort J (2012) Substructure-based annotation of high-resolution multistage MSn spectral trees. Rapid Commun Mass Spectrom 26(20):2461–2471. doi:10.1002/rcm.6364
44. MAGMa+. https://github.com/savantas/MAGMa-plus. Accessed 8 Dec 2016
45. MetFrag Command Line. http://c-ruttkies.github.io/MetFrag/projects/metfragcl/. Accessed 8 Dec 2016
46. Royal Society of Chemistry: ChemSpider. http://www.chemspider.com/
47. Interactive Heat Map of CASMI 2016 Challenges Negative Mode. http://www.casmi-contest.org/2016/heatmapNegCat2.html. Accessed 8 Dec 2016
48. Interactive Heat Map of CASMI 2016 Challenges Positive Mode. http://www.casmi-contest.org/2016/heatmapPosCat2.html. Accessed 8 Dec 2016
49. McEachran AD, Sobus JR, Williams AJ (2016) Identifying "known unknowns" using the US EPA's CompTox Chemistry Dashboard. submitted
50. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, Hanson AD, Fiehn O, Tyo KE, Henry CS (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J Cheminform 7(1):1
51. Böcker S, Rousu J, Schymanski E (2016) Computational metabolomics (Dagstuhl Seminar 15492). Dagstuhl Rep 5(11):180–192. doi:10.4230/DagRep.5.11.180
52. Schymanski EL, Gallampois CMJ, Krauss M, Meringer M, Neumann S, Schulze T, Wolf S, Brack W (2012) Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. Anal Chem 84:3287–3295
53. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS (2016) CERAPP: Collaborative estrogen receptor activity prediction project. J Environ Health Perspect 124(7):1023–1033
54. CASMI 2014 challenges. http://www.casmi-contest.org/2014/results-cat2.shtml. Accessed 8 Dec 2016
55. Kerber A, Laue R, Meringer M, Rücker C, Schymanski E (2014) Mathematical chemistry and chemoinformatics: structure generation, elucidation and quantitative structure–property relationships. Walter de Gruyter, Berlin

Data Sharing and Data Standards

9

My contributions in the area of Data Sharing in metabolomics include developer feedback and implementing support for these formats in e.g. the metabolite profiling tools and MassBank **[MCS10, HRS19]**, [DCN11, CMB12, GJS14]. I implemented support for the formats in e.g. the metabolite profiling tools and worked on the implementation of data models together with students [KN06, GN07, Neu07].

With Daniel Schober (PostDoc in my group) we contributed best practices in creation of ontologies [SMM13, SWJ14] and coordinated the development of a data standard for NMR **[SJW18]**.

In the area of experimental metadata, I contributed early feedback during the development of ISA-Tab with example datasets, feedback and software patches [RSBM10, GBNM14] which is used in the MetaboLights repository **[HSC13]**, and since many years I am the most active submitter. I prepared and uploaded MTBLS2, MTBLS10, MTBLS74, MTBLS160, MTBLS169, MTBLS188, MTBLS291, MTBLS297, MTBLS338, MTBLS341, MTBLS381, MTBLS389, MTBLS441, MTBLS433, MTBLS544, MTBLS671, MTBLS687 and MTBLS1430. This expertise was important for leading work packages in the EU FP7 project COSMOS [SNS15] and the H2020 project PhenoMeNal [PWW18]. For several years I am contributing to the development of mzTab [GJS14, HRS19], a standard for reporting of results in proteomics and metabolomics. For reference spectra, we co-developed the **SP**ectra**L** ha**ASH** (SPLASH), a data-driven identifier for spectra **[WMM16]**.

With this expertise, I was active in standards-developments and promotion [KFS10, SRSF12], and coordinated a review paper on practical standards in metabolomics [RSSA16]. These efforts continue within the German de.NBI network and the European ELIXIR infrastructure [vRBC17].

⌘ *Author's Choice*

© 2011 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

mzML—a Community Standard for Mass Spectrometry Data*

Lennart Martens‡§, Matthew Chambers¶, Marc Sturm‖, Darren Kessner, Fredrik Levander‡‡, Jim Shofstahl§§, Wilfred H. Tang¶¶, Andreas Römpp‖‖, Steffen Neumann,[a] Angel D. Pizarro,[b] Luisa Montecchi-Palazzi,[c] Natalie Tasman,[d] Mike Coleman,[e] Florian Reisinger,[c] Puneet Souda,[f] Henning Hermjakob,[c] Pierre-Alain Binz,[g] and Eric W. Deutsch[h,i]**

Mass spectrometry is a fundamental tool for discovery and analysis in the life sciences. With the rapid advances in mass spectrometry technology and methods, it has become imperative to provide a standard output format for mass spectrometry data that will facilitate data sharing and analysis. Initially, the efforts to develop a standard format for mass spectrometry data resulted in multiple formats, each designed with a different underlying philosophy. To resolve the issues associated with having multiple formats, vendors, researchers, and software developers convened under the banner of the HUPO PSI to develop a single standard. The new data format incorporated many of the desirable technical attributes from the previous data formats, while adding a number of improvements, including features such as a controlled vocabulary with validation tools to ensure consistent usage of the format, improved support for selected reaction monitoring data, and immediately available implementations to facilitate rapid adoption by the community. The resulting standard data format, mzML, is a well tested open-source format for mass spectrometer output files that can be readily utilized by the community and easily adapted for incremental advances in mass spectrometry technology. ***Molecular & Cellular Proteomics 10: 10.1074/mcp.R110.000133, 1–7, 2011.***

Mass spectrometry (MS)[1] has recently emerged as a major discovery tool in the life sciences (1). This analytical technique is used to analyze the molecular composition of a biological sample by ionizing the sample or analyte molecules and then measuring the mass-to-charge ratios of the resulting ions. The data from an MS experiment consist of mass spectra that are used to identify, characterize, and quantify the abundance of the molecules of interest. The resulting MS spectra, along with their associated metadata (*e.g.* experimental protocol, MS instrumentation, operational parameters, etc.), are then semi-automatically processed by specialized software packages to identify or quantify the sampled ions. The inherent variability introduced by using different instruments, instrument software, and experimental conditions, however, affects the downstream ability to analyze, integrate, and compare data sets originating from different MS experiments.

Indeed, with the ever-increasing use of mass spectrometry, two issues have arisen in terms of handling MS data: (i) the necessity to share data throughout the scientific community in order to facilitate integration and comparison (2), and (ii) the importance of utilizing open and readily accessible standard formats that verifiably capture a consistent amount of crucial information. The importance of addressing these issues has been further emphasized in prominent journal editorials (3–4). Data repositories have since been created to allow data to be shared, including Tranche (5), GPMDB (6), PRIDE (7), and PeptideAtlas (8), among others (9), and various proposed standard formats for MS data (10–14) were developed. Other formats such as JCAMP-DX

From the ‡Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium §Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium ¶Vanderbilt University, Nashville, TN, 37232, USA ‖Eberhard Karls University, 72074, Tübingen, Germany **University of Southern California, Los Angeles, CA, 90089, USA ‡‡Department of Immunotechnology and CREATE Health, Lund University, 22362, Lund, Sweden §§Thermo Fisher Scientific, San Jose, CA, 95134, USA ¶¶Agilent Technologies, Santa Clara, CA, 95051, USA ‖‖Justus Liebig University, 35390 Giessen, Germany [a]Leibniz Institute of Plant Biochemistry, 06120 Halle, Germany [b]University of Pennsylvania, Philadelphia, PA, 19104, USA [c]EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB101SD, UK [d]Insilicos LLC, Seattle, WA, 98109, USA [e]Stowers Institute, Kansas City, MO, 64110, USA [f]University of California, Los Angeles, Los Angeles, CA, 90095, USA [g]Geneva Bioinformatics (GeneBio) SA, 1206 Geneva, Switzerland and Swiss Institute of Bioinformatics, Geneva, Switzerland [h]Institute for Systems Biology, Seattle, WA, 98103, USA

⌘ *Author's Choice*—Final version full access.

Received, April 29, 2010 and in revised form, July 26, 2010

Published, MCP Papers in Press, August 17, 2010, DOI 10.1074/mcp.R110.000133

[1] The abbreviations used are: MS, mass spectrometry; HUPO, Human Proteome Organization; PSI-MS, Proteomics Standards Initiative working group for mass spectrometry standards; LC-MS/MS, liquid chromatography-tandem mass spectrometry; CV, controlled vocabulary.

This is an open access article under the CC BY license.

Want to cite this article? Please look on the last page for the proper citation format.

DOI:10.1074/mcp.R110.000133

(http://www.acornnmr.com/JCAMP.htm; www.jcamp.org), which was designed for IR spectrometry and adapted to NMR and mass spectrometry, and NetCDF are quite variably implemented, difficult to validate, and cannot encode extensive metadata in a standard fashion and therefore have not gained much use for proteomics applications and other complex MS analyses. Analytical Information Markup Language (AnIML; http://animl.sourceforge.net/), which aims to encompass several analytical platforms, including eventually mass spectrometry, is still being designed. For mass spectrometry-based proteomics workflows, mzXML (13) and mzData (14) have been the most widely used open formats for several years.

However, each of these initial efforts to develop an open, vendor-neutral XML data format to store MS information was undertaken with a different underlying purpose. One format, mzData, was developed by HUPO-PSI as a data exchange and archive standard (14, 15), and was implemented as such in PRIDE (16). The other format, mzXML, was developed at the Institute for Systems Biology in an effort to streamline their data processing software (17), and became a popular *de-facto* standard format. These two formats also differed in their underlying philosophies regarding flexibility. mzData utilized a controlled vocabulary that could be frequently updated as the technology advanced. In contrast, mzXML had a strict schema that used enumerated attributes to describe the auxiliary information, such that support for new annotations required revisions to the schema and software updates.

Although each of the proposed formats satisfied the requirements of openness and accessibility, the multiplicity of the formats proved to be confusing and distracting to scientists and computer programmers alike. In order to resolve this situation, the teams that developed mzData and mzXML, along with many other researchers and developers from academia, industry, and vendors joined forces in the Human Proteome Organization (HUPO) Proteomics Standards Initiative working group for mass spectrometry standards (PSI-MS), and set out to create a single MS data standard that would build on the strengths of the previous efforts. The challenge in creating the new unified output format, called mzML, was therefore the resolution of the opposing philosophies of mzXML and mzData, while retaining the best technical attributes of these two formats.

History—In 2006, the unification process was initiated at a PSI workshop based on the guiding design principles determined by members representing instrument and software vendors, data repositories, end users, and the teams that built the mzXML and mzData standards. The designers of mzML focused on four key objectives: (i) creation of a simple format, (ii) elimination of alternate ways to encode the same information, (iii) support for all the features of both mzXML and mzData, and (iv) validation through implementation prior to release. Taken together, these goals would lead to a single unified format that could support the current capabilities of mzXML and mzData and that could be easily supported by vendors and current software, with further enhancements to be considered in future releases. In order to facilitate swift adoption and uniform implementation of the new standard format, the participants of PSI-MS also created open source tool sets that enabled developers as well as end users to immediately pick up the format without having to write their own software.

Progress on the format was made at regular PSI workshops as well as special workshops dedicated to mzML. In June 2008, the mzML 1.0 standard format was released (18, 19). However, despite the rather rigorous review process (20), several shortcomings became apparent as vendors quickly moved to implement the new format, most notably insufficient support for precursor ion scans and neutral loss scans, and a severe file size inflation problem for Selected Reaction Monitoring runs (all of which represented novel features that had been absent from the precursor formats). These deficiencies, along with several other minor issues, were remedied by the PSI-MS working group in collaboration with the implementers that had detected the issues. As a result, mzML version 1.1.0 was released in June 2009, with the expectation that this new version will remain stable for quite some time.

Design—In addition to incorporating the best technical attributes of the predecessor formats, several key innovations were introduced in mzML. First, in order to support new hybrid instruments such as the LTQ Orbitrap and LTQ FT, mzML can specify multiple operational configurations for an instrument, and link individual spectra to a specific configuration. Another new feature is the ability to capture Selected Reaction Monitoring data efficiently, through the newly introduced chromatogram elements. More detailed improvements are also found in mzML, such as the ability to encode isolation window size, enabling gas phase fractionation/MS^e data to be correctly annotated, and accommodating the presence of multiple precursor ions within a typical liquid chromatography (LC)-MS/MS isolation window (21). Associated with mzML comes a rich, schema-linked controlled vocabulary (CV) that allows accurate and unambiguous annotation of metadata. In addition, mzML comes with a set of semantic validation rules. These rules are encoded in a mapping XML document according to the PSI Validator framework (22)(see http://www.psidev.info/validator) and have been implemented in two independent mzML validator applications (see http://www.psidev.info/index.php?q=node/390).

The full technical details of the mzML standard are available online, together with complete specification documentation, graphical depictions of its structure, and various example files at http://www.psidev.info/index.php?q=node/257. Next we will highlight the primary technical aspects of the mzML standard and discuss current implementations.

All of the information from a single MS run, including the spectra and associated metadata, is contained within the

DOI:10.1074/mcp.R110.000133

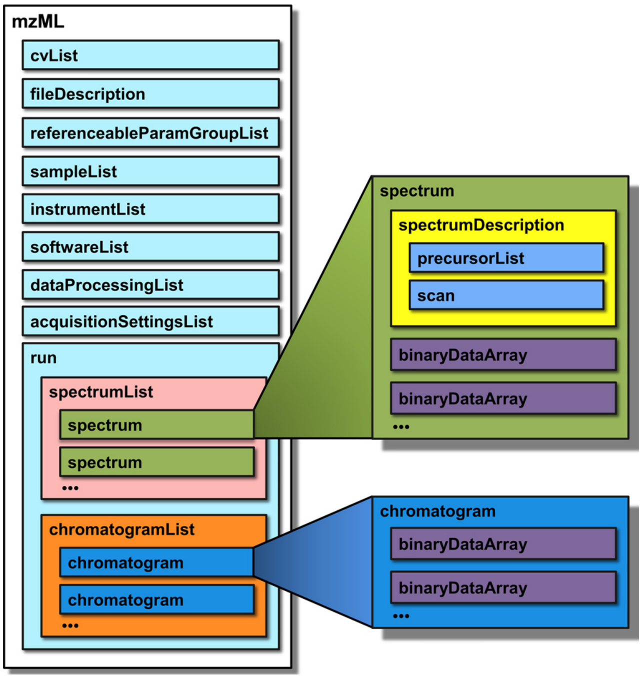


FIG. 1. **A schematic representation of mzML, showing key elements of the format.** Each rectangle represents an XML element. See main text for a full description.

mzML file. Like its predecessors, mzML is encoded in XML. An XML schema definition defines the format structure, and many industry-standard tools are readily available to validate whether an XML document conforms to its XML schema definition.

The overall mzML file structure (Fig. 1) is as follows (elements presented top-to-bottom): <cvList> contains information about the controlled vocabularies referenced in the rest of the mzML document; <fileDescription> contains basic information on the type of spectra contained in the file; <referenceableParamGroupList> is an optional element that of groups of controlled vocabulary terms that can be referenced as a unit throughout defines a list the document; <sampleList> can contain information about samples that are referenced in the file; <instrumentConfigurationList> contains information about the instrument that generated the run; <softwareList> and <dataProcessingList> provide a history of data processing that occurred after the raw acquisition; <acquisitionSettingsList> is an optional element that stores special input parameters for the mass spectrometer, such as inclusion lists. These elements are followed by the acquired spectra and chromatograms. Both spectral and chromatographic data are represented by binary format data encoded into base 64 strings, rather than human-readable ASCII text for enhanced fidelity and efficiency when dealing with profile data. This design choice does not enjoy unanimous approval, but has been agreed upon by the majority of designers.

In order to enable fast access to the file, mzML was designed with a standardized but optional mechanism for random access indexing, in the same way as mzXML. This enables programs to directly locate a specific spectrum within the file during processing, rather than having to read the file sequentially. Although there is debate about whether to include a random access index because of the possibility of index corruption, years of experience with mzXML have demonstrated that these problems are rare and are outweighed by the benefits of having an index. To compensate for the possibilities of an error in the index, reader software can easily be written to verify the offsets and automatically rebuild the index if there is an error. To make the index completely optional, mzML was designed so that the primary document does not have an index, but the document can still be enclosed in a wrapper schema that has an index. Thus, an mzML file may contain either a plain or indexed mzML document and reader software is designed to handle either case transparently.

Finally, although the open and standardized XML formatting provides clear advantages, it also implicitly requires a certain verbosity that enlarges the size of the data files by as much as a factor of 10 for profile-mode spectra without compression when compared with the original raw files. However, enabling in-line zlib compression typically reduces the files by a factor of 2 below the uncompressed form. Further, because any remaining size increase is primarily because of the presence of XML tags, standard and commonly used compression tools can be used to reduce this size overhead for storage and transport of mzML files. Compression factors for mzML files can vary by compression algorithm, but GZIP, zip, and 7zip (all freely available compression algorithms, supported on many platforms) provide size reductions of another factor of 2, thus essentially offsetting the size increase initially seen in uncompressed files. Profile mode spectra often undergo peak picking (if desired by the user) during conversion to mzML and therefore lead to smaller files than the original. A typical ion trap file with already centroided peaks in the original file becomes 1.8× larger with mzML without compression, or just 1.3× larger when using in-line zlib compression, and 0.5× (*i.e.* half as large as the original) when using both in-line compression and total file compression with the gzip algorithm. Some applications are able to work directly with gzipped mzML files, thus providing an overall savings in disk space, assuming the original files are archived elsewhere. Nonetheless, overall additional disk space costs associated with using standard formats are typically much lower than the human costs associated with trying to work with multiple proprietary formats.

Controlled Vocabulary—In an effort to prevent encoding the same information in slightly different ways and to provide support for new technologies with mzML, we have designed the format to encode most of the metadata in <cvParam> elements, which provide a reference to a specific concept within the PSI MS controlled vocabulary (CV). These CV terms have explicit and detailed definitions, including the data type and type of units required. The controlled vocabulary is adjustable and new CV terms can be added without modification of the mzML schema. Whenever an implementer requires a new term to describe a new concept, the proposed term and definition can be mailed to the PSI-MS vocabulary list (psidev-ms-vocab@lists.sourceforge.net), where the addition can be discussed and then added to the CV within hours or days. Additionally, other CVs can also be used to annotate specific elements; the NEWT ontology for species can for instance be used to annotate the sample.

Semantic Validation—To enforce the use of CV terms, a semantic validator was released with the data format. Semantic validation provides a simple yet powerful means to assess the completeness and semantic correctness of the metadata in an mzML file, automatically spotting errors such as the absence of a required binary array type annotation, the incorrect annotation of an ionization source with a detector-specific CV term, or the use of two conflicting CV terms where only one can be valid at a time. We have made the validator available as a webpage with file uploading, or as a standalone tool for local validation. Furthermore, because the mzML format is designed to support full MIAPE (10, 23) compliance, automated semantic analyses can be carried out *a priori* by any consumer of an mzML data file to ascertain the presence of the required minimal information. An additional benefit is that the metadata can be customized for different types of data, so that different types of spectra can be encoded using the same tags, but with different metadata.

mzML Development Process—Development of the mzML format has followed the overall HUPO PSI community standard development process, which in turn is largely based on the highly successful open source software development model. A centralized group of core volunteers takes care of coordinating the efforts of the many enthusiastic community members that contribute their time and expertise at different times, and a full record of the entire process is maintained through an online mailing list that is directly accessible to all. This development model has been proven to be (perhaps paradoxically) extremely robust as compared with more tightly organized and coordinated projects. Indeed, even though the core development team of mzML has changed substantially over the years, this never impacted the development of the standard proper.

The mzML standard is furthermore deemed quite future proof as it has been developed with change in mind. The required flexibility of the format comes primarily from its mixed structure—certain aspects of the data are rather rigidly defined in the XML format specification, such as the necessity to include an instrument description. Yet the actual form that this description takes is quite open, and not defined by the XML schema. As an example, consider the "source" element for an instrument. The different types of sources are defined solely through controlled vocabulary parameters, and if a new source is invented tomorrow, a simple update to the CV will automatically enable mzML files to communicate the use of this new source. Furthermore, because CV terms are linked through defined relationships, this new source term will be immediately recognizable to existing software as describing a source, because it will have an "is a" relationship to the metaterm "ion source." This approach is employed in virtually every element in mzML, making the format extremely flexible without requiring any updates to either XML schema or software parsers. Changes need thus only be made to the CV, which is a simple text file that is made available in a version control system online, and that can be updated and read on-the-fly. Indeed, because the first public release of mzML, numerous updates have already been introduced to the controlled vocabulary without effecting any downstream changes on the XML schema or the existing software.

Implementations—Because of the broad community participation in PSI-MS, there are several implementations of the mzML format in software tools, legacy data converters, and programming libraries for a variety of languages (see http://www.psidev.info/index.php?q=node/257 for a current summary). In fact, the wide variety of software that uses mzML continues to grow and is one of the strengths of mzML. The ProteoWizard software project (24–25) has provided the framework for testing and reference implementation of mzML in its final stages of development. It consists of a set of open-source, cross-platform tools and libraries written in C++ for proteomic data analyses. The libraries provide a well-tested framework that unifies data file access and performs standard chemistry and LCMS dataset computations, making ProteoWizard an ideal library to include in any software project that needs to add mzML read or write support. ProteoWizard is available under a very permissive license, which allows the library to be used in commercial software without affecting the license terms of that software. The ProteoWizard library is already used by several unrelated software projects to provide mzML support. The Proteowizard "msconvert" tool can convert many different vendor formats to mzML, as well as convert mzXML files into mzML.

OpenMS (26), an open-source C++ library for mass spectrometry, also provides classes for reading and writing mzML which can be easily integrated in other software tools. Additionally, it supports both XSD validation and semantic validation of mzML files. This functionality of OpenMS was used to implement an off-line tool for validation of mzML files which is part of TOPP - The OpenMS Proteomics Pipeline (27). Similarly, the NCBI C++ toolkit and the jmzML Java toolkit (28) provide libraries for reading and writing mzML. Because these libraries are already available to simplify addition of mzML support, several software applications are already being distributed with mzML 1.1 support. These include search engines and postprocessing software such as X!Tandem (29), Myrimatch (30), the Trans-Proteomic Pipeline (TPP) (31–33), and the Proteios Software Environment (34). Most vendors

DOI:10.1074/mcp.R110.000133

have committed to provide mzML support in the next release of their software.

The widespread support for mzML in existing, commercial tools, along with the availability of several production-grade open source software packages and libraries in a variety of programming languages, ensures that data encoded in the mzML format is readily accessible to any interested end user or software developer.

Example Usages—Because the main advantages of open data standards over closed, proprietary formats are interoperability and portability, we have chosen two corresponding use cases in the field of mass spectrometry-based proteomics to illustrate some of the usages of the mzML data standard. First, many laboratories employ multiple instruments from different vendors for their analyses. Although this heterogeneity in instrumentation confers the important advantage of providing complementary strengths of the different machines, it also creates a logistical problem at the level of data processing. The various proprietary data formats employed by each instrument to report its data, are essentially tied to these specific instruments - even different models from the same vendor can deliver incompatible output files. As a result, the development of software that can operate on data from any instrument, such as the tools in the Trans-Proteomic Pipeline, becomes quite difficult indeed. This in fact was one of the main reasons why the original mzXML format was developed as part of the Trans-Proteomic Pipeline: to unify the various vendor formats in a common, open data structure that maintains sufficient amounts of data to reliably support various kinds of downstream processing, including identification and quantification of proteins. As a direct descendant from mzXML, mzML provides these same benefits, allowing data from many instruments to be transformed (using the freely available ProteoWizard or TPP tools) into the common mzML format, which is in turn read and interpreted homogenously by all downstream data processing software applications.

A second important use case of standard data formats concerns the dissemination of data to the wider scientific community, an endeavor that is very deeply ingrained in the life sciences (35). If data were disseminated in proprietary formats, three problems would occur (discussed in detail in (36)): (i) referees wishing to evaluate (privately) deposited data during peer review would have difficulties accessing, interpreting, and validating the data and derived conclusions unless they happened to own the same instruments and software compatible with the format, (ii) after publication of the data, interested consumers would face similar difficulties in accessing and processing the data, and (iii) over a relatively short time span, all data would become unreadable, as the required vendor-specific software will no longer be supported or available. By employing an open, XML (and therefore ultimately text-based) format such as mzML, these three key issues are implicitly circumvented.

Both of these examples, of course, rely on the availability of software supporting the format, but as can be seen from the previous section, many actively supported free and open source implementations in a variety of programming languages and for a variety of platforms are already available for mzML today, and many other implementations are underway or will be available with their next software release. Finally, it should be noted that the two use cases are in fact connected: by switching to mzML as the format for within-lab data processing and analysis, the step to disseminate in mzML becomes effectively trivial.

Integration with Other Standards in the Life Sciences—The data accommodated by mzML will most likely not stand alone in a modern-day workflow. Preceded by sample treatment and sample separation (often through chromatography), mass spectrometry data is then usually further processed to identify or quantify the recorded signals. As such, it is important to note that HUPO PSI has also released standards for protein separation including gel based and column chromatography based methods (http://www.psidev.info/index.php?q=node/83), for identification of molecules from mass spectra (http://www.psidev.info/index.php?q=node/319), and for the annotation of modifications on proteins (http://www.psidev.info/index.php?q=node/319). Furthermore, the overall integration of standardized data and metadata across domains in the life sciences is being actively undertaken by the Reporting Structure for Biological Investigations (RSBI) working group of the MGED Society (http://www.mged.org), which has culminated in the ISA-TAB format (37). Minimal information assurance in all the relevant formats on the other hand is coordinated through the MIBBI project (38).

CONCLUSION

In 2009, three years after its conception, mzML 1.1 was released and has proven to be a solid format that can easily accommodate incremental advances in mass spectrometry technology, while providing a good foundation for extension to accommodate encoding of data from new technologies. An existing set of software libraries that support mzML will enable quick adoption of the format. However, because the precursor formats are also highly capable, the incentive to migrate existing workflows is low, and the adoption of mzML in practice will be gradual. An initial wave of implementations necessitated a revision of 1.0 to 1.1, but since the release of 1.1, there have not been any significant changes necessary. It is therefore expected that 1.1 will remain stable for quite some time. The involvement of instrument vendors in PSI-MS further ensures that mzML export will become available on instrument software by default.

Like all PSI standards, mzML 1.1 has gone through a formal review process called the PSI document process (20), which consists of three review periods managed by the PSI Editor: an internal review, an external review by invited experts, and a public review stage. As such, we believe that mzML 1.1 can

DOI:10.1074/mcp.R110.000133

now readily be utilized by the community at large, providing a single, open, and accessible community standard format for mass spectrometer output files. With CV annotations, semantic validation, and MIAPE compliance as part of the design of the standard, unambiguous reporting of metadata will thus become standard practice, ensuring that mzML can be used as a highly reliable data exchange format. The PSI-MS working group will meanwhile continue to refine the controlled vocabulary and coordinate software development surrounding mzML to ensure that mzML stays up-to-date with the progress of the field.

Acknowledgments—We would like to thank all individuals who have contributed in PSI meetings, mailing lists, and elsewhere with ideas, comments, and implementation efforts on mzML, especially Ruedi Aebersold, Rolf Apweiler, Ron Beavis, Benito Cañas, Mike Coleman, David Creasy, Eva Duchoslav, Jimmy Eng, Jayson Falkner, Lola Gutierrez, Jari Häkkinen, David Horn, Phil Jones, Marius Kallhardt, Jim Langridge, Kent Laursen, Parag Mallick, Ruth McNally, Alberto Medina, Luis Mendoza, Lars Nilse, Erik Nilsson, Sandra Orchard, Norman Paton, Patrick Pedrioli, Rune Philosof, Brian Pratt, Howard Read, Sean Seymour, David Shteynberg, David Sparkman, Chris Taylor, and Trish Wheztel. We thank Julie Bletz, Terry Farrah, and Gordon Sun for assistance with preparation of the manuscript. L.M. and F.R. would like to thank Rolf Apweiler for his support, and L.M. would additionally like to thank Joël Vandekerckhove for support.

* L.M. is supported by the "ProDaC" grant LSHG-CT-2006-036814 of the European Union, F.R. by the Wellcome Trust [grant number WT085949MA], and E.W.D. by the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179 and by the Duchy of Luxembourg.

We declare the following competing financial or commercial interests:

Several authors work for commercial entities whose software supports the mzML format.

Author contributions: E.W.D. is the chair, P.A.B. is the co-chair, and L.M. is the secretary of PSI-MS WG. All authors actively contributed to the creation and implementation of the standard format. All authors have agreed to all the content in the manuscript, including the data as presented.

[i] To whom correspondence should be addressed: Institute for Systems Biology, 1441 N 34th St, Seattle, WA 98103, E-mail: edeutsch@systemsbiology.org.

REFERENCES

1. Editors (2007) Mind the technology gap. *Nat. Methods* **4,** 765
2. Prince, J. T., Carlson, M. W., Wang, R., Lu, P., and Marcotte, E. M. (2004) The need for a public proteomics repository. *Nature Biotechnology* **22,** 471–472
3. Editors, (2008) Thou shalt share your data. *Nat. Methods* **5**, 209
4. Editors, (2007) Democratizing proteomics data. *Nat Biotechnol* **25,** 262
5. Falkner, J. A., and Andrews, P. C. (2007) Tranche: Secure Decentralized Data Storage for the proteomics community. *Journal of Biomolecular Techniques* **18,** 3
6. Craig, R., Cortens, J. P., and Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res* **3,** 1234–1242
7. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005) PRIDE: the proteomics identifications database. *Proteomics* **5,** 3537–3545
8. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res* **34**, D655–658.
9. Mead, J. A., Bianco, L., and Bessant, C. (2009) Recent developments in public proteomic MS repositories and pipelines. *Proteomics* **9,** 861–881
10. Taylor, C. F., Binz, P. A., Aebersold, R., Affolter, M., Barkovich, R., Deutsch, E. W., Horn, D. M., Huhmer, A., Kussmann, M., Lilley, K., Macht, M., Mann, M., Muller, D., Neubert, T. A., Nickson, J., Patterson, S. D., Raso, R., Resing, K., Seymour, S. L., Tsugita, A., Xenarios, I., Zeng, R., and Julian, R. K., Jr. (2008) Guidelines for reporting the use of mass spectrometry in proteomics. *Nat Biotechnol.* **26,** 860–861
11. McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J., Venable, J., Graumann, J., Johnson, J. R., Cociorva, D., and Yates, J. R., 3rd (2004) MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom* **18**, 2162–2168
12. Orchard, S., Montechi-Palazzi, L., Deutsch, E. W., Binz, P. A., Jones, A. R., Paton, N., Pizarro, A., Creasy, D. M., Wojcik, J., and Hermjakob, H. (2007) Five years of progress in the Standardization of Proteomics Data 4(th) Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23–25, 2007 Ecole Nationale Superieure (ENS), Lyon, France. *Proteomics* **7,** 3436–3440
13. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **22,** 1459–1466
14. mzData, *http://psidev.info/index.php?q=node/80#mzdata.*
15. Orchard, S., Zhu, W., Julian, R. K., Jr., Hermjakob, H., and Apweiler, R. (2003) Further advances in the development of a data interchange standard for proteomics data. *Proteomics* **3,** 2065–2066
16. Jones, P., Cote, R. G., Martens, L., Quinn, A. F., Taylor, C. F., Derache, W., Hermjakob, H., and Apweiler, R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res* **34**, D659–663
17. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1,** 2005.0017
18. Deutsch, E. (2008) mzML: a single, unifying data format for mass spectrometer output. *Proteomics* **8,** 2776–2777
19. Deutsch, E. W. (2010) Mass spectrometer output file format mzML. *Methods Mol. Biol*. **604,** 319–331
20. Vizcaino, J. A., Martens, L., Hermjakob, H., Julian, R. K., and Paton, N. W. (2007) The PSI formal document process and its implementation on the PSI website. *Proteomics* **7,** 2355–2357
21. Luethy, R., Kessner, D. E., Katz, J. E., Maclean, B., Grothe, R., Kani, K., Faca, V., Pitteri, S., Hanash, S., Agus, D. B., and Mallick, P. (2008) Precursor-ion mass re-estimation improves peptide identification on hybrid instruments. *J Proteome Res* **7,** 4031–4039
22. Montecchi-Palazzi, L., Kerrien, S., Reisinger, F., Aranda, B., Jones, A. R., Martens, L., and Hermjakob, H. (2009) The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics* **9,** 5112–5119
23. Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., Julian, P. K., Jr., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J., Leitner, A.,Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., 3rd, and Hermjakob, H. (2007) The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol.* **25,** 887–893
24. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–2536
25. ProteoWizard, *http://proteowizard.sourceforge.net.*
26. Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) OpenMS-An open-source software framework for mass spectrometry. *BMC Bioinformatics* **9**, 163
27. Kohlbacher, O., Reinert, K., Gropl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007) TOPP-The OpenMS proteomics pipeline.

DOI:10.1074/mcp.R110.000133

Bioinformatics **23,** e191–197

28. Cote, R. G., Reisinger, F., and Martens, L. (2010) jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* **10**, 1332–1335
29. Bjornson, R. D., Carriero, N. J., Colangelo, C., Shifman, M., Cheung, K. H., Miller, P. L., and Williams, K. (2008) X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *J Proteome Res* ***7,*** 293–299
30. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **6,** 654–661
31. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1,** 2005.0017
32. Pedrioli, P. G. (2010) Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol Biol* **604,** 213–238
33. 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., and Aebersold, R. (2010) A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10,** 1150–1159
34. Hakkinen, J., Vincic, G., Mansson, O., Warell, K., and Levander, F. (2009) The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J Proteome Res* **8,** 3037–3043
35. Vizcaíno, J. A., Mueller, M., Hermjakob, H., and Martnes, L. (2009) Charting online OMICS resources: a navigational chart for clinical researchers. *Proteomics Clinical Applications* **3,** 18–29
36. Martens, L., Nesvizhskii, A. I., Hermjakob, H., Adamski, M., Omenn, G. S., Vandekerckhove, J., and Gevaert, K. (2005) Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* **5,** 3501–3505
37. Sansone, S. A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., Fostel, J., Garrow, A. G., Gilbert, J., Goodsaid, F., Hardy, N., Jones, P., Lister, A., Miller, M., Morrison, N., Rayner, T., Sklyar, N., Taylor, C., Tong, W., Warner, G., and Wiemann, S. (2008) The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?". *OMICS* **12,** 143–149
38. Taylor, C. F., Field, D., Sansone, S. A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P. A., Bogue, M., Booth, T., Brazma, A., Brinkman, R. R., Michael Clark, A., Deutsch, E. W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J. M., Hardy, N. W., Hermjakob, H., Julian, R. K., Jr., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Novere, N. L., Leebens-Mack, J., Lewis, S. E., Lord, P., Mallon, A. M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J. M., Robertson, D. G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R. H., Schober, D., Smith, B., Snape, J., Stoeckert, C. J., Jr., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., and Wiemann, S. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* **26**, 889–896

In order to cite this article properly, please include all of the following information: Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011) mzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* 10(1):R110.000133. DOI: 10.1074/mcp.R110.000133.

DOI:10.1074/mcp.R110.000133

Nucleic Acids Research Advance Access published October 29, 2012

Nucleic Acids Research, 2012, ***1–6***
doi:10.1093/nar/gks1004

MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data

Kenneth Haug[1], Reza M. Salek[1,2,3], Pablo Conesa[1], Janna Hastings[1], Paula de Matos[1], Mark Rijnbeek[1], Tejasvi Mahendraker[1], Mark Williams[1], Steffen Neumann[4], Philippe Rocca-Serra[5], Eamonn Maguire[5], Alejandra González-Beltrán[5], Susanna-Assunta Sansone[5], Julian L. Griffin[2,3] and Christoph Steinbeck[1,*]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, [2]MRC HNR, Elsie Widdowson Laboratory, Fulbourn Road, Cambridge CB1 9NL, [3]Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1GA, UK, [4]Department of Stress- and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany and [5]Oxford e-Research Centre, University of Oxford, 7 Keble Road, Oxford OX1 3QG, UK

Received August 15, 2012; Revised September 22, 2012; Accepted October 1, 2012

ABSTRACT

MetaboLights (http://www.ebi.ac.uk/metabolights) is the first general-purpose, open-access repository for metabolomics studies, their raw experimental data and associated metadata, maintained by one of the major open-access data providers in molecular biology. Metabolomic profiling is an important tool for research into biological functioning and into the systemic perturbations caused by diseases, diet and the environment. The effectiveness of such methods depends on the availability of public open data across a broad range of experimental methods and conditions. The MetaboLights repository, powered by the open source ISA framework, is cross-species and cross-technique. It will cover metabolite structures and their reference spectra as well as their biological roles, locations, concentrations and raw data from metabolic experiments. Studies automatically receive a stable unique accession number that can be used as a publication reference (e.g. MTBLS1). At present, the repository includes 15 submitted studies, encompassing 93 protocols for 714 assays, and span over 8 different species including *human, Caenorhabditis elegans, Mus musculus and Arabidopsis thaliana*. Eight hundred twenty-seven of the metabolites identified in these studies have been mapped to ChEBI. These studies cover a variety of techniques, including NMR spectroscopy and mass spectrometry.

INTRODUCTION

Metabolomics is the systematic study of the small molecular metabolites in a cell, tissue, biofluid or cell culture media that are the tangible result of cellular processes or responses to an environmental stress (1,2). The identification and quantification of such metabolites provide unique insights into the metabolic processes that are taking place in the cellular environment. Metabolic profiles taken from body fluids have the potential to act as biomarkers for many different diseases, an approach that has already shown value in, for example, heart disease and diabetes (3), the effects of diet (4) and interactions with the environment (5). Metabolomics technologies yield many insights into basic biological research in areas such as systems biology and metabolic modeling (6), pharmaceutical research (7), nutrition (8) and toxicology (9). However, to harness the full potential of metabolomics, researchers needs access to data and knowledge to compare, contrast and make inferences from the results they obtain in their experiments (10). The metabolome is the total complement of metabolites present in a biological sample under given genetic, nutritional or environmental conditions. Since such conditions can vary dramatically, it is clear that databases will need to collect numerous experiments together for a given species to accurately reflect the underlying diversity and complexity. In recent years, several instrument or species-specific dedicated metabolomics reference databases have been created. Examples include the Human Metabolome Database [HMDB, http://www.hmdb.ca, (11)], the Biological

Downloaded from http://nar.oxfordjournals.org/ at Inst f Pflanzenbiochemie on October 30, 2012

---

*To whom correspondence should be addressed. Tel: +44 1223 492640; Fax: +44 1223 494468; Email: steinbeck@ebi.ac.uk

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.

DOI:10.1093/nar/gks1004

Magnetic Resonance Data Bank [BMRB, http://www.bmrb.wisc.edu, (12)], METLIN [http://metlin.scripps.edu, (13)], LIPIDMAPS [http://www.lipidmaps.org, (14)] and more general databases such as KNApSAck (http://kanaya.aist-nara.ac.jp/KNApSAcK/). However, the various metabolomics communities worldwide have not had a global open repository to share experimental data and associated metadata across species and platforms. MetaboLights will (i) provide a single point of access to worldwide data and knowledge in metabolomics, (ii) facilitate the development and adoption of a common data sharing format, (iii) ensure data traceability and reproducibility and (iv) progressively promote interoperability across existing resources.

MetaboLights consists of two distinct layers: a *repository*, enabling the metabolomics community to share findings, data and protocols for any form of metabolomics study, and a *reference layer* of curated knowledge about metabolites (forthcoming). MetaboLights is not intended to replace specialist resources but is specifically designed to build on prior art and extensively collaborate with the existing databases to ensure that data are exchanged and that assimilation efforts target gaps in worldwide available knowledge. We are dedicated to close collaboration with all major parties involved in the creation of this prior art, such as the Metabolomics Society, Metabomeeting and the Metabolomics Standards Initiative (MSI) (15). MetaboLights is working towards the setup of formal data sharing agreements with major resources such as the HMDB, the Golm Metabolome Database (16), MetabolomeExpress (17) and the Riken Metabolomics Platform (18). MetaboLights contains references to identified metabolites in existing databases, such as HMDB and ChEBI (19), and does not duplicate compound information residing in these external databases. Rather, it uses programmatic access to retrieve relevant data to display a unified metabolite-centric view to our users. In the future, such metabolite-centric views will be extended to show metabolites in the context of pathways, harnessing the Reactome database of biochemical pathways (20). In this article, we report on the structure and content of the MetaboLights repository and describe on-going work in the development of the reference layer.

DATABASE DESCRIPTION

The MetaboLights repository can be accessed at http://www.ebi.ac.uk/metabolights and http://metabolights.org.

Database content

We store and display an extensive set of associated information for studies in MetaboLights. This includes submitter and author information, publication references, the study design, protocols applied, names of data files included, platform information and metabolite information. The metabolite information includes a description, external database identifiers, formula and intensity or concentration, and where the metabolite was identified in the sample.

At present, the repository includes 15 submitted studies, of which 10 are publicly visible. These studies encompass 93 protocols for 714 assays, and span over 8 different types of organism including human, *Caenorhabditis elegans, Mus musculus and Arabidopsis thaliana*. Eight hundred twenty-seven of the metabolites identified in these studies have been mapped to ChEBI and 136 to HMDB. Thirty-eight users are currently registered.

Technical architecture

The MetaboLights repository is based on open source freely available software and tools. The web application runs on an Apache Tomcat server and the database backend is an Oracle database, but other standard SQL databases like MySQL and PostgreSQL can be used.

At the core of the database implementation is the ISA framework (21). The main database schema is powered by the ISA BioInvestigation Index (BII), which contains user information and all searchable metadata for the studies. Currently, there are 72 tables in this database schema. Any data-files that are associated with a study are stored on a traditional file system, and only their reference is stored in the database. Each study has a separate folder on the file systems containing the study metadata and associated files. This ensures a relative small database schema, but individual studies can be very large depending on the size of attached data files.

Searching for data

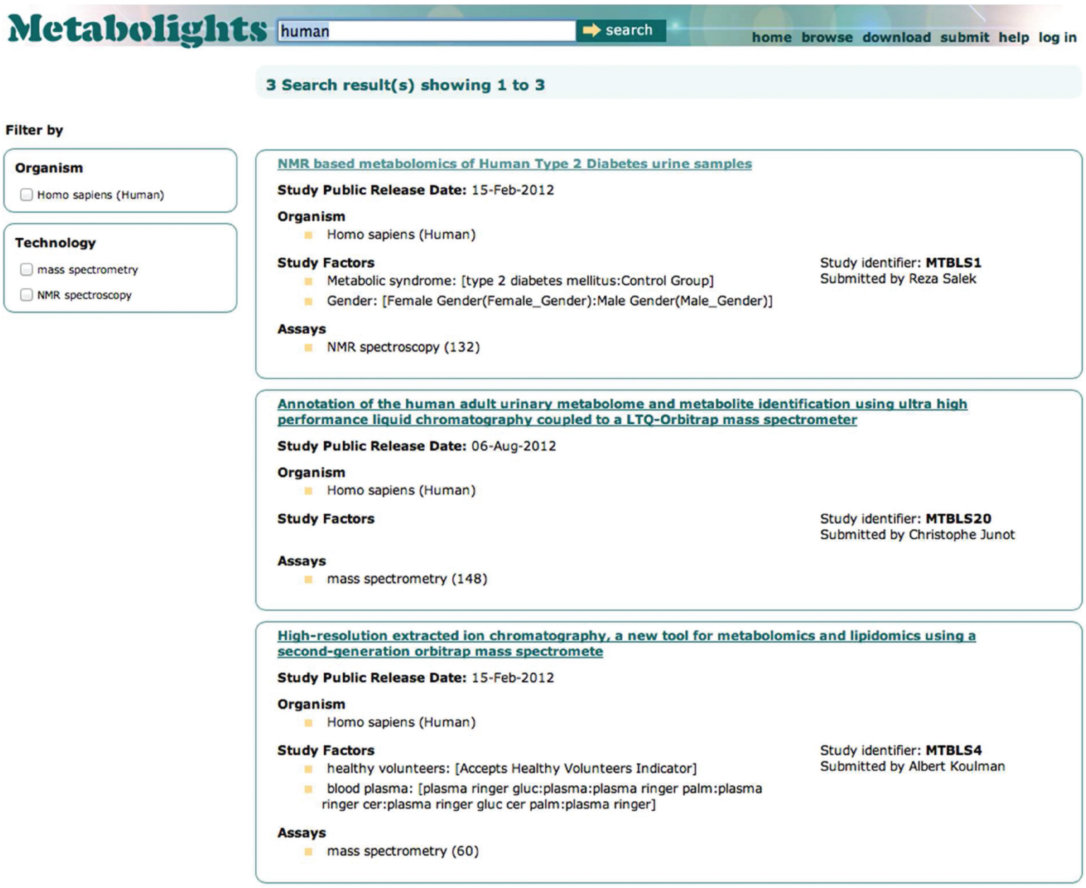
The online search facility provides the ability to search using free text through most of the underlying data fields, including the study description, study title, protocols, metabolites and authors. Currently, we support free-text searching and you can combine multiple search terms, for example ‘human urine’ will give you all studies where you find the terms ‘human’ and ‘urine’ are used. The search result page, as illustrated in Figure 1, shows general study information like the submitter of the study, the study title, public release date, organism(s), study design and platform.

It is possible to further refine the search result using ‘search facets’. Search facets give the user the ability to limit the search results to a selection of species, platform and metabolite. For example, if you select a specific organism from the filter, the search results are limited to show only studies containing this organism. The search mechanism in MetaboLights is implemented using a text index (Lucene index) so no direct backend database queries are performed during a general search. This ensures a fast search facility.

Figure 1 shows the search results page when searching for ‘human’ across all of MetaboLights. To see the details of a study, the user can simply click on the study title. Example of what is displayed in the study details are in Supplementary Figures S1–S4. These images show screenshots of the web interface of MetaboLights with study data loaded for an NMR-based metabolomics study, MTBLS1. The Study details page consists of four tabs. The first tab (Supplementary Figure S1) shows information about the submitters, the relevant dates, study title and

Downloaded from http://nar.oxfordjournals.org/ at Inst f Pflanzenbiochemie on October 30, 2012

DOI:10.1093/nar/gks1004



3 Search result(s) showing 1 to 3

Filter by

Organism
☐ Homo sapiens (Human)

Technology
☐ mass spectrometry
☐ NMR spectroscopy

NMR based metabolomics of Human Type 2 Diabetes urine samples
Study Public Release Date: 15-Feb-2012
Organism
- Homo sapiens (Human)

Study Factors
- Metabolic syndrome: [type 2 diabetes mellitus:Control Group]
- Gender: [Female Gender(Female_Gender):Male Gender(Male_Gender)]

Assays
- NMR spectroscopy (132)

Study identifier: **MTBLS1**
Submitted by Reza Salek

Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a LTQ-Orbitrap mass spectrometer
Study Public Release Date: 06-Aug-2012
Organism
- Homo sapiens (Human)

Study Factors

Assays
- mass spectrometry (148)

Study identifier: **MTBLS20**
Submitted by Christophe Junot

High-resolution extracted ion chromatography, a new tool for metabolomics and lipidomics using a second-generation orbitrap mass spectromete
Study Public Release Date: 15-Feb-2012
Organism
- Homo sapiens (Human)

Study Factors
- healthy volunteers: [Accepts Healthy Volunteers Indicator]
- blood plasma: [plasma ringer gluc:plasma:plasma ringer palm:plasma ringer cer:plasma ringer gluc cer palm:plasma ringer]

Assays
- mass spectrometry (60)

Study identifier: **MTBLS4**
Submitted by Albert Koulman

Figure 1. Searching for 'human' in the MetaboLights web application.

description, organisms, study design, publications and the experimental factors. The next tab (Supplementary Figure S2) details the protocols used during this study, from how the sample was collected through to the metabolite identification. Next, we have the data tab (Supplementary Figure S3). Here, we show data files for this study, detailed for technology platform used and experimental factors. Finally, we have metabolite identification (Supplementary Figure S4). Each identified metabolite has an external database reference, for example a ChEBI or HMDB identifier. Metabolites identified with a ChEBI accession show additional molecule description. The identified metabolite tab details which sample the compound was identified in. Unknown compounds are listed without a database reference.

Browsing data

Users can browse studies in MetaboLights using the 'browse' link. This will give a complete list of all the public studies currently available. If the user is registered and currently logged in to MetaboLights, additional private studies may be displayed. These private studies are either under the users control or have been directly shared from other users. To limit the number of studies in the browsing list, the user can activate the same facets available for a general search.

Downloads and programmatic access

MetaboLights software components are open source and all data are free to download and use for any purpose. All public studies are downloadable as ISA-Tab (22) metadata files with associated data files directly from the online study details page, and from the MetaboLights download page http://www.ebi.ac.uk/metabolights/download. A direct bulk download using ftp is available from ftp://ftp.ebi.ac.uk/pub/databases/metabolights/, organized into sub-folders for public studies. There are no web services for programmatic access available at present. However, this functionality is scheduled for a future release of the repository.

Submitting data

MetaboLights accepts experimental descriptions in ISA-Tab format, which can be created by the ISAcreator editor tool. MetaboLights also offers different templates for the ISAcreator tool to accommodate the description of different types of metabolomics experiments. ISAcreator is a standalone Java desktop application that enables researchers to report experimental information, associate raw and processed data files, and submit the collated information to the MetaboLights database. Building on the OSGI plugin architecture, the ISAcreator has been

Downloaded from http://nar.oxfordjournals.org/ at Inst f Pflanzenbiochemie on October 30, 2012

DOI:10.1093/nar/gks1004

extended to create a 'Metabolite Identification' add-on to capture relevant information for all small molecules identified in a study, with a link to a relevant chemical database (Figure 2). MetaboLights also accepts studies that have unknown or incomplete metabolite identification. This information has the potential to facilitate the identification of unknown metabolites in the future.

Currently, we accept all data formats for 'raw' instrumental data, converted open source file formats and any processed data, but we strongly recommend that processed data should be made available in open formats, such as mzML (23) for MS data.

MetaboLights implements metadata guidelines according to the recommendations of the Metabolomics Standards Initiative (MSI). The MSI defined a set of metabolomics reporting standards by harnessing and coordinating the efforts of several pre-existing international initiatives. MSI developed checklists and standards that have subsequently been adopted by the community, including minimum metadata reporting recommendations (24).

To facilitate high quality data submissions for NMR or MS experiments, there is a guided submission process to help meet MSI recommendations and extensively use community-developed controlled vocabularies and ontologies. ISAcreator also provides advanced mechanisms for mapping to and uploading information from existing spreadsheets. Figure 3 illustrates the ISA components in a typical data creation scenario.

An R package has been developed to facilitate data analysis (Supplementary Method). The Risa module, available in the next BioConductor release, includes functionality to process mass spectrometry data relying on the xcms package (25), and to save analysis results back to ISA archives.

Installing a local copy of the MetaboLights repository

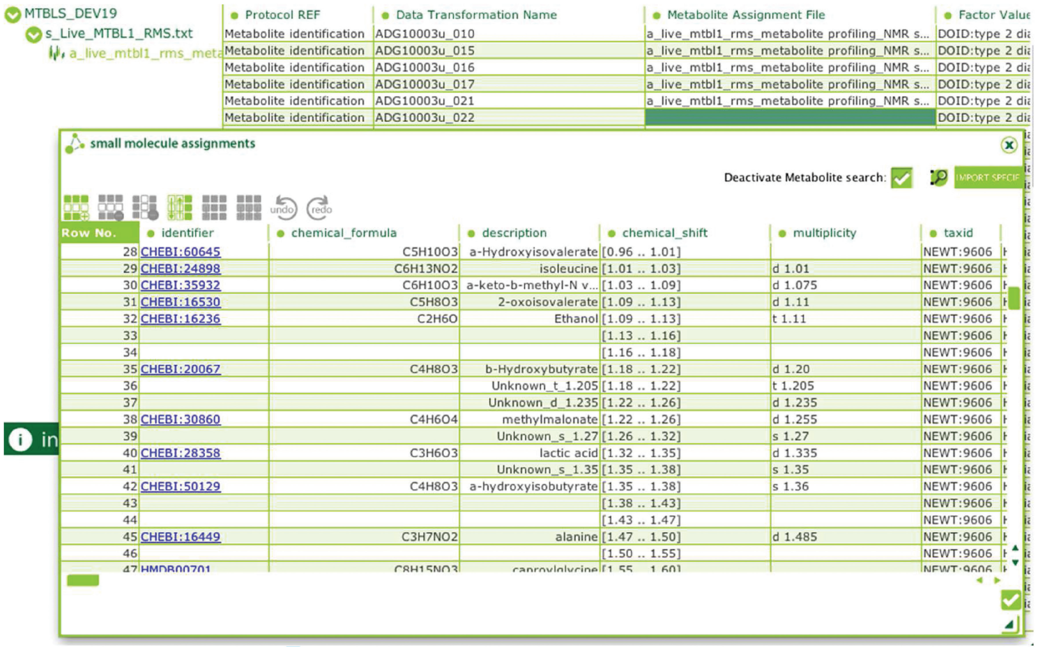
To install MetaboLights locally, you require a SQL database (MySQL, PostgreSQL or Oracle), a subversion client (svn) and an Apache Tomcat server. The MetaboLights Repository source code can be found at http://sourceforge.net/projects/metabolomes, here you will also find more details regarding how to install MetaboLights locally. The ISAcreator Metabolite Identification plugin can be found at: https://github.com/EBI-Metabolights/ISAcreatorPlugins. The ISA framework is also open source and is available at: https://github.com/ISA-tools. Figure 4 shows the principal components of a local MetaboLights repository installation.

Access and privacy policy

MetaboLights grants free access and reuse of the public data it stores to everyone. Only registered users can upload and share study data. To facilitate deposition of research data not yet publicly visible, the submitter can set a data embargo for a period of up to 60 months, which can be lifted on results publication or extended upon request. Submitters can also request for access to their private data to be granted to specific other registered users. This feature may be particularly useful in facilitating collaborations and the peer review process.

Feedback

To facilitate user feedback, we have created a SourceForge tracker for logging issues, available at http://sourceforge.net/projects/metabolomes. There is also an online contact form, http://www.ebi.ac.uk/metabolights/contact, and contact email address, metabolights-help@ebi.ac.uk.



Protocol REF	Data Transformation Name	Metabolite Assignment File	Factor Value
Metabolite identification	ADG10003u_010	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 dia
Metabolite identification	ADG10003u_015	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 dia
Metabolite identification	ADG10003u_016	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 dia
Metabolite identification	ADG10003u_017	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 dia
Metabolite identification	ADG10003u_021	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 dia
Metabolite identification	ADG10003u_022		DOID:type 2 dia

Row No.	identifier	chemical_formula	description	chemical_shift	multiplicity	taxid
28	CHEBI:60645	C5H10O3	a-Hydroxyisovalerate	[0.96 .. 1.01]		NEWT:9606
29	CHEBI:24898	C6H13NO2	isoleucine	[1.01 .. 1.03]	d 1.01	NEWT:9606
30	CHEBI:35932	C6H10O3	a-keto-b-methyl-N v...	[1.03 .. 1.09]	d 1.075	NEWT:9606
31	CHEBI:16530	C5H8O3	2-oxoisovalerate	[1.09 .. 1.13]	d 1.11	NEWT:9606
32	CHEBI:16236	C2H6O	Ethanol	[1.09 .. 1.13]	t 1.11	NEWT:9606
33				[1.13 .. 1.16]		NEWT:9606
34				[1.16 .. 1.18]		NEWT:9606
35	CHEBI:20067	C4H8O3	b-Hydroxybutyrate	[1.18 .. 1.22]	d 1.20	NEWT:9606
36			Unknown_t_1.205	[1.18 .. 1.22]	t 1.205	NEWT:9606
37			Unknown_d_1.235	[1.22 .. 1.26]	d 1.235	NEWT:9606
38	CHEBI:30860	C4H6O4	methylmalonate	[1.22 .. 1.26]	d 1.255	NEWT:9606
39			Unknown_s_1.27	[1.26 .. 1.32]	s 1.27	NEWT:9606
40	CHEBI:28358	C3H6O3	lactic acid	[1.32 .. 1.35]	d 1.335	NEWT:9606
41			Unknown_s_1.35	[1.35 .. 1.38]	s 1.35	NEWT:9606
42	CHEBI:50129	C4H8O3	a-hydroxyisobutyrate	[1.35 .. 1.38]	s 1.36	NEWT:9606
43				[1.38 .. 1.43]		NEWT:9606
44				[1.43 .. 1.47]		NEWT:9606
45	CHEBI:16449	C3H7NO2	alanine	[1.47 .. 1.50]	d 1.485	NEWT:9606
46				[1.50 .. 1.55]		NEWT:9606
47	HMDB00701	C8H15NO3	caproylglycine	[1.55 .. 1.60]		NEWT:9606

Figure 2. Part of Study MTBLS1 in ISAcreator with the Metabolite Identification Plugin active.

Downloaded from http://nar.oxfordjournals.org/ at Inst f Pflanzenbiochemie on October 30, 2012

DOI:10.1093/nar/gks1004

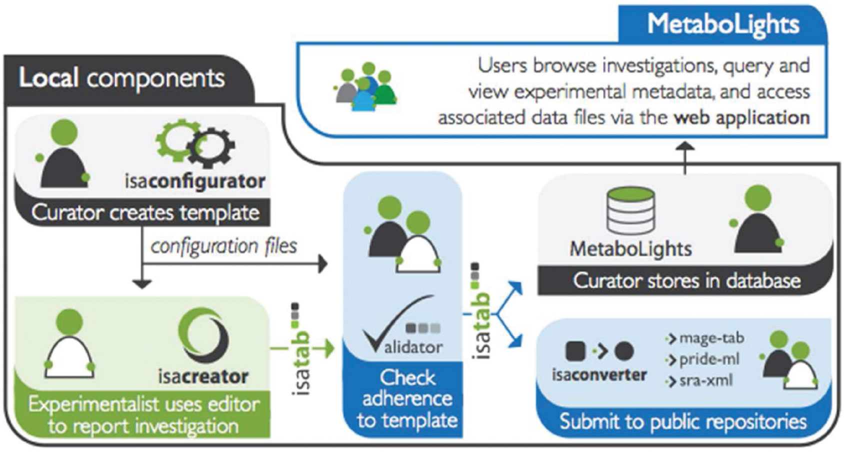

MetaboLights
Users browse investigations, query and view experimental metadata, and access associated data files via the web application
Local components
isaconfigurator
Curator creates template
configuration files
isacreator
Experimentalist uses editor to report investigation
isatab
validator
Check adherence to template
isatab
MetaboLights
Curator stores in database
isaconverter
mage-tab
pride-ml
sra-xml
Submit to public repositories

Figure 3. A typical workflow, using the ISA framework, for reporting information and submitting it to the MetaboLights database.

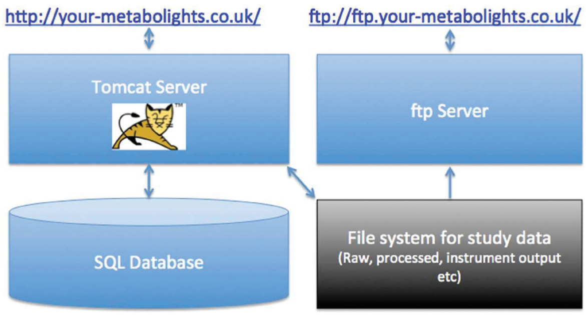

http://your-metabolights.co.uk/
ftp://ftp.your-metabolights.co.uk/
Tomcat Server
ftp Server
SQL Database
File system for study data
(Raw, processed, instrument output etc)

Figure 4. Simple technical architecture of a local MetaboLights repository, including web (http) and ftp access.

DISCUSSION

The MetaboLights repository was launched on the 28 June 2012 at the 8th International Conference of the Metabolomics Society in Washington, DC, USA. The repository is now accepting study submissions from a growing number of active users worldwide with submission privileges. For the latest statistics on current studies and submitters, please see http://www.ebi.ac.uk/metabolights/stats.

The requirement by a growing number of publishers and funding agencies to deposit data associated with journal publications to public repositories is expected to motivate a substantial number of future submissions. As more datasets become available, Metabolights will become an invaluable resource for those wishing to develop new algorithms for the processing of metabolomic data. The creation of a long-term institution-backed, as it will be maintained by EBI after the grant ends, public repository such as MetaboLights at EMBL-EBI allows laboratories across the globe to collaborate on projects through data sharing, and thereby to begin to collaboratively generate the large datasets needed to address how the environment, genome and diet influence the metabolome of a species.

Future work

The MetaboLights team is now actively specifying the *MetaboLights Reference Layer*, which will be launched in Summer 2013. The Reference Layer will be a comprehensive knowledge base organized around a metabolite-centric view, and will include elements such as reference spectra of various types, biological reference data, protocols, cross-references to other resources and advanced search and download functionality. There will be comprehensive manually curated data, including chemical structures and characteristics from ChEBI, metabolic pathways, reference spectroscopy and chromatography. Furthermore, there will be information about the reference biology, metabolites and their occurrence and concentration in species, organs, tissues and cellular compartments in various conditions, both healthy and diseased. Publication references and protocols will also be available. This will enable experimentalists to get a comprehensive Metabolomic view on known metabolites.

We are also substantially enhancing our online help capabilities with online video instructions as well as detailed scenarios for completing new studies for submission. A new section with 'Gold Standard Studies' will be included for easy reference. These studies can be used as templates for similar experiments.

In October 2012, the European COordination of Standards in MetabOlomicS (COSMOS) consortium, comprising 14 European partners, will start its work on Metabolomics data standardization, publication and dissemination workflows. The MetaboLights database is a key component in this effort. It is the aim of the COSMOS project to develop efficient policies to ensure that Metabolomics data are encoded in open standards, tagged with a community-agreed and complete set of metadata, supported by a communally developed set of open source data management and capturing tools, disseminated in open-access databases adhering to these standards, supported by vendors and publishers, who require deposition upon publication, and properly interfaced with data in other biomedical and life science

Downloaded from http://nar.oxfordjournals.org/ at Inst f Pflanzenbiochemie on October 30, 2012

DOI:10.1093/nar/gks1004

e-infrastructures [such as ELIXIR (26), BioMedBridges (http://www.biomedbridges.eu), EU-OPENSCREEN (http://www.eu-openscreen.de) and BBMRI (http://bbmri.eu)].

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–4 and Supplementary Methods.

ACKNOWLEDGEMENTS

The MetaboLights project team thanks the following persons for their invaluable contributions: Rafael Alcántara, Masanori Arita, Mike Beale, Nick Bond, Kees van Bochove, Jildau Bouwman, Scarlet Brockmoeller, Steve Bryant, Hong Cao, Juan Castrillo, Jenny Cham, Cecilia Castro, Yajing Chu, Tim Ebbels, Michael Eiden, Oliver Fiehn, Andrew Gibbs, Roy Goodacre, Martin Hornshaw, Jan Hummel, Albert Koulman, Peter Meadows, Pablo Moreno, Theo Reijmers, Francis Rowland, Linda Scoriels, Mark Seymour, Tim Smith, Anthony Taylor, Chris Taylor, Michael Wakelam, Jane Ward and David Wishart.

FUNDING

The development of MetaboLights is funded by the BBSRC [BB/I000933/1]; ISA framework by the BBSRC [BB/I025840/1, BB/I000771/1, BB/J020265/1 to S.A.S.]; NERC, EU [EC 312941 to S.A.S.]; University of Oxford e-Research Centre [to S.A.S.]. COSMOS is funded by European Commission Grant [EC312941]. Funding for open access charge: BBSRC [BB/1000933/1].

Conflict of interest statement. None declared.

REFERENCES

1. Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.
2. German,J.B., Hammock,B.D. and Watkins,S.M. (2005) Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, **1**, 3–9.
3. Pearson,H. (2007) Meet the human metabolome. *Nature*, **446**, 8.
4. Cheng,K.K., Benson,G.M., Grimsditch,D.C., Reid,D.G., Connor,S.C. and Griffin,J.L. (2010) Metabolomic study of the LDL receptor null mouse fed a high-fat diet reveals profound perturbations in choline metabolism that are shared with ApoE null mice. *Physiol. Genomics*, **41**, 224–231.
5. Veldhoen,N., Ikonomou,M.G. and Helbing,C.C. (2012) Molecular profiling of marine fauna: integration of omics with environmental assessment of the world's oceans. *Ecotoxicol. Environ. Saf.*, **76**, 23–38.
6. Kell,D.B. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.*, **7**, 296–307.
7. Xu,E.Y., Schaefer,W.H. and Xu,Q.W. (2009) Metabolomics in pharmaceutical research and development: metabolites, mechanisms and pathways. *Curr. Opin. Drug Discovery Dev.*, **12**, 40–52.
8. Gibney,M.J., Walsh,M., Brennan,L., Roche,H.M., German,B. and van Ommen,B. (2005) Metabolomics in human nutrition: opportunities and challenges. *Am. J. Clin. Nutrition*, **82**, 497–503.
9. Kaddurah-Daouk,R., Kristal,B.S. and Weinshilboum,R.M. (2008) Metabolomics: a global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.*, **48**, 653–683.
10. Goodacre,R., Vaidyanathan,S., Dunn,W.B., Harrigan,G.G. and Kell,D.B. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, **22**, 245–252.
11. Wishart,D.S., Tzur,D., Knox,C., Eisner,R., Guo,A.C., Young,N., Cheng,D., Jewell,K., Arndt,D., Sawhney,S. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.
12. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
13. Smith,C.A., O'Maille,G., Want,E.J., Qin,C., Trauger,S.A., Brandon,T.R., Custodio,D.E., Abagyan,R. and Siuzdak,G. (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monit.*, **27**, 747–751.
14. Sud,M., Fahy,E., Cotter,D., Brown,A., Dennis,E.A., Glass,C.K., Merrill,A.H., Murphy,R.C., Raetz,C.R.H., Russell,D.W. *et al.* (2007) Lmsd: lipid maps structure database. *Nucleic Acids Res.*, **35**, D527–D532.
15. Sansone,S.A., Fan,T., Goodacre,R., Griffin,J.L., Hardy,N.W., Kaddurah-Daouk,R., Kristal,B.S., Lindon,J., Mendes,P., Morrison,N. *et al.* (2007) The metabolomics standards initiative. *Nat. Biotechnol.*, **25**, 846–848.
16. Kopka,J., Schauer,N., Krueger,S., Birkemeyer,C., Usadel,B., Bergmuller,E., Dormann,P., Weckwerth,W., Gibon,Y., Stitt,M. *et al.* (2005) Gmd@Csb.Db: the Golm metabolome database. *Bioinformatics*, **21**, 1635–1638.
17. Carroll,A.J., Badger,M.R. and Harvey Millar,A. (2010) The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics*, **11**, 376.
18. Akiyama,K., Chikayama,E., Yuasa,H., Shimada,Y., Tohge,T., Shinozaki,K., Hirai,M.Y., Sakurai,T., Kikuchi,J. and Saito,K. (2008) PRIMe: a web site that assembles tools for metabolomics and transcriptomics. *In Silico Biol.*, **8**, 339–345.
19. de Matos,P., Alcantara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical entities of biological interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
20. Vastrik,I., D'Eustachio,P., Schmidt,E., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S., Matthews,L. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
21. Sansone,S.A., Rocca-Serra,P., Field,D., Maguire,E., Taylor,P., Hofmann,O., Fang,H., Neumann,S., Tong,W., Amaral-Zettler,L. *et al.* (2012) Toward interoperable bioscience data. *Nat. Genet.*, **44**, 121–126.
22. Rocca-Serra,P., Brandizi,M., Maguire,E., Sklyar,N., Taylor,C., Begley,K., Field,D., Harris,S., Hide,W., Hofmann,O. *et al.* (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, **26**, 2354–2356.
23. Martens,L., Chambers,M., Sturm,M., Kessner,D., Levander,F., Shofstahl,J., Tang,W.H., Rompp,A., Neumann,S., Pizarro,A.D. *et al.* (2011) mzML—a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, **10**, R110.000133.
24. Goodacre,R., Broadhurst,D., Smilde,A.K., Kristal,B.S., Baker,J.D., Beger,R., Bessant,C., Connor,S., Calmani,G., Craig,A. *et al.* (2007) Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, **3**, 231–241.
25. Smith,C.A., Want,E.J., O'Maille,G., Abagyan,R. and Siuzdak,G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.
26. Crosswell,L.C. and Thornton,J.M. (2012) ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.*, **30**, 241–242.

Downloaded from http://nar.oxfordjournals.org/ at Inst f Pflanzenbiochemie on October 30, 2012

DOI:10.1093/nar/gks1004

SPLASH, A hashed identifier for mass spectra

Authors: Gert Wohlgemuth* (1), Sajjan S. Mehta (1), Ramon F. Mejia (2), Steffen Neumann (3), Diego Pedrosa (1), Tomáš Pluskal (4), Emma L. Schymanski* (5), Egon L. Willighagen (6), Michael Wilson (7), David S. Wishart (7), Masanori Arita (2, 8), Pieter C. Dorrestein (9, 10), Nuno Bandeira (9, 11, 12), Mingxun Wang (11, 12), Tobias Schulze (13), Reza M. Salek (14), Christoph Steinbeck (14), Venkata Chandrasekhar Nainala (14), Robert Mistrik (15), Takaaki Nishioka (16), Oliver Fiehn* (1,17)

Affiliations:

1. University of California Davis, West Coast Metabolomics Center, Genome Center, 451 Health Sciences Drive, Davis, CA 95616, USA
2. RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, 230-0045, Japan
3. Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany
4. Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA
5. Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland
6. Maastricht University, Department of Bioinformatics - BiGCaT, NUTRIM, P.O. Box 616, UNS 50 Box 19, NL-6200 MD Maastricht, The Netherlands
7. University of Alberta, Department of Computing Science, Edmonton, AB, Canada T6G 2E8
8. National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
9. Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, United States
10. Departments of Pharmacology and Pediatrics, School of Medicine, UC San Diego, La Jolla, United States
11. Computer Science and Engineering, UC San Diego, La Jolla, United States
12. Center for Computational Mass Spectrometry, UC San Diego, La Jolla, United States
13. UFZ Helmholtz Centre for Environmental Research GmbH, Department of Effect-Directed Analysis, Permoserstrasse 15, 04318 Leipzig, Germany
14. European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
15. HighChem Ltd., Leškova 11, Bratislava 81104, Slovakia
16. Kyoto University, Graduate School of Agriculture, Kitashirakawa Oiwake-cho, Kyoto 606-8502, Japan
17. King Abdulaziz University, Biochemistry Department, Jeddah, Saudi Arabia

*Corresponding authors.

Keywords: Mass spectrometry, Identifier, Database access

To the editor:

Over the past few years, as the use of mass spectrometry (MS) has increased, multiple spectral libraries, databases and software frameworks have been created to enable sharing and searching of MS data. However, finding all the spectra that correspond to a specific compound or peptide across different databases continues to be a challenge. A spectral identifier that improves the searchability and exchange of mass spectra, as well as provenance and duplicate detection, would address these issues and enhance searchability.

MassBank[2] has been the source of data for other open libraries such as the Global Natural Products Social Molecular Networking[3] (GNPS) and Human Metabolome Database[4] (HMDB) libraries and the MetaboLights reference layer[5]. In turn, HMDB and community-contributed spectra from GNPS have

Author manuscript. Published in final edited form as: Nat Biotechnol. 2016 November 08; 34(11): 1099–1101. doi:10.1038/nbt.3689.

DOI:10.1038/nbt.3689

also been imported into MassBank of North America[2c] (MoNA), while GNPS searches public MS data against the above-mentioned libraries as well as the NIST spectral library[6]. The mzCloud[7] library contains some spectra generated from the same raw data that was used to create MassBank records. As these examples show, the complexity and the cross-import of data is increasing, together with the number of mass spectra, such that these different resources can now contain identical or near identical spectra under different accession numbers. For example, the library entries PR100026 (MassBank, MoNA), 5464 (HMDB), and CCMSLIB00000222858 (GNPS) all refer to exactly the same mass spectrum of caffeine, originally sourced from MassBank. As the different libraries focus on different compound domains[1], complete coverage of substances requires the use of several resources, some of which are commercial (e.g., NIST and mzCloud).Mass spectra are highly variable, with one to potentially thousands of mass-to-charge (*m/z*) and intensity entries per spectrum, presenting a challenge in the design of an optimal identifier. However, other life science databases have faced a similar need. For databases with chemical structures, the InChI code and the hashed InChIKey[8,9] of fixed length, which have been broadly adopted as chemical identifiers, can be easily stored in databases, compared across resources and, for InChIKeys, searched on general-purpose search engines[10]. A hash is a one-directional mapping between a long, potentially complex object and a typically much shorter hash string with a fixed length of characters and numbers. For chemicals, the InChIKey is much easier to search than the (generally) much longer InChI, which contains special characters. While it is not possible to obtain the original object back purely from the hash value, hash keys provide easy access to the original data within a data collection.

We designed the SPLASH (SPectraL hASH) as an unambiguous, database-independent spectrum identifier that fulfills the criteria outlined above and offers some additional functionality. Inspired by the broad applicability of the InChIKey across cheminformatics and like the InChIKey (which encodes skeleton, stereochemistry, and charge), SPLASH contains separate blocks that define different layers of information, separated by dashes. As an example, the full SPLASH of the caffeine spectrum above is “`splash10-0002-0900000000-b112e4e059e1ecf98c5f`”. The first block is the SPLASH identifier, the second and third are summary blocks, while the fourth is the hash block.

To calculate a SPLASH, spectra are converted into a canonical text representation: the intensities are normalized to an integer value between 0 and 100, with *m/z* values given in exactly 6 decimal places. To ensure consistent handling between different software and implementations, entries with zero intensities are included, but empty (“N/A”) values are eliminated prior to creating the SPLASH. The first block (“`splash10`”) encodes the SPLASH identifier, starting with letters for semantic web compatibility, followed by a number representing the measurement type (1 for MS, 2 and above for other data types to be included in the future) and the version, starting at 0, to allow for future specification updates. Thus, `splash10` is a SPLASH identifier for MS, version 0.

Both the second and third blocks are spectral summaries, which serve to prefilter and restrict searches. In the second and third blocks, intensities are summed over fixed (but different) bin sizes and wrapped over 10 bins. The wrapped bin (zero-based) index for a given ion is computed as floor (m/z ÷ BinSize) modulo 10. This wrapping allows accommodation of all mass spectral ranges. The second block (e.g., “`0002`” for caffeine) is formed using a reduced spectrum (the top 10 or fewer ions greater than 10% of the base peak). This reduced spectrum is summed over bins of 5 Da. Each bin is then scaled to a single-digit integral value in base 3 (0-2), and the resulting length 10 histogram is converted to a base 36 number, resulting in a 4-digit block. In the third block (e.g., “`0900000000`”) the intensities are summed over 100 Da bin sizes, each bin is then scaled to a single-digit, integral base 10 digit (0-9).

The fourth block (e.g., “`b112e4e059e1ecf98c5f`”) is a hash of the full spectrum in Secure Hash Algorithm[11] SHA256 (numbers and lowercase letters only), calculated in hexadecimal notation and truncated to 20 characters. The full spectrum string of *m/z* and relative intensity pairs are sorted by ascending *m/z* and then by descending intensity. The *m/z* value is multiplied by 10^6, cast to a long (64-bit) integer, and joined with the normalized intensity as strings separated by a colon. The resulting ion pairs are then joined, delimited by a single space. Specification document and reference implementations have been created for several programming environments (Python, Scala, C++, C#, R, Ruby, and Java) under a BSD-3 license as well as a REST interface; additional information is available at http://splash.fiehnlab.ucdavis.edu/.

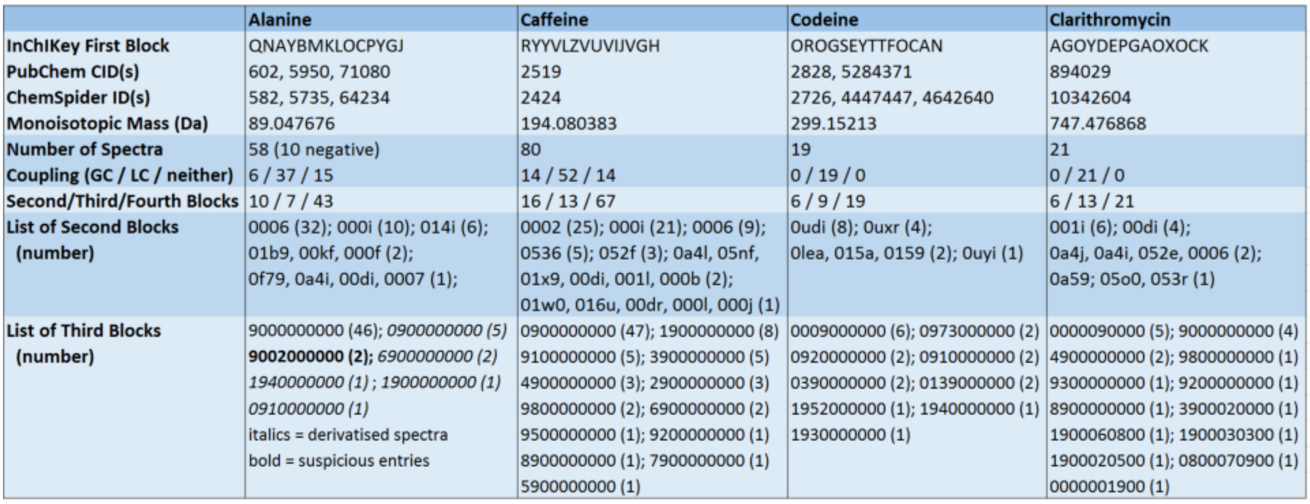
The SPLASH concept was developed and refined on a dataset of 563,902 mass spectra from MassBank[2], GNPS[3], HMDB[4], ReSpect[12], FiehnLib[13] and NIST 14[6]; all but the NIST spectra (which cannot be released publically) are available on MoNA (http://mona.fiehnlab.ucdavis.edu/). This dataset is a mix of many types of mass spectra and the SPLASH was designed to account for this, plus be easily searchable in general-purpose search engines, offer a unique identifier (through the hash) and basic pre-filter and similarity functionality (through the second and third blocks).

Ensuring all these features are present in one short text string requires compromise; the SPLASH is not intended to replace more sophisticated database-specific functions, but does offer simple cross-database functionality. The second block was chosen from 136 different potential block formats as the best short, web search-compatible way to reduce the mass spectral search space. In order to determine the best performing second block, we queried a subset of 19,435 spectra against the full 563,902 dataset. The second block that we selected for use reduced the search space by 94% or above (36,107 spectra or less) in all cases, while returning 87% of all spectra within a similarity score of 700 (using the NIST cosine similarity score[6,14]) of the queried spectra. In contrast, other tested formats for this block returned more spectra (maximum 93.4%), but too many spectra (up to 100,000 or 1 in 5 spectra) remained in the search space so that the search space reduction was insufficient. The third block provides a visual summary (shown in Table 1 for selected compounds) and a simple text-based summary and basic similarity search, even in search engines or spreadsheets. More information on the most common second and third blocks, as well as the most common combinations and the approximate distribution of substances (not all spectra are annotated with structures) is given in **Table 2**.

While the mapping from object to hash should ideally be unique, hash collisions (where two totally different objects have the same hash, or fourth block of the SPLASH) may occur, depending on the hash algorithm and length of the hash string. Testing the fourth block for hash collisions on the full dataset of 563,902 spectra revealed that identical SPLASHes only arose from mass spectra containing a single ion of the same mass, where the SPLASH is identical by definition due to intensity normalization. The theoretical probability for a collision[15] with any given hash is approximately 10^{-31} for a database containing 10^9 spectra and is further reduced by the presence of two preceding spectral summary blocks. Thus, the SPLASH fulfills its role as a unique identifier while offering simple summary and searching functionality.

Author manuscript. Published in final edited form as: Nat Biotechnol. 2016 November 08; 34(11): 1099–1101. doi:10.1038/nbt.3689.

Table 1: *SPLASH statistics for selected compounds. Data for alanine shows how derivative spectra and suspicious database entries can be detected with the third block (see bold, italic entries), the lower two rows show the variety of different spectra per compound. The combination of second and third blocks is selective, e.g. 0a41-1940000000 and 01ea-1940000000 for alanine and codeine.*

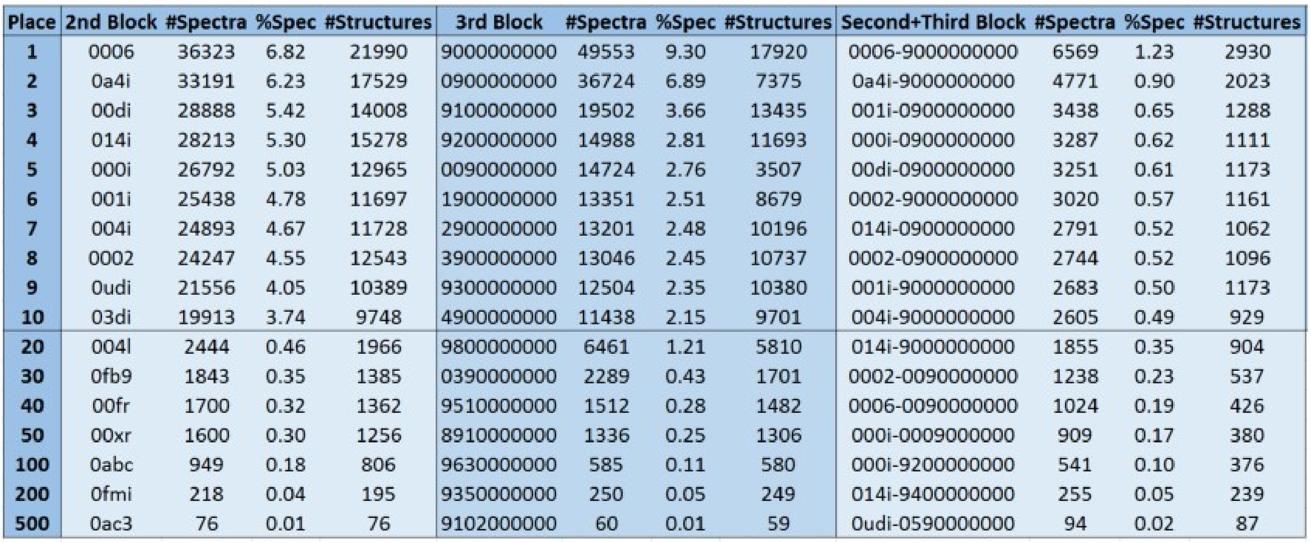


	Alanine	**Caffeine**	**Codeine**	**Clarithromycin**
InChIKey First Block	QNAYBMKLOCPYGJ	RYYVLZVUVIJVGH	OROGSEYTTFOCAN	AGOYDEPGAOXOCK
PubChem CID(s)	602, 5950, 71080	2519	2828, 5284371	894029
ChemSpider ID(s)	582, 5735, 64234	2424	2726, 4447447, 4642640	10342604
Monoisotopic Mass (Da)	89.047676	194.080383	299.15213	747.476868
Number of Spectra	58 (10 negative)	80	19	21
Coupling (GC / LC / neither)	6 / 37 / 15	14 / 52 / 14	0 / 19 / 0	0 / 21 / 0
Second/Third/Fourth Blocks	10 / 7 / 43	16 / 13 / 67	6 / 9 / 19	6 / 13 / 21

List of Second Blocks (number)	0006 (32); 000i (10); 014i (6);
	01b9, 00kf, 000f (2);
	0f79, 0a4i, 00di, 0007 (1);
		0002 (25); 000i (21); 0006 (9);
		0536 (5); 052f (3); 0a4l, 05nf,
		01x9, 00di, 001l, 000b (2);
		01w0, 016u, 00dr, 000l, 000j (1)
			0udi (8); 0uxr (4);
			0lea, 015a, 0159 (2); 0uyi (1)
				001i (6); 00di (4);
				0a4j, 0a4i, 052e, 0006 (2);
				0a59; 05o0, 053r (1)

List of Third Blocks (number)	9000000000 (46); *0900000000 (5)* **9002000000 (2);** *6900000000 (2)* *1940000000 (1) ; 1900000000 (1)* *0910000000 (1)* italics = derivatised spectra bold = suspicious entries	0900000000 (47); 1900000000 (8) 9100000000 (5); 3900000000 (5) 4900000000 (3); 2900000000 (3) 9800000000 (2); 6900000000 (2) 9500000000 (1); 9200000000 (1) 8900000000 (1); 7900000000 (1) 5900000000 (1)	0009000000 (6); 0973000000 (2) 0920000000 (2); 0910000000 (2) 0390000000 (2); 0139000000 (2) 1952000000 (1); 1940000000 (1) 1930000000 (1)	0000090000 (5); 9000000000 (4) 4900000000 (2); 9800000000 (1) 9300000000 (1); 9200000000 (1) 8900000000 (1); 3900020000 (1) 1900060800 (1); 1900030300 (1) 1900020500 (1); 0800070900 (1) 0000001900 (1)

Table 2: *The number of spectra and substances (estimated by first block of the InChIKey) with the "most common" second, third and second+third SPLASH blocks, calculated on a de-duplicated dataset of 532,675 spectra. The number of structures is an estimate; missing structure information was filled in automatically using the Chemical Translation Service (http://cts.fiehnlab.ucdavis.edu/). The place indicates how common the combination is (1 = most common, 200 = 200th most common)*



Place	2nd Block	#Spectra	%Spec	#Structures	3rd Block	#Spectra	%Spec	#Structures	Second+Third Block	#Spectra	%Spec	#Structures
1	0006	36323	6.82	21990	9000000000	49553	9.30	17920	0006-9000000000	6569	1.23	2930
2	0a4i	33191	6.23	17529	0900000000	36724	6.89	7375	0a4i-9000000000	4771	0.90	2023
3	00di	28888	5.42	14008	9100000000	19502	3.66	13435	001i-0900000000	3438	0.65	1288
4	014i	28213	5.30	15278	9200000000	14988	2.81	11693	000i-0900000000	3287	0.62	1111
5	000i	26792	5.03	12965	0090000000	14724	2.76	3507	00di-0900000000	3251	0.61	1173
6	001i	25438	4.78	11697	1900000000	13351	2.51	8679	0002-9000000000	3020	0.57	1161
7	004i	24893	4.67	11728	2900000000	13201	2.48	10196	014i-0900000000	2791	0.52	1062
8	0002	24247	4.55	12543	3900000000	13046	2.45	10737	0002-0900000000	2744	0.52	1096
9	0udi	21556	4.05	10389	9300000000	12504	2.35	10380	001i-9000000000	2683	0.50	1173
10	03di	19913	3.74	9748	4900000000	11438	2.15	9701	004i-9000000000	2605	0.49	929
20	004l	2444	0.46	1966	9800000000	6461	1.21	5810	014i-9000000000	1855	0.35	904
30	0fb9	1843	0.35	1385	0390000000	2289	0.43	1701	0002-0090000000	1238	0.23	537
40	00fr	1700	0.32	1362	9510000000	1512	0.28	1482	0006-0090000000	1024	0.19	426
50	00xr	1600	0.30	1256	8910000000	1336	0.25	1306	000i-0009000000	909	0.17	380
100	0abc	949	0.18	806	9630000000	585	0.11	580	000i-9200000000	541	0.10	376
200	0fmi	218	0.04	195	9350000000	250	0.05	249	014i-9400000000	255	0.05	239
500	0ac3	76	0.01	76	9102000000	60	0.01	59	0udi-0590000000	94	0.02	87

The SPLASH has already been implemented in MassBank[2], MoNA[2c], GNPS[3], HMDB[4], MetaboLights[5] and mzCloud[6], as well as software tools including MZmine[16], MS-DIAL[17], RMassBank[18], BinBase[19], Bioclipse[20] and the Mass Spectrometry Development Kit (MSDK)[21].

The format of the SPLASH allows direct access to spectra on database websites and searching using general purpose search engines. Spectral libraries with more restrictive licenses (e.g. mzCloud and possibly NIST) could also use the SPLASH to provide summarized information about their spectra. SPLASH enables an easier calculation of spectral overlap between libraries, to detect and remove exact duplicate spectra and perform provenance operations. Through the second and

Author manuscript. Published in final edited form as: Nat Biotechnol. 2016 November 08; 34(11): 1099–1101. doi:10.1038/nbt.3689.

DOI:10.1038/nbt.3689

third blocks, SPLASH empowers quick searches for similar spectra within or between libraries, using a variety of search methods. The SPLASH algorithm has been kept independent of metadata, similar to the InChIKey, because an extension to include and distinguish metadata (such as analytical conditions or chemical information) would rapidly become complex and reduce the applicability of the identifier. Instead, the SPLASH is designed to facilitate quick queries and subsequent metadata retrieval.

The widespread adoption of the SPLASH as a standard spectral identifier could impact spectral exchange and searchability and enables enhanced searchability and data processing across mass spectrometry platforms.

Conflict of Interest

Pieter C. Dorrestein is on the scientific advisory board to Sirenas Marine Biosciences.

Acknowledgements

GW, SSM, DP, and OF were supported by National Institute of Health U24 DK097154 and National Science Foundation MCB 1139644, MW, PCD and NB by the National Institutes of Health 5P41GM103484 for the Center for Computational Mass Spectrometry; ES by SOLUTIONS (European Union's Seventh Framework Programme Grant Agreement No. 603437); PCD by the European Union's Horizon2020 program under the Grant agreement No. 634402 (METASPACE); RFM was funded by the Database Integration Coordination Program of the National Bioscience Database Center, Japan. European MassBank is supported by the NORMAN Association (France) and hosted by the Helmholtz Centre for Environmental Research, Discussions with anonymous parties, T. Hofstetter and the reviewer feedback are gratefully acknowledged.

References

1. Vinaixa, M., *et al. TrAC-Trends Anal. Chem.*, **78**, 23-35 (2016).

2. Horai, H. *et al. J. Mass Spectrom.*, **45,** 703-714 (2010). (a) http://www.massbank.jp (accessed 8 June 2016), (b) http://massbank.eu/MassBank/ (accessed 8 June 2016), (c) http://mona.fiehnlab.ucdavis.edu/ (accessed 8 June 2016).

3. Wang, M. *et al. Nat. Biotech.*, **34**, 828-837 (2016).4. Wishart, D.S. *et al. Nucleic Acids Res.* **41** D801-807 (2013).

5. Haug, K. *et al. Nucl. Acids Res.* 1-6 (2012) doi:10.1093/nar/gks1004

6. Stein S.E. *et al.* NIST Mass Spectral Search Program and NIST/EPA/NIH Mass Spectral Library version 2.2, June 2014. National Institute of Standards and Technology, U.S. Secretary of Commerce, USA.

7. mzCloud https://www.mzcloud.org/ (accessed 8 June 2016).

8. Heller, S.R. *et al. J. Chem. Inf.* **5** 7 (2013).

9. Heller, S.R. *et al. J. Chem. Inf.* **7** 23 (2015).

10. Southan, C. *J. Chem. Inf.* **5** 10 (2013).

11. National Institute of Standards and Technology, *Secure Hash Standard*, FIPS PUB 180-4, http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf (accessed 8 June 2016).

12. Sawada *et al. Phytochemistry*. **82** 38-45 (2012).

Author manuscript. Published in final edited form as: Nat Biotechnol. 2016 November 08; 34(11): 1099–1101. doi:10.1038/nbt.3689.

DOI:10.1038/nbt.3689

13. Kind, T. *et al*. *Anal. Chem.* **81**:24 10038–10048 (2009).

14. Stein, S.E. and Scott, D.R. *J. Am. Soc. Mass Spectrom.* **5** 859-866 (1994).

15. Preshing, J. http://preshing.com/20110504/hash-collision-probabilities/ (accessed 8 June 2016).

16. Pluskal, T. *et al. BMC Bioinformatics,* **11**:395 (2010).

17. Tsugawa, H. *et al. Nature Methods*, **12** 523–526 (2015).

18. Stravs, M.A. *et al, J. Mass Spectrom.* **48**:1 89-99 (2013).

19. Skogerson, K. *et al. BMC Bioinformatics*, **12** 321 (2011).

20. Spjuth, O. *et al. BMC Bioinformatics*, **8** 59 (2007).

21. Mass Spectrometry Development Kit (MSDK) https://msdk.github.io/ (accessed 8 June 2016).

Author manuscript. Published in final edited form as: Nat Biotechnol. 2016 November 08; 34(11): 1099–1101. doi:10.1038/nbt.3689.

DOI:10.1038/nbt.3689

This is an open access article published under a Creative Commons Attribution (CC-BY) License, which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

ACS AuthorChoice

analytical chemistry

Perspective

Cite This: Anal. Chem. 2018, 90, 649−656

pubs.acs.org/ac

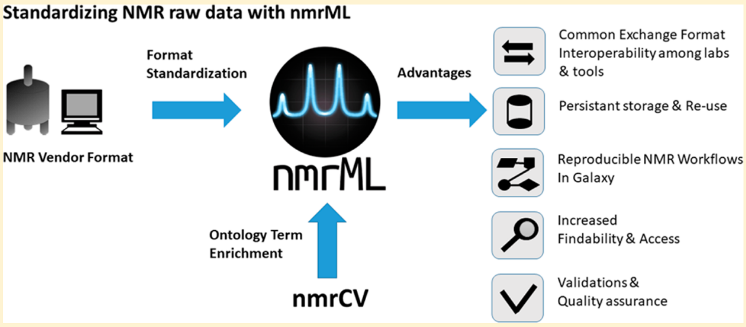
Downloaded via 217.234.218.81 on December 25, 2019 at 10:36:58 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data

Daniel Schober,[*,†] Daniel Jacob,[‡] Michael Wilson,[§] Joseph A. Cruz,[§] Ana Marcu,[§] Jason R. Grant,[§] Annick Moing,[‡] Catherine Deborde,[‡] Luis F. de Figueiredo,[∥] Kenneth Haug,[∥] Philippe Rocca-Serra,[⊥] John Easton,[#] Timothy M. D. Ebbels,[⊗] Jie Hao,[⊗] Christian Ludwig,[$] Ulrich L. Günther,[×] Antonio Rosato,[○] Matthias S. Klein,[¶] Ian A. Lewis,[¶] Claudio Luchinat,[○] Andrew R. Jones,[∇] Arturas Grauslys,[∇] Martin Larralde,[+] Masashi Yokochi,[◆] Naohiro Kobayashi,[◆] Andrea Porzel,[&] Julian L. Griffin,[%] Mark R. Viant,[■] David S. Wishart,[§] Christoph Steinbeck,[∥] Reza M. Salek,[*,∥] and Steffen Neumann[†]

[†]Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany
[‡]INRA, Univ. Bordeaux, UMR1332 Fruit Biology and Pathology, Metabolome Facility of Bordeaux Functional Genomics Center, MetaboHUB, IBVM, Centre INRA Bordeaux, 71 av Edouard Bourlaux, F-33140 Villenave d'Ornon, France
[§]Departments of Computing Sciences and Biological Sciences, University of Alberta, Edmonton, Canada T6G 2E8
[∥]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.
[⊥]University of Oxford, e-Research Centre1, 7 Keble Road, Oxford OX1 3QG, U.K.
[#]School of Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.
[⊗]Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, SW7 2AZ, U.K.
[$]Institute of Metabolism and Systems Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.
[×]Institute of Cancer and Genomic Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.
[○]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Florence, Italy
[¶]Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
[∇]Institute of Integrative Biology, University of Liverpool, Bioscience Building, Crown Street, Liverpool L69 7ZB, U.K.
[+]Ecole Normale Supérieure Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan, France
[◆]Institute for Protein Research (IPR), Osaka University, 3-2 Yamadaoka, Suita-shi, Osaka, 565-0871, Japan
[&]Department of Bioorganic Chemistry, Leibniz Institute of Plant Biochemistry, 06120 Halle (Saale), Germany
[%]Department of Biochemistry, University of Cambridge, Downing Site, Cambridge CB2 1QW, U.K.
[■]School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.

Supporting Information


Standardizing NMR raw data with nmrML
NMR Vendor Format
Format Standardization
nmrML
Advantages
Ontology Term Enrichment
nmrCV
Common Exchange Format Interoperability among labs & tools
Persistant storage & Re-use
Reproducible NMR Workflows In Galaxy
Increased Findability & Access
Validations & Quality assurance

ABSTRACT: NMR is a widely used analytical technique with a growing number of repositories available. As a result, demands for a vendor-agnostic, open data format for long-term archiving of NMR data have emerged with the aim to ease and encourage

continued...

Received: July 18, 2017
Accepted: October 16, 2017
Published: October 16, 2017

© 2017 American Chemical Society

DOI:10.1021/acs.analchem.7b02795

sharing, comparison, and reuse of NMR data. Here we present nmrML, an open XML-based exchange and storage format for NMR spectral data. The nmrML format is intended to be fully compatible with existing NMR data for chemical, biochemical, and metabolomics experiments. nmrML can capture raw NMR data, spectral data acquisition parameters, and where available spectral metadata, such as chemical structures associated with spectral assignments. The nmrML format is compatible with pure-compound NMR data for reference spectral libraries as well as NMR data from complex biomixtures, i.e., metabolomics experiments. To facilitate format conversions, we provide nmrML converters for Bruker, JEOL and Agilent/Varian vendor formats. In addition, easy-to-use Web-based spectral viewing, processing, and spectral assignment tools that read and write nmrML have been developed. Software libraries and Web services for data validation are available for tool developers and end-users. The nmrML format has already been adopted for capturing and disseminating NMR data for small molecules by several open source data processing tools and metabolomics reference spectral libraries, e.g., serving as storage format for the MetaboLights data repository. The nmrML open access data standard has been endorsed by the Metabolomics Standards Initiative (MSI), and we here encourage user participation and feedback to increase usability and make it a successful standard.

Nuclear magnetic resonance (NMR) spectroscopy is an important analytical tool in organic chemistry, biochemistry, natural products research, structural biology, and metabolomics. Recently the need for an open NMR data standard covering the free induction decay (FID) to support data reproducibility has been acknowledged.[1] As instrument vendors typically provide the data processing software and produce evolving data formats together with the instrument hardware, developers of third party NMR analysis software often need to devote considerable effort into reading and writing these vendor-specific formats. This applies both to commercial software and to community developed open-source tools such as the BATMAN R package,[2] Bayesil,[3] NMRProcFlow,[4] rNMR[5] and MetaboLab.[6] With the recent termination of the Agilent/Varian NMR spectrometer range, the question of long-term readability of discontinued vendor formats has become paramount for a growing NMR community. Data in proprietary formats can age quickly, and NMR data stored in such formats can become obsolete, making valuable results inaccessible and irreproducible in the long term. Also, spectra processing and quantification tools would benefit from a standardized storage format for processed NMR data, i.e., serving workflow systems. For NMR data repositories such as MetaboLights,[7] Metabolomics WorkBench,[8] Human Metabolome Database HMDB,[9] and BioMagResBank,[10] key questions regarding long-term data persistence, i.e., on sustainability, usability, and accessibility are arising.

Currently, the most widely used open data exchange format for NMR data is JCAMP-DX version 6.0,[11] but due to the broad scope and complexity of this format, many different vendor-dependent variants exist. Coordinated updating for all variants, in order to reflect the state of the art in NMR methodology, is rarely seen in this 30 year old format. This variability can lead to incompatibilities between different software packages, and as a result no content-based (semantic) validation of JCAMP-DX is available. While JCAMP-DX is likely to remain in use for NMR data capture for many years, it is clear that alternative approaches, such as XML or JavaScript Object Notation (JSON) with peer-maintained ontologies, would be beneficial.

The first efforts toward establishing an XML-based open NMR standard and controlled vocabulary were discussed in 2007 by the ontology working group[12] of the Metabolomics Standards Initiative (MSI)[13] and a consortium of U.K. universities discussing minimal reporting guidelines.[14] In 2011, a series of initiatives by members of the NMR-based metabolomics and biomolecular NMR communities were launched to explore the creation of a new community standard for NMR data exchange and storage. This included meetings attended by NMR stakeholders including metabolomics database representatives and vendors. This initiative and subsequent meetings were then taken over by the COSMOS (COordination of Standards in MetabOlomicS) EU FP7 consortium,[15] aiming to coordinate the establishment of a persistent NMR data format and open source data analysis tools for the NMR community. The main goals were

(1) *Data sharing in an open vendor-agnostic manner*, so that users, tool developers, and public repositories can import or export data to support integrated (meta-)analysis and secondary data usage.

(2) *Search and retrieval of relevant results*, minimizing alternate ways of encoding the same information, so that data sets with a similar setup can be identified and compared.

(3) *Spreading best practices and evaluation of the results*, whereby the data quality can be assessed in light of intelligibility and completeness along minimum information standards supported by automatic validation aids.

(4) *Improved data persistence and traceability over time*, delivering a self-describing easy-to-use yet robust raw data storage format to support long-term archiving.

From such efforts, it was decided that the new data format would be called nmrML (for NMR Markup Language) and it should

(1) Be compatible with existing vendor formats (Varian/Agilent, Bruker, JEOL) and partially compatible with certain variants of JCAMP-DX.

(2) Be XML-based, so as to be similar to established XML formats by the Proteomics Standard Initiatives (PSI), i.e., mzML for mass spectrometry.[16]

(3) Support the use of controlled vocabularies/ontologies to annotate spectral data and metadata with standardized community descriptors, which can be maintained in a decentralized peer production manner.

(4) Initially focus on the capture of small molecule NMR data with macromolecular NMR data being addressed in succession; but be flexible enough to be expanded in scope.

(5) Be easy to understand and integrate into existing open analysis and processing software.

(6) Contain sufficient spectrometer data, acquisition, and processing metadata to permit the reconstruction of the NMR spectrum and experiment.

(7) Capture coarse-grained spectral assignment data for molecule identification and quantification in chemical mixtures. Capture fine-grained assignment and chemical structure data of pure-compound spectra for use in organic synthesis and natural product studies, medicinal chemistry, and reference NMR spectral libraries.

Under these development constraints, members of the nmrML COSMOS team created the nmrML data standard, the necessary software support, and fostered support from databases to both accept and display nmrML data. Figure 1 summarizes available nmrML compliant tools and functionalities in support

DOI:10.1021/acs.analchem.7b02795

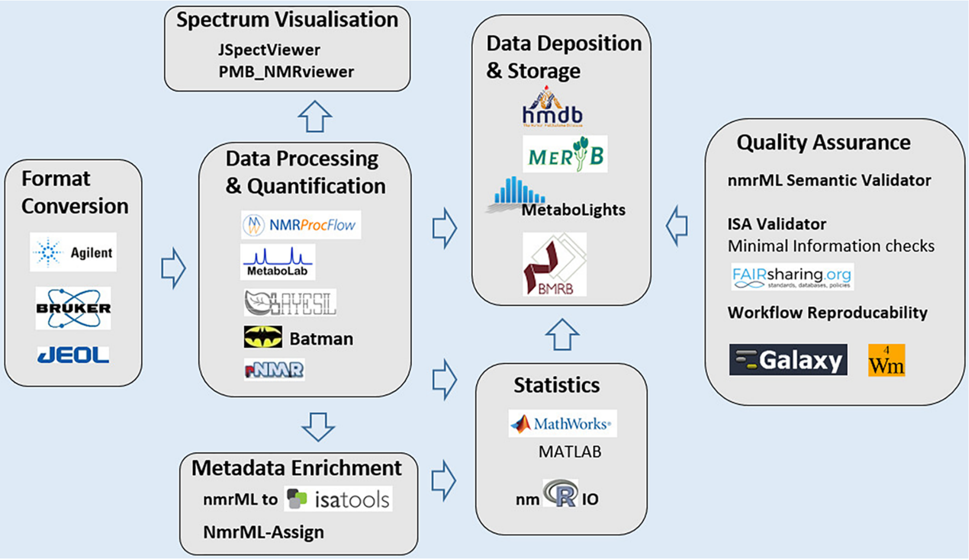


Figure 1. A prototypical metabolomics workflow for NMR data processing and storage is shown and nmrML-aware tools supporting each workflow step are illustrated. Vendor to nmrML converters, NMR data processing, and visualization tools as well as public repositories that accept nmrML as standard data format are highlighted. Parsers for MATLAB and R, which make nmrML data accessible to statistics tools, and content validators that assist in data quality control and workflow reproducibility are shown. Many of our tools already run in Galaxy-based workflow management environments. Bruker logo reprinted with kind permission from Bruker BioSpin Group, Copyright 2017. JEOL logo reprinted with kind permission from JEOL (Germany) GmbH, Copyright 2017. Agilent Technologies Corporate Signature Copyright 1999 Agilent Technologies, Inc. Reproduced with permission, Courtesy of Agilent Technologies, Inc. MATLAB is a registered trademark and reprinted with kind permission from The MathWorks, Inc.

of a typical NMR data handling workflow for a metabolomics or similar experiment.

■ MATERIALS AND METHODS

The nmrML format specification is composed of an XML Schema Definition (XSD) and an accompanying controlled vocabulary called nmrCV. Leveraging on existing efforts, the nmrML development started by updating a predecessor XSD developed at The Metabolomics Innovation Centre (TMIC, http://www.metabolomicscentre.ca/exchangeformats.htm) in Edmonton, Canada, with additional elements and structures from a BML-NMR XSD developed at the University of Birmingham.[17] Both of these efforts were integrated, expanding the TMIC predecessor, as it was already capturing the basic raw data and had the controlled vocabulary (CV) reference mechanism in place. The nmrML CV referencing mechanism and basic XML architecture was inspired by mass spectrometry markup language (mzML), the PSI standard mass spectrometry data format used in proteomics and metabolomics.[16] The mzML community standard captures raw MS spectral data, instrument parameters, experiment metadata, and peak assignment, as well as compound quantitation data. Given the similarity in data capture, storage, and retrieval between modern MS and NMR experiments, many of the successful features found in mzML were transferred and adapted to nmrML. The NMR.owl CV by the MSI,[12] and a parallel TMIC effort NMR CV, developed to serve the TMIC XSD, were integrated. The merged nmrCV organizes common and essential NMR terms into a simple is-a class hierarchy (taxonomy). The nmrML 1.0.0 format presented here is the outcome of these integration efforts and will serve as the MSI recommended common data standard and terminology for open access NMR data. While the nmrML.xsd mostly covers raw data, it also provides for some NMR data elements computed by open access NMR processing and quantification tools. Development was coordinated via mailing lists, video conferences, and during multiple workshops and hackathons. The choice of XML was motivated by technical maturity, flexibility and universality of XML in both capturing and presenting scientific data. There is an abundant XML expertise to leverage on, as XML resides at the base of the semantic Web stack. The appearance of all knowledge capture XML elements can be controlled via the XSD (mandatory vs optional) and hence allows for content completeness checks. We implemented converter Web services to generate valid nmrML from vendor raw data files. Links to nmrML compliant databases as well as NMR processing and spectrum visualization software are provided in Table 1. Format parsers, application program interfaces (APIs), and validation Web services have been set up. All code libraries, an issue tracker as well as a file versioning and release policy are available on the developer's GitHub pages at https://github.com/nmrML/nmrML.

■ RESULTS

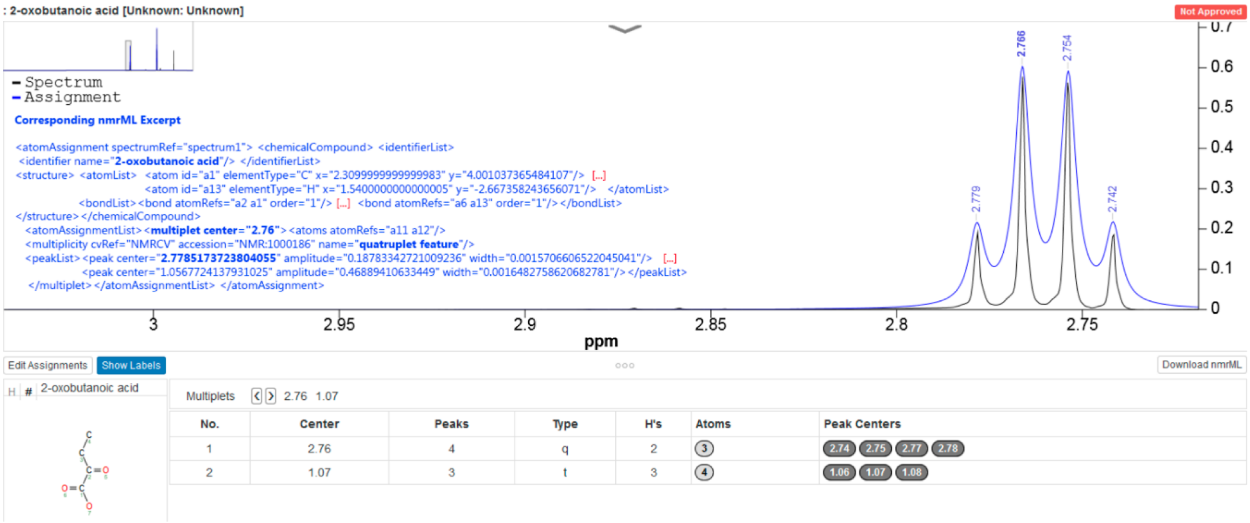
The nmrML core specification, including the XSD and nmrCV, can be found at http://nmrml.org. The referenced nmrCV.owl currently contains over 600 terms and is indexed under the National Center for Biomedical Ontology (NCBO) Bioportal ontology library.[18] Our documentation Web site (http://nmrml.org/examples) provides tutorial material and videos, code examples for single compound reference spectra, as well as mixed-compound 1D and 2D NMR spectra.

nmrML Architecture. The nmrML XSD element hierarchy contains multiple sections that organize the information that can appear in an nmrML XML data file in a community-agreed and self-explanatory way. This facilitates understanding of the format by both humans and by data processing software alike. The current top level XSD structure provides high-level base elements for the grouping and capture of NMR data, describing the

Table 1. Non-exhaustive list of nmrML compatible open source software, clustered by tool category

tool category	tool name	key functions	URL	developer
format converters	nmrML converter (Java)	converts vendor to nmrML format (*recommended*)	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Java	Institut National de la Recherche Agronomique (INRA), France
	nmrML converter (Python)	converts vendor to nmrML format	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/python/pynmrml	The Metabolomics Innovation Center (TMIC), Canada
	nmrML to ISA converter	generates prepopulated ISA files from nmrML files	https://github.com/ISA-tools/nmrml2isa	EMBL-EBI, United Kingdom
	BMSxNmrML	converts BMRB metabolomics entries to nmrML format	http://bmrbdep.pdbj.org/en/bmsxnmrml.html	Institute for Protein Research (IPR), Japan
parsers	MATLAB parser	MATLAB functions parsing and decoding nmrML files, and also writing MATLAB data into nmrML format.	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Matlab	Imperial College London (ICL), United Kingdom
	nmRIO	R package for parsing and decoding nmrML files	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/R/nmRIO	Leibniz Institute of Plant Biochemistry (IPB), Germany
	nmrGlue	modules for working with NMR data in Python; for processing, analyzing, and inspecting NMR data	https://github.com/jjhelmus/nmrglue	Argonne National Laboratoriy (ANL), IL, USA
data validators	nmrML semantic validator	XML Schema compliance and rule-based validation of CV term usage	http://nmrml.org/validator	Leibniz Institute of Plant Biochemistry (IPB), Germany
spectrum viewers	JSpectraViewer (JSV)	interactive 1D NMR Spectral viewer used in tools such as Bayesil and nmrML-Assign	http://nmrml.bayesil.ca	The Metabolomics Innovation Center (TMIC), Canada
NMR processing, and identification/quantification tools	NMRProcFlow	interactive 1D NMR spectral viewer, spectral processing and quantification tool dedicated to metabolomics	http://nmrprocflow.org	Institut National de la Recherche Agronomique (INRA), France
	Bayesil	automated compound identification, quantification and annotation from 1D NMR spectra	http://bayesil.ca, http://tmic.bayesil.ca	The Metabolomics Innovation Center (TMIC), Canada
	nmrML-Assign	nmrML conversion, annotation and peak assignment to compounds for reference 1D NMR spectra	http://nmrml.bayesil.ca	The Metabolomics Innovation Center (TMIC), Canada
	Batman	Bayesian deconvolution and automated quantification of metabolites from 1D NMR spectra	http://batman.r-forge.r-project.org	Imperial College London (ICL), United Kingdom
	rNMR	region-of-interest based NMR spectra quantification from 1D and 2D NMR spectra	http://rnmr.nmrfamde.wisc.edu	University of Calgary (U of C), Canada
workflow tools	PhenoMeNal app library	lists containerized nmrML aware tools to build Galaxy NMR workflows	http://portal.phenomenal-h2020.eu/app-library	EMBL-EBI, United Kingdom
	SOMA:tameNMR	NMR data processing and analysis via Galaxy Workflows	https://github.com/pgb-liv/tameNMR	University of Liverpool (UoL), United Kingdom

DOI:10.1021/acs.analchem.7b02795



Corresponding nmrML Excerpt

```
<atomAssignment spectrumRef="spectrum1"> <chemicalCompound> <identifierList>
 <identifier name="2-oxobutanoic acid"/> </identifierList>
<structure> <atomList>  <atom id="a1" elementType="C" x="2.3099999999999983" y="4.001037365484107"/> [...]
                        <atom id="a13" elementType="H" x="1.5400000000000005" y="-2.667358243656071"/>   </atomList>
         <bondList><bond atomRefs="a2 a1" order="1"/> [...] <bond atomRefs="a6 a13" order="1"/></bondList>
</structure></chemicalCompound>
  <atomAssignmentList><multiplet center="2.76"><atoms atomRefs="a11 a12"/>
  <multiplicity cvRef="NMRCV" accession="NMR:1000186" name="quatruplet feature"/>
  <peakList><peak center="2.7785173723804055" amplitude="0.18783342721009236" width="0.0015706606522045041"/>  [...]
                <peak center="1.0567724137931025" amplitude="0.46889410633449" width="0.0016482758620682781"/></peakList>
   </multiplet></atomAssignmentList> </atomAssignment>
```

No.	Center	Peaks	Type	H's	Atoms	Peak Centers
1	2.76	4	q	2	3	2.74 2.75 2.77 2.78
2	1.07	3	t	3	4	1.06 1.07 1.08

Figure 2. Assignment of an identified molecule in a single compound spectrum, generated in nmrML-Assign and displayed using the JSpectraViewer (JSV). An uploaded raw FID for the 2-oxobutanoic acid reference compound was automatically processed with Bayesil. The resulting interactive JSV spectrum then allows the assignment of peaks to specific atoms, using the nmrML-Assign tool. The assignment metadata is then saved in the nmrML format (see https://github.com/nmrML/nmrML/tree/master/examples/reference_spectra_examples/hmdb). An excerpt view of the corresponding nmrML code (blue code inset) is shown for the quadruplet assignment (Multiplet no. 1) of the second peak (bold code). The corresponding HMDB entry is available from http://www.hmdb.ca/metabolites/HMDB00005, with the ^{1}H spectrum found at http://www.hmdb.ca/spectra/nmr_one_d/1024.

nmrML version, the sources of the controlled vocabularies or ontologies used for metadata annotation, the data depositor contact, source files/formats, software lists, the instrument configuration, sample information (e.g., solvent and reference standards), acquisition settings, and data processing information. This is followed by the spectral FID raw data, as a base64-encoded binary. In addition to such a "minimal" nmrML data file, additional information such as molecule identification/spectral assignment metadata and quantification data can also be included. For example, if the NMR data is for a pure reference compound or a newly isolated/synthesized single chemical, the nmrML file can include data on the chemical structure and corresponding atom-specific peak feature assignments (see example generated by nmrML-Assign in Figure 2 or http://nmrml.org/examples/3). If the NMR data is for a complex mixture, consisting of many different compounds from an analytical setting, the nmrML file can include data on peak positions, integrated peak areas, and putative peak assignments, together with relative or absolute concentrations of some or all of the compounds but no annotation of individual peak features to atom environments (see http://nmrml.org/examples/4). Code examples of a minimal nmrML data XML file as well as for the expanded metadata case are provided on the examples page (http://nmrml.org/examples).

The nmrML structure consists of an XSD that allows it to reference a dedicated NMR controlled vocabulary (nmrCV). The XSD defines the allowed XML structure, whereas the controlled vocabulary provides the terminology to describe the NMR data in detail using standardized textual values for XML-defined tags. In areas where the terminology is likely to change faster than the nmrML XSD can be updated, the representation is branched out from XSD to CV-usage. This approach can accommodate rapid technology/terminology changes in a flexible way, as the CV can be maintained externally by a larger NMR user peer group: for example, terms for new NMR probes can be represented in a nmrML file by requesting the addition of corresponding new CV terms in the nmrCV, without the need for a full XSD and any subsequent software revisions. The combined usage of XML and a separate CV also allows multiple validation levels to be established (see below). The CV referencing mechanism is explained in detail on the documentation pages.

Tools Compatible with nmrML. We have created Web-based easy-to-use tools to make nmrML more accessible to the broader organic chemistry and metabolomics communities. To ensure that nmrML will be broadly adopted by life sciences and chemical researchers, these tools cover a large fraction of a typical NMR data acquisition, processing and storage workflow to generate, convert, process, validate, and publish nmrML files (Figure 1). Additionally, we have worked closely with open source and commercial tool developers to encourage nmrML format support and adoption. We have summarized efforts already leveraging on the nmrML format in Table 1.

nmrML Converters, Parsers, and Validators. Format converters translate the exchange syntax from vendor raw data formats into XSD-compliant nmrML by means of mappings from Bruker "acqus" or Varian "procpar" raw files to nmrML elements and CV terms. An extensive parameter mapping table is available in the documentation pages. A comprehensive JAVA-based converter automatically generates valid nmrML files from Bruker, Agilent/Varian, and JEOL raw files. It is also available as a Web service (http://nmrml.org/converter) and Docker container. It can be run in batch mode for high-throughput batch conversion of multiple zipped raw data. A Python-based converter that uses the nmrGlue API[19] to access the vendor parameters is also available. Also an nmrML2ISA parser,[20] written in Python, has the ability to read experimental NMR data and metadata from nmrML data files and passing it over to an autogenerated ISA-Tab[21] assay file, i.e., defining a basic metadata backbone ISA-Tab format, i.e., for submission to the MetaboLights repository.[7] In addition, nmrML bindings for multiple programming languages

DOI:10.1021/acs.analchem.7b02795

(Java, Python, Ruby) as well as for widespread data analysis tools like the R statistics package, MATLAB, and open source NMR tools exist. A parser called nmRIO makes nmrML content available to R-based tools such as Batman and rNMR for higher level analysis. A MATLAB parser renders nmrML data available to the MATLAB tool for further statistical processing. An nmrML semantic validator allows the revisal of the correct implementation of manually populated or enriched nmrML files, with regard to XML schema compliance, CV term usage, and allowed term cardinalities. At the core, the XML syntax and structural validity of nmrML XML instances, such as XML element and attribute position, order, and cardinality, can be checked by any validating XML parser against the nmrML.xsd, which defines these allowed elements and their expected characteristics. On the next higher layer, so-called mapping rules can enforce semantic validity[22] of the ontological descriptions used, by testing which CV terms are allowed in which elements. The elements with their allowed CV descriptor hierarchies are outlined in a mapping rule file. The OpenMS/Topp-based[23] nmrML validator (http://nmrml.org/validator) checks that these higher level semantic criteria are being met in a given XML instance. For example, a validation rule file can enforce minimal reporting guidelines such as the MSI-sanctioned Core Information for Metabolomics Reporting (CIMR; http://mibbi.sourceforge.net/projects/CIMR.shtml, accessed November 27, 2017). These validation scenarios make nmrML more easily accessible to quality assurance than JCAMP-DX or other more verbose and equivocal formats that do not rely on controlled vocabularies.

nmrML Data Processors and Viewers. The following tools facilitate NMR data processing and compound identification. nmrML-Assign (http://nmrml.bayesil.ca) is a JavaScript Web application based on Bayesil that allows users to upload vendor formatted 1D NMR raw data or nmrML and to then interactively add compound identification metadata (see Figure 2, Example 3). The Bayesil-generated interactive spectrum allows assigning peaks to specific atoms in a proposed molecule after the Bayesil Web service[3] was used to upload a chemical structure and perform a spectral prediction to help with the assignment process. The assigned atoms are displayed on both the spectrum and the molecule image. Once the assignment process is complete, the annotated file can be saved as enriched nmrML file, which can then be reloaded and interactively viewed and edited or submitted to HMDB. nmrML-Assign works both with ^{1}H and ^{13}C NMR spectra in Bruker or Agilent/Varian format. Bayesil also allows users to upload 1D spectra of biological mixtures (e.g., serum, plasma, cerebrospinal fluid) as shown in Example 4 on our Web site and to perform an automated assignment and quantification of all visible peaks.

The Batman R package estimates metabolite relative concentrations from spectral data and automatically assigns them to metabolites from a target list. Batman can access nmrML data and is using the nmRIO parser. rNMR[5] is a region-of-interest rather than peak-list-based software for visualizing, assigning, and quantifying metabolites in complex 1D and 2D NMR data. The upcoming version of rNMR will read nmrML files directly and can convert them into its native data format. NMRProcFlow is a pipeline tool for the reproducible processing and visualization of 1D NMR data in metabolomics. It allows to pipe processed NMR data as tabular data matrix to statistics workflow tools like biostatflow.org. It relies on the NMR spectra viewer (https://github.com/nmrML/nmrML/tree/master/tools/Visualizers/PMB_NMRviewer), as its design acknowledges iterative parameter adjustments by means of repeated visual inspection by the user.

nmrML Compatible Databases. A principal objective behind the establishment of nmrML is to ensure data continuity and persistence in NMR repositories and reference libraries. Several key NMR experiment and reference databases now support the upload, storage, display, and download of nmrML data. HMDB, with more than 1500 1D ^{1}H and ^{13}C NMR spectra collected at 500 and 600 MHz ("Human Metabolome Database: Database Statistics", http://www.hmdb.ca/statistics, accessed May 15, 2017), describes more than 1000 reference spectra for pure compounds in the Human Metabolome Library (HML, http://www.hmdb.ca/hml). More than 600 metabolites in HMDB now include NMR reference spectra with complete spectral assignments. These metabolites have 1D NMR annotated spectra available and are downloadable in the nmrML format. Other databases such as DrugBank,[24] YMDB[25] and ECMDB[26] plan to support nmrML compatible reference spectra in the future. BMRB entries are available in XML and RDF, as common open representations of NMR-STAR data format.[27] BMRB has archives of time-domain data and fully assigned nmrML files are accessible, which were generated from BMRB/XML files via the BMSxNmrML converter (see Table 1). In addition to the growing collection of reference spectral libraries, the open access NMR data repository MetaboLights[7] has experimental NMR data archival, which now accepts nmrML data from depositors and allows one to extract basic ISA-Tab metadata from it (see above). It now handles nmrML data from biological mixtures as well as from pure reference compounds. The MeRy-B[28] plant metabolomics NMR knowledge base accepts both JCAMP-DX and nmrML format with the plan to fully adopt nmrML in order to leverage ontological spectra preprocessing terms embedded within nmrML. Work is underway to have the Metabolomics WorkBench[8] accept nmrML data as part of the international MetabolomeXchange initiative (metabolomexchange.org/).

Pipelines and Workflow Support. With the recent push to standardize and facilitate the access to data processing workflows,[29] devoted workflow environments such as Galaxy[30] have gained more weight, the intent here being transparency, traceability, and reproducibility of pipeline-generated data and audit. Galaxy-based metabolomics analysis pipelines are emerging[31] and some are in development for NMR data, such as W4M-NMR[31] (http://workflow4metabolomics.org/the-nmr-workflow) and SOMA:tameNMR (https://github.com/pgb-liv/tameNMR). The NMR processing tool NMRProcFlow[4] uses nmrML as its native spectral data format and containerization of modules for workflow integration is progressing. To foster nmrML as input format for Galaxy workflow pipelines, the PhenoMeNal projects App library portal (http://portal.phenomenal-h2020.eu/app-library already provides nmrML-aware tools (like the nmrML converter) as containers for NMR workflow integration.

■ DISCUSSION

This Perspective describes the first iteration of nmrML (version 1.0.0). We have designed and developed a flexible, open standard data format called the NMR Markup Language (nmrML) for capturing and disseminating NMR data for small molecules. This represents a community-driven effort that involved extensive consultations and many metabolomics, NMR spectroscopy, chemoinformatics, and computing science laboratories from across Europe and North America. Further enhancements are planned for nmrML, and these will include extensions to *n*D NMR data and the inclusion of macromolecular data in the XML and additional terms in nmrCV. Currently, only basic processed data is captured, e.g., for molecule identification and

DOI:10.1021/acs.analchem.7b02795

quantification, and the latter is equivalent to what mzTab stores for MS data and what is captured in mzIdentML[32] and mzQuantML.[33] The introduction of nmrML hence brings NMR spectroscopy in alignment with existing data standardization efforts in metabolomics, such as mzML for mass spectrometry and will ultimately contribute to cross-technology and multiple omics data comparison. We hope further tools like XEASY[34] for macromolecular NMR analysis and NMRPipe[35] for *n*D NMR will leverage on nmrML in the future. MetaboLab[6] provides high-throughput preprocessing for MATLAB driven NMR statistics and is currently implementing an nmrML parser for standardized data import. In addition, further metadata will be added to nmrML, i.e., as required to store *n*D spectra. In addition to the persistent data storage/exchange standard and CV, we have also described and developed database support and software tools that make use of nmrML. These tools include nmrML viewers, nmrML data converters, processors, and annotators, and these will facilitate the widespread adoption of nmrML and permit the facile generation of nmrML data from proprietary vendor formats. Bruker Corp. indicated willingness to incorporate the nmrML converter into their TopSpin software as nmrML file format export option. Although the benefits to individual users will become more evident as more software supports this open standard, users can already store and archive their NMR data in a persistent format, which stays readable in the long term. Users can extract NMR metadata into the ISA Tab format,[20] e.g., easing submissions to public databases such as MetaboLights. Data in local institutional repositories will gain value through eased reanalysis of old data with future state-of-the-art tools. Furthermore, users can integrate their data into workflow management systems, which eases repetitive processing tasks. Reproducibility and trustability of data is further increased by community data validation, e.g., in terms of minimal information coverage, and will result in increased data quality. The use of nmrML validators will allow users to check nmrML files with regard to consistency and content completeness. Together with ISA-Tab metadata validation, this will greatly contribute to overall quality assurance and traceability of NMR data. The nmrML standard also enables easier multicenter collaborations, e.g., allowing for an interoperable data exchange format when communicating with a regional NMR metabolomics center. It also eases comparison of results among different laboratories, e.g., for the purpose of standardization or SOP development. On the tool developers side, nmrML can save programmers' time and effort to write multiple parsers for all vendor formats. Given cross-communication between the MSI, PSI, and other standardization governance bodies, harmonized data standards will ease community integration, i.e., bridging over different technologies, e.g., by allowing MS and NMR data comparisons or even multiple omics investigations. This would pave the way for more integrative systems biology approaches.

Overall the nmrML specification and the expandable nmrCV will allow for a detailed standardized description of NMR workflow functionalities. The use of nmrML in workflow tools like tameNMR and the reuse of containerized workflow components in recombinable app libraries will allow NMR data processing to be more traceable and rerunnable in different (local or cloud) environments. The capture of selected basic metadata within the same nmrML file as the data eases pipeline development as it reduces the complexity of file tracking in Galaxy, as data moves between modules.

A recent survey (http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Deliverable8.1.pdf) on data standards usage among the metabolomics community indicated that 13.5% of NMR practitioners are already using nmrML, about the same number of people indicating that they use JCAMP.

Further testing of the current XSD with diverse experimental configurations is required to increase coverage, fitness of purpose, and future flexibility. We hence welcome any community feedback and engagement via our email list at https://groups.google.com/forum/?hl=en#!forum/nmrml/join to improve and evaluate this first nmrML release. Remarks, suggested changes, and extension requests should be sent to info@nmrml.org or via our Git issue tracker. By standardizing data descriptions, nmrML and its accompanying nmrCV will help make NMR data *Findable, Accessible, Interoperable, and Reusable,* FAIR.[36] This is particularly relevant in light of the recent push by funding bodies to have scientists conduct and publish more reproducible research.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b02795.

Serialization of the nmrML.xsd as Schema description (PDF)

Documentary material with FAQ and nmrML tutorial (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: dschober@ipb-halle.de. Phone: +49 (0)345 5582 1476.
*E-mail: reza.salek@ebi.ac.uk. Phone: +44 (0)1223 48 4502.

ORCID

Daniel Schober: 0000-0001-8014-6648
Antonio Rosato: 0000-0001-6172-0368
Julian L. Griffin: 0000-0003-1336-7744

Author Contributions

D.S. drafted the manuscript, coordinated the nmrML updates, contacted NMR data repository developers and created tutorial material. D.S., D.J., M.W., and P.R.-S. implemented the nmrCV. D.S. and S.N. set up the semantic validator and created the mapping files. M.W. created and updated the nmrML.xsd, set up the Git and nmrML.org home pages, and wrote the Python parser. D.J. updated the nmrML.xsd and created and maintains the JAVA converter. D.J. created example nmrML files and coordinated NMRProcFlow interactions. J.A.C. created the predecessor nmrML XSD and an nmrCV predecessor. A. Marcu and J.R.G. deployed the nmrML Assign Web server, added assigned spectra for HMDB compounds, and coordinated Bayesil interactions. A. Moing and C.D. supervised the Java converter development, MeryB example generation, and provided feedback as wet-lab NMR and metabolomics database users. L.F.d.F., K.H., and R.M.S. helped integrating nmrML format with the MetaboLights repository. L.F.d.F. contributed to the initial version of the JAVA converter and created example nmrML files. P.R.-S. worked on and advised on CV and ontology reuse and coordinated the nmrML to ISA converter development with M.L. and R.M.S. T.M.D.E. and J.H. contributed the MATLAB parser and BATMAN advice. C. Ludwig, J.E., and U.L.G. were instrumental in aligning the initial XSD to BML-NMR repository needs. C. Luchinat and A.R. were initial driving forces overseeing the overall NMR data standards coordination. M.S.K. and I.A.L. worked toward rNMR integration. A.R.J. and A.G. were contributing to the NMR workflow tool tameNMR.

DOI:10.1021/acs.analchem.7b02795

M.L. implemented the nmrML2ISA converter. M.Y. developed an nmrML converter for BMRB with guidance from N.K. A.P. tested the vendor to NMR parameter mappings. J.L.G., M.R.V., and D.S. contributed to the first round of nmrCV and CIMR MSI development. D.S.W. initiated the project and hosted the first round of XSD development and HMDB oversight. C.S. and R.M.S. initiated and coordinated the COSMOS EU project. R.M.S. advised on NMR and database issues and created example data sets. R.M.S. contributed to the nmrCV and nmrML development and MSI approval. S.N. contributed the nmrRIO parser, alignments to the semantic validator and helped with the Git. All authors contributed to, reviewed, and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was financed via the EU FP7 Project COSMOS Grant EC312941, Genome Canada and the Canadian Institutes of Health Research. I.A.L. and M.S.K. (rNMR) are supported by Alberta Innovates−Health Solutions (AIHS, Translational Health Chair) and the Natural Sciences and Engineering Research Council (NSERC, Discovery Grant 04547). D.J., A. Moing, and C.D. thank MetaboHUB ANR-11-INBS-0010 for financing. A.R.J. and A.G. acknowledge funding from the UK's Biotechnology and Biological Sciences Research Council (BBSRC) Grant BB/M020282/1. R.M.S. acknowledges BBSRC Grant BB/M027635/1 and MRC UK MEDical BIOinformatics partnership, Grant MR/L01632X/1. M.R.V. acknowledges BBSRC Grant BB/M019985/1, D.S., C.S., S.N., P.R.-S., and R.M.S. acknowledge the PhenoMeNal European Commission's Horizon2020 program, Grant 654241. We thank the late Ivano Bertini for his initial vision towards this standardization effort. The nmrML logo is copyrighted by nmrML.org.

■ REFERENCES

(1) Bisson, J.; Simmler, C.; Chen, S.-N.; Friesen, J. B.; Lankin, D. C.; McAlpine, J. B.; Pauli, G. F. *Nat. Prod. Rep.* **2016**, *33* (9), 1028−1033.

(2) Hao, J.; Astle, W.; De Iorio, M.; Ebbels, T. M. D. *Bioinformatics* **2012**, *28* (15), 2088−2090.

(3) Ravanbakhsh, S.; Liu, P.; Bjorndahl, T. C.; Bjordahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; Greiner, R.; Wishart, D. S. *PLoS One* **2015**, *10* (5), e0124219.

(4) Jacob, D.; Deborde, C.; Lefebvre, M.; Maucourt, M.; Moing, A. *Metabolomics* **2017**, *13* (4), 36.

(5) Lewis, I. A.; Schommer, S. C.; Markley, J. L. *Magn. Reson. Chem.* **2009**, *47* (S1), S123−S126.

(6) Ludwig, C.; Günther, U. L. *BMC Bioinf.* **2011**, *12*, 366.

(7) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. *Nucleic Acids Res.* **2013**, *41*, D781−D786.

(8) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44* (D1), D463−D470.

(9) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. *Nucleic Acids Res.* **2013**, *41*, D801−D807.

(10) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; et al. *Nucleic Acids Res.* **2008**, *36*, D402−D408.

(11) Davies, A. N.; Lampen, P. *Appl. Spectrosc.* **1993**, *47* (8), 1093−1099.

(12) Sansone, S.-A.; Schober, D.; Atherton, H. J.; Fiehn, O.; Jenkins, H.; Rocca-Serra, P.; Rubtsov, D. V.; Spasic, I.; Soldatova, L.; Taylor, C.; Tseng, A.; Viant, M. R. *Metabolomics* **2007**, *3* (3), 249−256.

(13) Fiehn, O.; Robertson, D.; Griffin, J.; van der Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L. W.; Goodacre, R.; Hardy, N. W.; Taylor, C.; et al. *Metabolomics* **2007**, *3* (3), 175−178.

(14) Rubtsov, D. V.; Jenkins, H.; Ludwig, C.; Easton, J.; Viant, M. R.; Günther, U.; Griffin, J. L.; Hardy, N. *Metabolomics* **2007**, *3* (3), 223−229.

(15) Salek, R. M.; Neumann, S.; Schober, D.; Hummel, J.; Billiau, K.; Kopka, J.; Correa, E.; Reijmers, T.; Rosato, A.; Tenori, L.; et al. *Metabolomics* **2015**, *11* (6), 1587−1597.

(16) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; et al. *Mol. Cell. Proteomics* **2011**, *10* (1), R110.000133.

(17) Ludwig, C.; Easton, J. M.; Lodi, A.; Tiziani, S.; Manzoor, S. E.; Southam, A. D.; Byrne, J. J.; Bishop, L. M.; He, S.; Arvanitis, T. N.; Günther, U. L.; Viant, M. R. *Metabolomics* **2012**, *8* (1), 8−18.

(18) Whetzel, P. L.; Noy, N. F.; Shah, N. H.; Alexander, P. R.; Nyulas, C.; Tudorache, T.; Musen, M. A. *Nucleic Acids Res.* **2011**, *39* (Suppl), W541−W545.

(19) Helmus, J. J.; Jaroniec, C. P. *J. Biomol. NMR* **2013**, *55* (4), 355−367.

(20) Larralde, M.; Lawson, T. N.; Weber, R. J. M.; Moreno, P.; Haug, K.; Rocca-Serra, P.; Viant, M. R.; Steinbeck, C.; Salek, R. M. *Bioinformatics* **2017**, *33*, 2598−2600.

(21) Rocca-Serra, P.; Brandizi, M.; Maguire, E.; Sklyar, N.; Taylor, C.; Begley, K.; Field, D.; Harris, S.; Hide, W.; Hofmann, O.; et al. *Bioinformatics* **2010**, *26* (18), 2354−2356.

(22) Montecchi-Palazzi, L.; Kerrien, S.; Reisinger, F.; Aranda, B.; Jones, A. R.; Martens, L.; Hermjakob, H. *Proteomics* **2009**, *9* (22), 5112−5119.

(23) Bertsch, A.; Gröpl, C.; Reinert, K.; Kohlbacher, O. *Methods Mol. Biol.* **2011**, *696*, 353−367.

(24) Wishart, D. S. *Pharmacogenomics* **2008**, *9* (8), 1155−1162.

(25) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; Wilson, M.; Wishart, D. S. *Nucleic Acids Res.* **2012**, *40*, D815−D820.

(26) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. *Nucleic Acids Res.* **2013**, *41*, D625−D630.

(27) Yokochi, M.; Kobayashi, N.; Ulrich, E. L.; Kinjo, A. R.; Iwata, T.; Ioannidis, Y. E.; Livny, M.; Markley, J. L.; Nakamura, H.; Kojima, C.; Fujiwara, T. *J. Biomed. Semantics* **2016**, *7* (1), 16.

(28) Ferry-Dumazet, H.; Gil, L.; Deborde, C.; Moing, A.; Bernillon, S.; Rolin, D.; Nikolski, M.; de Daruvar, A.; Jacob, D. *BMC Plant Biol.* **2011**, *11*, 104.

(29) Weber, R. J. M.; Lawson, T. N.; Salek, R. M.; Ebbels, T. M. D.; Glen, R. C.; Goodacre, R.; Griffin, J. L.; Haug, K.; Koulman, A.; Moreno, P.; et al. *Metabolomics* **2017**, *13* (2), 12.

(30) Goecks, J.; Nekrutenko, A.; Taylor, J.; The Galaxy Team. *Genome Biol.* **2010**, *11* (8), R86.

(31) Giacomoni, F.; Le Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; Goulitquer, S.; Thévenot, E. A.; Caron, C. *Bioinformatics* **2015**, *31* (9), 1493−1495.

(32) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; et al. *Mol. Cell. Proteomics* **2012**, *11* (7), M111.014381.

(33) Walzer, M.; Qi, D.; Mayer, G.; Uszkoreit, J.; Eisenacher, M.; Sachsenberg, T.; Gonzalez-Galarza, F. F.; Fan, J.; Bessant, C.; Deutsch, E. W.; Reisinger, F.; Vizcaino, J. A.; Medina-Aunon, J. A.; Albar, J. P.; Kohlbacher, O.; Jones, A. R. *Mol. Cell. Proteomics* **2013**, *12* (8), 2332−2340.

(34) Bartels, C.; Xia, T. H.; Billeter, M.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **1995**, *6* (1), 1−10.

(35) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6* (3), 277−293.

(36) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. *Sci. Data* **2016**, 3, 160018.

DOI:10.1021/acs.analchem.7b02795

Downloaded via 217.234.218.81 on December 25, 2019 at 10:38:54 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

This is an open access article published under a Creative Commons Attribution (CC-BY) License, which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

ACS AuthorChoice

analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 3302−3310

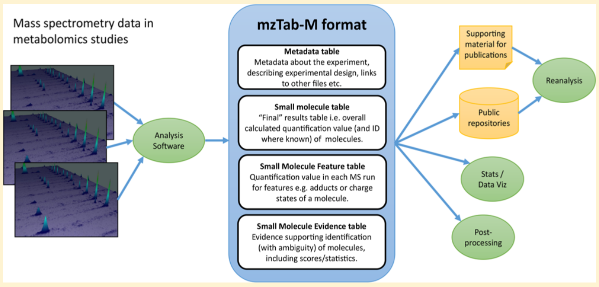
pubs.acs.org/ac

mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics

Nils Hoffmann,[†] Joel Rein,[‡] Timo Sachsenberg,[§] Jürgen Hartler,[∥,⊥] Kenneth Haug,[#] Gerhard Mayer,[∇] Oliver Alka,[§] Saravanan Dayalan,[○] Jake T. M. Pearce,[◆] Philippe Rocca-Serra,[¶] Da Qi,[%,&] Martin Eisenacher,[∇] Yasset Perez-Riverol,[#] Juan Antonio Vizcaíno,[#] Reza M. Salek,[*,@] Steffen Neumann,[*,+,=] and Andrew R. Jones[*,&]

[†]Leibniz-Institut für Analytische Wissenschaften-ISAS-e.V., Otto-Hahn-Straße 6b, 44227 Dortmund, Germany
[‡]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom
[§]Applied Bioinformatics Group, Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany
[∥]Institute of Computational Biotechnology, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
[⊥]Center for Explorative Lipidomics, BioTechMed-Graz, 8010 Graz, Austria
[#]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
[∇]Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, Universitätsstraße 150, D-44801 Bochum, Germany
[○]Metabolomics Australia, The University of Melbourne, Parkville, Victoria 3010, Australia
[◆]MRC-NIHR National Phenome Centre, Department of Surgery & Cancer, Imperial College London, London SW7 2AZ, United Kingdom
[¶]University of Oxford, e-Research Centre, 7 Keble Road, Oxford OX1 3QG, United Kingdom
[%]BGI-Shenzhen, Shenzhen, 518083, People's Republic of China
[&]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom
[@]International Agency for Research on Cancer, 150 cours Albert Thomas, 69008 Lyon, France
[+]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, 06120 Halle, Germany
[=]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig Deutscher, Platz 5e, 04103 Leipzig, Germany

Supporting Information


Mass spectrometry data in metabolomics studies
Analysis Software
mzTab-M format
Metadata table
Metadata about the experiment, describing experimental design, links to other files etc.
Small molecule table
"Final" results table i.e. overall calculated quantification value (and ID where known) of molecules.
Small Molecule Feature table
Quantification value in each MS run for features e.g. adducts or charge states of a molecule.
Small Molecule Evidence table
Evidence supporting identification (with ambiguity) of molecules, including scores/statistics.
Supporting material for publications
Public repositories
Reanalysis
Stats / Data Viz
Post-processing

ABSTRACT: Mass spectrometry (MS) is one of the primary techniques used for large-scale analysis of small molecules in metabolomics studies. To date, there has been little data format standardization in this field, as different software packages export results in different formats represented in XML or plain text, making data sharing, database deposition, and reanalysis highly challenging. Working within the consortia of the Metabolomics Standards Initiative, Proteomics Standards Initiative, and the Metabolomics Society, we have created mzTab-M to act as a common output format from analytical approaches using MS on small molecules. The format has been developed over several years, with input from a wide range of stakeholders. mzTab-M is a simple tab-separated text format, but importantly, the structure is highly standardized through the design of a detailed specification document, tightly coupled to validation software, and a mandatory controlled vocabulary of terms to populate it. The format is able to represent final quantification values from analyses, as well as the evidence trail in terms of features measured directly from MS (e.g., LC-MS, GC-MS, DIMS, etc.) and different types of approaches used to identify molecules. mzTab-M allows for ambiguity in the identification of molecules to be communicated clearly to readers of the files (both people and software). There are several implementations of the format available, and we anticipate widespread adoption in the field.

■ INTRODUCTION

It is now commonplace for high-throughput quantitative technologies to be used for analysis of biological, biomedical,

Received: September 21, 2018
Accepted: January 28, 2019
Published: January 28, 2019

© 2019 American Chemical Society

DOI:10.1021/acs.analchem.8b04310

and environmental samples. Technologies include those for measurements of gene expression using microarrays or RNA sequencing (transcriptomics), proteins by mass spectrometry (MS, proteomics), and MS or nuclear magnetic resonance (NMR) spectroscopy for measuring small molecules/metabolites (metabolomics) and lipids (lipidomics). These methods can provide the source data for systems biology/medicine investigations into the complex network of interactions that reflect both their functional and dysfunctional states, as well as reflect nutritional and environmental impacts. There is now an accepted principle in scientific research that data should be made openly and easily accessible to allow groups other than the initial data generators to verify the findings or search for new interpretations. Such guidelines are now commonly referred to as the "FAIR" principles, data being findable, accessible, interoperable, and reusable.[1] Furthermore, data from omics experiments are typically expensive to generate and often have potential uses beyond their initial purpose, including in meta-analyses, in data integration, or for testing and assisting in the development of new software. In omics research, there is always some heterogeneity in the approaches taken in different laboratories, such as different instrument platforms or analysis software, which usually have their own file formats. To allow data sets to be open for reuse generally requires the formulation of nonproprietary data formats, or more ideally, agreed data standards to which different producers of data must adhere. Without agreed standards (or ubiquitous formats originating from one package), data reuse is highly challenging, since informatics groups would need to write file format converters for every possible source of data, as well as keep these converters updated whenever data-producing software makes a format change. This scenario makes development of analysis software or a specific usage of public databases very challenging.

In a typical MS-based metabolomics/lipidomics pipeline, samples are analyzed by liquid or gas chromatography, coupled to MS (LC-MS/GC-MS), or by direct infusion (DIMS). Measurement of molecular intensity is typically done via software that detects features formed from isotopic patterns (or single peaks) along the time axis. For LC-MS, ionization can be performed in either positive or negative mode to produce protonated or deprotonated ions. It is also common for ion adducts to be formed, including metal adducts (Na^+, K^+), which have the same time elution profile but different m/z values. Many software packages perform adduct grouping, such that quantification values are reported both for individual features, as well as for the summed abundance across different adduct forms assumed to have come from the same starting molecule. For quantification across different samples, software may perform retention time alignment to ensure that the same features are quantified in each sample. In GC-MS, analysis is performed on volatile molecules and, in some cases, a derivatization step is applied to increase the volatility of compounds of interest.

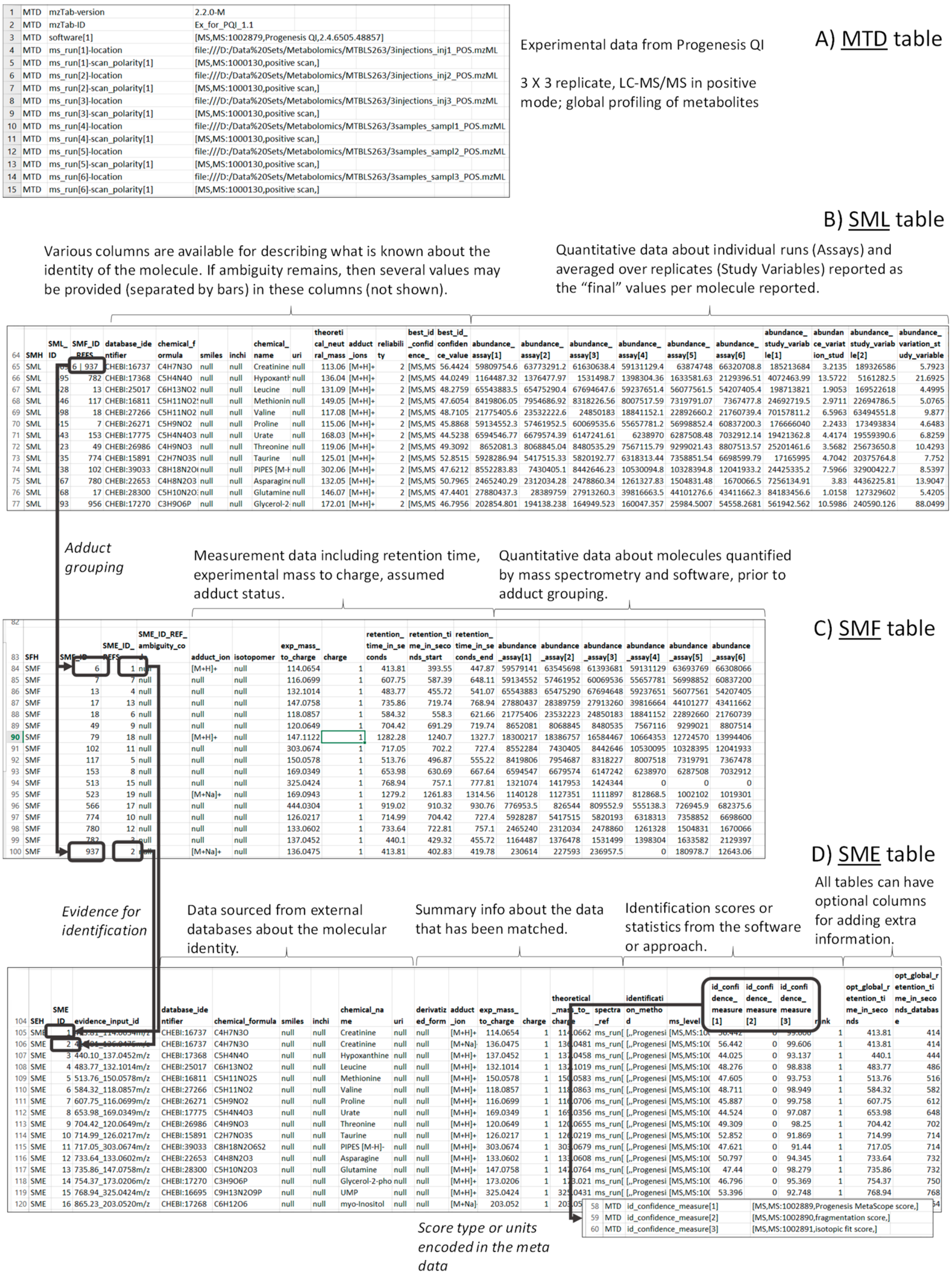
Molecular identification remains challenging in metabolomics. Typically, some combination of the following steps can assist with identification via searching a pre-existing library or database: accurate neutral mass, the relative abundance of isotopomers, the retention time, masses of fragmentation products (MS/MS and MS^n spectra), collisional cross section for platforms with ion mobility, etc. (see the reviews in refs 2−4 for more details). In the case of MS/MS and MS^n fragmentation, the spectra can be compared against an in-house spectral library or databases storing reference spectra for molecules including Metlin,[5] The Human Metabolome Database,[6] Global Natural Products Social Molecular Networking[7] (GNPS), MassBank,[8] and others (see ref 9 for a review) or analyzed by in silico identification software.[10]

It is common in all approaches for many molecules to remain unannotated or for ambiguity to remain: i.e., software provides a list of possible molecules for each MS feature. Reporting standards and guidelines on these certainties have been developed in several communities.[11,12] Following quantification (and identification), statistical analysis usually proceeds via univariate approaches, e.g. to find differentially expressed molecules between conditions, or multivariate/machine learning approaches to explore structure within the data and find molecules that can separate sample groups and thus act as potential biomarkers.

There exists a wide range of software, both free and commercial, for processing MS data for metabolomics/lipidomics.[13,14] Most software produces output data in a unique file format, annotated to different levels of detail, often with the description of preprocessing procedures followed implicit rather than specified, making it highly challenging to compare or integrate the results of different pipelines. For public data sharing, there are several databases that host data sets in support of publications or community data sets, including the European Bioinformatics Institute (EMBL-EBI) MetaboLights database[15] and the NIH Metabolomics Workbench.[16]

In this work, we describe a data standard for MS-based metabolomics analytical pipelines, called mzTab-M, which captures the downstream results of analysis (i.e., excluding raw data), suitable for statistical analysis, result visualization, or submission to a public repository in support of a publication. The standard has been developed in a joint and open process between members of the Metabolomics Standards Initiative (MSI),[17] the Metabolomics Society Data Standards Task Group, and the Proteomics Standards Initiative (PSI), which had originally developed the mzTab format on which it is based.[18] There are several related and complementary efforts, which include efforts to define minimum reporting requirements for different aspects of metabolomics.[11,19] There is also general agreement among standards groups (MSI, Metabolomics Society) to promote the use of the PSI's mzML format for raw data storage.[20] mzML is an XML-based standard for MS data, either for profile data as recorded directly from the instrument or for centroided data (peak picked in the m/z domain). The freely available ProteoWizard software embeds software libraries from several vendors of MS instruments, enabling the conversion of vendor raw files into mzML.[21] For NMR metabolomics, the recently released nmrML standard follows a design principle similar to that of mzML, capturing NMR spectra and some metadata within an XML-based standard.[22] For the description of study design, experimental metadata, and sample processing parameters, the ISA framework,[23] while generally applicable to all types of experimental design, has been particularly taken up by the metabolomics field. The PSI previously developed the mzTab format (version 1.0) to act as a simple format for quantified and/or identified peptides and proteins in MS workflows.[18] mzTab version 1.0 also has a section to allow small-molecule data to be captured. However, the data model was rather simple and did not cover some important use cases for metabolomics/lipidomics and, as a result, it has not been extensively used for small molecules or lipids. The development of mzTab-M has thus branched off from the original mzTab format development, and we report it here as a new standard for metabolomics called mzTab-M ("version 2.0" to differentiate it

DOI:10.1021/acs.analchem.8b04310



Experimental data from Progenesis QI

3 X 3 replicate, LC-MS/MS in positive mode; global profiling of metabolites

1	MTD	mzTab-version	2.2.0-M
2	MTD	mzTab-ID	Ex_for_PQI_1.1
3	MTD	software[1]	[MS,MS:1002879,Progenesis QI,2.4.6505.48857]
4	MTD	ms_run[1]-location	file:///D:/Data%20Sets/Metabolomics/MTBLS263/3injections_inj1_POS.mzML
5	MTD	ms_run[1]-scan_polarity[1]	[MS,MS:1000130,positive scan,]
6	MTD	ms_run[2]-location	file:///D:/Data%20Sets/Metabolomics/MTBLS263/3injections_inj2_POS.mzML
7	MTD	ms_run[2]-scan_polarity[1]	[MS,MS:1000130,positive scan,]
8	MTD	ms_run[3]-location	file:///D:/Data%20Sets/Metabolomics/MTBLS263/3injections_inj3_POS.mzML
9	MTD	ms_run[3]-scan_polarity[1]	[MS,MS:1000130,positive scan,]
10	MTD	ms_run[4]-location	file:///D:/Data%20Sets/Metabolomics/MTBLS263/3samples_sampl1_POS.mzML
11	MTD	ms_run[4]-scan_polarity[1]	[MS,MS:1000130,positive scan,]
12	MTD	ms_run[5]-location	file:///D:/Data%20Sets/Metabolomics/MTBLS263/3samples_sampl2_POS.mzML
13	MTD	ms_run[5]-scan_polarity[1]	[MS,MS:1000130,positive scan,]
14	MTD	ms_run[6]-location	file:///D:/Data%20Sets/Metabolomics/MTBLS263/3samples_sampl3_POS.mzML
15	MTD	ms_run[6]-scan_polarity[1]	[MS,MS:1000130,positive scan,]

Various columns are available for describing what is known about the identity of the molecule. If ambiguity remains, then several values may be provided (separated by bars) in these columns (not shown).

Quantitative data about individual runs (Assays) and averaged over replicates (Study Variables) reported as the “final” values per molecule reported.

64	SMH	SML_ID	SMF_ID_REFS	database_identifier	chemical_formula	smiles	inchi	chemical_name	uri	theoretical_neutral_mass	adduct_ions	reliability	best_id_confidence_	best_id_confidence_value	abundance_assay[1]	abundance_assay[2]	abundance_assay[3]	abundance_assay[4]	abundance_assay[5]	abundance_assay[6]	abundance_study_variable[1]	abundance_variation_stud	abundance_study_variable[2]	abundance_variation_study_variable
65	SML	[illegible]	6 | 937	CHEBI:16737	C4H7N3O	null	null	Creatinine	null	113.06	[M+H]+	2	[MS,MS	56.4424	59809754.6	63773291.2	61630638.4	59131129.4	63874748	66320708.8	185213684	3.2135	189326586	5.7923
66	SML	95	782	CHEBI:17368	C5H4N4O	null	null	Hypoxanth	null	136.04	[M+H]+	2	[MS,MS	44.0249	1164487.32	1376477.97	1531498.7	1398304.36	1633581.63	2129396.51	4072463.99	13.5722	5161282.5	21.6925
67	SML	28	13	CHEBI:25017	C6H13NO2	null	null	Leucine	null	131.09	[M+H]+	2	[MS,MS	48.2759	65543883.5	65475290.4	67694647.6	59237651.4	56077561.5	54207405.4	198713821	1.9053	169522618	4.4995
68	SML	46	117	CHEBI:16811	C5H11NO2	null	null	Methionin	null	149.05	[M+H]+	2	[MS,MS	47.6054	8419806.05	7954686.92	8318226.56	8007517.59	7319791.07	7367477.8	24692719.5	2.9711	22694786.5	5.0765
69	SML	98	18	CHEBI:27266	C5H11NO2	null	null	Valine	null	117.08	[M+H]+	2	[MS,MS	48.7105	21775405.6	23532222.6	24850183	18841152.1	22892660.2	21760739.4	70157811.2	6.5963	63494551.8	9.877
70	SML	15	7	CHEBI:26271	C5H9NO2	null	null	Proline	null	115.06	[M+H]+	2	[MS,MS	45.8868	59134552.3	57461952.5	60069535.6	55657781.2	56998852.4	60837200.3	176666040	2.2433	173493834	4.6483
71	SML	43	153	CHEBI:17775	C5H4N4O3	null	null	Urate	null	168.03	[M+H]+	2	[MS,MS	44.5238	6594546.77	6679574.39	6147241.61	6238970	6287508.48	7032912.14	19421362.8	4.4174	19559390.6	6.8259
72	SML	23	49	CHEBI:26986	C4H9NO3	null	null	Threonine	null	119.06	[M+H]+	2	[MS,MS	49.3092	8652081.3	8068845.04	8480535.29	7567115.79	9299021.43	8807513.57	25201461.6	3.5682	25673650.8	10.4293
73	SML	35	774	CHEBI:15891	C2H7NO3S	null	null	Taurine	null	125.01	[M+H]+	2	[MS,MS	52.8515	5928286.94	5417515.33	5820192.77	6318313.44	7358851.54	6698599.79	17165995	4.7042	20375764.8	7.752
74	SML	38	102	CHEBI:39033	C8H18N2O	null	null	PIPES [M-H	null	302.06	[M+H]+	2	[MS,MS	47.6212	8552283.83	7430405.1	8442646.23	10530094.8	10328394.8	12041933.2	24425335.2	7.5966	32900422.7	8.5397
75	SML	67	780	CHEBI:22653	C4H8N2O3	null	null	Asparagine	null	132.05	[M+H]+	2	[MS,MS	50.7965	2465240.29	2312034.28	2478860.34	1261327.83	1504831.48	1670066.5	7256134.91	3.83	4436225.81	13.9047
76	SML	68	17	CHEBI:28300	C5H10N2O	null	null	Glutamine	null	146.07	[M+H]+	2	[MS,MS	47.4401	27880437.3	28389759	27913260.3	39816663.5	44101276.6	43411662.3	84183456.6	1.0158	127329602	5.4205
77	SML	93	956	CHEBI:17270	C3H9O6P	null	null	Glycerol-2-	null	172.01	[M+H]+	2	[MS,MS	46.7956	202854.801	194138.238	164949.523	160047.357	25984.5007	54558.2681	561942.562	10.5986	240590.126	88.0499

Adduct grouping

Measurement data including retention time, experimental mass to charge, assumed adduct status.

Quantitative data about molecules quantified by mass spectrometry and software, prior to adduct grouping.

82																	
83	SFH	SMF_ID	SME_ID_REFS	SME_ID_REF_ambiguity_code	adduct_ion	isotopomer	exp_mass_to_charge	charge	retention_time_in_seconds	retention_time_in_seconds_start	retention_time_in_seconds_end	abundance_assay[1]	abundance_assay[2]	abundance_assay[3]	abundance_assay[4]	abundance_assay[5]	abundance_assay[6]
84	SMF	6	1	null	[M+H]+	null	114.0654	1	413.81	393.55	447.87	59579141	63545698	61393681	59131129	63693769	66308066
85	SMF	7	7	null	null	null	116.0699	1	607.75	587.39	648.11	59134552	57461952	60069536	55657781	56998852	60837200
86	SMF	13	4	null	null	null	132.1014	1	483.77	455.72	541.07	65543883	65475290	67694648	59237651	56077561	54207405
87	SMF	17	13	null	null	null	147.0758	1	735.86	719.74	768.94	27880437	28389759	27913260	39816664	44101277	43411662
88	SMF	18	6	null	null	null	118.0857	1	584.32	558.3	621.66	21775406	23532223	24850183	18841152	22892660	21760739
89	SMF	49	9	null	null	null	120.0649	1	704.42	691.29	719.74	8652081	8068845	8480535	7567116	9299021	8807514
90	SMF	79	18	null	[M+H]+	null	147.1122	1	1282.28	1240.7	1327.7	18300217	18386757	16584467	10664353	12724570	13994406
91	SMF	102	11	null	null	null	303.0674	1	717.05	702.2	727.4	8552284	7430405	8442646	10530095	10328395	12041933
92	SMF	117	5	null	null	null	150.0578	1	513.76	496.87	555.22	8419806	7954687	8318227	8007518	7319791	7367478
93	SMF	153	8	null	null	null	169.0349	1	653.98	630.69	667.64	6594547	6679574	6147242	6238970	6287508	7032912
94	SMF	513	15	null	null	null	325.0424	1	768.94	757.1	777.81	1321074	1417953	1424344	0	0	0
95	SMF	523	19	null	[M+Na]+	null	169.0943	1	1279.2	1261.83	1314.56	1140128	1127351	1111897	812868.5	1002102	1019301
96	SMF	566	17	null	null	null	444.0304	1	919.02	910.32	930.76	776953.5	826544	809552.9	555138.3	726945.9	682375.6
97	SMF	774	10	null	null	null	126.0217	1	714.99	704.42	727.4	5928287	5417515	5820193	6318313	7358852	6698600
98	SMF	780	12	null	null	null	133.0602	1	733.64	722.81	757.1	2465240	2312034	2478860	1261328	1504831	1670066
99	SMF	782	3	null	null	null	137.0452	1	440.1	429.32	455.72	1164487	1376478	1531499	1398304	1633582	2129397
100	SMF	937	2	[illegible]	[M+Na]+	null	136.0475	1	413.81	402.83	419.78	230614	227593	236957.5	0	180978.7	12643.06

Evidence for identification

Data sourced from external databases about the molecular identity.

Summary info about the data that has been matched.

Identification scores or statistics from the software or approach.

All tables can have optional columns for adding extra information.

104	SEH	SME_ID	evidence_input_id	database_identifier	chemical_formula	smiles	inchi	chemical_name	uri	derivatized_form	adduct_ion	exp_mass_to_charge	charge	theoretical_mass_to_charge	spectra_ref	identification_method	ms_level	id_confidence_measure[1]	id_confidence_measure[2]	id_confidence_measure[3]	rank	opt_global_retention_time_in_seconds	opt_global_retention_time_in_seconds_database
105	SME	1	[illegible]	CHEBI:16737	C4H7N3O	null	null	Creatinine	null	null	[M+H]+	114.0654	1	114.0662	ms_run[	[,,Progenesi	[MS,MS:10	[illegible]	0	[illegible]	1	413.81	414
106	SME	2	[illegible]	CHEBI:16737	C4H7N3O	null	null	Creatinine	null	null	[M+Na]	136.0475	1	136.0481	ms_run[	[,,Progenesi	[MS,MS:10	56.442	0	99.606	1	413.81	414
107	SME	3	440.10_137.0452m/z	CHEBI:17368	C5H4N4O	null	null	Hypoxanthine	null	null	[M+H]+	137.0452	1	137.0458	ms_run[	[,,Progenesi	[MS,MS:10	44.025	0	93.137	1	440.1	444
108	SME	4	483.77_132.1014m/z	CHEBI:25017	C6H13NO2	null	null	Leucine	null	null	[M+H]+	132.1014	1	132.1019	ms_run[	[,,Progenesi	[MS,MS:10	48.276	0	98.838	1	483.77	486
109	SME	5	513.76_150.0578m/z	CHEBI:16811	C5H11NO2S	null	null	Methionine	null	null	[M+H]+	150.0578	1	150.0583	ms_run[	[,,Progenesi	[MS,MS:10	47.605	0	93.753	1	513.76	516
110	SME	6	584.32_118.0857m/z	CHEBI:27266	C5H11NO2	null	null	Valine	null	null	[M+H]+	118.0857	1	118.0863	ms_run[	[,,Progenesi	[MS,MS:10	48.711	0	98.949	1	584.32	582
111	SME	7	607.75_116.0699m/z	CHEBI:26271	C5H9NO2	null	null	Proline	null	null	[M+H]+	116.0699	1	116.0706	ms_run[	[,,Progenesi	[MS,MS:10	45.887	0	99.758	1	607.75	612
112	SME	8	653.98_169.0349m/z	CHEBI:17775	C5H4N4O3	null	null	Urate	null	null	[M+H]+	169.0349	1	169.0356	ms_run[	[,,Progenesi	[MS,MS:10	44.524	0	97.087	1	653.98	648
113	SME	9	704.42_120.0649m/z	CHEBI:26986	C4H9NO3	null	null	Threonine	null	null	[M+H]+	120.0649	1	120.0655	ms_run[	[,,Progenesi	[MS,MS:10	49.309	0	98.25	1	704.42	702
114	SME	10	714.99_126.0217m/z	CHEBI:15891	C2H7NO3S	null	null	Taurine	null	null	[M+H]+	126.0217	1	126.0219	ms_run[	[,,Progenesi	[MS,MS:10	52.852	0	91.869	1	714.99	714
115	SME	11	717.05_303.0674m/z	CHEBI:39033	C8H18N2O6S2	null	null	PIPES [M-H]-	null	null	[M+H]+	303.0674	1	303.0679	ms_run[	[,,Progenesi	[MS,MS:10	47.621	0	91.44	1	717.05	714
116	SME	12	733.64_133.0602m/z	CHEBI:22653	C4H8N2O3	null	null	Asparagine	null	null	[M+H]+	133.0602	1	133.0608	ms_run[	[,,Progenesi	[MS,MS:10	50.797	0	94.345	1	733.64	732
117	SME	13	735.86_147.0758m/z	CHEBI:28300	C5H10N2O3	null	null	Glutamine	null	null	[M+H]+	147.0758	1	147.0764	ms_run[	[,,Progenesi	[MS,MS:10	47.44	0	98.279	1	735.86	732
118	SME	14	754.37_173.0206m/z	CHEBI:17270	C3H9O6P	null	null	Glycerol-2-pho	null	null	[M+H]+	173.0206	1	173.021	ms_run[	[,,Progenesi	[MS,MS:10	46.796	0	95.369	1	754.37	750
119	SME	15	768.94_325.0424m/z	CHEBI:16695	C9H13N2O9P	null	null	UMP	null	null	[M+H]+	325.0424	1	325.0431	ms_run[	[,,Progenesi	[MS,MS:10	53.396	0	92.748	1	768.94	768
120	SME	16	865.23_203.0520m/z	CHEBI:17268	C6H12O6	null	null	myo-Inositol	null	null	[M+Na]	203.052	1	203.05	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

58	MTD	id_confidence_measure[1]	[MS,MS:1002889,Progenesis MetaScope score,]
59	MTD	id_confidence_measure[2]	[MS,MS:1002890,fragmentation score,]
60	MTD	id_confidence_measure[3]	[MS,MS:1002891,isotopic fit score,]

Score type or units encoded in the meta data

Figure 1. Overall structure of an mzTab-M file. (A) Metadata about the experiment, describing experimental design (study variables and assays), links to other files, etc.. (B) The small molecule (SML) table, capturing “final” results table: i.e., overall calculated quantification value (and identity where known) of a metabolite. (C) Quantification value in each (aligned) MS run for MS^1 features: e.g., mapped to individual adducts or charge states of a molecule. (D) Evidence supporting identification (with ambiguity if needed) for molecules, using CV terms for scores/statistics where available.

from mzTab version 1.0). It follows design principles similar to those of mzTab 1.0, but it is not backward compatible.

■ METHODS

The mzTab-M format was designed in a process that was open to any interested parties. All associated materials and code for processing and validating files are fully open source and are hosted on GitHub (https://github.com/HUPO-PSI/mzTab). mzTab-M started from the design of mzTab version 1.0 format but was further developed to support the specific needs of metabolomics (see the Supporting Information for more details on the relationship). The development took place via face-to-

DOI:10.1021/acs.analchem.8b04310

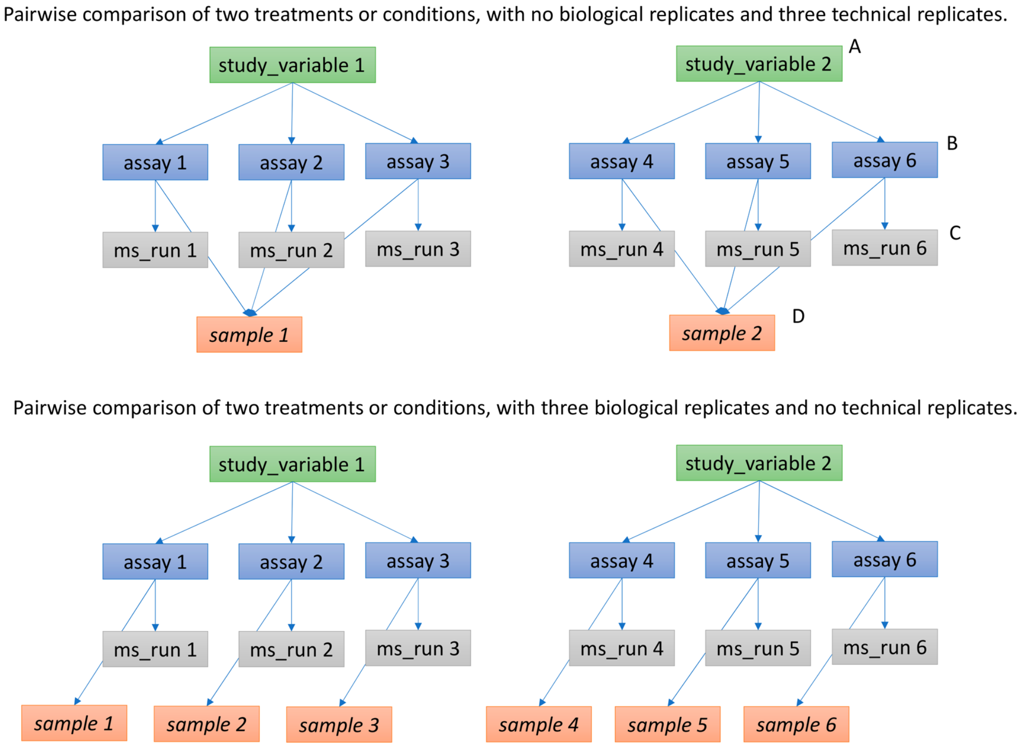


Figure 2. Simple experimental designs in mzTab-M can be represented using a combination of the elements study_variable (SV), assay, ms_run, and sample. Quantitative values can be reported in files for SVs and assays. (A) SV is intended to capture different groups of replicates, which might have resulted from different levels of a given variable: e.g. control versus treated (represented as 2 SVs) and *n* time points over a treatment course (as *n* SVs). (B) assay captures a measurement made about a molecule (small molecule/lipid) where multiple assays within the same SV are taken to be replicates of some kind (biological or technical). (C) ms_run captures a single run on an MS instrument. (D) Samples are optional in mzTab, since the quantitative software may often be unaware of the biological samples that have been analyzed. If that information is available, references from assay to the same (technical, upper half) or different (biological, bottom half) samples are used to describe the type of replication performed.

face workshops and regular conference calls. The specifications have been submitted to a formal document process for anonymous review, overseen by an editor commissioned jointly between the PSI and the Metabolomics Society. The mzTab-M format is defined by the specification document and example files that demonstrate how to encode certain features (see the GitHub repository). The specification document describes the overall structure of the format: what tables must be present, what columns and rows must be present in those tables, and what terminology is allowed as data values. For various aspects of metadata, the standard enforces (and can be checked by validation software) that controlled vocabulary (CV) terms are used (e.g. for names of software, databases, parameters, statistics, etc.), which can be sourced from the PSI-MS CV[26] (https://www.ebi.ac.uk/ols/ontologies/ms), as well as other CVs where appropriate.

■ RESULTS AND DISCUSSION

mzTab-M has been designed to act as a single data format for metabolomics and lipidomics, including an appropriate level of detail about the "final" results: i.e., molecules quantified across samples. The format also contains the ability to represent "intermediate" or supporting data, including the evidence trail for identifications from software (scores or statistics), as well as the quantification values derived directly from MS: i.e., prior to any adduct grouping or summarization steps. The format is represented as tab-separated text, meaning that it can be loaded directly into a spreadsheet editor or into statistical software such as R or SPSS for downstream analysis and visualization, without any need for coding, and can thus replace the use of tables (e.g., in pdf or Excel format) of supplementary data in support of publications. It is also relatively straightforward for informatics groups to develop software to add support for the standard to existing software.

The mzTab-M format consists of four cross-referenced data tables (Figure 1): metadata (MTD), small molecule (SML), small molecule feature (SMF) and the small molecule evidence (SME). The MTD and SML tables are mandatory, and for a file to contain any evidence about how molecules were quantified or identified by software, all four tables must be present. The tables must follow the order MTD, SML, SMF, and SME, with a blank line separating each table. The structure of each table, in terms of the rows and columns that must be present, is tightly specified, as explained in the following sections and formally in the mzTab-M specification document.

Metadata (MTD) Table. The metadata table has multiple rows and exactly three columns (Figure 1A). Each row must contain (1) "MTD", (2) a parameter name, and (3) the parameter value. The types of parameters that must or may be present are described in the specification document, and allowed values from CVs are defined in a mapping file. The MTD table must report at least a simple specification of the experimental design, in terms of the number of different measurements (i.e., usually the count of MS runs) and the groupings of those MS

DOI:10.1021/acs.analchem.8b04310

runs (i.e., experimental factors or groups) over which statistical analysis may be done. These values then inform the number of columns present in SML and SMF tables for which (relative or absolute) quantitative values are reported. The following concepts are specified in the MTD table so that they can be referenced and reused elsewhere in the file:

- Assay: the application of a measurement about the sample (in this case through MS), producing values about small molecules or lipids. One assay is typically mapped to one ms_run element (see below), although the differentiation between assay and ms_run is present to provide a mechanism for grouping multiple MS runs together if the sample has been fractionated and different fractions run on the instrument to increase coverage. The MTD table gives the count of assays with locally unique identifiers, so that they can be referenced by other elements.
- ms_run: an MS run is effectively one run on an MS instrument (e.g., by LC-MS, GC-MS, DIMS, etc.) and can be referenced from assay elements in different contexts. When an ISA-Tab document from mzTab-M is referenced, ms_run should be matched with the ISA "Assay Name" values found in an ISA "Assay Table" file (https://isa-specs.readthedocs.io/en/latest/isatab.html).
- Sample: a biological material that has been analyzed, to which descriptors of species, cell/tissue type, etc. can be attached. Samples are not mandatory, since some software packages that will produce mzTab-M files cannot determine what type of sample was analyzed (e.g., whether biological or technical replication was performed), although it is noted that, without such annotations, downstream statistical analysis of the results will often not be possible.
- study_variable: a "study_variable" (SV) element represents a grouping of replicates for which a quantitative value can be reported, for example following averaging of values from individual assays. More accurately, a "study_variable" element usually represents a *level* of some particular experimental variable, such as the value of time within a time course, dose of a drug, intervention performed on samples, etc. In other contexts, this concept is named differently: e.g., "Factor Value" in ISA format.

Clear definitions of biological and technical replicates are difficult to provide, as the commonly used terminologies are somewhat dependent upon the biological domain. However, we use the following general definitions in mzTab-M: biological replicates represent cases when different samples are analyzed by MS, and technical replicates represent cases where the same samples are analyzed multiple times by MS. As illustrated in Figure 2, a simple form of the experimental design can be captured in mzTab-M using a combination of assay, "study_variable", and "sample". In a complex, nested design, linkages between different study variables are not explicitly modeled but captured through the annotated values, as shown in the Supporting Information.

The MTD table also has the (optional) capability to capture additional metadata that can be useful to interpret the study, such as limited details about the sample processing steps performed, the MS instrument, software and parameters, contact details for the study producers, etc. However, it is acknowledged that other formats may capture such details, such as referenced mzML (including instrument information and parameters), other MS data file formats, or ISA-Tab files (containing experimental design and sample processing), which may be more appropriate locations for such information.

Small Molecule (SML) Table. The small molecule (SML) table (Figure 1B) is intended to capture the "final" results of the study in terms of molecules that have been quantified (with identification data, where available). If different adduct forms or fragments of a molecule have been observed as different MS features, it is common that feature grouping is performed, and the SML table should contain the final quantitative values after summarization. Thus, SML could be viewed as the equivalent of tabular results presented in a paper for the molecules quantified in different samples. For survey-type data, it is also possible to report quantities as "null", while still reporting identification evidence, as supported by the SME table.

The header row has "SMH" in the first column, followed by an ordered set of column headers. After the header row, each row reports one molecule, with the first cell containing "SML", followed by the data values for each specified column. The columns include a unique local identifier for the molecule (*SML_ID*), followed by a cell (*SMF_ID_REFS*) containing references to features in the SMF table. The referenced features are the different adduct forms or in-source fragments of the molecular features actually detected by MS. The next set of columns provides different ways to identify the molecule (*database_identifer, chemical_formula, smiles, inchi, chemical_name, uri, theoretical_neutral_mass*; see section *Identification evidence and ambiguity* below).

The following columns report quantitative data for the *n* assays (in *n* columns, where *n* is the count of assays reported in MTD) and the *m study_variable* groups (in *m* columns) e.g. as an average (e.g., mean) across assay values within that *study_variable*. A value can also be provided for the variability in the *study_variable* quantification value reported e.g. a standard error value. A parameter in MTD specifies how to interpret the quantitative values in these columns in terms of a data type exported from a specific piece of software or where appropriate, absolute values with units.

At the right-hand end of the SML table (and SMF and SME tables), it is possible to include user-specified (optional) columns, with a method for annotating that the columns refer to the entire molecule, or the measurement of the molecule in particular *assays* or *study_variables*. The user-specified columns thus make mzTab-M extensible to support custom data types not covered in the core model.

Small Molecule Feature (SMF) Table. The SMF table contains data on what features were actually measured by the instrument and quantified by software (Figure 1C). The header row of the table has "SFH" in the first cell, followed by a set of columns. Each row of the table is one MS feature recorded across different runs, starting with the code "SMF". It is assumed that an alignment process has taken place so that the same feature has been seen across different runs, with missing values handled as appropriate (see specification document for guidance on encoding nonaligned workflows). The next column (*SME_ID_REFS*) is for referencing down to the final table: Small Molecule Evidence (SME) via a set of identifier references, as well as a code telling the file reader how to interpret multiple references (*SME_ID_REF_ambiguity_code*), explained in Identification Evidence and Ambiguity.

The SMF table next contains information about the type of adduct and charge state observed, the experimental *m*/*z* value, the retention time of the feature (in a master or averaged run), and a method for optionally specifying if a given isotopomer has

DOI:10.1021/acs.analchem.8b04310

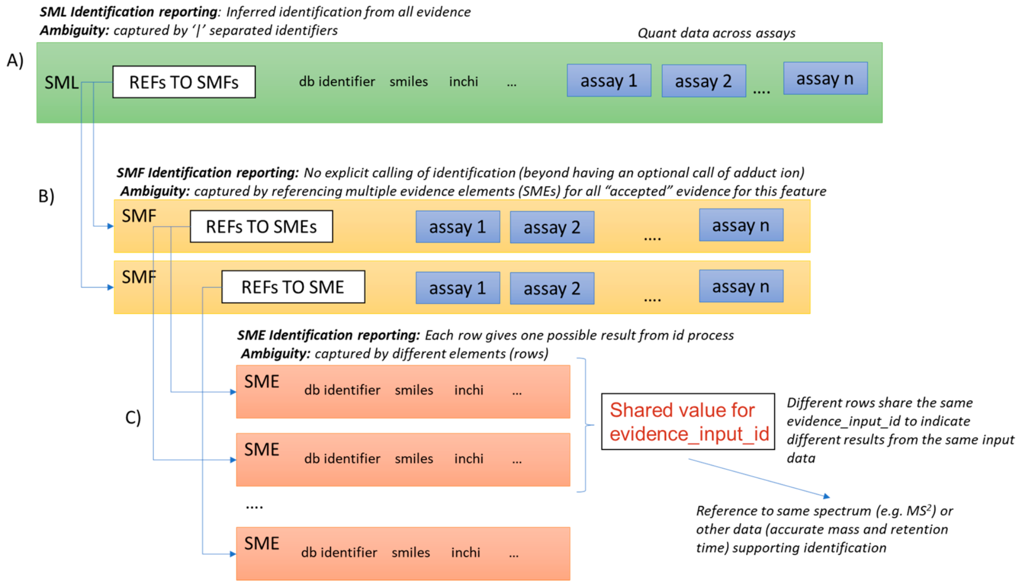

SML Identification reporting: Inferred identification from all evidence
Ambiguity: captured by '|' separated identifiers
Quant data across assays
A)
SML
REFs TO SMFs
db identifier
smiles
inchi
...
assay 1
assay 2
....
assay n
SMF Identification reporting: No explicit calling of identification (beyond having an optional call of adduct ion)
Ambiguity: captured by referencing multiple evidence elements (SMEs) for all "accepted" evidence for this feature
B)
SMF
REFs TO SMEs
assay 1
assay 2
....
assay n
SMF
REFs TO SME
assay 1
assay 2
....
assay n
SME Identification reporting: Each row gives one possible result from id process
Ambiguity: captured by different elements (rows)
C)
SME
db identifier
smiles
inchi
...
SME
db identifier
smiles
inchi
...
....
SME
db identifier
smiles
inchi
...
Shared value for evidence_input_id
Different rows share the same evidence_input_id to indicate different results from the same input data
Reference to same spectrum (e.g. MS2) or other data (accurate mass and retention time) supporting identification

Figure 3. (A) The summary level (SML) reports the final assumed identifications, allowing for ambiguity by including "|" separated results in the relevant columns. (B) The feature level (SMF) does not explicitly report identifications but references down to the SME level. Ambiguity is propagated via referencing multiple SME rows with different identification results. (C) One SME row represents a single possible identification from some input evidence. Multiple identifications from the same input data share the same value for *evidence_input_id*. Ambiguity can be captured by different rows for the same input data.

been quantified (+1 or +2 peak, ^{13}C peak, etc.) as used in some isotopic labeling/flux studies. The following columns represent the quantitative data within each of the *n* assays recorded in the MTD section. For SML, a parameter in MTD also describes how to interpret the quantitative values recorded.

Small Molecule Evidence (SME) Table. The SME table represents strands of potentially heterogeneous types of evidence supporting the identification of a molecule (Figure 1D). Each row contains the result of one identification process (library search, pattern match, manual curation, etc.). The header of the table starts with "SEH" followed by a set of columns. The second column is a local identifier for a row of evidence (*SME_ID*), followed by a local identifier for the input data to the process (*evidence_input_id*). *evidence_input_id* is needed for the cases where different rows of evidence are reported for the same input data (MS^2 spectrum, accurate mass + retention time, isotope pattern, etc.). They can be linked by sharing the same value for *evidence_input_id*. As in the SML table, a set of columns exists to specify the molecular identity from a variety of sources or identifier types (*database_identifer, chemical_formula, smiles, inchi, chemical_name, uri*). The experimental m/z value of the feature, the charge, and the theoretical m/z value (e.g., from a database) can be recorded, along with scores or confidence measures coming from the software used to support the identification. If a fragmentation spectrum has been used, there is a mechanism for referencing the exact spectrum in the source file (e.g., mzML file) and the MS level of the input data to the identification process.

Identification Evidence and Ambiguity. Small-molecule identification is a well-known challenge in MS metabolomics, and even more so in MS lipidomics, where complete structural elucidation of molecules is often not possible. Different levels of "identification" might be possible, ranging from having the accurate mass only, the chemical formula, a list of possible identifiers to molecules in a database (with the same or different formula), or a complete molecular structure resolved: e.g., if a complementary technique such as NMR has been used. mzTab-M has been designed to accommodate all the different possibilities in a simple yet flexible structure (Figure 3). For further details on how identifications of lipids and other compound classes can be represented see the Supporting Information.

In a row of the final results (SML table), the export software can include one or more identifiers from external databases: e.g., "CHEBI:16811" where the prefix is defined in MTD as referencing the ChEBI database[24] (with a URL) and the identifier is the ChEBI unique identifier (in this case for methionine). Similarly, the specification allows for the chemical formula in standard notation, simplified molecular-input line-entry system (SMILES[25]), or InChi[26] to be provided. In all cases, if ambiguity has not been resolved, then a Pipe "|" separated list of identifiers can be provided in the same cell. There are several measures for describing the confidence of identification, including the use of reliability codes such as those developed by the MSI[11,12] and the score or confidence measures from identification software where available.

To trace the evidence source, references via the features (SMF table) and on to the SME table should be provided. In the case that adduct grouping (i.e., multiple SMF rows) has been performed prior to identification, then the different SMF rows will reference the same SME rows. At the SME level, if there are different rows from the same input data (e.g., different database matches), then it is expected that the SMF element(s) references multiple SME elements that share the same value for *evidence_input_id*. It is also possible to report different evidence streams to support identification, such as searches in different libraries. As such, SMF rows can reference multiple SME elements carrying different values of *evidence_input_id*. Given that these two cases would both result in multiple SME identifiers referenced from an SMF row, an extra code can be

DOI:10.1021/acs.analchem.8b04310

provided at the SMF level (*SME_ID_REF_ambiguity_code*) containing values to differentiate whether ambiguity has been resolved or still remains (see the specification document for a full description).

Using CVs and File Validation. mzTab-M extensively uses CVs to provide unambiguous terms for annotation. For parameters relating to MS and associated processing, CV terms should generally be sourced from the PSI MS CV.[27] Several other CVs are recommended for describing details about sample types, species taxonomy, sample preparation, etc. (see the specification document). To ensure that valid CV terms are used, we have extended the concept of the PSI semantic validation framework.[28] The framework includes a mapping file that states the groups of CV terms allowed at each position in mzTab-M, enabling the list of terms to expand over time, without changes in the standard or software. New terms can be added straightforwardly by making a request on a mailing list: e.g., for a term describing new software, scores, or statistics. A crucial part of the standard is therefore a validator to ensure that files exported from different packages fulfill the rules defined in the specification, so that they can be read without error by other software. We have developed validation software for mzTab-M, available from jmzTab (project: https://github.com/lifs-tools/jmzTab-m), which checks not only that the structure of the file is correct but also that valid and correct CV terms have been used throughout.

Implementation in Software and Databases. The specifications have been verified by both PSI and MSI formal review processes, from which the stable version (mzTab-M 2.0) has been released. It is not expected that there will be changes to the format for several years to allow implementations to be developed. A reference implementation with parser, writer, and validator (in jmzTab-m) has been developed in Java (as for mzTab 1.0[29]). jmzTab-m provides an OpenAPI 2.0 compatible API model that serves as the basis for automatic model generation in a wide number of programming languages (C++, JavaScript, R, Python), reducing the burden of implementation. The library provides parsing, validation, and writing of mzTab-M files and object models. A web-based application (https://apps.lifs.isas.de/mztabvalidator/) provides a user-friendly user interface to perform standard and semantic validation and to display validation results. Additional implementations are under development in software including XCMS,[30] Progenesis QI (Waters), Lipid Data Analyzer,[31] OpenMS,[32] and MetaboLights.[15] Over the coming years, we will be promoting the implementation of the standard in a wide variety of both open-source and commercial software to act as a universal standard for metabolomics and lipidomics.

CONCLUSIONS

We have developed mzTab-M for metabolomics data representation and sharing. The standard has been developed in an open process with widespread consultation of different approaches taken in the field and involvement of software teams from academic research groups as well as industry. The standard has undergone a rigorous peer review process by both the MSI and PSI to ensure that the resulting standard is of high quality and is stable. The standard is expected to remain stable for several years, except for improvements to documentation and extensions to the CV, allowing research groups and commercial developers to invest time in the implementation. We also encourage other groups interested in standardizing omics data, particularly those using MS (e.g. glycomics), to adopt the mzTab model/design, CV infrastructure, and associated software.

ASSOCIATED CONTENT

Supporting Information

Supplementary File 1. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b04310.

> Additional details on the relationship between mzTab-M and mzTab 1.0, on how lipid identifications can be reported in mzTab-M, and on the representation of complex experimental designs in mzTab-M (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail for R.M.S.: Salekr@iarc.fr.
*E-mail for S.N.: sneumann@ipb-halle.de.
*E-mail for A.R.J.: andrew.jones@liverpool.ac.uk.

ORCID

Jürgen Hartler: 0000-0002-1095-6458
Reza M. Salek: 0000-0001-8604-1732
Andrew R. Jones: 0000-0001-6118-9327

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

N.H. acknowledges funding by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen, the Regierende Bürgermeister von Berlin-inkl. Wissenschaft und Forschung, and the Bundesministerium für Bildung und Forschung (de.NBI/BMBF 031L0108A). S.N. acknowledges funding by the Bundesministerium für Bildung und Forschung (de.NBI/BMBF 031L0107). S.N., P.R.-S., K.H., and R.M.S. acknowledge funding from PhenoMeNal (H2020-EU.1.4.1.3 grant no. 654241). A.R.J. acknowledges funding from the BBSRC (BB/M020282/1, BB/L024128/1, BB/K01997X/1). J.T.M.P. was supported by the Medical Research Council and National Institute for Health Research UK through funding for the Imperial College MRC-NIHR Phenome Centre, MRC grant number MC-PC-12025. T.S. and O.A. acknowledge funding from de.NBI/BMBF (031A535A) GM from de.NBI/BMBF (031A534A). J.A.V. acknowledges funding from the BBSRC (BB/K01997X/1) and NIH (R24 GM127667-01. M.E.'s funding is related to PURE (Protein research Unit Ruhr within Europe), a project of Northrhine-Westphalia, a federal state of Germany.

REFERENCES

(1) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Sci. Data* **2016**, *3*, 160018.

(2) Holcapek, M.; Liebisch, G.; Ekroos, K. *Anal. Chem.* **2018**, *90*, 4249−4257.

(3) Blazenovic, I.; Kind, T.; Ji, J.; Fiehn, O. *Metabolites* **2018**, *8*, 31.

DOI:10.1021/acs.analchem.8b04310

(4) Bingol, K.; Bruschweiler-Li, L.; Li, D.; Zhang, B.; Xie, M.; Bruschweiler, R. *Bioanalysis* **2016**, *8*, 557−573.

(5) Guijas, C.; Montenegro-Burke, J. R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A. E.; Wolan, D. W.; Spilker, M. E.; Benton, H. P.; Siuzdak, G. *Anal. Chem.* **2018**, *90*, 3156−3164.

(6) 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. *Nucleic Acids Res.* **2009**, *37*, D603−610.

(7) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W. T.; Crusemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C. C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; P, C. A. B.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodriguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrov, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Muller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. *Nat. Biotechnol.* **2016**, *34*, 828−837.

(8) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(9) Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. *TrAC, Trends Anal. Chem.* **2016**, *78*, 23−35.

(10) Nguyen, D. H.; Nguyen, C. H.; Mamitsuka, H. *Briefings Bioinf.* **2018**, bby066−bby067.

(11) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W. M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. *Metabolomics* **2007**, *3*, 211−221.

(12) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. *Environ. Sci. Technol.* **2014**, *48*, 2097−2098.

(13) Misra, B. B. *Electrophoresis* **2018**, *39*, 909−923.

(14) Spicer, R.; Salek, R. M.; Moreno, P.; Cañueto, D.; Steinbeck, C. *Metabolomics* **2017**, *13*, 106.

(15) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; Gonzalez-Beltran, A.; Sansone, S. A.; Griffin, J. L.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41*, D781−786.

(16) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44*, D463−D470.

(17) MSI Board Members. *Nat. Biotechnol.* **2007**, *25*, 846−848.

(18) Griss, J.; Jones, A. R.; Sachsenberg, T.; Walzer, M.; Gatto, L.; Hartler, J.; Thallinger, G. G.; Salek, R. M.; Steinbeck, C.; Neuhauser, N.; Cox, J.; Neumann, S.; Fan, J.; Reisinger, F.; Xu, Q. W.; Del Toro, N.; Perez-Riverol, Y.; Ghali, F.; Bandeira, N.; Xenarios, I.; Kohlbacher, O.; Vizcaino, J. A.; Hermjakob, H. *Mol. Cell. Proteomics* **2014**, *13*, 2765−2775.

(19) Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; Ebbels, T.; Kell, D. B.; Manetti, C.; Newton, J.; Paternostro, G.; Somorjai, R.; Sjöström, M.; Trygg, J.; Wulfert, F. *Metabolomics* **2007**, *3*, 231−241.

(20) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, 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. *Mol. Cell. Proteomics* **2011**, *10*, R110.000133.

(21) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. *Nat. Biotechnol.* **2012**, *30*, 918−920.

(22) Schober, D.; Jacob, D.; Wilson, M.; Cruz, J. A.; Marcu, A.; Grant, J. R.; Moing, A.; Deborde, C.; de Figueiredo, L. F.; Haug, K.; Rocca-Serra, P.; Easton, J. M.; Ebbels, T. M. D.; Hao, J.; Ludwig, C.; Gunther, U. L.; Rosato, A.; Klein, M. S.; Lewis, I.; Luchinat, C.; Jones, A. R.; Grauslys, A.; Larralde, M.; Yokochi, M.; Kobayashi, N.; Porzel, A.; Griffin, J.; Viant, M. R.; Wishart, D. S.; Steinbeck, C.; Salek, R. M.; Neumann, S. *Anal. Chem.* **2018**, *90*, 649.

(23) Sansone, S.-A.; Rocca-Serra, P.; Field, D.; Maguire, E.; Taylor, C.; Hofmann, O.; Fang, H.; Neumann, S.; Tong, W.; Amaral-Zettler, L.; Begley, K.; Booth, T.; Bougueleret, L.; Burns, G.; Chapman, B.; Clark, T.; Coleman, L.-A.; Copeland, J.; Das, S.; de Daruvar, A.; de Matos, P.; Dix, I.; Edmunds, S.; Evelo, C. T.; Forster, M. J.; Gaudet, P.; Gilbert, J.; Goble, C.; Griffin, J. L.; Jacob, D.; Kleinjans, J.; Harland, L.; Haug, K.; Hermjakob, H.; Sui, S. J. H.; Laederach, A.; Liang, S.; Marshall, S.; McGrath, A.; Merrill, E.; Reilly, D.; Roux, M.; Shamu, C. E.; Shang, C. A.; Steinbeck, C.; Trefethen, A.; Williams-Jones, B.; Wolstencroft, K.; Xenarios, I.; Hide, W. *Nat. Genet.* **2012**, *44*, 121.

(24) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. *Nucleic Acids Res.* **2016**, *44*, D1214−1219.

(25) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(26) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. *J. Cheminf.* **2013**, *5*, 7.

(27) Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A. R.; Binz, P.-A.; Deutsch, E. W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; Orchard, S.; Antonio Vizcaíno, J.; Hermjakob, H.; Stephan, C.; Meyer, H. E.; Eisenacher, M. *Database* **2013**, *2013*, bat009.

(28) Montecchi-Palazzi, L.; Kerrien, S.; Reisinger, F.; Aranda, B.; Jones, A. R.; Martens, L.; Hermjakob, H. *Proteomics* **2009**, *9*, 5112−5119.

(29) Xu, Q.-W.; Griss, J.; Wang, R.; Jones, A. R.; Hermjakob, H.; Vizcaíno, J. A. *Proteomics* **2014**, *14*, 1328−1332.

(30) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(31) Hartler, J.; Triebl, A.; Ziegl, A.; Trotzmuller, M.; Rechberger, G. N.; Zeleznik, O. A.; Zierler, K. A.; Torta, F.; Cazenave-Gassiot, A.; Wenk, M. R.; Fauland, A.; Wheelock, C. E.; Armando, A. M.; Quehenberger, O.; Zhang, Q.; Wakelam, M. J. O.; Haemmerle, G.; Spener, F.; Kofeler, H. C.; Thallinger, G. G. *Nat. Methods* **2017**, *14*, 1171−1174.

(32) Rost, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.;

DOI:10.1021/acs.analchem.8b04310

Schilling, O.; Choudhary, J. S.; Malmstrom, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, *13*, 741−748.

DOI:10.1021/acs.analchem.8b04310

Software environments and biological applications

10

All the previously mentioned developments can advance biomedical research, allowing more efficient data processing, but also to answer novel questions that were not possible before. My contributions in the area of experimental metabolomics started with data processing of Arabidopsis metabolite profiles in [BvRLS08], and network based data analysis in Nicotiana with Emmanuel Gaquerel **[GKN13]**. An integrated pipeline of our metabolite profiling and identification tools [SGDN13] was implemented by Jan Stanstrup during his stay in my group.

Together with Ivo Grosse I supervised the PhD student Diana Trutschel, who analysed and simulated implications of sources of variances and statistical power **[TSGN14]** on experimental design. My PhD student Susann Mönchgesang combined different types of -omics data for improved biological interpretation, e.g. transcriptomics and proteomics in [HMN16]. The combination of metabolomics data with SNP variants in a natural variation population of Arabidopsis which helped to identify metabolic pathways and link several metabolites and biosynthetic enzymes **[MSS16]**, [MST16].

In the rather recent field of applying metabolomics in ecology, I am working withing the iDiv and with my PostDoc Kristian Peters on this topic [PWW18]. The conducted studies **[PGBN18a, PGBN18b]**, [PTD19] included fully reproducible workflows from raw data to statistical analysis and chemical classification. These analyses were performed using the cloud and workflow technologies developed in the scope of the PhenoMeNal project **[PBB19]**.

The application of computational mass spectrometry tools is not limited to biomedical research, but has also in several cases been applied in environmental research in the FP7 project SOLUTIONS [BAS15, AAAA15] coordinated by Werner Brack (UFZ, Leipzig), where my group is one of the smaller project partners and closely collaborating with Emma L. Schymanski and Juliane Hollender (Eawag, CH).

Metabolomics
DOI 10.1007/s11306-013-0504-2

ORIGINAL ARTICLE

Computational annotation of plant metabolomics profiles via a novel network-assisted approach

Emmanuel Gaquerel · Carsten Kuhl · Steffen Neumann

Received: 31 October 2012 / Accepted: 31 January 2013
© Springer Science+Business Media New York 2013

Abstract Mass spectrometry (MS) has become the analytical method of choice in plant metabolomics. Nevertheless, metabolite annotation remains a major challenge and implies the integration of structural searches in compound libraries with biological knowledge inferred from metabolite regulation studies. Here we propose a novel integrative approach to process and exploit the rich structural information contained in in-source fragmentation patterns of high-resolution LC–MS profiles. In this approach, a correlation matrix is first calculated from individual mass features extracted by xcms processing. Mass feature co-regulation patterns corresponding to metabolite in-source fragmentation are then detected and assembled into compound spectra using the R package CAMERA and processed for in silico fragment-based structure elucidation using MetFrag. We validate the performance of this approach for the rapid annotation of the twelve largest compound spectra, including four *O*-acyl sugars and six 17-hydroxygeranyllinalool diterpene glycosides in metabolic profiles of insect-attacked *Nicotiana attenuata* leaves. Additionally, we demonstrate the power of refining MetFrag metabolite annotations based on co-regulation patterns between known and unknown compounds in the correlation matrix and proposed structural annotations on two previously un-characterized *O*-acyl sugars. In summary, this novel approach facilitates compound annotation from in-source fragmentation patterns using correlation between intensities of mass features of one or several metabolites. Additionally, this analysis provides further support that insect herbivory activates major metabolic reconfigurations in *N. attenuata* leaves.

Keywords *Nicotiana* · Metabolomics · Mass spectrometry · Network analysis · Metabolite annotation

1 Introduction

Plants continuously adjust their metabolism to cope with environmental stresses. In the case of plant responses to insect herbivory, investigations have traditionally been dominated by the analysis of single metabolites or metabolic routes that are thought by researchers to be relevant for resistance. In recent years, significant instrumental and conceptual improvements in mass spectrometry (MS) and nuclear magnetic resonance (NMR) have led to advances in untargeted metabolomics and allowed researchers to obtain a broader view of the changes in the small molecule signature induced by different stresses, including the attack of insects by (Macel et al. 2010).

Nicotiana attenuata, a wild tobacco native species from the Great Basin desert in the USA, is one of the few non-model plant for which such metabolomics approaches have been combined to different system-level analyses in order to

---

Electronic supplementary material The online version of this article (doi:10.1007/s11306-013-0504-2) contains supplementary material, which is available to authorized users.

E. Gaquerel (✉)
Department of Molecular Biology, Max Planck Institute for Chemical Ecology, Hans Knöll Str. 8, 07745 Jena, Germany
e-mail: egaquerel@ice.mpg.de

C. Kuhl · S. Neumann
Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, 06120 Halle, Germany
e-mail: ckuhl@ipb-halle.de

S. Neumann
e-mail: sneumann@ipb-halle.de

Published online: 23 February 2013

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

understand the mechanisms of anti-herbivore defense activation and efficiency under natural conditions (Halitschke et al. 2003; Giri et al. 2006; Gaquerel et al. 2009; 2010; Kim et al. 2011). Several major metabolic shifts detectable by MS based metabolomics are induced in this plant following the attack of herbivores, including those connected with the biosynthesis of the neurotoxin nicotine (Steppuhn et al. 2004), which functions synergistically with anti-digestive plant proteins (Steppuhn and Baldwin, 2007). Herbivore feeding also activates the *de novo* production of 17-hydroxygeranyllinalool (HGL) diterpene glycosides (Heiling et al. 2010) and of phenolic derivatives (Onkokesung et al. 2012). *O*-acyl sugars produced in trichomes are defensive metabolites constitutively present on the leaf surface and have been analyzed by MS approaches (Weinhold and Baldwin 2011).

Non-targeted MS-based metabolomic experiments involve the automated extraction of mass features (distinct *m/z* values measured at a given retention time) and the statistical interpretation of the influence of experimental conditions on the relative intensity of these features. This task can be performed by multiple vendor softwares or Open Source systems such as xcms (Smith et al. 2006), turning MS into the method of choice for the rapid discovery of stress biomarkers (reviewed in Allwood and Goodacre, 2010). Nevertheless, in spite of its high resolution and large range of applications, modern MS based metabolomics still suffers from a lack of comprehensive databases and efficient workflows for the annotation and identification of the compound structures corresponding to mass features of interest (Neumann and Bocker, 2010). For this reason, structure identification represents the key bottleneck in untargeted metabolomics studies, as discussed in the review on metabolite identification in mass spectrometry-focused metabolomics (Dunn et al. 2012).

In *N. attenuata*, only 15 % of the *m/z* features differentially regulated after simulating insect feeding could be assigned to structural elements of previously known and putative metabolites using a comprehensive strategy combining N atom labeling and MS2 experiments (Gaquerel et al. 2010). A recent development to circumvent the limitations of LC–MS databases available for compound annotation consists in searching broad or targeted compound libraries not only with the predicted exact mass and molecular formula of one compound, but also with fragment information from in-source fragments or targeted MS2 measurements (Hill et al. 2008; Wolf et al. 2010). MetFrag is an application that performs in silico fragmentation and ranks the candidate based on the molecular fragments that best explain the measured fragment peaks (Wolf et al. 2010). In silico structure analysis, which also includes predicting metabolite fragmentation trees from MS2 data (Rasche et al. 2011; 2012), represents, in addition with expert knowledge, a promising strategy to speed up the interpretation of MS data. However, these different computational approaches do not take biological and biochemical knowledge into account to provide constraints and narrow down the number of candidates.

The study of correlation relationships among metabolite levels has been used to facilitate the survey of metabolome organization and hypothesis generation on both pathway and metabolite identity. Typical metabolite profiling studies show a few but significantly high correlation values when measurements are repeated across samples harvested in a time course manner (Allen et al. 2010). Correlations between metabolites can inform about metabolic links between pairs or groups of metabolites but are not necessarily in agreement with known or postulated biochemical pathways. Small fluctuations in the level of one metabolite may translate into significant modifications of an organism's metabolic network (Szymanski et al. 2009). By contrast, strong and perturbation-persistent correlations between metabolite levels can originate from direct enzymatic conversions and from indirect transcriptional controls. Metabolic network neighborhood provides powerful hints towards the annotation of an unknown metabolite when few of the correlating metabolites share significant structural similarities. For instance, Hirai et al. (2004, 2005) detected shared regulation among precursors and glucosinolate metabolites after nitrogen and sulfur deficiency and integrated these co-expression data with gene expression profiles to predict novel glucosinolates biosynthetic genes using self-organizing maps (SOM). However, no study has tested how such conceptual 'guilt-by-association' approaches could help prioritizing structure elucidation of unknown metabolites based on their co-regulation with known pathways or sets of metabolites. Most of the procedures developed for gene-to-gene co-expression networks can be transposed to the analysis of metabolite expression data (Breitling et al. 2006; Jourdan et al. 2008), but in the case of MS-based metabolite profiling, each metabolite is often represented by more than one *m/z* feature—including isotopic peaks, different adducts and fragments generated during in-source fragmentation—which complicates network computation and interpretation. An intermediate processing step is required to combine metabolite-derived *m/z* features into clusters, defined by *m/z* feature co-occurrence, isotopic relationships and chromatographic correlations. CAMERA is a recently developed MS interpretation program that performs both reconstruction of metabolite-specific in-source compound spectra and their annotation (Kuhl et al. 2012).

In this study, we propose a novel integrative approach to combine the interpretation of feature correlation relationships visualized using network representations with the rich

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

structural information provided by in-source fragmentation patterns in order to refine in silico metabolite annotation. We tested the efficiency of this work-flow for the annotation of metabolic responses activated by mechanically wounding *N. attenuata* leaves and applying *Manduca sexta* larva oral secretions, a treatment known to recapitulate most of the large-scale changes occurring in the secondary metabolic network of this plant during the feeding of this insect. To this end, we constructed a correlation matrix based on temporal changes in the intensity of individual mass features as extracted by xcms and then integrated co-regulation information inferred from the list of correlating metabolites and CAMERA annotations as constraints during MetFrag metabolite annotation. Today, compound annotation is often performed on accurate mass MS1 data, or requires the additional acquisition of MS2 fragmentation spectra. Our annotation is based on deconvoluted MS1 compound spectra with in-source fragment ions, but where isotopic and multi-charged ions are removed (hereafter referred to as MS1.5 data) workflow facilitates the interpretation of biologically meaningful clusters of molecular fragments in co-regulation networks, and speeds up initial compound annotation or class assignment based on initial MS profiling data.

2 Materials and methods

2.1 Plant growth and treatment

We used an isogenic line of *Nicotiana attenuata* obtained after 30 generations of inbreeding from field-collected seeds. Seeds were germinated as described in Gaquerel et al. (2010). All plants were grown in the glasshouse in 1 L individual pots at 26–28 °C under 16 h of light supplied by Philips Sun-T Agro 400- or 600 W sodium lights (Philips, Turnhout, Belgium).

Metabolic changes induced during *Manduca sexta* feeding were reproduced by producing with a fabric pattern wheel three rows of punctures onto each side of the midvein of five fully expanded leaves per plant (5 biological replicates) and directly applying 1:1 diluted *M. sexta* oral secretions. Treated leaves from the same plant were harvested and flash frozen 0, 1, 2, 4, 14 and 24 h after elicitation. 100 mg of ground leaf tissue was weighed and transferred to a Fast Prep tube containing 0.9 g of Fast Prep matrix (BIO 101, Vista, USA). 1 mL extraction buffer per 100 mg tissue [50 mM acetate buffer, pH 4.8, containing 40 % methanol spiked with reserpine (600 ng/mL), atropine (200 ng/mL)] was added and samples were homogenized. After centrifugation (13,200 rpm, 20 min, 4 °C) the supernatant was collected in a fresh 1.5 mL Eppendorf tube, centrifuged again and 100 µL of the supernatant was transferred to a HPLC vial.

2.2 Metabolite profiling by HPLC-ESI/TOF–MS

Two microliters of the leaf extract were separated using a HPLC 1,100 Series system (Agilent, Palo Alto, USA). The column used was a 150 mm × 2 mm i.d., 3 µm, Phenomenex Gemini NX RP-18 column with a 2 mm × 4 mm i.d. guard column of the same material. The following binary gradient was applied: 0 to 2 min isocratic 95 % A (deionized water, 0.1 % [v/v] acetonitrile [Baker, HPLC grade] and 0.05 % formic acid), 5 % B (acetonitrile, 0.05 % formic acid); 2 to 30 min linear gradient to 80 % B; isocratic for 5 min. Flow rate was 200 µL/min.

Eluted compounds were detected by a micrOTOF mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with an electrospray ionization source in positive and negative ion mode. Typical instrument settings were as follows: capillary voltage 4,500 V, capillary exit 130 V, dry gas temperature 200 °C, dry gas flow of 8 L/min. Ions were detected from *m/z* 200 to 1,400 at a scan rate of 1 Hz. Mass calibration was performed using sodium formate clusters (10 mM solution of NaOH in 50/50 % v/v isopropanol/water containing 0.2 % formic acid).

2.3 Mass spectrometry data processing

Raw data were exported as netCDF files and processed with the *m/z* feature detection and retention time alignment BioConductor package xcms (v1.30, http://www.bioconductor.org/packages/release/bioc/html/xcms.html), using following parameters for feature detection: method = "centWave", ppm = 20, snthresh = 10, peakwidth = (20,50). For alignment, we repeated the xcms *group* method twice and performed a retention time correction in between with the following parameters: minfrac = 0.5, bw1 = 60, bw2 = 25, mzwid1 = 0.05, mzwid2 = 0.02, span = 1, extra = 0, missing = 0. Areas of missing features were estimated using the *fillPeaks* method.

Annotation of compound spectra and corresponding ion species was performed with the BioConductor package CAMERA (v1.9.8, http://www.bioconductor.org/packages/release/bioc/html/CAMERA.html). Compound spectra were built with CAMERA according to the retention time similarity, the presence of detected isotopic patterns and to the strength of the correlation values among extracted ion chromatograms (EICs) of co-eluting *m/z* features. CAMERA grouping and correlation methods were used with default parameters except the threshold for EIC correlation (cor_eic_th) that was increased to 0.85. Clustered features were annotated based on the match (±5 ppm) of calculated *m/z* differences versus an ion species and neutral loss transitions rule set. Mass differences corresponding to NH_3CH_2, $NH_3CH_3H_2$ and $NH_3C_2H_4$ neutral losses were added to the default rule set of CAMERA.

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

If one or more mass differences can be assigned to a compound spectrum, CAMERA will also deduce the neutral exact mass [M] of the underlying compound.

To facilitate sharing and re-analysis of the metabolite profiling data presented in this study, the raw data netCDF files and metadata annotation in the ISAtab format (Sansone et al. 2012) were submitted to the EBI's MetaboLights repository (Haug et al. in press) with the accession MTBLS10 (http://www.ebi.ac.uk/metabolights/MTBLS10).

2.4 Correlation analysis, network visualization and interpretation

Combined xcms and CAMERA output matrix was exported into Excel (v14.1). Based on the ion species annotation, +1, +2 and +3 isotopic peaks were excluded to reduce information redundancy. Zero values remaining after the estimation of missing features' peak areas were replaced by 1/5 of the minimum area measured for a given median *m/z* feature and data were normalized sample-wise using the 75-percentile procedure. Only *m/z* features detected in at least 4 out of 5 biological replicates in at least one time point group were retained for univariate statistics and ab initio correlation weighted network reconstruction.

Networks were visualized using the correlation-weighted layout algorithm in Cytoscape (v2.8.2, http://www.cytoscape.org/; Shannon et al. 2003), a software for visualization of biochemical networks. The Cytoscape plugin MetaNetter (v2.1) was used for inference and visualization of networks from high-resolution mass spectrometry metabolomic data (Jourdan et al. 2008). This open source plugin has been designed to interpret metabolomics experiments by allowing user-friendly correlation matrix computation and *m/z* feature annotation using a list of *m/z* values calculated for candidate metabolites. MetaNetter requires a list of masses (one mass per line) followed by tabulated quantitative data for each mass (*m/z* x area matrices) that was extracted from the xcms/CAMERA report table. In this study, we evaluated the influence of data reduction on the interpretation of network topology by comparing networks obtained for (i) the full *m/z* peak list information and (ii) only a subset of CAMERA-annotated pseudo-molecular ions. Network edges were created using the zero-order Pearson correlation values calculated within MetaNetter using $r = 0.75$ as threshold. Similarly to the CAMERA annotation tool, the ab initio mapping function of MetaNetter detects feature pairs with an *m/z* distance corresponding to known mass differences (called transformation in the MetaNetter nomenclature) within 5 ppm mass accuracy. We used the Cytoscape VizMapper tools to define node colors according to CAMERA grouping, manual compound class annotation, and assigned node shapes according to adduct annotation and edges line styles according to specific neutral losses within CAMERA compound spectra or inter-metabolite *m/z* differences.

2.5 Metabolite annotation

2.5.1 *In silico metabolite annotation with batch queries*

The in silico metabolite annotation tool MetFrag is available both as web application (http://msbi.ipb-halle.de/MetFrag) and as a local application that processes text-based query files and saves results as an SD file of ranked candidate structures. A MetFrag MS1.5 query file was created for every CAMERA-deduced molecular mass hypothesis [M] if the associated compound spectrum contained at least one fragment peak, *i.e.* a *m/z* feature with a lower *m/z* value than [M]. Annotated isotopic peaks and multiply-charged ions such as $[M+2H]^{2+}$ were excluded. The functionality to create MetFrag batch query files is now available in the CAMERA package, starting with version 1.5.2. Extracted MS1.5 compound spectra usually contain a subset of the information that can be expected in a full tandem mass spectrum, but may also include false positives, i.e. unrelated features. Compounds within the injection peak were not considered for in silico analysis.

An example query file is included in the supporting information. Candidate metabolite structures were obtained from a local PubChem mirror (dated 2010-09-06) and an in-house compound structure database (Gaquerel et al. 2010). The combined database was searched using the estimated compound mass [M] within the default 10 ppm error of MetFrag (Wolf et al. 2010). Molecular structures of candidate metabolites and corresponding query files were processed with a locally installed MetFrag (v1.1) using the following parameters: mzabs = 0.01, mzppm = 10, TreeDepth = 2, charge = positive, bio = true (only CHNOPS atoms). Especially within the large lists of ranked candidate structures proposed by MetFrag, the correct solution is not always ranked first. We therefore performed a structural clustering to derive consensus structures among high-scoring candidates, see below.

The SmartFormula 3D algorithm, which verifies that elemental formulas of fragment ions are subsets of the precursor elemental composition, was used to support the annotation of sodium containing fragments within MS1.5 and MS2 spectra of unknown metabolites. For the processing of the MS1.5 and MS2 spectra of two previously unknown *O*-acyl sugar metabolites with MetFrag, we adjusted the *m/z* values of sodium containing fragments to their corresponding protonated forms (subtraction of 21.9825 Da). This step is necessary since MetFrag assumes protonated spectra.

2.5.2 *Relative ranking position*

For several previously known compounds we report the performance of the MS1.5 based annotation, giving both

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

the absolute and a relative rank. The relative ranking position (RRP) among MetFrag candidates was calculated to compare the ranking independently from the length of the candidate lists. A score of 0 corresponds to the correct compound on first rank, whereas a score of 1 correspond to the last rank position.

$$RRP = \frac{1}{2} \times \left(1 + \frac{BC - WC}{TC - 1}\right)$$

where TC is the number of total candidates, BC is the number of better candidates with a higher MetFrag score, WC is the number of worse candidates with a lower MetFrag score.

2.5.3 *Candidate structure clustering*

For each MetFrag result, we performed a chemical structure clustering analysis. As preprocessing we pruned the MetFrag result list, by filter out non-explaining candidates. For result lists with hundreds of candidates, we retained only the best candidates according to the 90th percentile of explained peaks. After filtering a MetFrag result, we calculated for all structures with the R package rcdk (Guha, 2007) an extended binary fingerprint, which encodes 1,024 different chemical properties like a C–C Bond. A property is set to 1, when a structure contains it, otherwise set to 0. Afterwards we calculated the Tanimoto similarity between each fingerprint. The Tanimoto similarity between a fingerprint a and b is defined as C/(A+B–C), where A is the number of 1 in structure a, B is the number of 1 in structure b and C is the number of 1, which both fingerprints have at the same positions. Afterwards we perform a hierarchical clustering using the Tanimoto similarity as distance between two structures (Supplementary Fig. 2). The resulting dendrogram tree was cut at a height of 0.2 (corresponding to a chemical similarity of 80 %), and for each of the resulting clusters, the maximum common substructure was calculated. All calculations were performed in R v2.14 using package rcdk (v3.1.4, http://www.jstatsoft.org/v18/i05/) (Steinbeck et al. 2003). Complete chemical clustering results are presented as Supplementary material.

2.5.4 *Metabolite annotation nomenclature*

Metabolite identifiers are presented in Supplementary File 5. Targeted analyses on metabolites presented in Table 1 have previously been published (Gaquerel et al. 2010; Heiling et al. 2010; Weinhold and Baldwin 2011). Annotation/identification level for each of these metabolites according to the four levels of the metabolite annotation nomenclature proposed by the Metabolome Standard Initiative (Sansone et al. (2007) and employed as described in Matsuda et al. (2010) are summarized in Gaquerel et al. (2010). Briefly, each HGL-diterpene glycosides reported in Table 1—nicotianosides I, II, III, IV, lyciumoside IV, attenoside—was purified, analyzed by MS2 high resolution HPLC-ESI/TOF–MS and its structure identified—identified/level 1—by NMR (Heiling et al. 2010). In Gaquerel

Table 1 Evaluation of the MetFrag annotation process with MS1.5 compound spectra for 12 previously identified *Nicotiana attenuata* metabolites

Annotation	*m/z*	Rt [s]	Database hits	MetFrag	RRP
O-acyl sugars					
AS #1	594.2889	636.4219	514/2	33/298	0.1077
AS #2	608.3046	657.654	432/3	71/256	0.2745
AS #3	622.3203	680.7421	300/3	5/178	0.0226
AS #4	678.3464	774.2913	237/2	4/129	0.0234
HGL-DTGs					
Nicotianoside I	862.4213	586.2605	83/1	13/55	0.2222
Nicotianoside II	948.4227	599.4421	46/1	1/31	0
Nicotianoside III	922.4766	574.3238	56/3	7/43	0.1429
Nicotianoside IV	1024.474	565.2787	27/3	2/20	0.0526
Attenoside	938.4729	557.0429	63/1	5/42	0.0976
Lyciumoside IV	776.419	574.3238	109/2	1/63	0
Others					
Nicotine (CHEBI:18723)	162.1144	51.33913	862/5	253/685	0.3684
Rutin (CHEBI:28527)	610.1535	492.6253	648/6	1/344	0
Overall				Median: 5/96	Mean: 0.1394

Database: PubChem plus in-house; MetFrag results: Cluster rank/total cluster number

17-HGL-DTGs 17-hydroxygeranyllinalool-diterpene glycosides, *AS O*-acyl sugars

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

et al. (2010), the well-established identity of rutin and nicotine—identified/level 1—was further supported by MS1 and MS2 UHPLC-ESI/TOF–MS of authentic compounds. In Matsuda et al. (2010), level 3 of identification corresponding to 'characterized' compounds is defined as 'based upon characteristic physicochemical properties of a chemical class of compounds or by spectral similarity to known compounds of a chemical class'. Fragmentation patterns of *O*-acyl sugars characterized in *N. attenuata* exhibit typical losses of acetylated and non-acetylated fructose, previously observed for *O*-acyl sugars described in other Solanaceae. Short fatty acid chains involved in *N. attenuata* *O*-acyl sugars were identified, after trans-methylation, by GC–MS and comparison with authentic compounds (Weinhold and Baldwin 2011). However, the position of these acyl moieties could not be inferred from the MS2 analyses conducted on *O*-acyl sugars (Weinhold and Baldwin 2011), which justifies the classification of these compounds as level 3 of annotation.

3 Results and discussion

3.1 *Nicotiana attenuata* responses to herbivory as case study for network-assisted computational annotation

The methodology (Fig. 1) presented in this article is based on a combination of mass spectrometry data processing tools, correlation analysis and representation as networks, and in silico metabolite annotation. Significant patterns of biological co-regulation between metabolites as those used in this analysis can more easily be inferred when metabolic systems are subjected to intense perturbations. We used as case study co-expression networks constructed from time-course metabolomics analysis of *N. attenuata*, an ecological model plant. Rosette stage *N. attenuata* plants mount a strong and specific metabolic counter-response when attacked by insects or treated with insect-derived elicitors (Keinanen et al. 2001; Gaquerel et al. 2010). In order to survey the temporal nature of these metabolic changes, mechanically wounded leaves of *N. attenuata* were treated with *Manduca sexta* oral secretions, harvested over a 72 h time period and profiled by HPLC-ESI/TOF–MS in positive ionization mode. In agreement with previous analyses performed using the same experimental design (Gaquerel et al. 2010), pronounced alterations of total ion chromatogram profiles were observed post-elicitation. Non-targeted processing of the raw data files and subsequent pairwise statistical analysis showed that a cumulative proportion of 40.6 % of total reproducibly detected *m/z* features were more than 1.5-fold differentially regulated ($P < 0.05$, for details see Supplementary File 1) for at least one of the harvesting time points compared to untreated leaf samples harvested at the initial stage of the time-course experiment. Changes with the greatest amplitude were detected 14 and 72 h post-elicitation. Supplementary File 1 summarizes the regulation and CAMERA-based annotation of reproducibly detected *m/z* features along the time-course analysis.

One of the central goals of this study was to demonstrate the value of rapid and proper annotation of in-source fragmentation clusters from profiling data to make sense of a metabolomics data set. We first manually annotated major in-source compound spectra using both CAMERA and knowledge from previous studies on this model system. Consistently, *m/z* features associated with previously identified and predicted 17-hydroxygeranyllinalool (HGL) diterpene glycosides, phenolic derivatives and *O*-acyl sugars accounted for a large proportion of the variance within the data-set and their intensity was differentially affected by the herbivory-mimicking treatment (Supplementary File 1). Many of these metabolites are only present in few plant relatives and some have been recently structurally elucidated and therefore are not present in compound databases. The sparsity of compound databases for such secondary metabolites represents an ideal platform to prove the value of fragment computational analysis (Fig. 1) as presented in this article to annotate some of these metabolites or formulate hypothesis on some of their structural features.

3.2 Network visualization of mass feature correlation supports CAMERA-reconstruction of in-source fragmentation clusters and informs on metabolite links

Correlations between metabolite levels calculated from time series represent a well-recognized source of information that can be used to assess the modularity of a metabolic system (e.g. Fukushima et al. 2011). In the case of MS-based metabolic profiling, in addition to the links between metabolites originating from connected metabolic pathways, strong and persistent correlations are observed among *m/z* signals derived from the same molecule (Breitling et al. 2006; Draper et al. 2009; Gaquerel et al. 2010). The degree and intensity of correlation-based connectivities among *m/z* features of a data-set can therefore be used, in addition to the chromatography-based clustering performed by CAMERA, to distinguish features derived from a specific compound spectrum from those corresponding to differentially regulated co-eluting compounds. To that end, correlations above 0.85 were visualized using the correlation-weighted network layout in Cytoscape (Fig. 2a). In this layout, network edges are inversely proportional to the strength of the correlation

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

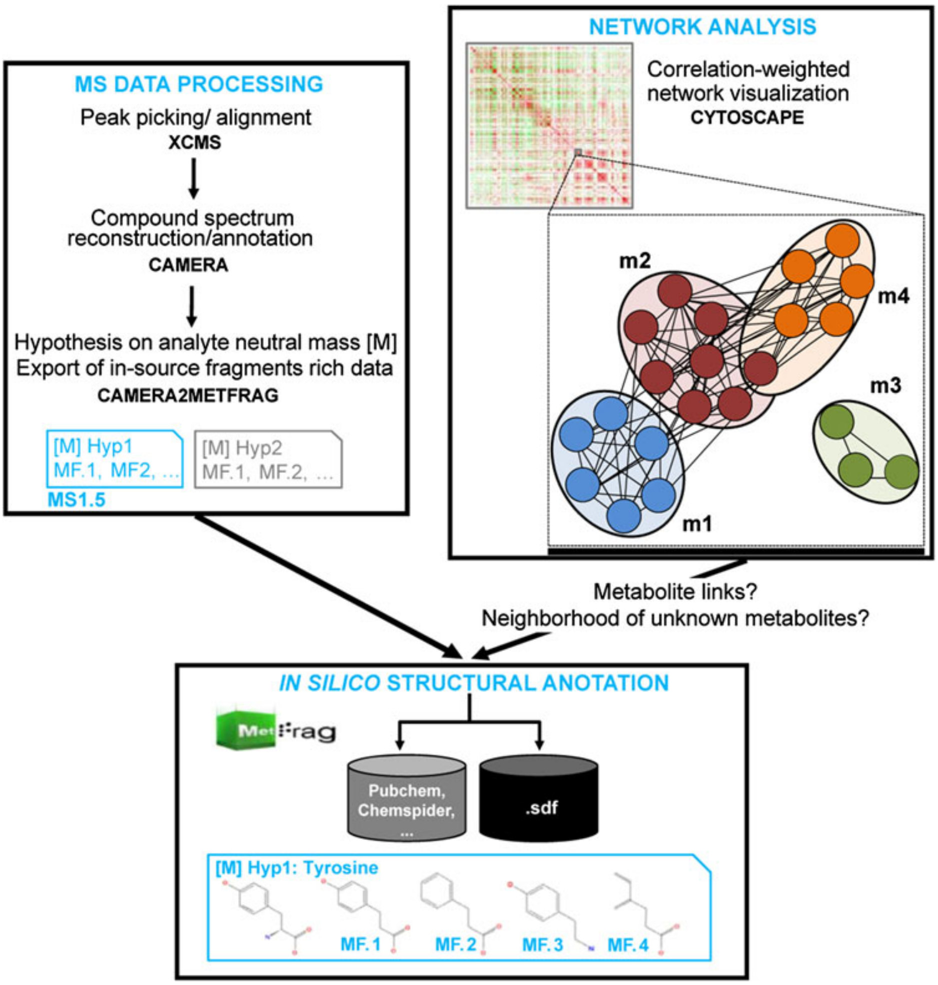


Fig. 1 Workflow for combined network and fragment analysis. Non-targeted processing ('MS DATA PROCESSING') of mass spectrometry data by xcms and CAMERA generates isotopic/adduct/neutral loss-annotated m/z feature x intensity matrices that are used for correlation-weighted network visualization using Cytoscape ('NETWORK ANALYSIS'). Annotated compound spectra (MS1.5) corresponding to each neutral mass hypothesis [*M*] proposed by CAMERA and containing more than molecular fragment (*MF*) are extracted using a custom *R* script and analyzed for in silico structural annotation in MetFrag using public and in-house library ('IN SILICO STRUCTURAL ANNOTATION'). Combining these two approaches facilitates the survey of metabolite co-regulation (e.g. m1 to m4), including known-to-unknown metabolite correlation. The immediate metabolic neighborhood of unknown metabolites is included to increase the coverage of potentially structurally related metabolites during fragment analysis by MetFrag

between mass features. This allows localizing groups of *m/z* features sharing high correlation within the network representation. Main modules overlapped with sets of *m/z* features differentially regulated over the time series, reinforcing the value of using network representations to uncover hidden patterns of independent regulation structuring this large data-set.

The interpretation of groups of highly correlated *m/z* features—densely connected modules of the network—is more intuitive when meta-information is mapped onto nodes and edges of the network. Mapping compound spectra information obtained from CAMERA onto the network using the VizMapper tool from Cytoscape supported that in most cases densely connected clusters represented in-source fragments. Most intra-compound spectrum *m/z* pairs shared a correlation coefficient above 0.85 (data not shown). We accordingly observed that nodes with highest connectivity indexes were located among features within a compound spectrum (data not shown). As expected, partial networks representing correlations between m/z features of a given compound spectrum show high connectivity. Thus, Fig. 2c shows the partial network for co-expressed *m/z* features within the compound spectrum extracted and annotated by CAMERA for rutin.

The visualization of CAMERA annotations onto the network also allowed the rapid assessment of the performance of CAMERA chromatographic peak grouping, based on the inspection of the densely connected modules. For example, the

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

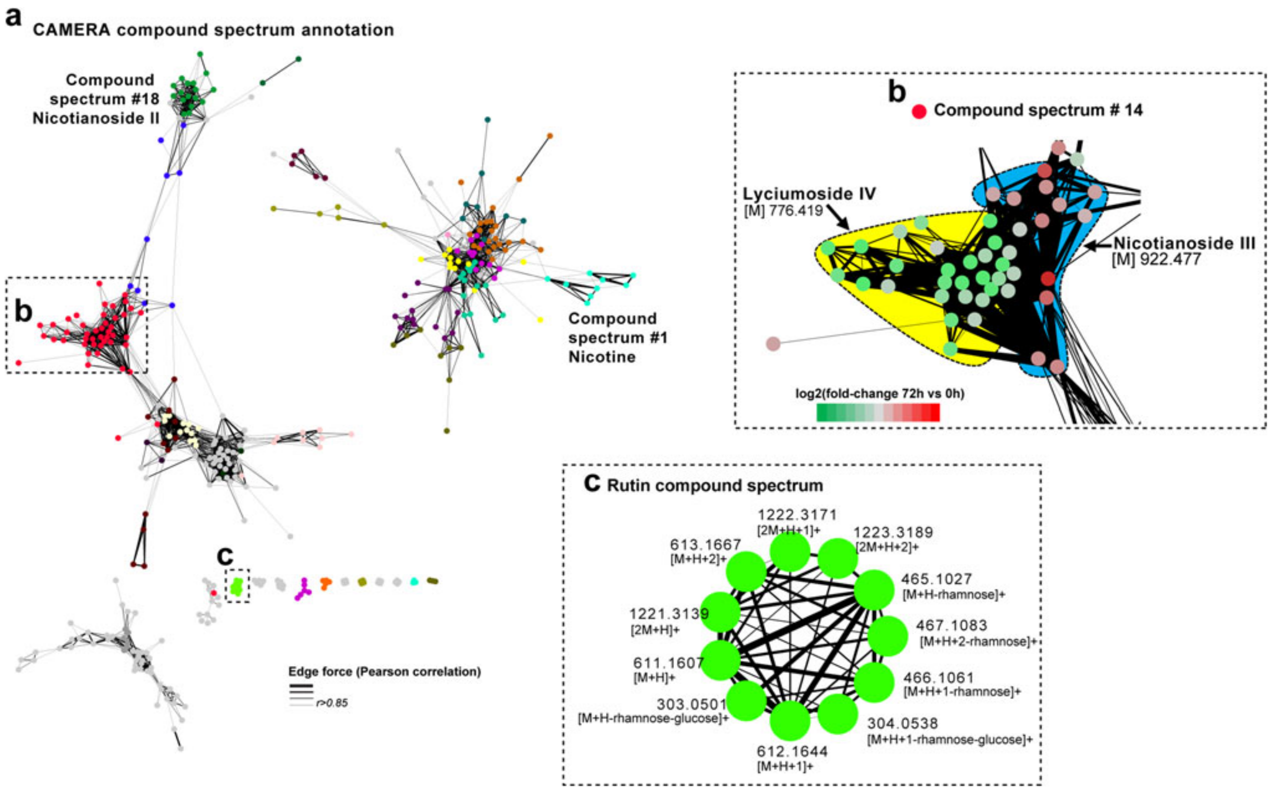

a
CAMERA compound spectrum annotation
Compound spectrum #18 Nicotianoside II
b
Compound spectrum #1 Nicotine
c
Edge force (Pearson correlation)
r>0.85
b
Compound spectrum # 14
Lyciumoside IV
[M] 776.419
Nicotianoside III
[M] 922.477
log2(fold-change 72h vs 0h)
c
Rutin compound spectrum
1222.3171 [2M+H+1]+
1223.3189 [2M+H+2]+
613.1667 [M+H+2]+
465.1027 [M+H-rhamnose]+
1221.3139 [2M+H]+
467.1083 [M+H+2-rhamnose]+
611.1607 [M+H]+
466.1061 [M+H+1-rhamnose]+
303.0501 [M+H-rhamnose-glucose]+
304.0538 [M+H+1-rhamnose-glucose]+
612.1644 [M+H+1]+

Fig. 2 *m/z* feature-based correlation analysis and network representation of *N. attenuata* induced responses. **a** Correlation coefficients calculated between xcms/CAMERA-processed *m/z* features were visualized using the correlation-weighted network layout of Cytoscape. Only correlation coefficients $r > 0.85$ were visualized as edges. Densely connected regions of the network overlapped with sets of *m/z* features differentially regulated over the time series (Supplementary File 1 and Fig. 1). The network representation, when colored according to CAMERA's clustering, allows distinguishing among correlation modules between intra-metabolite spectra and inter-metabolite links. The 25 largest compound spectra constructed by CAMERA are mapped onto the network using different node colors. Compound spectra 1 and 18 corresponding respectively to nicotine and nicotianoside II in-source fragmentation patterns are annotated. **b** Magnification on compound spectrum #14. This module, which includes a total of 80 features eluting within a retention time window of 6s, appears to be composed of two distinct 17-hydroxygeranyllinalool diterpene glycosides (HGL-DTG), namely nicotianosides III and lyciumoside IV. **c** Network constructed for the rutin compound spectrum using Cytoscape circular network layout. Edge length is not weighted with this layout. Edge *thickness* denote for the correlation strength

compound spectrum #14 shown in Fig. 2b, which includes a total of 80 features eluting within a retention time window of 6 s, appeared to be composed of two distinct HGL-diterpene glycosides, namely nicotianoside III and lyciumoside IV. This shows the value of combining chromatography- and regulation-based information to facilitate the deconvolution of spectra especially in dense regions of a chromatogram. This evaluation was extended to each spectrum retrieved by CAMERA prior to performing metabolite annotation searches.

3.3 In silico annotation of compound spectra

As a next step, we tentatively annotated compound spectra in the network. CAMERA has been shown to facilitate the rapid annotation of molecular masses [M] based on adduct mass differences and isotopic cluster relationships. We therefore used CAMERA annotations as primary data for metabolite identity searches. However, as already shown by multiple other studies (Kind and Fiehn 2006), querying compound databases like PubChem solely with the molecular mass or even elemental compositions usually returned too many candidate structures (ranging typically from dozens to thousands, Table 1). We developed a novel workflow for compound annotation that exploits the rich fragment information contained in each compound spectrum. To this end, we extracted for each [M] hypothesis, based on CAMERA and network clustering results, the in-source fragments as additional structural hints from the corresponding compound spectrum, and removed uninformative *m/z* signals like adducts and isotopic peaks. Resulting deconvoluted compound spectra derived from in-source ionization, hereafter referred to as MS1.5 data, were used for *in silico* structural searches with MetFrag using a custom R script. The functionality to create MS1.5 MetFrag batch query files is now available in the CAMERA package, starting with version 1.5.2. MetFrag allows

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

comparing observed peaks with fragments generated *in silico* from candidate compound structures retrieved from public databases, according to the molecular mass [M] or elemental composition of a given unknown metabolite (Wolf et al. 2010).

In-source fragmentation depends heavily on compound structure and instrumental settings; the performance of MetFrag for this type of data had not been thoroughly tested before. We therefore conducted an evaluation of the improvement of using MS1.5 data compared with searches in compound libraries solely based on [M] or elemental composition (Table 1). Several *N. attenuata* metabolites previously identified to the Metabolome Standard Initiative annotation levels one or three (Gaquerel et al. 2010; Heiling et al. 2010; Weinhold and Baldwin 2011) were used for the validation of this pipeline. From 43 non-trivial compound spectra that contained at least one fragment peak after filtering annotated isotopic and adduct peaks, we obtained 81 MS1.5 batch query files (due to the different [M] hypotheses formulated by CAMERA). Noteworthy, abundant and well characterized herbivory-inducible *N. attenuata* metabolites, such as most phenolic derivatives (Gaquerel et al. 2010; Onkokesung et al. 2012), whose compound spectra, obtained in positive ionization mode, did not contain other fragment peaks than isotopic and multi-charge ion peaks were not included in this analysis. Each MS1.5 query file was processed by MetFrag against a mirror of the PubChem library, which was supplemented with entries of an in-house database (Gaquerel et al. 2010) to raise the coverage of plant metabolites.

Table 1 summarizes the MetFrag results for MS1.5 data of previously identified *N. attenuata* metabolites. The complete result list for all 43 compound spectra is included as Supporting Information to this article. In this proof-of-concept study, the rank of the correct compound among the MetFrag candidates ranged from 1 (nicotianoside II, lyciumoside IV and rutin) to 253 (nicotine). To take into account the different number of candidates retrieved by MetFrag when interpreting rank positions, we calculated the relative ranking positions (RRP, see Methods) of each identified metabolite. The RRP of the correct structure among MetFrag results ranged from 0 (correct candidate ranked first) to 0.369 for nicotine and 0.1394 on average. These results indicate the superiority of our approach, compared with searches solely based on [M] and elemental formula, for fast metabolite re-annotation.

In the case of nicotine, the rank of the correct compound was beyond 250, which is a very poor result. The nicotine MS1.5 spectrum with six fragment peaks is presented as Supplemental information (Supplementary Note 1, “Query Spectrum for nicotine” and Supplementary Fig. 1). MetFrag explained only two of them, and those were also present in more than half of the other candidate structures, indicating that these in-source fragments were too unspecific. Consistent with this explanation, we observed that performing MetFrag processing on high quality MS2 data measured for nicotine improved neither the absolute (rank = 253) nor the relative rank positions (RRP = 0.375) of nicotine. By clear contrast, MS1.5 spectra were sufficient to identify and even discriminate among HGL-diterpene glycoside molecules: the median absolute rank for HGL-diterpene glycosides was 3.5, the mean relative rank position 0.086 and fragment elemental formulae proposed by MetFrag were in perfect agreement with those published in Gaquerel et al. (2010) and Heiling et al. (2010). HGL-diterpene glycoside compound spectra all share diagnostic fragments at *m/z* 271.24 and 289.25 that correspond to the aglycone after successive glucose and rhamnose losses. Correct structural annotations were obtained for these *m/z* signals by MetFrag.

Typically, there are dozens to hundreds of structure candidates for each query compound. In order to condense these result-lists and to highlight the maximum common substructures (or consensus motifs) among high-scoring candidates, we performed a structural clustering for each MS1.5 result list. Structural clustering based on molecule fingerprints is also a convenient means of identifying backbones shared by different molecules (Schymanski et al. 2012) (Supplementary Fig. 2). The correct metabolite is not guaranteed to be ranked on the first place after MetFrag processing, but the top candidates typically include multiple structurally-related metabolites. We therefore postulate that this clustering analysis provides important hints for structural motifs when interpreting MetFrag results, even in the case of unknowns not present in the compound database. Figure 3 presents predictions and consensus structure reconstructions obtained using MetFrag followed by structural clustering for the MS1.5 spectra corresponding to rutin, the major flavonoid glycosides in *N. attenuata* leaves. After filtering fingerprints of the best 66 candidate structures from MetFrag were grouped as two main clusters and three singletons, with a Tanimoto distance above 0.2. Both clusters exhibit the flavone backbone of rutin. It is also noteworthy that the rutinose (6-*O*-L-rhamnosyl-D-glucose) moiety of rutin was contained in numerous of MetFrag candidates, although not consistently well positioned. Structural clustering results for each of Table 1 entries are available as Supplemental information (Supplementary File 2).

3.4 Metabolite co-regulation informs compound class prediction of unknown structures during in silico fragment analysis

The annotation strategy explained above solely depends on the interpretation of individual compound spectra, where

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

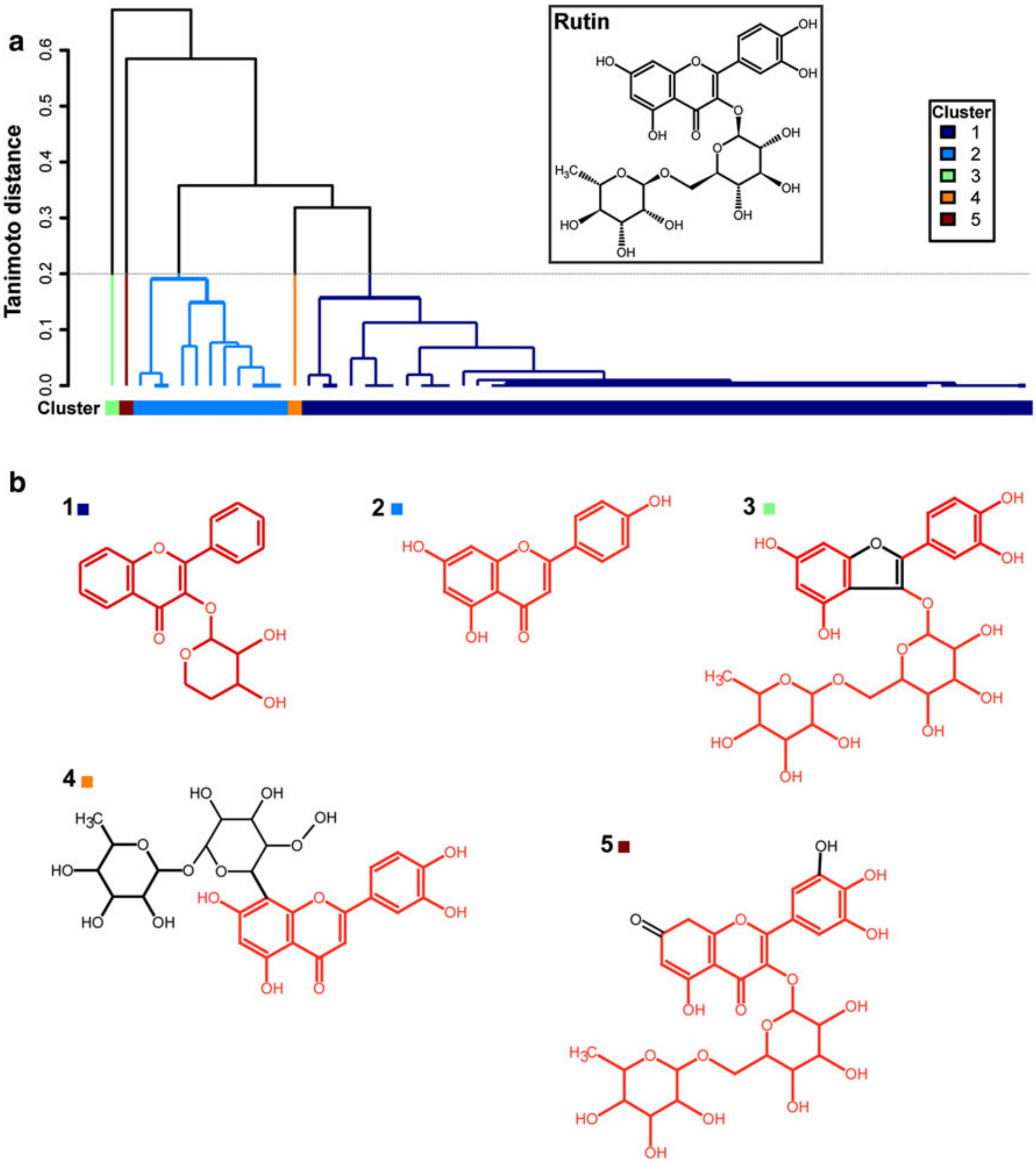


Fig. 3 Chemical similarity clustering annotates the flavone backbone of rutin, *Nicotiana attenuata*'s major leaf flavonoid. **a** Dendrogram representation of the chemical similarity between 66 top-scoring MetFrag candidates retrieved from the processing of rutin MS1.5 spectrum. Chemical similarity clustering analyzes the chemical resemblance, calculated as Tanimoto similarity between binary fingerprints of metabolite structures. **b** Maximum common substructures extracted for the two clusters (1 and 2) and three singletons (3, 4 and 5) separated by a Tanimoto distance of 0.2 (corresponding to a chemical similarity of 80 %). Both clusters (1 and 2) exhibit the flavone backbone of rutin (highlighted in *red*) (Color figure online)

no a priori knowledge about the potential compound class is available. We next analyzed whether metabolite links in the network provide additional information to improve the annotation. Considering the clear compound-class-based cluster demarcation achieved by the correlation analysis, our assumption is that metabolite co-regulation reflects to a large extent compound class membership (Fig. 4) and in turn that unknown metabolites sharing tight co-regulation with known metabolites might be biosynthetically related.

Two unknown metabolites sharing high correlation with several characterized metabolites leaped out as interesting candidates (Fig. 4). In the network visualization, the *m/z* signals derived from those metabolites were located in the *O*-acyl sugars enriched sub-network and annotated by

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

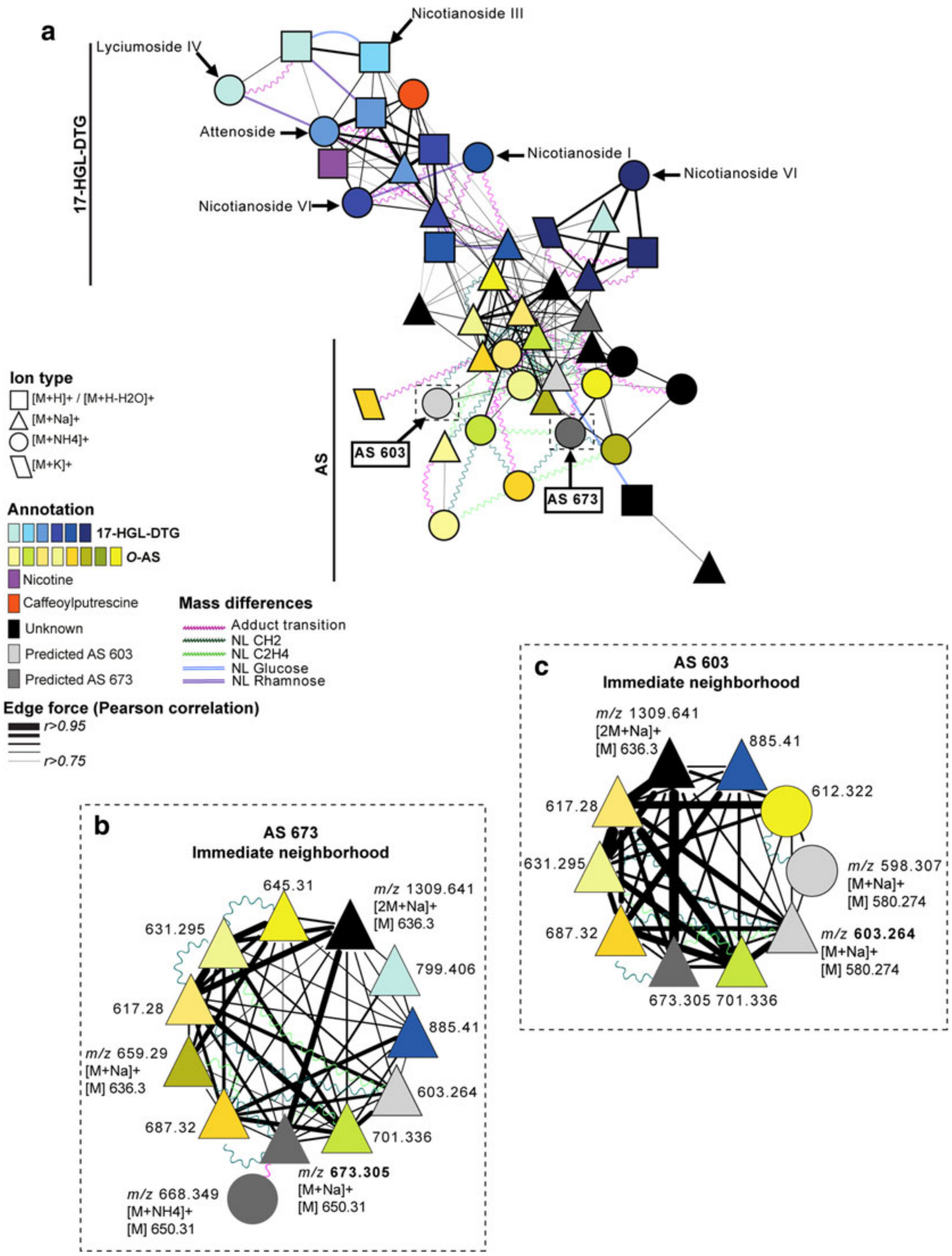


Fig. 4 Metabolite co-regulation informs compound class prediction of unknown pseudo-molecular ions. **a** To direct and prioritize the annotation process of unknown neighbors of known metabolites, we computed a new correlation ($r > 0.75$, correlation-weighted layout) network solely based on precursor *m/z* signals identified for each of the compound spectra processed with MetFrag. Taking benefit of the reduced chemical redundancy, adduct mass differences or specific neutral losses were identified by implementing CAMERA rules into the MetaNetter plugin of Cytoscape and mapped as waved edges onto the network. Different adduct types are depicted as different node shapes. Node colors correspond to different compound classes. Considering the clear compound-class-based cluster demarcation within the correlation network, we assumed that *m/z* features derived from unknown metabolites and sharing a high connectivity index with previously characterized metabolites likely belong to the same compound class as known metabolites. Such phenomenon was especially observed for the subnetwork enriched in *O*-acyl sugar (AS)-derived pseudo-molecular m/z features. **b**, **c** Networks built for the first neighbors of *m/z* 603.264 (AS 603) (b) and *m/z* 673.305 (AS 673) **c** predicted as sodium adducts of non-previously reported *O*-acyl sugars spectrum using Cytoscape circular network layout. Edge length is not weighted with this layout. Edge *thickness* denote for the correlation strength. Structural predictions for AS 603 MS 1.5 and AS 673 MS2 are presented in Fig. 5

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

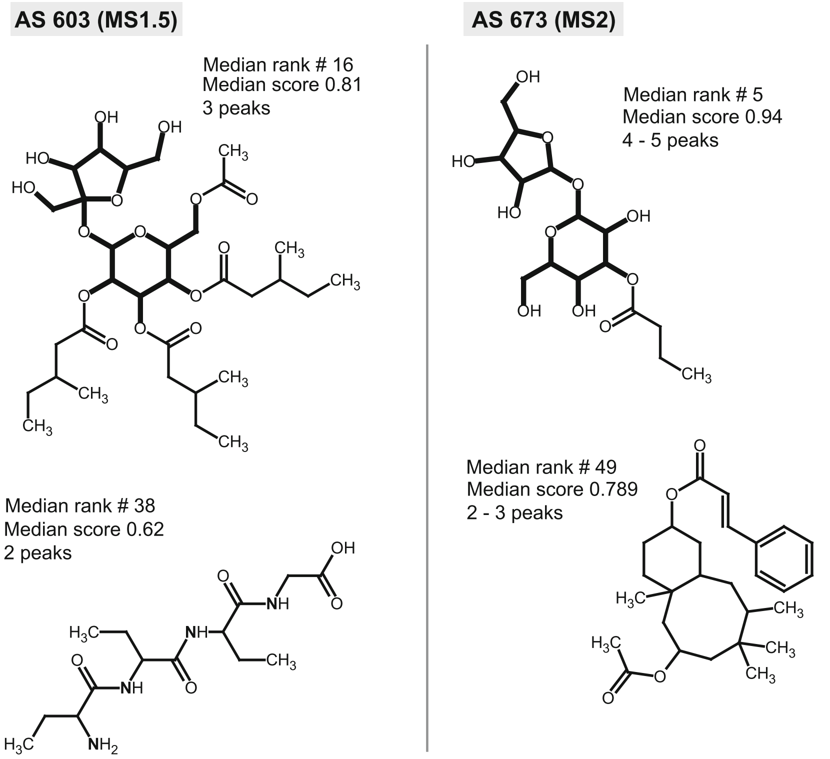


Fig. 5 Examples of maximum common substructures generated for AS 603 MS1.5 and AS 673 MS2. AS 603 and AS 673 correspond to two unknown metabolites predicted to belong to the *O*-acyl sugar compound class based on the high co-regulation shared with known AS (Fig. 4). An MS1.5 extracted for AS 603 and an MS2 spectrum measured for AS 673 precursor were processed in MetFrag after merging candidate structures sharing a Pearson correlation $r > 075$ with AS 603 and AS 673—immediate network neighborhood—in order to enrich MetFrag fragment analysis with compound features structurally-related to AS. Maximum common substructures retrieved for each cluster and singleton separated by a Tanimoto distance of 0.2 (corresponding to a chemical similarity of 80 %, Supplementary Figs 4, 5 and 6) were manually inspected. Based on elemental composition analyses presented in Supplementary Fig. 3, nitrogen-containing candidate structures can easily be ruled out during manual inspection of the consensus structure list. MetFrag retrieves per default structures containing only CHNOPS atoms. In both cases (MS1.5 and MS2 spectra), consensus structures exhibiting the *O*-acyl sugar backbone (*bold*) appear among top-ranking candidates with three to five peaks explained at a 10 ppm threshold and a median score greater than 0.8

CAMERA as sodium adducts with *m/z* 673.3031 and 603.2639 (Supplementary Fig. 3). Mapping known mass differences/biochemical transformations between both candidates and neighboring, known *O*-acyl sugar compounds showed that these metabolites share known neutral mass shifts (CH_2, C_2H_4, $C_6H_{10}O_4$, …) and also high correlation, two indicators for both biosynthetic relationships and co-regulation (Supplementary File 3). Hence, we predict that both correspond to two not previously reported *O*-acyl sugars and therefore named them accordingly AS 673 and AS 603. Network visualizations only provide graphic representations of particular relationships calculated for features of interest and their topology is *per se* strongly influenced by the type of layout used for their representation. To detect molecular fragments corresponding to these two unknown metabolites and co-regulated compound spectra, we listed *m/z* features correlated (Pearson correlation $r > 0.75$) with the pseudo-molecular ions of AS 673 and AS 603 and called each set of correlated *m/z* features the immediate metabolic neighborhood of these two metabolites (Supplementary File 3).

The MS1.5 spectrum of AS 603 contains 9 query peaks (Supplementary File 2). PubChem returned more than 600 candidates matching within 10 ppm the [M] hypothesis for AS 603, but we had previously observed that *O*-acyl sugars are practically absent from this database (Table 1). Our in-house library contained only one compound with a corresponding mass. To confirm our *O*-acyl sugar working hypothesis for AS 603, we merged all 2,122 candidates of the 10 connected metabolites (Supplementary File 3) and processed them with MetFrag (Supplementary Fig. 4). Inclusion of candidates from the immediate neighborhood provides potentially related structures, which is especially helpful when compound databases have only limited coverage for the required compound class. Obviously, the inclusion of compound spectra corresponding to correlated *m/z* features

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

increased the number of candidates by five-fold, but more than two-thirds of those had a MetFrag score of zero. Again, we performed a chemical clustering of the high-scoring candidates (Supplementary File 4 and Supplementary Fig. 5). The reduced list of 20 candidate structures obtained after consensus structure generation can be more easily screened manually for particular structural motifs. The first candidate structure supporting the *O*-acyl sugar hypothesis appears at rank #10 (Fig. 5). Several of the other high-scoring structures returned from this analysis could potentially be ruled out based on their elemental composition, on their plant metabolite likeness or after expert analysis of their fragmentation pattern. The *O*-acyl sugar consensus candidate structure deduced by this analysis should represent the starting point for additional *O*-acyl sugar targeted studies involving expert knowledge and established analytics such as NMR for final structure identification.

The second unknown compound of interest was named as putative AS 673. No in-source fragments were detected for this metabolite, likely due to its relatively low abundance. We therefore applied the strategy described above to the targeted tandem mass spectrum (Supplementary Fig. 3). Here, MetFrag retrieved several *O*-acyl sugar compounds with very good scores among the 2,414 candidates of the 11 merged immediate neighbors (Supplementary Files 3 and 4). The maximum common substructure of the corresponding cluster (cluster #2) contains the *O*-acyl sugar backbone (Fig. 5), reinforcing the prediction that AS 673 is an *O*-acyl sugar (Supplementary Fig. 6). Individual members of this cluster have their short chain fatty acid residues located on different positions (Supplementary File 4). As for AS 603, candidate consensus structures provide initial hypotheses to be tested in further analytical experiments. Pubchem, the compound library used in this study, is extremely large (more than 30 million entries) but its coverage of plant metabolites is rather limited. Plant-specific public or in-house libraries may alternatively be used as upstream databases for MetFrag analysis. We notably tested KNApSAcK, a database of species-metabolite relationships, for structure annotation. However, the number of candidates retrieved from KNApSAcK via the MS1.5 MetFrag approach was too low and did not provide enough resolution to perform chemical similarity analysis (data not shown). Several attempts have been carried out in recent years to evaluate the likeness of a structure to be a metabolite. Classifiers and molecular representations used to build metabolite likeness models could for instance be integrated to our pipeline to discriminate metabolite from non-metabolite structures among MetFrag candidates. For the two unknown compound spectra analyzed, approximately one-fourth of the structural hypotheses inferred from the clustering analysis can be falsified based on the presence of halogen elements (Figs. 4, 5). This proportion increased up to 85 % if nitrogen-containing structures are filtered out as well, knowing that AS 603 and AS 673 likely do not contain nitrogen elements. Nevertheless, filtering-out nitrogen containing structures may also significantly affect the structural resolution required for consensus backbone prediction during chemical similarity clustering.

4 Concluding remarks

Correlation analysis from high-resolution MS metabolic profiles is an extremely powerful approach to analyze dependencies in the response of metabolites to stress conditions. Our study suggests that intra- and inter-compound spectra correlations that can be graphically distinguished by mapping CAMERA annotations onto network representations can additionally be used for evaluating in silico metabolite annotation. The MS1.5 MetFrag approach introduces an additional level of confidence compared to simple accurate mass searches for metabolite annotation, because additional molecule fragmentation is included for comparison of compound spectra. This scoring strategy allows the annotation of known metabolites with a median relative rank of 5 and a RRP of 0.1394. Additionally, we discuss the value of extending MetFrag analyses to highly correlated compound spectra—represented as immediate neighbors in the network—to circumvent the limited metabolic coverage of many chemical database and to develop initial hypotheses on unknown metabolites (here two previously uncharacterized *O*-acyl sugars).

The pipeline on MS1.5 data presented here is *per se* restricted to the analysis of metabolites exhibiting sufficient in-source fragmentation for the ionization and extraction conditions tested and therefore many previously characterized *N. attenuata* metabolites not fulfilling this condition were not included in this analysis. Non-common secondary metabolites strongly induced in *N. atttenuata* leaves during insect herbivory such as HGL-diterpene glycosides are here valuable examples for demonstrating the performance of our approach for the annotation of compounds not included in conventional databases. Importantly, this pipeline solely based on open-access programs is equally applicable to high-resolution MS metabolic profiles obtained from other experimental systems, including human metabolome data. For these reasons, the performance of MS1.5 for compound annotation and class prediction based on metabolite co-regulation speaks for the value of combining hints derived from different resources for structure elucidation.

Acknowledgments We thank Prof. Ian T. Baldwin and Prof. Dierk Scheel for critical comments on the manuscript, Dr. Matthias Schöttner for analytical advices and for the measurement of metabolomic profiles, Sven Heiling for sharing his expertise on HGL-

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

diterpene glycoside metabolism and Alexander Weinhold for help with the interpretation of AS.

References

Allen, E., Moing, A., Ebbels, T. M., Maucourt, M., Tomos, A. D., Rolin, D., et al. (2010). Correlation Network Analysis reveals a sequential reorganization of metabolic and transcriptional states during germination and gene-metabolite relationships in developing seedlings of Arabidopsis. *BMC Systems Biology, 4*, 62.

Allwood, J. W., & Goodacre, R. (2010). An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis, 21*, 33–47.

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., & Barrett, M. P. (2006). Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics, 2*, 155–164.

Draper, J., Enot, D. P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., et al. (2009). Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics, 10*, 227.

Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, Hankemeier T, Goodacre R, Neumann S, Kopka J, Viant MR (2012) Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics. in press doi:10.1007/s11306-012-0434-4.

Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2011). Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Systems Biology, 5*, 1.

Gaquerel, E., Heiling, S., Schoettner, M., Zurek, G., & Baldwin, I. T. (2010). Development and validation of a liquid chromatography-electrospray ionization-time-of-flight mass spectrometry method for induced changes in *Nicotiana attenuata* leaves during simulated herbivory. *Journal of Agriculture and Food Chemistry, 58*, 9418–9427.

Gaquerel, E., Weinhold, A., & Baldwin, I. T. (2009). Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphigidae) and its natural host *Nicotiana attenuata*. VIII. An unbiased GCxGC-ToFMS analysis of the plant's elicited volatile emissions. *Plant Physiology, 149*, 1408–1423.

Giri, A. P., Wunsche, H., Mitra, S., Zavala, J. A., Muck, A., Svatos, A., et al. (2006). Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata*. VII. Changes in the plant's proteome. *Plant Physiology, 142*, 1621–1641.

Guha, R. (2007). Chemical informatics functionality in R. *Journal of Statistical Software, 18*(5), 1–16.

Halitschke, R., Gase, K., Hui, D., Schmidt, D. D., & Baldwin, I. T. (2003). Molecular interactions between the specialist herbivore *Manduca sexta* (lepidoptera, sphingidae) and its natural host *Nicotiana attenuata*. VI. Microarray analysis reveals that most herbivore-specific transcriptional changes are mediated by fatty acid-amino acid conjugates. *Plant Physiology, 131*, 1894–1902.

Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P, Maguire E, González-Beltrán A, Sansone SA, Griffin JL, Steinbeck C. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data Nucl Acid Res (in Press).

Heiling, S., Schuman, M. C., Schoettner, M., Mukerjee, P., Berger, B., Schneider, B., et al. (2010). Jasmonate and ppHsystemin regulate key Malonylation steps in the biosynthesis of 17-hydroxygeranyllinalool diterpene glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. *Plant Cell, 22*, 273–292.

Hill, D. W., Kertesz, T. M., Fontaine, D., Friedman, R., & Grant, D. F. (2008). Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Analytical Chemistry, 80*, 5574–5582.

Hirai, M. Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D. B., Yamazaki, Y., et al. (2005). Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *Journal of Biological Chemistry, 280*, 25590–25595.

Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., et al. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences USA, 101*, 10205–10210.

Jourdan, F., Breitling, R., Barrett, M. P., & Gilbert, D. (2008). MetaNetter: inference and visualization of high-resolution metabolomic networks. *Bioinformatics, 24*, 143–145.

Keinanen, M., Oldham, N. J., & Baldwin, I. T. (2001). Rapid HPLC screening of jasmonate-induced increases in tobacco alkaloids, phenolics, and diterpene glycosides in *Nicotiana attenuata*. *Journal of Agriculture and Food Chemistry, 49*, 3553–3558.

Kim, S. G., Yon, F., Gaquerel, E., Gulati, J., & Baldwin, I. T. (2011). Tissue specific diurnal rhythms of metabolites and their regulation during herbivore attack in a native tobacco *Nicotiana attenuata*. *PLoS One, 6*, e26214.

Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics, 7*, 234.

Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., & Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry, 84*, 283–289.

Macel, M., Van Dam, N. M., & Keurentjes, J. J. (2010). Metabolomics: The chemistry between ecology and genetics. *Molecular Ecology Resources, 10*, 583–593.

Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., et al. (2010). AtMetExpress development: A phytochemical atlas of Arabidopsis development. *Plant Physiology, 152*, 566–578.

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

Onkokesung, N., Gaquerel, E., Kotkar, H., Kaur, H., Baldwin, I. T., & Galis, I. (2012). MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A: Polyamine transferases in *Nicotiana attenuata*. *Plant Physiology, 158*, 389–407.

Rasche, F., Scheubert, K., Hufsky, F., Zichner, T., Kai, M., Svatos, A., et al. (2012). Identifying the unknowns by aligning fragmentation trees. *Analytical Chemistry, 84*, 3417–3426.

Rasche, F., Svatos, A., Maddula, R. K., Böttcher, C., & Böcker, S. (2011). Computing fragmentation trees from tandem mass spectrometry data. *Analytical Chemistry, 83*, 1243–1251.

Sansone, S. A., Fan, T., Goodacre, R., Griffin, J. L., Hardy, N. W., Kaddurah-Daouk, R., et al. (2007). The metabolomics standards initiative. *Nature Biotechnology, 25*, 846–848.

Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al. (2012). Toward interoperable bioscience data. *Nature Genetics, 44*, 121–126.

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

Schymanski, E. L., Gallampois, C. M., Krauss, M., Meringer, M., Neumann, S., Schulze, T., et al. (2012). Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Analytical Chemistry, 84*, 3287–3295.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*, 2498–2504.

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*, 779–787.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences, 43*, 493–500.

Steppuhn, A., & Baldwin, I. T. (2007). Resistance management in a native plant: nicotine prevents herbivores from compensating for plant protease inhibitors. *Ecology Letters, 10*, 499–511.

Steppuhn, A., Gase, K., Krock, B., Halitschke, R., & Baldwin, I. T. (2004). Nicotine's defensive function in nature. *PLoS Biology, 2*, E217.

Szymanski, J., Jozefczuk, S., Nikoloski, Z., Selbig, J., Nikiforova, V., Catchpole, G., et al. (2009). Stability of metabolic correlations under changing environmental conditions in Escherichia coli–a systems approach. *PLoS ONE, 4*, e7441.

Weinhold, A., & Baldwin, I. T. (2011). Trichome-derived O-acyl sugars are a first meal for caterpillars that tags them for predation. *Proceedings of the National Academy of Sciences USA, 108*, 7855–7859.

Wolf, S., Schmidt, S., Muller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics, 11*, 148.

DOI:10.1007/s11306-013-0504-2 Reprinted with permission. © 2013 Springer Science+Business Media

Metabolomics
DOI 10.1007/s11306-014-0742-y

ORIGINAL ARTICLE

Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data

Diana Trutschel · Stephan Schmidt · Ivo Grosse · Steffen Neumann

Received: 29 April 2014 / Accepted: 17 October 2014
© Springer Science+Business Media New York 2015

Abstract Univariate hypotheses tests such as Student's *t* test or variance analysis (ANOVA) can help to answer a variety of questions in metabolomics data analysis. The statistical power of these tests depends on the setup of the experiment, the experimental design and the analytical variance of the actual observations. In this paper, we demonstrate how a well-designed pilot study prior to an experiment with the aim to find differences between e.g. several genotypes, can help to determine the variance at multiple levels ranging from biological variance, sample preparation to instrumental variances. Next, we illustrate how these variances can be used to obtain several parameters (e.g. minimum statistically significant effect, number of required replicates and error probabilities) which influence the design of the actual study. In particular, we are going to sketch how technical replicates can improve the performance of a test, when they are correctly used in the statistical analysis, e.g. with a hierarchical model. Finally, we demonstrate the process of evaluating the trade-off between different experimental designs with different replication strategies. The choice of an experimental design beyond the gut feeling can be influenced by factors such as costs, sample availability and the accuracy of of the tests. We use metabolite profiles of the model plant *Arabidopsis thaliana* measured on an UPLC-ESI/QqTOF-MS as real-world dataset, but the approach is equally applicable to other sample types and measurement methods like NMR based metabolomics.

Keywords Metabolomics · Statistics · Variances · Hierarchical experiment design

1 Introduction

The aim of metabolomics is to obtain a snapshot of metabolite levels in biological samples. Identification and quantification of metabolites help to understand the metabolic state and metabolic changes e.g. in response to environmental stimuli. Mass spectrometry (MS) is an important analytical method in metabolomics experiments (Dunn 2008), which provides high level of sensitivity for quantification as well as structural hints towards identification (Dunn et al. 2013).

One aim of metabolomics is to detect differences between sample classes, e.g. the comparison of different genotypes (Broadhurst and Kell 2006). Commonly, Student's *t* test (Student 1908) is used as univariate hypothesis test in order to detect significant changes in measured data and to check whether different sample classes have the

Electronic supplementary material The online version of this article (doi:10.1007/s11306-014-0742-y) contains supplementary material, which is available to authorized users.

D. Trutschel (✉) · S. Schmidt · S. Neumann
Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany
e-mail: Diana.Trutschel@ipb-halle.de

S. Schmidt
e-mail: sschmidt@ipb-halle.de

S. Neumann
e-mail: sneumann@ipb-halle.de

I. Grosse
Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle, Germany
e-mail: ivo.grosse@informatik.uni-halle.de

I. Grosse
German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

Published online: 02 November 2014

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

same mean of feature intensities $\mu_1 = \mu_2$ or whether they differ significantly. ANOVA (Tutz et al. 1996) is used to compare the intensities among more than two sample classes. Another level of generalisation are multilevel mixed models, where observed data is approximated by linear regression models and thus both fixed and random effects can be modelled (Pinheiro and Bates 2014).

Recent papers have described the appropriate design of experiments in order to optimize the sample processing steps and metabolomics protocols (Danielsson et al. 2012; Eliasson et al. 2012). However, they did not reflect on the design of experiment in relation to biological questions. The Metabolomics Standards Initiative (MSI) has published recommendations for reporting statistical analyses of metabolite data (Goodacre et al. 2007) because the fact has been criticized that "only a small percentage of papers in metabolomics make much of importance of statistics" (Broadhurst and Kell 2006), especially concerning the appropriate experimental design for addressing biological questions. The main differences between univariate and multivariate statistical methods are discussed in Saccenti et al. (2013). The tools and challenges in metabolomics data analysis are reviewed in Hendriks et al. (2011), while Vinaixa et al. (2012) is focused on a workflow to apply univariate statistical methods. Here, we highlight the use of univariate methods in metabolomics experiments with a focus on replication types, to design a multi-level study as suggested in Hendriks et al. (2011). For both the Student's *t* test and ANOVA it is important to accurately estimate the mean and variance of the intensities. The uncertainty when determining these values directly influences experimental design decisions for a study, such as the number of samples, whether or not and how many technical replicates are required to assure the study's statistical validity. The choice of an experimental setup is also influenced by considering costs and experimental constraints.

In microarray analysis, suggestions for specific tests to cope with small sample sizes have been published for a long time, see e.g. (Baldi and Long 2001; Lönnstedt and Speed 2001). Moreover, the use of different replication types as well as the amount of samples in relation to statistical validity has been discussed for epidemiology studies (Donner and Klar 1996; Dreyhaupt et al. 2013). However, little attention has been paid to these issues in the field of metabolomics.

Depending on the experimental design, several sources of variance are present in metabolomics data that influence type and result of the hypothesis tests. Previous studies have analysed the total of variances observed for technical, preparation and biological replicates (Roepenack-Lahaye et al. 2004).

Here we present a detailed analysis of all variance levels. We suggest a pilot study with a hierarchical experiment design, which allows the usage of a nested linear regression model to obtain exact and unbiased estimates of individual variances at different levels using random effects on different levels. The metabolite intensities include the fixed effect we are interested in for the detection of biomarkers, and random effects that occur at different steps or levels during the experiment. Multilevel mixed models can capture both types of effects and their hierarchical structure (Davis 2002). In addition, mixed models can cope with uneven sample numbers, inhomogeneous variances, missing values and the structure of dependent observations. However, in this paper we restrict the discussion to the case of equal sample numbers and homogeneous variances, and focus on the effect from not-independent measurements resulting from the experimental design. We describe how these dependencies can be handled using the commonly used *t* test statistics. We are going to illustrate that a hierarchical *t* test correctly includes both biological and technical replicates without distorting the results. We also provide the information of the implementation for the general case, a hierarchical ANOVA, which is a restricted mixed model, to analyse such datasets from hierarchical experiments.

Additionally, we consider the impact of the respective number of replicates on the statistical power of the tests, which indicates statistical validity. Moreover, we provide functions to calculate quantities like the resulting power, required number of replicates or the minimal statistically significant effect for different combinations of replicates. We can associate costs related to different levels of replication which are not limited to actual expenses, but also human efforts, availability of samples or time constraints. The overall aim is to find a compromise between expenses and the quality of inference possible in a particular experiment. In addition to general information, we present an example with real-world data from metabolite profiles of *Arabidopsis thaliana*.

2 Materials and methods

In this section, we explain the hierarchical experiment design for our pilot study. Furthermore, we mention methods for calculating statistical power and confidence interval of means as indicators of expected quality of hypothesis testing.

The measured MS data are first preprocessed with feature detection algorithms to reduce the raw data to feature lists resembling metabolite abundances, and then with alignment algorithms to produce a single $M \times S$ matrix of

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

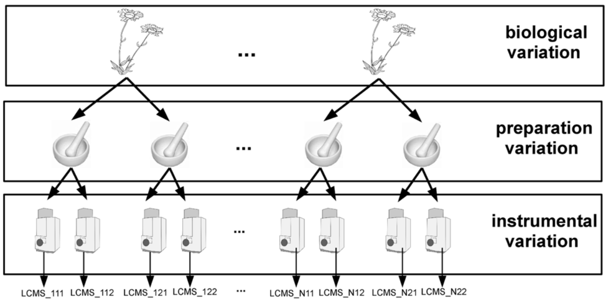


Fig. 1 Hierarchical experiment design. At all levels of variation replicates were prepared: to extract biological variation several plants were grown. From each plant, several extractions were performed, to assess the preparation variation. To identify the instrumental variation each extract was measured several times. The number of LC-MS datasets is the product of the number of plants N, extracts E per plant and injections I per extract

mass features observed across the samples. This matrix is the basis of the subsequent statistical analysis.

2.1 Pilot study to identify sources of variation in MS experiments

The hierarchical experiment design is the precondition to quantify sources of variation separately using a linear hierarchical regression model, which is a special case of a linear mixed model. We then perform a simulation to determine the number of observations in each level of the hierarchical experiment.

2.1.1 *Hierarchical experimental design*

A hierarchical experiment design, shown in Fig. 1, was used to quantify variation at different levels of the experiment. Three sources of variation in MS experiments have been considered: (i) instrumental variation, (ii) preparation variation (both will later be combined into technical variation) and (iii) biological variation, which in this case is variation between plants. On top of these, other levels like e.g. experimental design factors or environmental variation could be introduced, but this was not examined in this paper. For the quantification we prepared a hierarchical set of samples at different levels of variation. The total variation is the sum of all three variations.

2.1.2 *Sample preparation*

Arabidopsis thaliana Col-0 was used as plant material. The plants were grown on soil in a growth chamber under controlled conditions. In the following, we refer to individual plants (grown at the same time and under the same conditions) as biological replicates. The frozen leaf material of each plant was ground and weighed into two samples (preparation replicates) using a cryogenics robot[1] with a weighing error ≤ 5 %. Each extract was measured twice (instrument replicates) under identical conditions. Overall $N = 27$ plants, $E = 2$ preparations, and $I = 2$ LC-MS runs resulted in $N \times E \times I = 108$ LC/MS runs. Full details are available in supplemental material S1.

2.1.3 *Mass spectrometry analysis and data processing*

Metabolite intensities were recorded according to Böttcher et al. (2009). In brief, the chromatographic separation was performed on a *Waters Acquity* UPLC system coupled to a *Bruker micrOTOF-Q* mass spectrometer. Mass spectra were recorded in positive ion centroid mode with a scan rate of 3 Hz and a mass range of 100–1000 m/z. Full details are available in supplemental material S1. This experimental setup can routinely detect semi-polar plant metabolites from major biosynthetic classes including glucosinolates, indolic compounds, phenylpropanoids, benzenoids, flavonoids, terpenes and fatty acid derivatives (Böttcher et al. 2011). Processing of MS raw data, including peak picking and retention time correction, was performed with XCMS (Smith et al. 2006). All statistical calculations were performed in R (http://www.r-project.org/). An underlying assumption of the original Student's *t* test (and also ANOVA) is that the mean intensities are normally distributed. To transform the data towards more normally distributed values, all gathered metabolite intensities were logarithmized. The raw data files, the preprocessed peak matrix and the protocol descriptions have been submitted to the Metabolights repository (Haug et al. 2013), and are available under the accession number MTBLS74[2].

2.1.4 *Variance estimation*

Only the overall variance σ^2_{tot}, i.e. the sum of technical and biological variances, can be estimated directly from the dataset. To obtain an unbiased estimation at *individual* hierarchical levels (Fig. 1), we model the instrumental σ^2_{instr}, preparation σ^2_{prep} and biological variances σ^2_{biol} as random effects with a three-level linear regression model for each detected feature:

$$Y_{nei} = \mu + \beta_n + \gamma_{ne} + \delta_{nei} \qquad (1)$$

where Y_{nei} is the observed measurement of injection i of extraction e of plant n, μ the overall mean of population, β_n

---

[1] http://www.labman.co.uk/portfolio-type/ipb-cryogenic-grinder-and-feeder-system.

[2] http://www.ebi.ac.uk/metabolights/MTBLS74.

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

the independent random biological effect on plant n, γ_{ne} the independent random preparation effect on preparation e in plant n and δ_{nei} the independent random instrumental effect on injection i in preparation e in plant n. The random effects $\beta_n, \gamma_{ne}, \delta_{nei}$ are independent between each other. The unbiased estimator is explained in supplemental information S2. We used the data of the pilot study and the preprocessing as described in 2.1.1 and 2.1.3. In general, we report the average across all features, but in 3.1.1 we also discuss the proportion of biological variance to total variance $\frac{\sigma^2_{biol}}{\sigma^2_{tot}}$, also known as intra-class correlation (ICC).

2.1.5 *Confidence of variance estimation*

With the multilevel linear regression model and hierarchical experiment design we can estimate the variances of different levels, but we also want to ensure estimation with a sufficiently small error. Therefore, we need a certain number of observations in each variance level. We performed a repeated simulation of observations of hierarchical experiments to obtain the minimal number of observations to estimate variances in Algorithm 1, more details are provided in supplemental section S3. The better the estimation of a parameter, the more closer the estimator is to the true value. With this simulation we can calculate the variation of the estimated variances. We determine the 95 % -quantile of estimated variances in each level. The smaller the quantile, the better the estimation. Hence, the number of plants N, the number of preparations E, and the number of measurements I can be determined with Algorithm 1 if the maximal size of the 95 % -quantile of estimated variances in each level is given.

2.2 Hypotheses tests for differential metabolites and biomarker detection

Biomarker detection and the analysis for differential metabolites requires to detect intensity differences between classes of samples. We give a short explanation of hypothesis tests used here and the concept of power and the impact of the degrees of freedom on test statistics to assess their reliability.

2.2.1 *Hierarchical and non-hierarchical hypotheses tests*

If there are only two sample classes to compare, then the Student's t test can be used to find differences in means of observed intensities. For more than two sample classes, the ANalysis Of VAriances, short ANOVA, is used. This method produces an F-statistic to test the class means for equality using the ratio of the variance calculated among the means to the variance within the samples, shown in Table S1 in the supplemental section S6. Both tests are non-hierarchical models, and can not be applied directly to multilevel observations.

The hierarchical version of ANOVA, nested ANOVA, implicitly averages the technical replicates and can thus be applied to multiple levels with biological and technical replicates. For just two sample classes, a hierarchical Student's t test can also be derived as shown in Table S1 in the supplemental section S6. Both are special cases of multilevel linear mixed models (Raudenbush and Bryk 2002). Note that, if the technical replicates of each biological observation are averaged beforehand, the level of technical replicates is eliminated and the non-hierarchical test can be used.

2.2.2 *Statistical power and confidence interval of means*

The statistical power of a test is a measure of the expected quality of an experimental design (Snijders 2001). The α cut-off in hypotheses tests defines the maximum allowed probability of Type I errors, i.e. false positives, where a non-differential feature is incorrectly determined as differential. The statistical power is defined as $1-\beta$, where β is the probability of errors of type II, and hence $1-\beta$ is the minimum desired probability to detect the true positives among all differential features.

The power can be visualised as the area under the curve of the alternative hypothesis H1 in a range between $[t_\alpha, \infty)$ using a right-tailed test and t_α as the critical t value given the threshold α. The graphical representation of the statistical power calculation is shown in supplemental material S4. The power can also be calculated if both the distribution under the null hypothesis and under the alternative hypothesis are known.

In the case of the Student's t test, the shape of the distribution of both the null and the alternative hypothesis depends on the number of observations in the test, in our example the number of plants N in each sample class. The location of the distribution of the alternative hypothesis depends on variance σ^2 and the effect size δ, which is denoted by the difference in means $|\mu_1 - \mu_2|$, and also on the number of observations N in the sample classes. Another way to assess the test quality is the standard error of the difference in means (Holmes 2004). Under the assumption that the effect size is normally distributed with $|\mu_1 - \mu_2|$ and standard deviation σ, then the standard error of difference in means is denoted as $SE = \frac{\sigma}{\sqrt{2*N}}$. The standard error, and thus the variance and the number of observations, determine the size of the 95 % confidence interval $soc = 2 * t_{\alpha=0.5, DoF=2*(N-1)} * SE$ within the true value of difference in means is. The smaller the size of confidence interval the more likely the more precise the estimates accuracy.

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

If four of the five parameters (i) power $1-\beta$, where β is the probability of error type II, (ii) number of samples N, (iii) effect δ between two groups, and (iv)variance σ^2 are given, the missing parameter can be calculated (Broadhurst and Kell 2006). The R package `stats` provides the function `power.t.test` for the Student's *t* test.

In both the hierarchical and the non-hierarchical case, the distribution under the null hypothesis $\mu_1 = \mu_2$ is t-distributed with $DoF = 2 * (N-1)$ degrees of freedom. The distribution of the alternative is a non-central t-distribution with the same number of DoF and the non-centrality-parameter $ncp = \sqrt{\frac{N}{2}}\frac{\mu_1-\mu_2}{\sigma}$. So via ncp the distribution of the alternative hypothesis depends on the three parameters N, δ, σ^2, which determine the position of the non-central t-distribution.

Here, we are interested in the influence of different sources of variation, replication strategies and sample sizes have on the statistical power in multilevel models (Snijders 2005).

In the case of non-hierarchical experiments the variance is $\sigma^2 = \sigma^2_{bio} + \sigma^{\mathbf{2}}_{\mathbf{tech}}$, while in the case of hierarchical experiments with different levels of variances $\sigma^2 = \sigma^2_{bio} + \frac{\sigma^{\mathbf{2}}_{\mathbf{tech}}}{\mathbf{T}}$. T is the number of technical replicates for each biological sample and the technical variance is $\sigma^2_{tech} = \sigma^2_{prep} + \sigma^2_{instr}$, where each preparation is a technical replicate which is measured once.

Thus the distribution of the alternative hypothesis between hierarchical and non-hierarchical models are different because $\sigma^2_{tech} > \frac{\sigma^2_{tech}}{T}$, or in other words the 95 % confidence interval of true difference in means is smaller when calculated via a hierarchical model compared to a non-hierarchical model.

We have implemented power calculation for the hierarchical case in the R-function `power.hierarch.ttest()`. The example in the supplemental material vignette shows the usage. The function is analogous to `power.t.test`, but requires the individual variances σ^2_{tech} and σ^2_{bio} and the number of technical replicates T.

3 Results and discussions

In this section we quantify the variance levels of our study. We also investigate the quality of variance estimation to guide the choice of required observations in each variance level. Knowing the variances we can calculate the loss of power using the total variance instead of biological variance in Student's *t* test and discuss the precision of tests with regard to the confidence interval of the mean effect size. Furthermore, we give some advice on using technical replicates or not, and how to include them in the analysis.

3.1 Sources of variation in MS experiments

3.1.1 Quantify sources of variation

We have implemented the variance estimation for pilot studies following a hierarchical design as introduced in Sect. 2.1.4 in R. The user needs to supply the preprocessed mass feature intensity matrix, which can be obtained with XCMS as described in Sect. 2.1.3, together with a description matrix that assigns the individual samples and the corresponding replication level. A detailed example is available as supplemental vignette.

We have performed the pilot study for a typical *A. thaliana* metabolomics experiment, as described in Sect. 2.1.2. After the data preprocessing of the 108 samples, we obtained a 108×642 intensity matrix. The actual identity of our 642 features is for the remainder of this paper not relevant.

Using the methods implemented in R we determined the individual variances in the dataset. Figure 2 shows the estimated variances for all $S = 642$ features both at the individual levels and the total variance. Negative variances can occur in a few cases in the upper levels because the estimator is unbiased.

The mean values of all feature variances are $\sigma^2_{instr} = 0.043, \sigma^2_{prep} = 0.076, \sigma^2_{biol} = 0.172$. The level increases from technical to biological variation $\sigma^2_{instr} < \sigma^2_{prep} < \sigma^2_{biol}$ and the mean total variance $\sigma^2_{tot} = 0.291$ is the sum of these individual contributions.

These values will vary according to the actual pilot study. If the samples are obtained from a homogeneous culture of e.g. bacteria, the biological variance might be lower, while a more complex sample processing (including e.g. solid phase SPME cartridges, vacuum concentration and re-solving) could increase the preparation variation.

We can also derive the proportion of each variance source of the total variance, analogous to the intra-class correlation (ICC) definitions in Sampson et al. (2013). For example the mean proportion of plant variance on total variance is the average proportion across all S features in the data matrix. It is calculated as $ICC_{mean} = \frac{1}{S}\sum_i^S ICC_i$ with $ICC_i = \frac{{}_i\sigma^2_{biol}}{{}_i\sigma^2_{tot}}$ for each feature i. In relative numbers the average instrumental variance is 16.7 %, preparation variance is 29.1 %, and plant variance is 54.2 % of the total. We can also use the distribution of ICC_i of the individual features to illustrate the amount of features with a minimum ICC, as shown in the second graph in Fig. 2. In our case, half of the features have in ICC above 0.58.

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

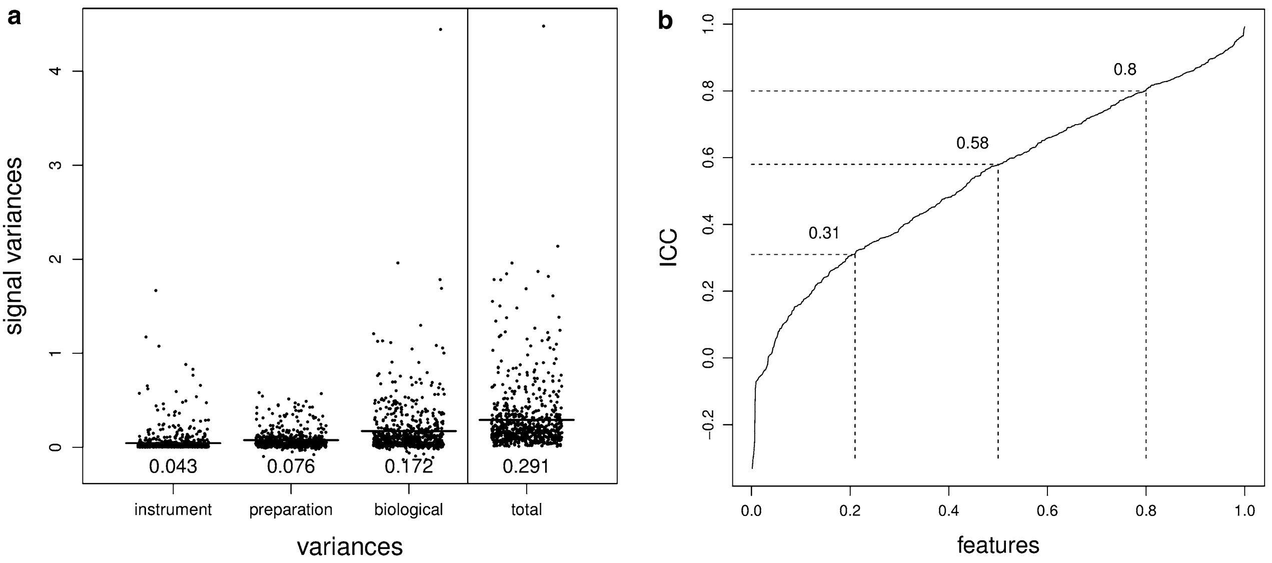


Fig. 2 The distribution of estimated variances of all measured features in leaf samples. **a** from left to right the estimated variances of all measured features $S = 642$ in leaf samples for $\sigma^2_{instr}, \sigma^2_{prep}, \sigma^2_{biol}$, and σ^2_{tot} are plotted. Each dot represents the estimated variance of one feature in the sample. The mean of all estimated feature variances for each variance level is given below and shown as black bar. **b** The cumulative distribution of ICC_i for all features i. E.g. 80 % of the features have an ICC above 0.31, half of the features have an ICC above 0.58, and even 20 % are above 0.8. The higher the proportion of features with a large ICC, the more important is a hierarchical experiment

In our pilot study we performed 108 LC/MS measurements altogether, but we will describe in the next subsection whether also a lower number of plants $N < 27$ would lead to sufficiently reliable variance estimates.

3.1.2 *Influence of replicate numbers on variance estimation quality*

The variation of an estimated parameter can be used as a measure of the quality of the estimation, because the less the estimator varies, the more accurate the estimator is. The estimation of variances described above results from the hierarchical design of the pilot study. We now determine how confident these estimates are, using the variance of the variance estimation depending on the number of replicates.

We simulate measurements in our hierarchical experiment design, drawing the intensities of one feature from a normal distribution with the mean and variance of our actual setup determined from the pilot study. With this simulated data, we can estimate the variances in each level. This simulation is repeated a large number of times, to determine the 95 % confidence interval of the variance estimation as a measurement of quality of estimation, see Algorithm 1 in supplemental section S3.

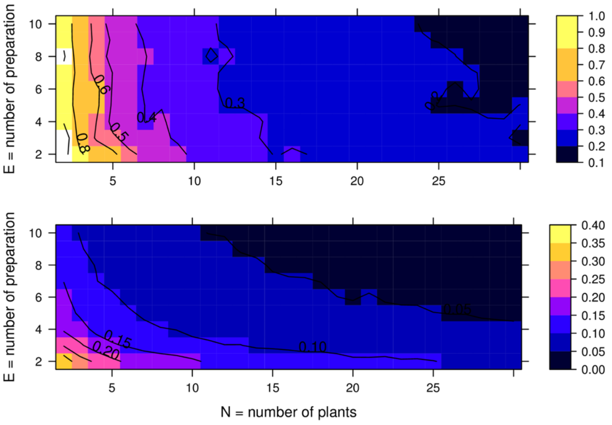


Fig. 3 95 %-confidence interval of estimated plant (upper) and preparation variance (lower). Using Algorithm 1 in supplementary material, we simulated data for $N = 2, 3, \ldots, 30$ plants and $E = 2, 3, \ldots, 12$. The quality of the estimation of plant variance in the upper plot is (mostly) independent of the number of extractions and only depends on N, while the estimation of preparation variance improves with both higher E and N. Generally, the quality of estimation is related to the product of observation numbers in the current level of preparations and the level of plants above

In Fig. 3 we show the width of the 95 %-confidence interval of estimated variances for a combination of simulated numbers of replicates in several levels. From the figure we can determine whether an increase in the number of preparations or in the number of biological replicates results in a more reliable variance estimation.

For the topmost level of biological variation σ^2_{biol}, we observe that the quality of the estimation depends almost

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

exclusively on the number of plant replicates (see top of Fig. 3). Because of the hierarchical design of the pilot study, we do not need a large number of preparation replicates from a single plant to reliably preparation variation estimate σ^2_{prep}.

We recommend to not acquire more than two technical (preparation or injection) replicates, and instead focus on the biological replicates, because the quality of preparation variance estimation is related to the product of the number of plant and preparation replicates. More generally, the quality of the variance estimation on any level is related to the product of observations in this level and the levels above.

In our case, if we want to estimate a plant variance with ± 0.15 and preparation variance with ± 0.075 confidence of 95 % it is sufficient to use $N = 18$ biological and $E = 2$ preparation replicates in the pilot study.

3.2 Hypotheses tests for differential metabolites and biomarker detection

3.2.1 *Treating with different types of replications*

In a typical metabolomics experiment, we need to detect statistically significant features. In the simplest case, we use a Student's *t* test between two sample classes, while ANOVA is used for more than two sample classes.

Because we are interested in the biological effects, a large number of biological replicates might be needed to accurately detect significant features. However, in reality several constraints might apply which limit the number of biological replicates, for example if only a finite number of samples is available. In that case, technical replicates can improve the accuracy of statistical tests.

The detection of differential features with the Student's *t* test has to be performed based on the biological replicates, rather than using technical replicates, or even worse, combining and treating technical and biological replicates as the same (Pavlidis et al. 2003; Johnson et al. 2007). Therefore, the measured biological and technical replicates must be treated separately in the hypothesis test: the Student's *t* test assumes that all samples are independent observations. Technical replicates of a sample are *not* independent from each other. This would violate the most important assumption and overestimates the degrees of freedom of the underlying hypothesis distribution. In general this lead to more false positives (Broadhurst and Kell 2006; Karp et al. 2005), as shown in Figure S3 in supplemental section S5. Considering the problem of non-independent observations scientists have to apply the correct analysis approach.

If a Student's *t* test is used for the statistics, the correct approach is to average the technical replicates (Broadhurst and Kell 2006; Horgan 2007). Averaging technical replicates will decrease the technical variance σ^2_{tech}. If technical replicates are measured from several preparations, then the technical variance decreases to $\frac{\sigma^2_{prep}+\sigma^2_{instr}}{E}$. If technical replicates are injection replicates from the same preparation, then the technical variance decreases to $\frac{\sigma^2_{instr}}{I} + \sigma^2_{prep}$. Thus, the observed total variance will be closer to the biological variance.

But are the technical replicates required in first place? The answer depends both on the achievable improvements in statistical power, but also on the actual costs and required efforts. If the estimated biological variance σ^2_{biol} is much greater than the technical variances σ^2_{prep} and σ^2_{instr}, doubling (or even tripling) the number of measurements will only gain little power, but significantly increase the effort required for data analysis and storage, while the same increase in power could also be achieved by increasing the number of biological replicates by some percentage.

If the technical variance σ^2_{tech} is too high, or if additional power resulting from technical replicates is required, they can be incorporated explicitly into a hierarchical type of ANOVA, also called nested variance analysis (Karp et al. 2005), or even more general, a multilevel mixed model, rather than using the simple averaging approach described above. In fact, the Student's *t* test can be interpreted as a special case of the general ANOVA, and this in turn as a special case of the the nested or hierarchical ANOVA (Ahrens 1967), which allows to explicitly consider different levels of replicates and thus variances, as described in Table 1 in supplemental section 6.

The R code in the supplemental information vignette provides the method `diffAnovData()` to detect significant features in experiments with both technical and biological replicates, using nested ANOVA. The usage is described in detail in the vignette itself.

3.2.2 *Example for experimental design and trade-off decisions*

Here we provide a discussion of the trade-off decisions using the variance levels we obtained for our analytical platform from the pilot study similar to the discussion of the influence of different sources of variability on the power of a test in Sampson et al. (2013). We use the following simple design as an example: two different sample classes, such as genotype wild-type (WT) and mutant (MT) are used. Using the variance estimation above, we obtain for our analytical setup $\sigma^2_{biol} = 0.172$ and $\sigma^2_{tech} = \sigma^2_{instr} + \sigma^2_{prep} = 0.119$ as the mean biological and technical variances of all features in our sample study.

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

Because the hierarchical *t* test separates the technical and biological variances, we implemented the new method `power.hierarch.ttest()` in the R code in supplemental information vignette for power analysis. Therefore we need the technical variance σ^2_{techn}, the biological variance σ^2_{bio} and number of technical replicates T of each biological replicate, in addition to the parameters of the Student's *t* test power analysis.

With the functions in our implementation, four questions relevant to the experimental design decision can be answered based on the variance estimation obtained in the pilot study. First, we are interested in the minimal number of biological replicates N required to detect differential features with a statistically significant effect of δ and at least a power of $1 - \beta$. Often, a power of more than 0.8 is deemed to be sufficient. If we want to be able to detect an effect of $\delta = 1$ with $M = 4$ measurements (technical replicates of each biological sample) and the observed variances $\sigma^2_{biol} = 0.172$ and $\sigma^2_{tech} = 0.119$ determined above in 3.1.1, a minimal number of $N = 5$ plants from MT and WT each is needed.

The effect δ (also called log fold-change) is the difference between the mean values $\delta = |\mu_1 - \mu_2|$ with μ being the arithmetic mean of the logarithmic data.

Secondly, we want to know how many measurements M of each biological replicate are required to detect differential features with a hierarchical *t* test if only a given number of biological samples are available. For example, $M = 27$ technical replicates are needed if we limit the number of biological replicates to $N = 4$, and leave the other parameters $\alpha, power, \delta, \sigma^2_{biol}$ and σ^2_{tech} as in the previous example. If the number of technical replicates is set to one, then the hierarchical test will reduce to the commonly used non-hierarchical Student's *t* test.

Thirdly, we determine the achievable power for a given number of samples N and measurements M. Common metabolomics experiment designs use e.g. four replicates per population (Böttcher et al. 2009), and two technical replicates are performed. For a given setup, a power of $1 - \beta = 0.69$ can be achieved for $N = 4$ and $M = 2$, so that 69 % of all differential features with a mean difference of $\delta = 1$ can be detected.

Finally, the question arises, which mean differences can be detected if at least 80 % true positive features are demanded. In this example of $N = 4$ samples and $M = 2$ measurements per sample, the real effect $\delta = 1.15$ is statistically significant.

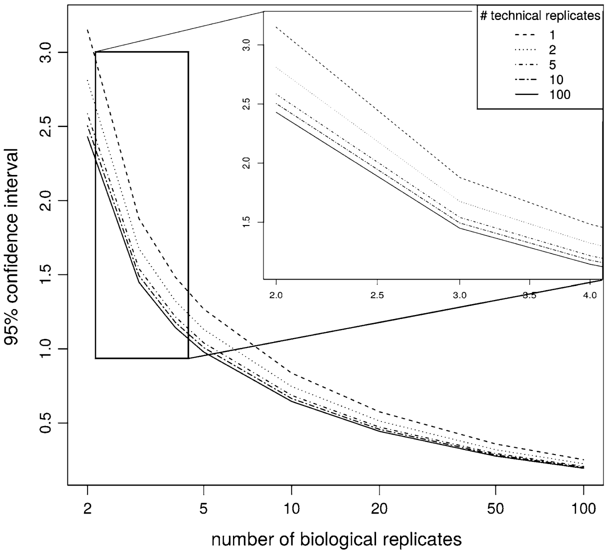


Fig. 4 Comparison of confidence interval sizes of fixed effect for different numbers of biological replicates (*x-axis*) and technical replicates (different line styles). The size of the 95 % confidence interval of the fixed effect (corresponding to $\alpha = 0.05$) is shown on the *y-axis*, assuming the variances obtained in the pilot study ($\sigma^2_{biol} = 0.172$ and $\sigma^2_{tech} = 0.119$)

Figure 4 provides a combined global view of the influence of replication on the achievable confidence interval of the mean difference. The smaller the confidence interval, the less uncertainty is in the results. We calculate the size of the confidence interval of mean differences, for each number of biological replicates $N = 2, \ldots, 100$ and technical replicates $M = 1, 2, 5, 10, 100$ per biological replicate for the type I error probability of $\alpha = 0.05$. The figure shows that the size of confidence interval decreases with increasing number of biological replicates, and also that additional technical replicates improve the results. But the gain from additional technical replicates are much smaller, and the practical effort for large numbers of technical replicates is in general not justified by the increase in detection ability. If the proportion of biological variation on total $\frac{\sigma^2_{bio}}{\sigma^2_{tot}}$, decreases, technical replicates will be more beneficial.

The experimentalists will have to decide whether the increased quality of the test justifies the added costs and the experimental effort when using more replicates. The costs can be interpreted as both actual costs, or as relative costs between biological and technical replicates.

We provide two further methods in the R code in supplemental information vignette to support this decision. First, `supportMat()` can be used to find all possible combinations of biological and technical replicates in a two-level hierarchical experiment design, given the parameters $\alpha, 1 - \beta, \delta, \sigma^2_{biol}$ and σ^2_{tech} and a maximum of possible number of biological and technical replicates. Given a ratio of the costs between biological and technical replicates, the second method `minCostPoss()` chooses the combination which has the lowest costs. This

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

Table 1 Optimal Experiment design with different relative cost ratios and desired effect detection.

$ Biol.	$ Techn.	δ	# Biol.	# Techn.
7	3	1.00	6	1
8	2	1.00	6	1
8	2	1.00	5	2
9	1	2.00	3	1
9	1	1.50	3	2
9	1	1.00	5	2
9	1	0.75	8	2
9	1	0.50	16	2
9	1	0.25	55	3
9	1	0.25	60	2

For a given setup with $\alpha = 0.05, 1 - \beta = 0.8$, the table lists for each cost relation between biological and technical replicates ($ biol., $ techn. in arbitrary units) and a given effect δ the cheapest possibility of number of replicates (# biol. and # techn.). Some cost relations have two designs with minimal costs, see row 2/3 or 9/10

comparison of costs can help to choose an efficient experimental design.

For a given setup with $\alpha = 0.05, 1 - \beta = 0.8$, estimated variances $\sigma^2_{biol} = 0.172, \sigma^2_{tech} = 0.119$ and maximal possible number of $N = 100$ biological and $M = 100$ technical replicates, two examples are given in Table 1: (a) for a minimum effect of $\delta \geq 1.00$ and different cost relations, (b) for a fixed cost relation of 9 : 1, but various minimum effects δ. The table shows the "cheapest" possibility of replicates for each cost ratio between biological and technical replicates and a given $\delta = 1.00$ in rows 1, 2, 3 and 6. Biological replicates will be more expensive than technical replicates until a ratio of 7:3, while at 8:2 the two choices 5 biological and 2 technical replicates or 6 biological replicates without technical replication deliver the same test quality at the same costs. For a fixed cost ratio, the dependency between different effect sizes and replicates can be compared. Table 1 shows for several effects the number of technical and biological replicates required to expect 80 % true differential features and only 5 % false positives at a cost ratio of 9 : 1 in rows 4–10. For a real effect of $\delta = 1.5$ or below, technical replicates and the hierarchical *t* test are superior (i.e. cheaper) than a normal *t* test without technical replication.

4 Conclusion

In mass spectrometry-based metabolomics there are several sources of variance. Based on a pilot study, we have shown that the hierarchical variance analysis is a method to quantify and separate these additive sources of variances. Such a pilot study is also a tool to determine the different sources of variance relative to the overall observed variance in an MS experiment and should be performed for each analytical setup and each organism or tissue type. Our proposed pilot study design is the most efficient to determine these variances. In our setup we found that the biological variance is larger than both the instrumental and preparation variance combined.

The statistical power depends on (1) the observed variance, and (2) the number of biological replicates and (3) the real effect that is relevant for the biological question and which is desired to be statistically significant. To decrease the influence of non-biological variance, technical replicates can be acquired and analysed with a hierarchical type of Student's *t* test, or having more than two classes with nested ANOVA, or in general with multilevel mixed models. In the supplemental material we have shown that the naïve use of a Student's *t* test for both technical and biological replicates yields false positives due to an overestimation of the degrees of freedom. In scientific publications it is thus very important to clearly report the structure of the experiment, and whether samples are independent. This includes the types of replicates, to avoid that "pseudo replicates" are used. Only with such information it is possible to select the appropriate test statistics.

For large studies following the pilot experiment, an optimal experiment design is highly desired to save costs and effort, while maintaining a desired level of statistical power. We have shown how different cost ratios between technical and biological replicates can affect the overall design. It should be noted that costs reflect both the monetary as well as human and infrastructure resources required to perform the experiment.

We provide the R code for the estimation of variances and the calculation of costs and benefit (in terms of statistical power) under the GPL license to support researchers in the design of experiments.

Conflict of interest The authors declare that they have no conflict of interest.

Compliance with ethical requirements This article does not contain any studies with human or animal subjects.

References

Ahrens, Heinz. (1967). *Varianzanalyse*. Berlin: Akademieverlag WTB.

Baldi, P., & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t test and statistical inferences of gene changes. *Bioinformatics*, *17*(6), 509–519.

Böttcher, C., von Roepenack-Lahaye, E., & Scheel, D. (2011) Genetics and genomics of the Brassicaceae, crops and models (

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

Vol XII). In: Resources for metabolomics (p. 677). New York: Springer

Böttcher, C., Westphal, L., Schmotz, C., Prade, E., Scheel, D., & Glawischnig, E. (2009). The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana. *The Plant Cell Online*, *21*(6), 1830–1845.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding fals discoveries in metabolomics and related experiments. *Metabolomics*, *2*(2):171–196.

Danielsson, A. P. H., Moritz, T., Mulder, H., & Spegel, P. (2012). Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics*, *8*, 50–63.

Davis, C. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.

Donner, A. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, *49*(4), 435–439.

Dreyhaupt, Jens., Sufeida, Sabrina., & Muche, Rainer. Power- und Fallzahlabschätzungen für hierarchische und longitudinale Studien. In 17. Konferenz der SAS-Anwender in Forschung und Entwicklung. KSFE e.V., 03 (2013).

Dunn, W. B. (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, *5*(1), 011001. (24pp).

Dunn, W., Erban, A., Weber, R., Creek, D., Brown, M., Breitling, R., et al. (2013). Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, *9*, 44–66. doi:10.1007/s11306-012-0434-4.

Eliasson, M., Rännar, S., Madsen, R., Donten, M. A., Marsden-Edwards, E., Moritz, T., et al. (2012). Strategy for optimizing LC-MS data processing in metabolomics: A design of experiments approach. *Analytical Chemistry*, *84*(15), 6869–6876.

Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., et al. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, *3*(3), 231–241.

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., et al. (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, *41*(Database issue), D781–D786.

Hendriks, M. M. W. B., van Eeuwijk, F. A., Jellema, R. H., Westerhuis, J. A., Reijmers, T. H., Hoefsloot, H. C. J., et al. (2011). Data-processing strategies for metabolomics studies. *Trends in Analytical Chemistry*, *30*(10), 1685–1698.

Holmes, T. H. (2004). Ten categories of statistical errors: A guide for research in endocrinology and metabolism. *American Journal of Physiology–Endocrinology and Metabolism*, *286*(4), E495–E501.

Horgan, G. W. (2007). Sample size and replication in 2D gel electrophoresis studies. *Journal of Proteome Research*, *6*(7), 2884–2887.

Johnson, H. E., Lloyd, A. J., Mur, L. A., Smith, A. R., & Causton, D. R. (2007). The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. *Metabolomics*, *3*, 517–530.

Karp, N. A., Spencer, M., Lindsay, H., O'Dell, K., & Lilley, K. S. (2005). Impact of replicate types on proteomic expression analysis. *Journal of Proteome Research*, *4*(5), 1867–1871.

Lönnstedt, I., & Speed, T. (2001). Replicated microarray data. *Statistica Sinica*, *12*, 31–46.

Pavlidis, P., Li, Q., & Stafford, N. W. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics*, *19*(13), 1620–1627.

Pinheiro, J. C., & Bates, D. (2014). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: SAGE.

Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2013). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 1–14.

Sampson, J. N., Boca, S. M., Shu, X. O., Stolzenberg-Solomon, R. Z., Matthews, C. E., Hsing, A. W., et al. (2013). Metabolomics in epidemiology: Sources of variability in metabolite measurements and implications. *Cancer Epidemiology Biomarkers & Prevention*, *22*(4), 631–640.

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.

Snijders, T. A. B. (2001). Sampling, Chapter 11. In A. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 159–174). Longford: Wiley.

Snijders, Tom A. B., & Snijders, T. A. (2005). Power and sample size in multilevel linear models. *Encyclopedia of Statistics in Behavioral Science*, *3*, 1570–1573.

Student, (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Tutz, G., Fahrmeir, L., & Hamerle, A. (1996). *Multivariate statistische verfahren*. Berlin: Walter de Gryuter.

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J., & Yanes, O. (2012). A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites*, *2*(4), 775–795.

von Roepenack-Lahaye, E., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U., Wessjohann, L., et al. (2004). Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiology*, *134*(2), 548–559.

DOI:10.1007/s11306-014-0742-y Reprinted with permission. © 2015 Springer Science+Business Media

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Natural variation of root exudates in *Arabidopsis thaliana*-linking metabolomic and genomic data

Received: 14 February 2016
Accepted: 14 June 2016
Published: 01 July 2016

Susann Mönchgesang[*], Nadine Strehmel[*], Stephan Schmidt[*], Lore Westphal, Franziska Taruttis[†], Erik Müller, Siska Herklotz, Steffen Neumann & Dierk Scheel

Many metabolomics studies focus on aboveground parts of the plant, while metabolism within roots and the chemical composition of the rhizosphere, as influenced by exudation, are not deeply investigated. In this study, we analysed exudate metabolic patterns of *Arabidopsis thaliana* and their variation in genetically diverse accessions. For this project, we used the 19 parental accessions of the Arabidopsis MAGIC collection. Plants were grown in a hydroponic system, their exudates were harvested before bolting and subjected to UPLC/ESI-QTOF-MS analysis. Metabolite profiles were analysed together with the genome sequence information. Our study uncovered distinct metabolite profiles for root exudates of the 19 accessions. Hierarchical clustering revealed similarities in the exudate metabolite profiles, which were partly reflected by the genetic distances. An association of metabolite absence with nonsense mutations was detected for the biosynthetic pathways of an indolic glucosinolate hydrolysis product, a hydroxycinnamic acid amine and a flavonoid triglycoside. Consequently, a direct link between metabolic phenotype and genotype was detected without using segregating populations. Moreover, genomics can help to identify biosynthetic enzymes in metabolomics experiments. Our study elucidates the chemical composition of the rhizosphere and its natural variation in *A. thaliana*, which is important for the attraction and shaping of microbial communities.

In *Arabidopsis thaliana (A. thaliana)*, natural genetic variation has been intensively exploited to study a variety of traits related to plant development, stress response and nutrient content (for review, see Weigel[1]). Several publications have demonstrated that natural variation is a suitable basis for dissecting secondary metabolite pathways by using genetic mapping analyses. The genetics of glucosinolates and its link to pathogen and herbivore resistance have been investigated thoroughly[2–5]. A large variation of glucosinolates in leaves and seeds was observed for 39 genetically diverse Arabidopsis accessions[6]. Houshyani *et al.*[7] found that natural variation of the general metabolic response to different environmental conditions is not necessarily associated with the genetic similarity between nine accessions.

Many metabolomics studies focus on aboveground plant tissues. As a result, only limited information is available with regard to the metabolism of belowground parts of the plant.

Roots are crucial for the uptake of water and nutrients. For example, Agrawal *et al.*[8] utilized natural variation of *A. thaliana* to identify malic acid as a key mediator for nickel tolerance. To communicate with the belowground environment, plant roots also exude metabolites such as flavonoids, phenylpropanoids and glucosinolates[9], which can attract microorganisms or increase the resistance against pathogens[9–11]. These interactions take place in the rhizosphere, which is regarded as the space adjacent to roots[12]. As the properties of the rhizosphere differ strongly from the bulk soil in terms of microorganism abundance[13], as well as the qualitative and quantitative metabolic composition[14,15], investigations on root exudates are needed to assess the role of this microenvironment. Micallef *et al.*[16] demonstrated that the rhizobacterial community composition is influenced by varying exudation profiles.

Non-targeted metabolite profiling of secondary metabolites by liquid chromatography coupled to mass spectrometry (LC/MS) is an ideal analytical platform to link natural metabolite variation to biosynthetic pathways. It allows for the detection and quantification of semipolar compounds[17], when the resulting three-dimensional

Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany. [†]Present address: University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to D.S. (email: dierk.scheel@ipb-halle.de)

DOI:10.1038/srep29033

signals with a specific mass-to-charge (*m/z*) ratio, retention time (RT) and intensity, so-called features, can be annotated. Depending on the nature of the compound, they are more likely to be detected upon electrospray ionization in the positive (ESI(+)) or negative mode (ESI(−)).

Our approach to investigate natural genetic variation of secondary metabolism in root exudates focuses on 19 *A. thaliana* accessions, which show a large degree of geographic and phenotypic diversity (Supplementary Table S1) and were used to generate the Multiparent Advanced Generation Inter-Cross (MAGIC) lines[18]. Whole genome sequencing revealed that the parental accessions and the MAGIC lines represent most of genetic variability of *A. thaliana* and therefore provide a valuable resource for genetic and metabolic studies[19,20].

The aim of this study is to find out if the root exudate composition in *A. thaliana* is genetically determined. For this purpose, we analysed which metabolites show natural variation, if similar metabolic phenotypes share a genetic base, in particular, if certain characteristics can be traced back to single nucleotide polymorphisms and hence, directly link phenotype and genotype.

Results

Non-targeted metabolite profiling of root exudates reveals distinct metabolic phenotypes for 19 Arabidopsis accessions. A clustering analysis was performed to find similarities between the metabolic profiles and sequence polymorphisms of the 19 founder accessions of the MAGIC population of *A. thaliana*. The dendrograms calculated from the metabolic features show a clear separation of accessions in Fig. 1a for exudates measured in ESI(−) and Fig. 1b in ESI(+). At a correlation threshold of 0.95 (dashed line), seven and five clusters, respectively, were observed.

No-0 and Po-0 (blue) were found in the same cluster (cluster 1, ESI(−); cluster 5 ESI(+)) in both ion modes. Ct-1 and Edi-0 (purple) also displayed high similarity in their metabolic profiles. Sf-2 and Kn-0 (green) were in close proximity and would have been in the same clade when cutting the ESI(+) dendrogram at a different threshold. Similar metabolic phenotypes were also detected in the exudation patterns of Wu-0 and Tsu-0, and additionally Mt-0 (orange). These three accessions either clustered in dendrogram branch 2 (ESI(−)) or 3 (ESI(+)).

In both metabolic dendrograms, one Oy-0 sample was observed as an outlier, which did not cluster with the other replicates of Oy-0. For Hi-0 and Ws-0, mixed clusters were observed. The positive ion mode generally harboured more outliers. As obvious from the quality control plots in Supplementary Fig. S1, the outlying samples did not show any extreme deviations on the technical side and were therefore not excluded from further analysis[21].

For the analysis of genetic diversity, sequence polymorphisms in coding sequences (CDS) extracted from the 19 genomes project[22] were used for a genetic clustering (Fig. 1c). One large dendrogram branch (L*er*-0, Kn-0, Wil-2; Ws-0, Ct-1, No-0; Hi-0, Tsu-0, Mt-0, Wu-0, Col-0, Rsch-4) had less than 825,000 mismatches (dashed line) while the outliers Bur-0, Sf-2, and Can-0 had increasing numbers of polymorphisms. Oy-0 and Po-0 formed a small cluster and were found in proximity to Edi-0, Zu-0 and the large dendrogram branch.

The metabolic analysis was based on a non-targeted metabolite profiling approach considering metabolic features characterised only by their *m/z* ratios, RTs and intensities. These characteristics are not sufficient to investigate the underlying molecules, its biosynthetic pathway and its potential in plant signaling. Annotations and identifications of metabolites, as shown in the next paragraph, are required to interpret non-targeted metabolic profiles in the biological context.

Semipolar secondary metabolites are the major components of the exudation patterns. Only 25 and 22 of the metabolic signals (455 (ESI(−)), 475 (ESI(+), respectively) could be assigned to metabolites which have been previously described as exudate-characteristic for Col-0[15]. Differential metabolites were detected by a generalized Welch-test between the 19 accessions; their colour-coded intensity map is shown in Fig. 2. Chemically related compounds were placed in groups separated by horizontal spacing.

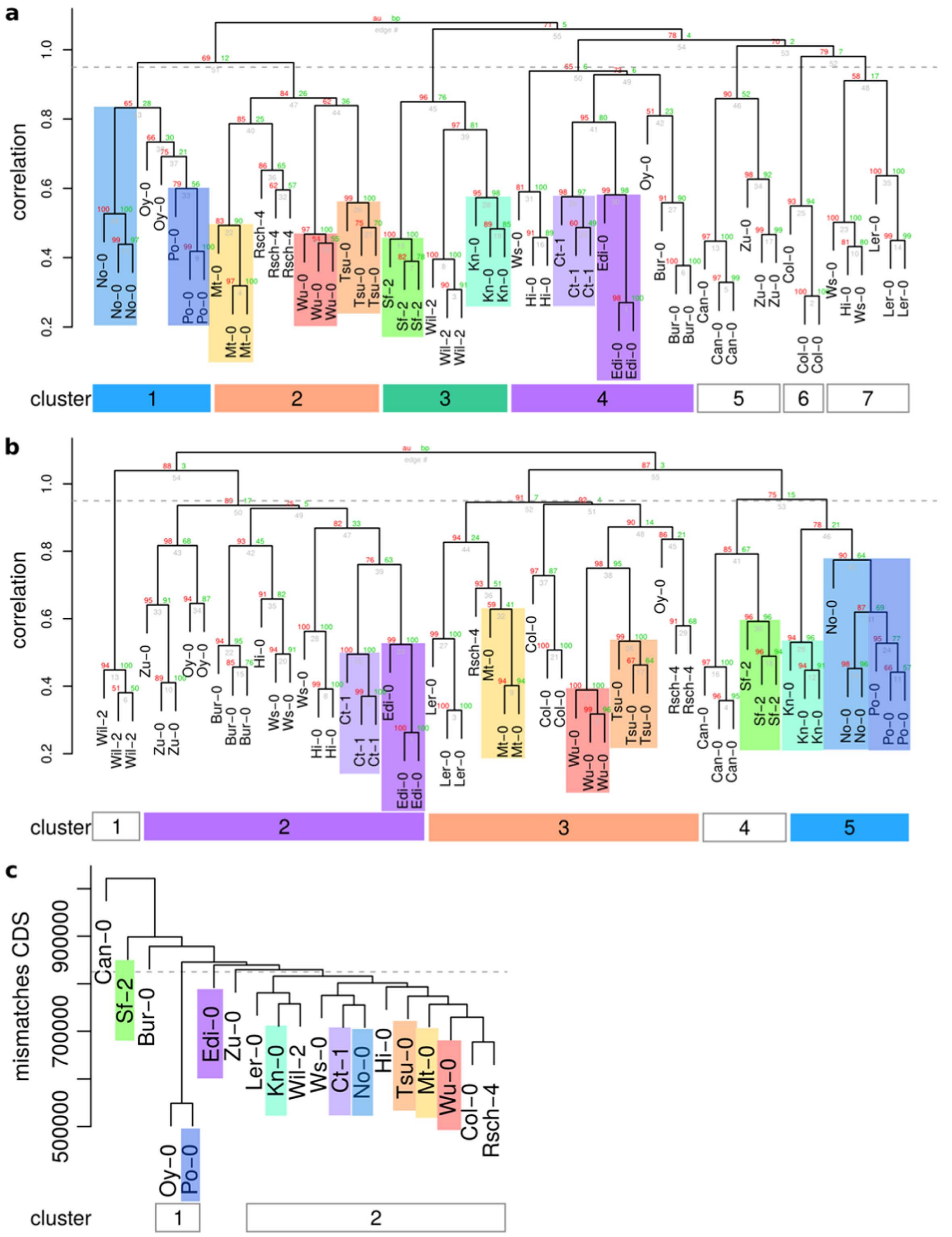
Among the differential metabolites, there were several compounds with an aromatic moiety, such as the nucleoside thymidine and the amino acids Phe and Tyr. The amino acid derivative hexahomo-Met *S*-oxide had low abundance in the exudates of Sf-2 and was enriched in Mt-0.

A range of glucosinolate degradation products was characteristic for the exudates of some accessions. Edi-0 had rather low levels of indolic compounds and the isothiocyanate hydrolysis product of 8-MeSO-Octyl glucosinolate. Wu-0 showed a clear absence of the neoglucobrassicin (1-MeO-I3M) hydrolysis product 1-methoxy-indole-3-ylmethylamine (1-MeO-I3CH_2NH_2), while Sf-2 was missing the malonyl-glucoside of 6-hydroxyindole-3-carboxylic acid (6-(Malonyl-GlcO)-I3CH_2CO_2H). An unknown indole derivative ($C_{10}H_9NO_3$) was highly abundant in the exudates of Ct-1 and Wil-2, and lowly abundant in Sf-2. Generally, large amounts of the glucosinolate precursor and hydrolysis products were detected in the exudates of L*er*-0, Mt-0 and Wil-2.

Plant hormone-derived metabolites also differed between the 19 accessions. Two salicylic acid (SA) catabolites, 2,3 and 2,5-dihydroxybenzoic acid (DHBA) pentosides, were highly abundant in Col-0, Kn-0, L*er*-0, Mt-0, Wil-2, Ws-0 and Wu-0. No preference for the 3′ or 5′ hydroxylated variant of DHBA was noticed, and both isomers correlated positively with a Pearson correlation of 0.91. 9,10-dihydrohydroxy jasmonic acid (JA) *O*-sulfate was another differential plant hormone catabolite in *A. thaliana* exudates with low levels in Bur-0, Can-0 and Zu-0 and high levels in Col-0, Kn-0, Po-0, Rsch-4 and Wu-0.

Among the phenylpropanoids, the coumarin scopoletin and its glycosides differed in the exudates of the 19 accessions. A hexose-pentose conjugate of scopoletin as well as three other glycosides ($C_4H_{10}O$ Hex-DeoxyHex, $C_{12}H_{16}O_5$ Hex, $C_7H_{14}O_4$ Malonyl-Hex) were among the differentially abundant metabolites which were described for Col-0 exudates[15].

DOI:10.1038/srep29033



4

Figure 1. Hierarchical clustering of metabolic features from (**a**) exudates ESI(−), (**b**) ESI(+) and of (**c**) genetic distances. (**a+b**) Features were obtained by UPLC/ESI(−)-QTOF-MS (a) or UPLC/ESI(+)-QTOF-MS (**b**) from exudate samples and differed from the blank (Welch test, $p < 0.05$). Intensities were corrected for batch effects using SVA and subjected to average linkage clustering with correlation as a distance measure. (**c**) Variant tables of the 19 genomes project were reduced to coding regions, as annotated by TAIR. The sum of all mismatches was used as a distance matrix for average linkage clustering. Dendrograms were cut at a correlation threshold of 0.95 (dashed line). As cluster numbers were not comparable, consistent clusters were coloured across ion modes as a visual guidance.

Other differential phenylpropanoids include the monolignol glucoside syringin as well as both isomers of the sulfated dilignol G(8-*O*-4)FA *O*-sulfate consisting of coniferyl alcohol (G) and ferulic acid (FA): it was present at high levels in Kn-0 and Wil-2 exudates. Two hydroxylated fatty acids also showed natural variation and were highly abundant in Mt-0.

Several isoforms of known glycosylated metabolites (e.g. kaempferol triglycosides with *m/z* 739.21) were detected at different RTs indicating differences in sugar conjugation. The investigation of these putatively annotated metabolites can be facilitated by exploring polymorphisms in genes encoding their biosynthetic enzymes.

DOI:10.1038/srep29033

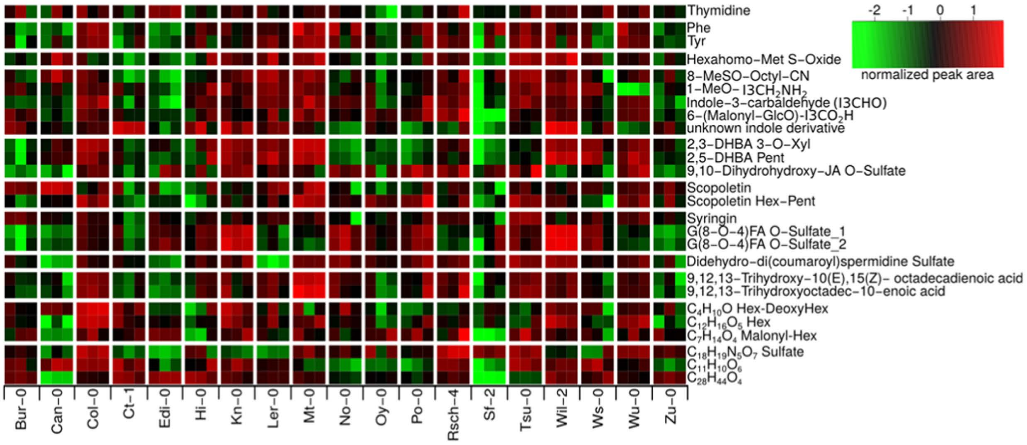


Figure 2. Colour-coded intensity matrix of differential metabolites occurring in exudates. Integrated peak areas were log-transformed and scaled to zero mean and standard variance. A Welch-test was used to find differentially abundant metabolites between the 19 accessions.

The absence of an indolic glucosinolate hydrolysis product and a hydroxycinnamic acid conjugate is genetically determined. Wiesner *et al.*[23] reported that the accession Wu-0 lacks the 1′-methoxylated indolic glucosinolate due to a premature stop codon in the *CYP81F4* gene[24]. Its frameshift mutation leads to a loss of function and subsequently to the absence of 1-MeO-I3M in roots and leaves[23], and also its amine, 1-MeO-I3CH$_2$NH$_2$, in the exudates of our hydroponic system.

To elucidate if further metabolite absences in the exudates like 1-MeO-I3CH$_2$NH$_2$ in Wu-0 can be traced back to a single gene, we developed a workflow to link genomic and metabolic patterns (Fig. 3). Features with the same absence pattern could be different molecular species of the same compound (adducts, isotopes, fragment or cluster ions). Alternatively, they may be different isomers from the same biosynthetic pathway with a common precursor.

Among the seven metabolic features with absence in two accessions, three were characteristic for Can-0 and *Ler*-0. The hydroxycinnamic acid polyamine derivative cyclic didehydro-di(coumaroyl)spermidine sulfate previously identified in Col-0[15] and also detected in other accessions was clearly absent in Can-0 and *Ler*-0 (Fig. 2). This compound with RT = 3.6 min was absent in the negative ion mode as [M-H]$^-$ adduct with *m/z* = 514.17 and [M-2H + Na + CH$_2$O$_2$]$^-$ adduct with *m/z* = 582.15. Another compound with *m/z* = 514.17 eluting at 4.2 min was also absent in Can-0 and *Ler*-0. Tandem mass spectrometry (MS/MS) analysis revealed a sulfur trioxide loss in the fragmentation pattern similar to the sulfated cyclic didehydro-di(coumaroyl)spermidine conjugate. Can-0 carries a premature stop codon in the gene AT2G25150 encoding spermidine dicoumaroyl transferase (SCT), whereas in *Ler*-0, a large deletion is present in the CDS of this gene[22]. Both accessions have no detectable levels of SCT transcript in their roots (Fig. 4a).

Thus, neither Can-0 nor *Ler*-0 possess SCT activity to most likely produce cyclic didehydro-di(coumaroyl) spermidine sulfate and its isomer. To further support the data observed with these two accessions, we analysed the exudates of the homozygous knockout line SALK_098927C (Col-0 background), which indeed did not display any peaks with *m/z* 514.17 ESI(−) at 3.6 min, as shown in Fig. 4b, and thus confirm our hypothesis.

The above results for the Wu-0 and Can-0/*Ler*-0 pattern showed the feasibility of such an association analysis to link compounds to their biosynthetic pathways. In specific cases, there is a direct connection between metabolic phenotype and genotype. Therein, metabolite variation among Arabidopsis accessions can be traced back to individual SNPs without trait segregation and QTL mapping.

Matching metabolic and genetic patterns can indicate compound class. Genetic alterations may be exploited to characterise so far unknown compounds which are part of related biosynthetic pathways[25]. MS/MS fragmentation facilitates the annotation of chemical substructures, which are often characteristic for a certain class of compounds. Knowledge about biosynthetic pathways can further support the assignment of unknown features to compound classes.

For the annotation of metabolites, collision-induced dissociation (CID-) MS was performed for 17 selected MS1 ESI(−) features obtained by the above described screening.

With the help of MS/MS spectra, nine out of 17 features were annotated and for five further features, the elemental composition was determined. An overview of compounds, fragment spectra and matching enzymes is given in Supplementary Table S5.

A compound (*m/z* 739.21, RT = 4.3 min) that was not found in the exudates of Wu-0 (Fig. 5a) was identified as a flavonoid with the same elemental composition ($C_{33}H_{40}H_{19}$) and fragment spectrum as kaempferol 3-*O*-Rha(1→2)Glc 7-*O*-Rha[15]. The RT shift indicates different glycosidic conjugation. This compound was identified as robinin (kaempferol 3-*O*-Rha-Gal 7-*O*-Rha) by an authentic standard having a galactose moiety instead of glucose in the diglycoside at the 3′ position (Fig. 5b). One out of the 16 premature stop codons characteristic for Wu-0 was present in AT2G22590.1, which encodes the UDP-glycosyltransferase (UGT) superfamily protein

DOI:10.1038/srep29033

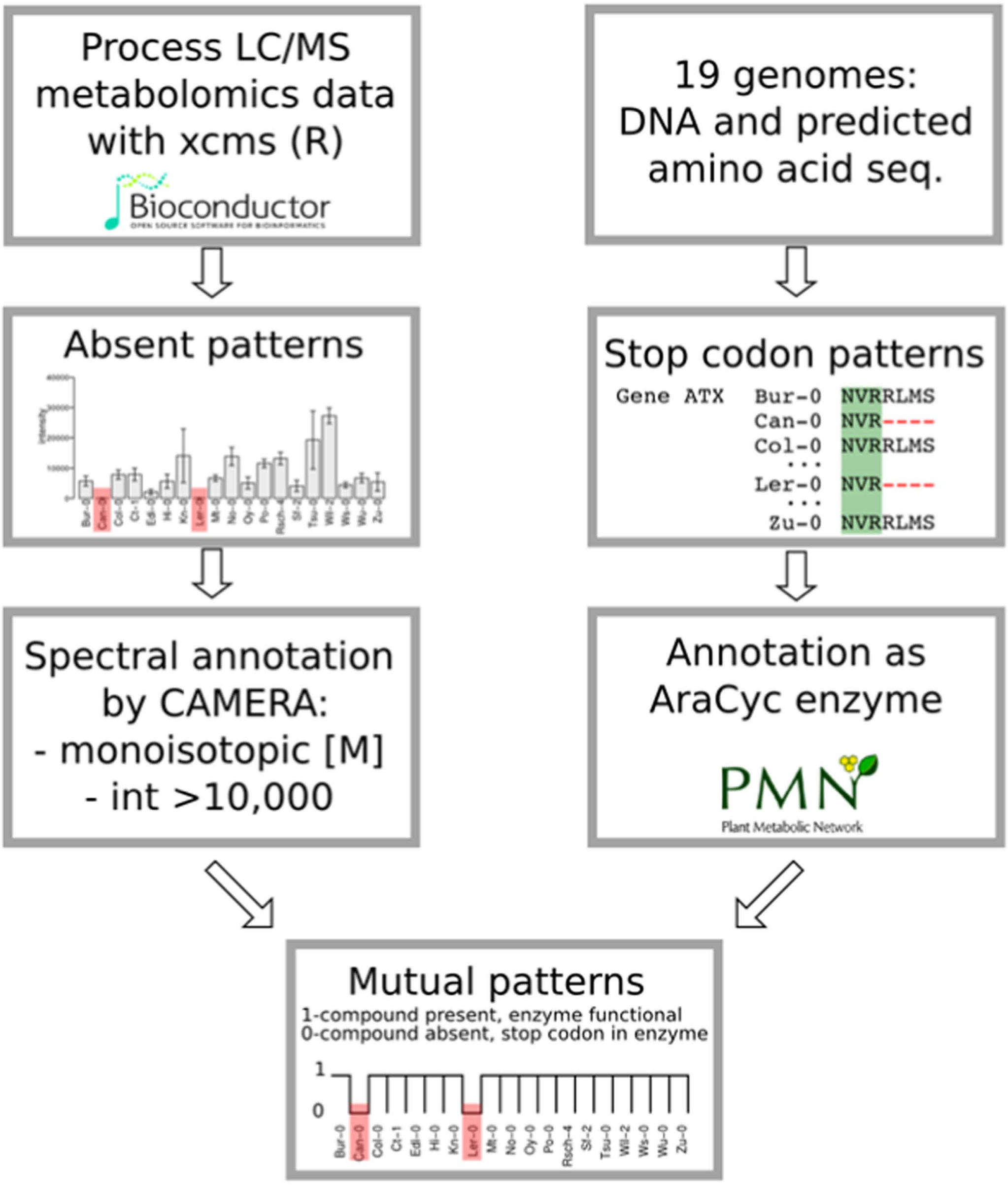


Figure 3. Workflow for matching metabolic patterns of absence with stop codons in genes annotated as AraCyc enzymes. For the metabolic data, 384 out of 455 metabolic features from the ESI(−) data set were absent in at least one accession. 38 of them were annotated as monoisotopic peak [M] by CAMERA. Approximately 32,000 stop codons were detected. 1,588 of AraCyc enzyme-encoding genes displayed a prematurely ended amino acid sequence possibly representing non-functional enzymes that can be causative for metabolite absence.

UGT91A1. This gene is coexpressed with the flavonol synthase 1 (FLS1, AT5G08640) and chalcone flavanone isomerase (TT5, AT3G55120) encoding genes that are annotated with the "flavonoid biosynthetic process" by Gene Ontology[26]. The exudates of the homozygous knockout line SALK_088702C (Col-0 background) were missing robinin and its UGT91A1 transcript levels in roots were diminished (Fig. 5c–e).

The hydroxylated fatty acid 9,12,13-trihydroxyoctadec-10-enoic acid (9,12,13-TriHOME, KEGG C14833) was not present in the exudates of Edi-0 and Zu-0 (Fig. 2). Its lack corresponds to a SNP pattern introducing a stop codon into a long-chain-alcohol *O*-fatty-acyltransferase gene (AT5G55360.1). The unsaturated variant 9,12,13-tri hydroxyoctadec-10(E),15(Z)-enoic acid, however, could be detected in Edi-0 and Zu-0 exudates, but not in the Ct-1 accession, and accordingly, pointed to different polymorphism patterns. Besides, similar intensity distributions of both hydroxylated fatty acids were found across the exudates of the 19 accessions (Fig. 2).

These examples show that the direct search for a metabolite-enzyme-connection can provide valuable insights into biosynthetic pathways but require careful examination of the resulting candidate genes.

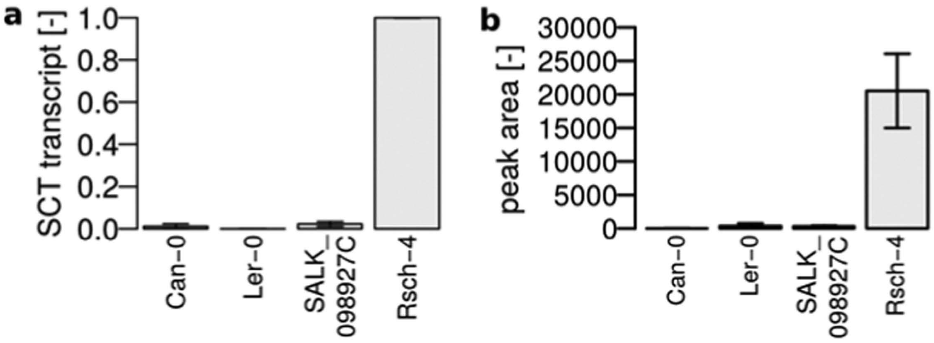


Figure 4. Natural and T-DNA insertion knockouts of SCT. (**a**) Relative transcript levels of SCT in root tissue as determined by qPCR, PP2A as reference, normalized to Rsch-4, mean ± s.e.m., n = 3. (**b**) Peak area counts of cyclic didehydro-di(coumaroyl)spermidine sulfate in exudates, mean ± s.e.m., n = 3.

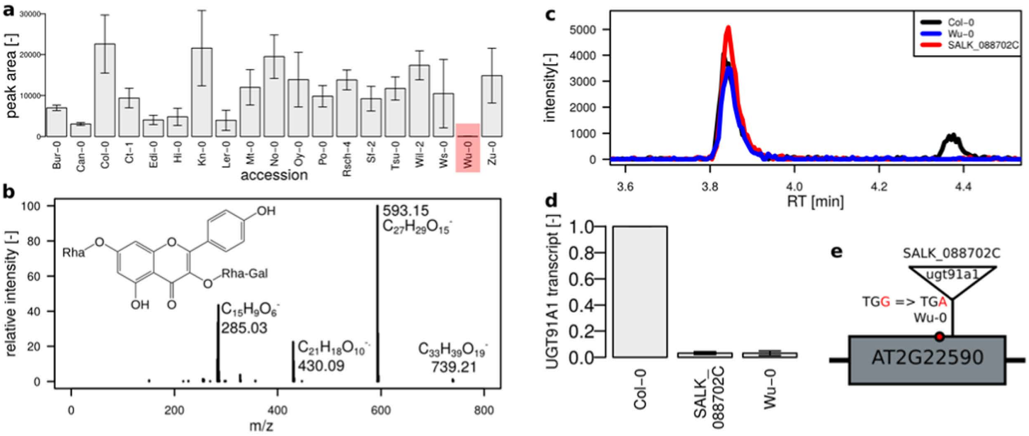


Figure 5. Robinin absence is linked to a stop codon in the UGT91A1 encoding gene. (**a**) Peak area counts, mean ± s.e.m. (n = 3) with absence in Wu-0 (highlighted in red) (**b**) MS/MS spectrum of robinin, 30 eV, (**c**) extracted ion chromatogram at *m/z* 739.21 with kaempferol 3-*O*-Rha(1→2)Glc 7-*O*-Rha eluting at 3.9 min and the galactose-conjugated robinin eluting at 4.3 min not detected in the natural knockout Wu-0 and T-DNA insertion line SALK_088702C, (**d**) relative transcript levels of UGT91A1 in roots as determined by qPCR, PP2A as reference, normalized to Col-0, mean ± s.e.m., n = 4, (**e**) schematic representation of the UGT91A1 gene (one exon) and the loss-of-function mutations in Wu-0 and SALK_088702C.

Discussion

This study showed how the exudation pattern of *A. thaliana* accessions is reflected by a genetic clustering of polymorphisms in their CDS. The previously reported similarity of the German and Norwegian accession Po-0 and Oy-0[22] was only observable at metabolic level in the ESI(−) dendrogram. The close relation was confirmed by the genetic clustering. However, we also observed closely related metabolic profiles of Po-0 with No-0 (blue), which has not been described before. Neither the metabolic proximity of Sf-2 and Kn-0 (green) nor of Ct-1 and Edi-0 (purple) were reflected by small genetic distances.

The similarity of the Wu-0, Tsu-0 and Mt-0 was present in both ESI dendrograms of the exudate analysis and seems to be genetically determined. The close genetic relation between the Japanese accession Tsu-0 and Mt-0 from Libya has already been reported by Nordborg *et al.*[19] as well as by De Pessemier *et al.*[27], and was confirmed for metabolic exudate and the CDS profiles (orange).

The clustering of metabolic profiles demonstrated that genetic variation between the 19 founder accessions of the Arabidopsis MAGIC population is indeed reflected in the exudate metabolome. This is in contrast to the previously reported only minor correlation between shoot metabolic and genetic similarity[7] of nine accessions, partially overlapping with the MAGIC founder lines. Compared to 149 SNPs that were used to estimate a genetic distance by Houshyani *et al.*[7], our analysis included 640,066 polymorphisms that were exclusively within CDS. The usage of SNPs in CDS ensures a comprehensive, but most direct genotype-phenotype-association, disregarding regulatory sequences. From hierarchical clustering, we can conclude that the three dendrograms reflect the genetic determination of the exudation profile of several Arabidopsis accessions. Both, the genetic and thus the metabolic profiles, may have been affected by selection processes at the collection sites[25]. Information on

DOI:10.1038/srep29033

environmental conditions, especially characteristic rhizosphere data of the original locations, would be of great interest, but unfortunately, these are not well documented[28].

In our study, a variety of glycosylated and sulfated compounds are the key metabolites that underlie natural variation in the exudates of the MAGIC parental lines. Scopoletin was found both as an aglycone and hexose-pentose conjugate. However, glucosinolates were only detected as degradation products (amines, carbaldehydes, isothiocyanates). Currently, we cannot elucidate whether glucosinolate exudation is initiated by myrosinase activation or if hydrolysis was caused by the sample preparation procedure.

Previously, hormones were described as constituents of root exudates[29]. Despite that, plant hormones were difficult to detect with the analytical method due to their low abundance. Plant hormone-derived metabolites were detected as glycosylated and sulfated in case of SA and JA, respectively. Natural variation is reflected by a great spectrum of glycosidic conjugation. This was shown for SA catabolites. SA was present in the exudates of Col-0 in the study of Strehmel *et al.*[15] but did not pass their stringent filtering criteria to be included in their exudate compound collection, while SA derivatives with 2,3 or 2,5- dihydroxy-substituted benzoic acid pentose conjugates passed the filter. As shown in Supplementary Fig. S2, high amounts of SA were found in Kn-0, Wil-2 and Wu-0, the lowest amount was present in Sf-2 exudates, one of the accessions with also low DHBA pentoside levels. Interestingly, solely pentosides but no hexosides of DHBA were detected in the root exudates of Col-0[15]. Li *et al.*[30] investigated the discrimination of hexose and pentose conjugation in 96 *A. thaliana* accessions. Combined QTL and association mapping pointed to a locus on chromosome 5 within proximity of a gene encoding a putative UGT with pentose specificity. The findings of this study support the previously reported low ratio of pentose-hexose conjugates for Edi-0[30]. Sf-2 was the accession with the lowest DHBA pentoside-hexoside ratio, which may be caused by a non-functional pentose-conjugating UGT and a background hexose-transferase activity that leads to a DHBA hexoside phenotype.

Chemically related compounds often derive from the same biosynthetic pathway. The characterisation of these metabolites might be facilitated by combining metabolic patterns with genomic data. Thus, an analysis workflow was developed which compares metabolite and sequence polymorphism patterns. In order to reduce the complexity, qualitative metabolic patterns were extracted and compared with the presence of premature stop codons in enzyme-encoding genes. The absence of a sulfated cyclic di(dehydrocoumaroyl)-spermidine was traced back to a single genomic alteration diminishing SCT activity in Can-0 and L*er*-0. These data support the hypothesis postulated by Strehmel *et al.*[15] that the cyclic conjugate is derived from di(coumaroyl)spermidine synthesized from spermidine and coumaroyl-CoA by SCT as illustrated in Fig. 6. A subsequent oxidative ring formation and sulfonylation led to sulfated cyclic di(dehydrocoumaryol)-spermidine[31]. Nevertheless, the coumaroyl spermidine transferase activity can hardly be inferred from the gene annotation as "HXXD-type acyl transferase family protein". This workflow furthermore pointed towards the substrate specificity of UGT91A1. Previous studies have shown that UGT91A1 is regulated by MYB transcription factors and speculated about its involvement in glycosylation of flavonols or flavonol glycosides[32]. We could show that in the absence of UGT91A1 enzymatic activity no galactose transfer to kaempferol 3-*O*-Rha 7-*O*-Rha (kaempferitrin) is catalysed to produce robinin. However, the presence of the glucose-substituted isomer kaempferol 3-*O*-Rha(1→2)Glc 7-*O*-Rha implies the involvement of a different UGT not accepting galactose but rather glucose as a substrate. We hereby found that UGT91A1 might have similar flavonoid substrate specificity as UGT73C6 and UGT78D1[33]. However, the patterns of two closely related hydroxylated fatty acids did not show mutual absences. Their intensity distributions were similar and point out the threshold issue in the absence definition. The SNP in AT5G55360 is likely to be a false positive candidate that needs to be excluded by a careful interpretation.

Future investigations will focus on the refinement of our approach by addressing the following points: i) When is a peak defined as absent? We relied on the decision of the peak-picking method centWave[34] in the xcms package[35]. If the algorithm found a peak at a particular *m/z* and RT in one accession but could erroneously not match its peak criterion in any replicates of another accession, the peak was defined as absent. ii) For a proof of concept, our workflow only included nonsense mutations in CDS of single genes. More complex studies would include amino acid exchanges in CDS, alterations in promoter regions as well as cases of gene function redundancies.

Linking stop codons with metabolite absences helps with the elucidation of secondary metabolite pathways but still requires fragment spectra to be interpreted manually and gene annotations have to be carefully checked for a possible involvement within the biosynthetic pathway of the metabolite. The connection has to be validated by knockout lines of the respective candidate genes.

Our study revealed natural variation in the root exudate composition of 19 genetically diverse accessions of *A. thaliana*. Combining nonsense mutations with metabolic patterns of the exudates facilitated to determine the genetic base of specific metabolite absences. Furthermore, the integration of sequence data can help to identify compound classes in metabolomics experiments. Our study can aid to further unravel biochemical and molecular processes in the rhizosphere by providing a metabolomics resource of root exudates (MetaboLights, accession number MTBLS160, http://www.ebi.ac.uk/metabolights/MTBLS160). Future investigations should aim at correlating metagenomics with exudation profiles in order to deduce characteristics that can be exploited to circumvent limiting abiotic factors and decrease the susceptibility towards biotic stresses.

Methods

Plant material. Seeds of the accessions Bur-0, Col-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, L*er*-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0 of *A. thaliana* (Supplementary Table S1) were obtained from the European Arabidopsis Stock Centre. The T-DNA insertion lines SALK_098927C and SALK_088702C were obtained from the SALK institute and Dr. Ralf Stracke (Bielefeld), respectively, and characterised as elaborated in the Supplementary Methods.

DOI:10.1038/srep29033

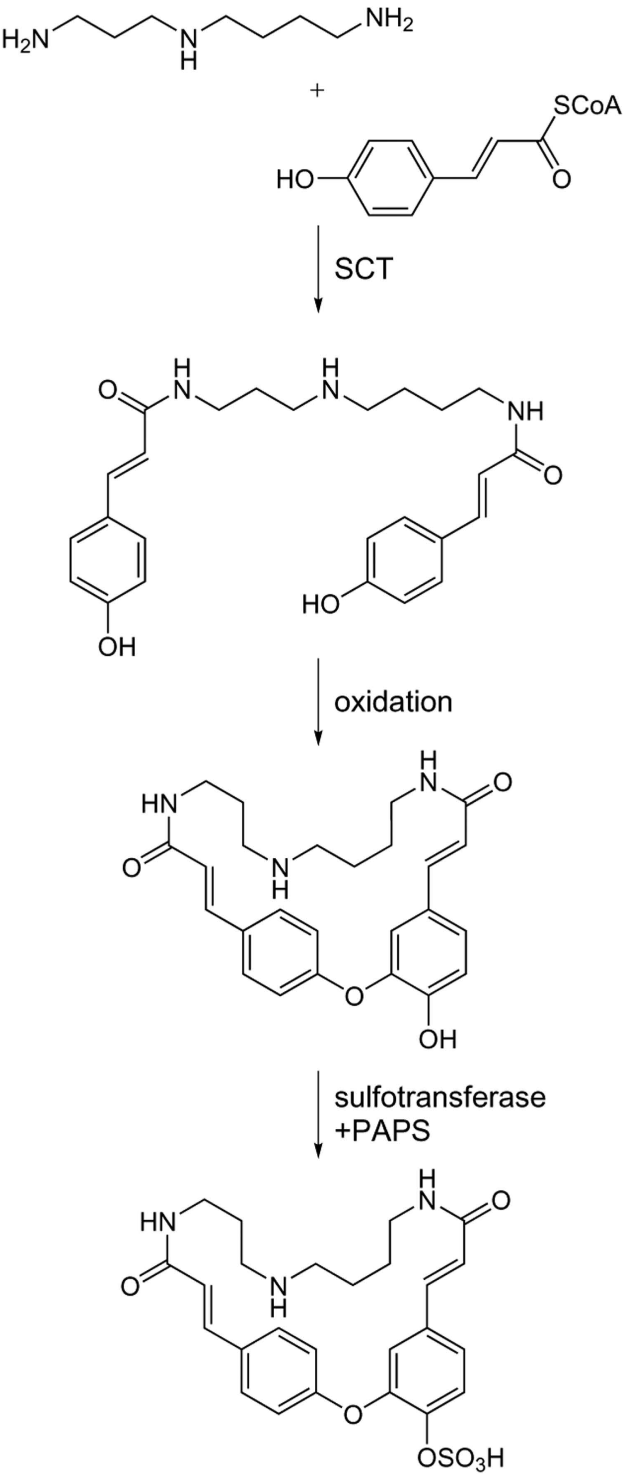


Figure 6. Biosynthetic pathway of cyclic didehydro-di(coumaroyl) spermidine sulfate. Di(coumaroyl) spermidine is synthesized by SCT[47] and subsequent oxidative ring closure and sulfonylation leads to cyclic didehydro-di(coumaroyl) spermidine sulfate, PAPS = 3′-phosphoadenosine-5′-phosphosulfate.

Plant cultivation. All seeds were surface-sterilized prior to plant cultivation. Then, all lines were cultivated in a hydroponic system with three independent biological experiments as previously described[15] and in the Supplement. Culture medium was used as a blank. Medium was collected after one-week-exudation (week 5–6) and resulted in 57 pooled root exudates (of four plants each).

DOI:10.1038/srep29033

Sample preparation. Root exudates were prepared according to Strehmel *et al.*[15] and as described in Supplementary Methods.

Non-targeted metabolite profiling analysis. Changes in metabolism were analysed by ultra-performance liquid chromatography coupled to electrospray ionization quadrupole time–of–flight mass spectrometry (UPLC/ESI-QTOF-MS) according to Böttcher *et al.*[36].

All mass spectra were acquired in centroid mode and recalibrated on the basis of lithium formate cluster ions.

A detailed description of plant cultivation, sample preparation and metabolite profiling can be found in Supplementary Methods.

Data analysis. Raw data files were converted to mzData using CompassXPort version 1.3.10 (Bruker Daltonics 4.0). Subsequently, the R package xcms version 1.41.0[35] was used for feature detection, alignment and filling of missing values. On this account, features were detected with the help of the centWave algorithm according to Tautenhahn *et al.*[34] (snthr = 5, scanrange = c(1,3060), ppm = 20, peak width = c(5,12)), matched across samples (xcms function group, minfrac = 0.75, bw = 2, mzwid = 0.05, max = 50), corrected for retention time shifts (method = "loess") and grouped again. Missing values were imputed with the xcms function fillPeaks which integrates raw chromatographic data. The data matrix was extracted using the diffreport function.

DataAnalysis 4.0 (Bruker Daltonics) was used for generation of extracted ion chromatograms, deconvolution of compound mass spectra and calculation of elemental compositions. For relative quantification of compounds extracted ion chromatograms from the non-targeted analysis were integrated with QuantAnalysis 2.0 (Bruker Daltonics) using the quantifier ions as listed in Supplementary Table S3. Peak areas were log-transformed and z-scaled to achieve normal distribution. Differential metabolites were detected by a generalized Welch-test between the 19 accessions (unequal variances, one-way layout, $p < 0.05$, corrected for multiple testing by Benjamini-Hochberg's method[37]).

All statistical procedures were performed with the R statistical language version 3.0.0[38] and the Bioconductor environment[39]. All data are available from the MetaboLights repository under the accession number MTBLS160 (see Supplementary Methods).

Hierarchical clustering. Before hierarchical clustering, remaining missing values were replaced with half of the minimum feature intensity. Feature intensities were logarithmized, z-transformed and checked for normality with a Kolmogorov-Smirnow test. Non-biological sources of variation were removed by surrogate variable analysis from the SVA package version 3.8.0[40]. In order to discriminate between experimental artifacts and metabolic features in the non-targeted analysis, a generalized Welch test (unequal variances, one-way layout) was applied to find differential features ($p < 0.05$, corrected for multiple testing by Benjamini-Hochberg's method[37]) between the 19 accessions and blank. As a post-hoc test, 2-sample Welch tests were used to find features that were differential ($p < 0.05$) from the blank in at least one accession. This resulted in 455 out of 1950 ESI(−) and 475 out of 3738 ESI(+) metabolic features used for hierarchical clustering. Hierarchical clustering was performed via multiscale bootstrap resampling with the R package pvclust version 1.2–2[41], which improves robustness by providing an approximately unbiased p-value (AU, red number in Fig. 1). Pearson correlation was used as distance measure and average linkage as a linkage method. Since the combination of both ion modes results in redundancy by compounds giving rise to several features, each mode was processed separately. Consistent clusters between the ESI(−) and ESI(+) mode were coloured.

Unspecific signals were more pronounced (87% vs. 75%) in ESI(+) vs. ESI(−). This had led to us to focus on ESI(−) in subsequent analyses.

Sequence analysis. Genetic distances were estimated from the variant tables available from the 19 genomes project[22]. Loci were reduced to CDS as annotated by the R packages Bsgenome.Athaliana.TAIR.TAIR9[42] and Genomic Ranges version 1.14.4[43]. For each variant locus, 19 × 19 comparisons were conducted. In order to construct a distance matrix, mismatches were penalized by increasing the distance by 1. The sum of matrices over all 6,400,466 loci was used as a distance matrix (Supplementary Table S2) for hierarchical clustering via the hclust package with average linkage.

Predicted amino acid sequences were processed with BioPerl (Bio::Tools::Run::Alignment::Clustalw, Bio::SeqIO, Bio::Seq, and Bio::AlignIO) and aligned with the Clustalw algorithm with ktuple = 2 and a BLOSUM scoring matrix. Multiple sequence alignments were evaluated for premature ending with the R packages Biostrings version 2.30.1 and plyr version 1.8.1.

Combination of metabolic and genetic patterns. A metabolic feature was defined as absent when below the limit of detection in all replicates of an accession. Applying this stringent definition, the peak list created from aligning all spectra from ESI(−) was screened for metabolic features with absence, thus reducing the number of features by 25% for exudates ESI(−). The distribution of absence across the 19 accessions is referred to as a pattern. The length of a pattern is the number of accessions that lack the same feature, i.e. a feature absent in Can-0 und Zu-0 is a pattern of length two. Out of the 455 metabolic features in the exudate data set (ESI(−)), 384 were missing in at least one accession. 46 were missing in exactly one accession (length = 1), 52 were absent in two accessions (length = 2) (see Supplementary Table S4). The R package CAMERA version 1.23.2[44] was used for annotation of adduct species and isotope information. In order to find an association between metabolic patterns of absence and its genetic background, features with a pattern of absence, a monoisotopic annotation by CAMERA and a minimal median intensity of 10,000 were evaluated. 31 features that passed the intensity threshold were matched with stop codon patterns resulting in 9/7/1 features of absence with length 1/2/3.

DOI:10.1038/srep29033

These matching features or their corresponding quasi-molecular ion were subjected to fragmentation by MS/MS with 10, 20 and 30 eV. Stop codon patterns were derived from multiple sequence alignments of AraCyc enzyme genes[45] (ftp.plantcyc.org/Pathways/BLAST_sets/aracyc_enzymes.fasta, Dec 2015) as annotated by TAIR10_functional annotations from TAIR.org[46].

References

1. Weigel, D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol*. **158,** 2–22 (2012).
2. Mithen, R., Clarke, J. H., Lister, C. & Dean, C. Genetics of aliphatic glucosinolates. III. Side chain structure of aliphatic glucosinolates in Arabidopsis thaliana. *Heredity (Edinb)*. **74,** 210–215 (1995).
3. Mithen, R. & Campos, H. Genetic variation of aliphatic glucosinolates in Arabidopsis thaliana and prospects for map based gene cloning. *Entomologica Experimentalis et Applicanta*. **53,** 202–205 (1996).
4. Magrath, R. *et al.* Genetics of aliphatic glucosinolates. I. Side chain elongation in Brassica napus and Arabidopsis thaliana. *Heredity (Edinb)*. **72,** 290–299 (1994).
5. Mitchell-Olds, T. & Pedersen, D. The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. *Genetics*. **149,** 739–747 (1998).
6. Kliebenstein, D. J. *et al.* Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol*. **126,** 811–825 (2001).
7. Houshyani, B. *et al.* Characterization of the natural variation in Arabidopsis thaliana metabolome by the analysis of metabolic distance. *Metabolomics*. **8,** 131–145 (2012).
8. Agrawal, B., Lakshmanan, V., Kaushik, S. & Bais, H. P. Natural variation among Arabidopsis accessions reveals malic acid as a key mediator of Nickel (Ni) tolerance. *Planta*. **236,** 477–489 (2012).
9. Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S. & Vivanco, J. M. The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu Rev Plant Biol*. **57,** 233–266 (2006).
10. Schmid, N. B. *et al.* Feruloyl-CoA 6′-Hydroxylase1-dependent coumarins mediate iron acquisition from alkaline substrates in Arabidopsis. *Plant Physiol*. **164,** 160–172 (2014).
11. van de Mortel, J. E. *et al.* Metabolic and transcriptomic changes induced in Arabidopsis by the rhizobacterium Pseudomonas fluorescens SS101. *Plant Physiol*. **160,** 2173–2188 (2012).
12. Hiltner, L. Über neue Erfahrungen und Probleme auf dem Gebiete der Bodenbakteriologie. *Arbeiten der Deutschen Landwirtschaft Gesellschaft*. **98,** 59–78 (1904).
13. Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E. & Schulze-Lefert, P. Structure and functions of the bacterial microbiota of plants. *Annu Rev Plant Biol*. **64,** 807–838 (2013).
14. Bressan, M. *et al.* Exogenous glucosinolate produced by Arabidopsis thaliana has an impact on microbes in the rhizosphere and plant roots. *ISME J*. **3,** 1243–1257 (2009).
15. Strehmel, N., Böttcher, C., Schmidt, S. & Scheel, D. Profiling of secondary metabolites in root exudates of Arabidopsis thaliana. *Phytochemistry*. **108C,** 35–46 (2014).
16. Micallef, S. A., Shiaris, M. P. & Colon-Carmona, A. Influence of Arabidopsis thaliana accessions on rhizobacterial communities and natural variation in root exudates. *J Exp Bot*. **60,** 1729–1742 (2009).
17. Böttcher, C., von Roepenack-Lahaye, E. & Scheel, D. Resources for metabolomics in *Genetics and Genomics of the Brassicaceae* Vol. 9 *Plant Genetics and Genomics: Crops and Models* (eds Bancroft, I. & Schmidt, R.) 469–503 (Springer, 2011).
18. Kover, P. X. *et al.* A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet*. **5,** e1000551 (2009).
19. Nordborg, M. *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol*. **3,** e196 (2005).
20. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*. **317,** 338–342 (2007).
21. Editorial. How robust are your data? *Nature Cell Biology*. **11** (2009).
22. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*. **477,** 419–423 (2011).
23. Wiesner, M., Schreiner, M. & Zrenner, R. Functional identification of genes responsible for the biosynthesis of 1-methoxy-indol-3-ylmethyl-glucosinolate in Brassica rapa ssp. chinensis. *BMC Plant Biol*. **14,** 124 (2014).
24. Pfalz, M. *et al.* Metabolic engineering in Nicotiana benthamiana reveals key enzyme functions in Arabidopsis indole glucosinolate modification. *Plant Cell*. **23,** 716–729 (2011).
25. Keurentjes, J. J. *et al.* The genetics of plant metabolism. *Nat Genet*. **38,** 842–849 (2006).
26. Obayashi, T. *et al.* ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res*. **35,** D863–869 (2007).
27. De Pessemier, J., Chardon, F., Juraniec, M., Delaplace, P. & Hermans, C. Natural variation of the root morphological response to nitrate supply in Arabidopsis thaliana. *Mech Dev*. **130,** 45–53 (2013).
28. Trontin, C., Tisne, S., Bach, L. & Loudet, O. What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants? *Curr Opin Plant Biol*. **14,** 225–231 (2011).
29. Ziegler, J. *et al.* Simultaneous analysis of apolar phytohormones and 1-aminocyclopropan-1-carboxylic acid by high performance liquid chromatography/electrospray negative ion tandem mass spectrometry via 9-fluorenylmethoxycarbonyl chloride derivatization. *J Chromatogr A*. **1362,** 102–109 (2014).
30. Li, X. *et al.* Exploiting natural variation of secondary metabolism identifies a gene controlling the glycosylation diversity of dihydroxybenzoic acids in Arabidopsis thaliana. *Genetics*. **198,** 1267–1276 (2014).
31. Negishi, M. *et al.* Structure and function of sulfotransferases. *Arch Biochem Biophys*. **390,** 149–157 (2001).
32. Stracke, R. *et al.* Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. *Plant J*. **50,** 660–677 (2007).
33. Jones, P., Messner, B., Nakajima, J., Schaffner, A. R. & Saito, K. UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in Arabidopsis thaliana. *J Biol Chem*. **278,** 43910–43918 (2003).
34. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*. **9,** 504 (2008).
35. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*. **78,** 779–787 (2006).
36. Böttcher, C. *et al.* The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana. *Plant Cell*. **21,** 1830–1845 (2009).
37. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).
38. R: A Language and Environment for Statistical Computing (2014).
39. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. **5,** R80 (2004).

DOI:10.1038/srep29033

40. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3,** 1724–1735 (2007).
41. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. **22,** 1540–1542 (2006).
42. BSgenome.Athaliana.TAIR.TAIR9: Full genome sequences for Arabidopsis thaliana (TAIR9).
43. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol*. **9,** e1003118 (2013).
44. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*. **84,** 283–289 (2012).
45. Mueller, L. A., Zhang, P. & Rhee, S. Y. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol*. **132,** 453–460 (2003).
46. Huala, E. *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*. **29,** 102–105 (2001).
47. Luo, J. *et al.* A novel polyamine acyltransferase responsible for the accumulation of spermidine conjugates in Arabidopsis seed. *Plant Cell*. **21,** 318–333 (2009).

Acknowledgements

We thank Dr. Ralf Stracke for providing seeds. This work was supported by the Joint Initiative for Research and Innovation of the Leibniz Association (SAW-2011-IPB-3 97, “Chemical Communication in the Rhizosphere”). Metabolomics expertise by Dr. Christoph Böttcher und expert technical assistance by Sylvia Krüger, Jessica Thomas and Susanne Kirsten are gratefully acknowledged. The publication of this article was funded by the Open Access fund of the Leibniz Association.

Author Contributions

D.S. designed the project. The hydroponic system was designed by D.S., N.S. and S.S. The genetic model was designed by S.M., E.M., S.S., L.W. and S.N. N.S., S.S. and S.M. performed measurement and general data analysis. Non-targeted profiling was analysed by N.S., F.T. and S.M.; the targeted analysis was performed by N.S., S.H. and S.M. S.N. submitted the data to MetaboLights. Expertise and proofreading was provided by L.W., D.S., N.S., S.S. and S.N. S.M. and N.S. structured and wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Mönchgesang, S. *et al.* Natural variation of root exudates in *Arabidopsis thaliana*-linking metabolomic and genomic data. *Sci. Rep.* **6**, 29033; doi: 10.1038/srep29033 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

DOI:10.1038/srep29033

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Data Descriptor: Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes

Received: 25 January 2018
Accepted: 23 July 2018
Published: 28 August 2018

Kristian Peters[1], Karin Gorzolka[1], Helge Bruelheide[2,3] & Steffen Neumann[1,3]

In Eco-Metabolomics interactions are studied of non-model organisms in their natural environment and relations are made between biochemistry and ecological function. Current challenges when processing such metabolomics data involve complex experiment designs which are often carried out in large field campaigns involving multiple study factors, peak detection parameter settings, the high variation of metabolite profiles and the analysis of non-model species with scarcely characterised metabolomes. Here, we present a dataset generated from 108 samples of nine bryophyte species obtained in four seasons using an untargeted liquid chromatography coupled with mass spectrometry acquisition method (LC/MS). Using this dataset we address the current challenges when processing Eco-Metabolomics data. Here, we also present a reproducible and reusable computational workflow implemented in Galaxy focusing on standard formats, data import, technical validation, feature detection, diversity analysis and multivariate statistics. We expect that the representative dataset and the reusable processing pipeline will facilitate future studies in the research field of Eco-Metabolomics.

Design Type(s)	time series design • database creation objective • process-based data analysis objective
Measurement Type(s)	metabolite profiling
Technology Type(s)	Ultra High-performance Liquid Chromatography/Tandem Mass Spectrometry
Factor Type(s)	Spatial Orientation • wetness of soil • degree of illumination • substrate type • season • scan polarity
Sample Characteristic(s)	Fissidens taxifolius • shoot system • Polytrichum strictum • Hypnum cupressiforme • Grimmia pulvinata • Plagiomnium undulatum • Rhytidiadelphus squarrosus • Calliergonella cuspidata • Brachythecium rutabulum • Marchantia polymorpha

[1]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany. [2]Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany. [3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. Correspondence and requests for materials should be addressed to K.P. (email: kpeters@ipb-halle.de).

DOI:10.1038/sdata.2018.179

Background & Summary

In Ecological Metabolomics (or short "Eco-Metabolomics"), metabolite profiles of organisms are studied in order to describe ecological processes such as biotic interactions or the impact of environmental changes on various biological species[1–3]. In contrast to biochemistry, wild non-model species are typically studied in their natural environment in ecology. This often involves different individuals of one or more species from populations growing under quite heterogeneous conditions when compared to the controlled conditions in greenhouses or growth chambers. As a result, metabolite profiles are highly variable when compared to each other. Moreover, profiles of non-model species contain a large number of novel compounds (so called "unknown unknowns") that are difficult to identify because of lacking reference compounds, which have so far been mostly elucidated in model organisms[3,4]. Furthermore, designing ecological experiments is often complex and involves multiple factors[5]. Thus, the metabolomics data processing pipeline needs to be adapted in order to deal with the particular hypotheses and idiosyncrasies of ecological experiments.

Here, we present a descriptor for a dataset that we consider representative for the research field of Eco-Metabolomics. Our study makes use of a field campaign with a two-factorial design (seasons and species), which includes (except *Marchantia polymorpha*) non-model species of bryophytes. In order to facilitate subsequent analysis, we kept the experiment design as simple as possible. The sampling was conducted on-site at the Botanical Garden of Martin Luther University Halle-Wittenberg once in each season over a period of one year (see below). Metabolite profiles were acquired using untargeted liquid chromatography coupled with mass spectrometry (LC/MS). Raw metabolite profiles are available in the metabolomics data repository MetaboLights[6] (Data Citation 1).

In biochemistry there are strict laboratory protocols that ensure reproducibility of the analytical methods, while in bioinformatics this function is accomplished by implementing reusable computational workflows[7,8]. Thus, in addition to the dataset we also address the typical bioinformatic challenges that come with Eco-Metabolomics experiments by implementing a reproducible and reusable computational workflow (Fig. 1). While the analysis and ecological interpretation of the study is described in Peters *et al.*[9], here we focus on the analytical and bioinformatic work that is required to create a computational processing pipeline that is reproducible and that can be reused by other subsequent studies.

We describe in detail the experimental methodology that was used to create the dataset as well as the methodology to make the computational workflow reproducible (to give identical results in different computational environments). By formalizing and validating the processes that led to the results[10,11], we expect that this approach can serve as a model for subsequent studies. We further expect that Eco-Metabolomics studies use our dataset and the computational workflow to foster reuse and improve future data processing pipelines.

Methods

These methods describe in detail the steps in producing the data, including full descriptions of the experimental design in our related work[9], data acquisition, computational processing, diversity analysis, biostatistics and bioinformatics procedures.

Sampling campaign

Samples of the nine moss species Brachythecium rutabulum (Hedw.) Schimp., Calliergonella cuspidata (Hedw.) Loeske, Fissidens taxifolius Hedw., Grimmia pulvinata (Hedw.) Sm., Hypnum cupressiforme Hedw. (H. lacunosum was not differentiated), Marchantia polymorpha L., Plagiomnium undulatum (Hedw.) T.J. Kop., Polytrichum strictum Menzies ex Brid. and Rhytidiadelphus squarrosus (Hedw.) Warnst. were collected in the Botanical Gardens of the Martin-Luther-University Halle-Wittenberg, Germany. Sampling was performed in summer (2016/08/08), autumn (2016/11/09), winter (2017/01/27) and spring (2017/05/11) at relatively stable weather conditions as it is known that short-term climatic fluctuations and rainfall can influence secondary metabolite content and ammonium uptake of bryophytes[12]. Thus, the bryophytes were only collected when there was sunshine at least two days prior to and during sampling. Furthermore, sampling was performed after mid-day between 13:00 and 15:00.

Sampling protocol

In each season, three composite samples of different individuals of each species were taken, leading to a total of $3 * 9 * 4 = 108$ samples. Only above-ground parts of the moss gametophytes such as leaves, branches, stems or thalloid parts were taken for sampling. From dioecious species such as *M. polymorpha*, *P. strictum* and *P. undulatum* female, male and sterile gametophytes were collected in a composite sample. Before sampling, visible archegonial and antheridial heads and any belowground parts such as rhizoids and rooting stems were removed with a sterile tweezer. The gametophytic moss parts were put in Eppendorf tubes and were frozen instantly on dry ice and later in the lab in liquid nitrogen.

Collecting ecological characteristics

In order to relate metabolomes of the bryophytes to ecology, several ecological characteristics were recorded on-site and compiled from literature. The on-site characteristics *type of substrate* with the nominal/categorical levels "soil", "rock with lean soil cover" and "rock"; *light conditions* with the ordinal

DOI:10.1038/sdata.2018.179

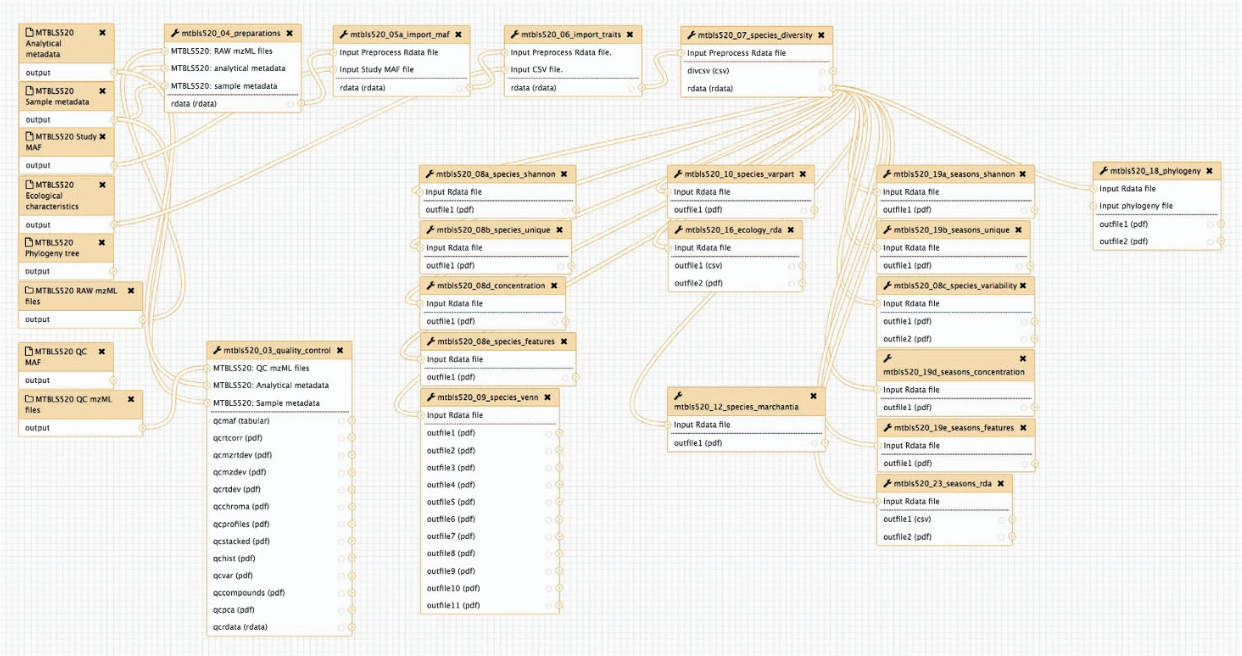


Figure 1. Computational workflow of the whole study (Data Citation 1) running in the Galaxy Workflow Management system. Each of the modules represent a particular step in the study of Peters *et al.*[9]. The modules have defined inputs, outputs and sets of parameters. The modules are connected to each other to give the resulting workflow. The function of the modules is explained in Table 1 (available online only).

levels "sunny", "half-shade" and "shade"; *moisture of the substrate* with the ordinal levels "dry", "fresh", "damp" and wet; and *exposition* with the nominal levels "North", "East", "South", "West", "Northeast", "Northwest", "Southeast" and "Southwest" were recorded when taking the samples in the field.

The nominal characteristics *growth form*, *habitat type*, *substrate* and *life strategy*, the ordinal life-history characteristics *spore size*, *gametangia distribution* and *sexual reproduction frequency*, as well as the ordinal Ellenberg indicator values (indices for *light*, *temperature*, *continentality*, *moisture*, *reaction*, *nitrogen* and *life-form*) were collected from the literature[13–17]. For an overview, please refer Table 1 (available online only) in Peters *et al.*[9] or the file m_characteristics.csv in the dataset (see Data Citation 1, and Table 1 (available online only)).

Extraction protocol and LC/MS analysis

Frozen moss samples were homogenized by adding 200 mg ceramic beads (0.5 mm diameter, Roth) and ribolysing (Precellys 24, 2 × 20 s at 6500 r.p.m., 5 min pause in liquid nitrogen). 1 ml ice-cold 80/20 (v/v) methanol/water spiked with internal standards 5 μM biochanin A (Sigma-Aldrich), 5 μM kinetin (Sigma-Aldrich) and 5 μM N-(3-indolylacetyl)-l-valine (Sigma-Aldrich) were added. Samples were vortexed and thawed while shaking for 15 min at 1,000 r.p.m. at room temperature followed by ultrasonification for 15 min and again 15 min shaking. After 15 min centrifugation at 13,000 r.p.m. 500 μl of supernatant were dried in a vacuum centrifuge at 40 °C and reconstituted in 80/20 (v/v) methanol/water with the volume adjusted to the initial fresh weight of the sample to a final concentration of 10 mg fresh weight per 100 μl extract.

Chromatographic separations were performed at 40 °C on an Acquity UPLC system (Waters) equipped with an HSS T3 column (100 × 1 mm, particle size 1.8 μm; Waters) applying the following binary gradient at a flow rate of 150 μL min^{-1}: 0 to 1 min, isocratic 95% A (water:formic acid: 99.9:0.1 [v/v]), 5% B (acetonitrile:formic acid: 99.9:0.1 [v/v]); 1 to 18 min, linear from 5 to 95% B; 18 to 20 min, isocratic 95% B. The injection volume was 2.0 μL (full loop injection).

Ultra-performance liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (UPLC/ESI-QTOF-MS) was performed using a high resolution MicrOTOF-Q II hybrid quadrupole time-of-flight mass spectrometer[18]. Data were acquired with the following MS instrument settings: nebulizer gas: nitrogen, 1.4 bar; dry gas: nitrogen, 6 L min^{-1}, 190 °C; capillary: 5000 V (+4000 V for negative mode); end plate offset: −500 V; funnel 1 radio frequency (RF): 200 Volts peak-to-peak (Vpp); funnel 2 RF: 200 Vpp; in-source collision-induced dissociation (CID) energy: 10 eV; hexapole RF: 100 Vpp; quadrupole ion energy: 3 eV (−5 eV for neg-mode); collision gas: nitrogen; collision energy: 7 eV (−7 eV for negative mode); collision cell RF: 250 Vpp (150 Vpp for negative mode); transfer time: 70 μs; prepulse storage: 5 μs; pulser frequency: 10 kHz; and spectra rate: 3 Hz. Mass spectra

DOI:10.1038/sdata.2018.179

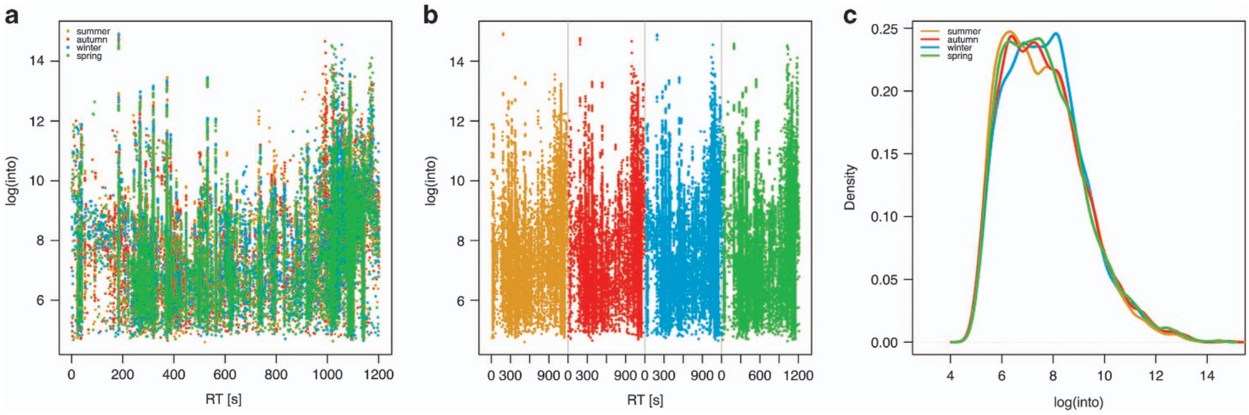


Figure 2. Plots of sets of the MM8 profiles to assess the performance of the technical setup. Green = spring, yellow = summer, red = autumn, blue = winter. n = 28. (**a**) Plot of the four sets of MM8 profiles against each other. X axis: Retention time [s]. Y axis: Logarithmic total ion current. (**b**) Stacked plot of the sets of MM8 profiles next to each other. X axis: Retention time [s]. Y axis: Logarithmic total ion current. (**c**) Density plot (histogram) of log intensities of the sets of MM8 profiles. X axis: Sample size. Y axis: Estimated kernel density.

were acquired in centroid mode. Calibration of the m/z scale was performed for individual raw data files on lithium formate cluster ions obtained by automatic infusion of 20 µL of 10 mM lithium hydroxide in isopropanol:water:formic acid, 49.9:49.9:0.2 (v/v/v) at the end of the gradient.

Quality control

In order to validate the instrument performance and to detect batch effects between the instrument runs, the following quality control (QC) protocol was realized. Samples with a lab-internal standard mix (MM8) were interspersed before and after 7 bryophyte samples in the MicrOTOF[18]. The following substances were used in the MM8: 2-Phenylglycine (Fluka), Kinetin (Roth), Rutin (Acros Organics), O-Methylsalicylic acid (Sigma), Phlorizin dihydrate (Sigma), N-(3-Indolyacetyl)-L-valine (Sigma), 3-Indolylacetonitrile (Fluka) and Biochanin A (Sigma). Substances in the MM8 were selected based on their ionization properties (ionization in both positive and negative mode and the differential adduct formation) and a wide coverage of known retention times throughout the gradient with our instrumental setup. Known ionization properties were used to detect shifts and effects in mass-to-charge ratios (m/z) and retention times (RT) of the respective batches and to validate RT correction made by XCMS (see below).

Raw data acquisition

Raw LC/MS data were converted to the open data format mzML[19] with the software CompassXPort 3.0.9 from Bruker Daltonics (available at http://www.bruker.com/service/support-upgrades/software-downloads.html). In compliance with the minimum information guidelines for Metabolomics studies[20], metadata were recorded to ISA-Tab format[21] using ISAcreator 1.7.10 (ref. 22) (available at https://github.com/ISA-tools/ISAcreator/releases) and uploaded together with the raw data to the metabolomics repository MetaboLights[6] (Data Citation 1). Profiles of positive mode were used for the data analyses as many important and known secondary metabolites classes in bryophytes such as flavonoids, phenylpropanoids, anthocyans, glycosides and previously characterized compounds such as Marchantins, Communins and Ohioensins ionize well in positive mode with our instrumental setup.

Peak detection

Chromatographic peak picking was performed in R 3.4.2 (available at https://cran.r-project.org) with the package XCMS 1.52.0 (ref. 23) using the centWave algorithm and the following parameters: ppm = 35, peakwidth = 4,21, snthresh = 10, prefilter = 5–50, fitgauss = TRUE, verbose.columns = TRUE. Grouping of chromatographic peaks was performed with two factors (in XCMS called “phenoData”): *seasons* with the levels summer, autumn, winter and spring; and *species* with the levels Brarut, Calcus, Fistax, Gripul, Hypcup, Marpol, Plaund, Polstr and Rhysqu. The following parameters were used for grouping: mzwid = 0.01, minfrac = 0.5, bw = 4. To improve subsequent data analyses, intensities in the peak table were log transformed before grouping. For further analysis, only features between the retention times 20 s and 1020 s were kept. Retention time correction was performed using the function retcor in XCMS using the parameters method = loess, family = gaussian, missing = 10, extra = 1, span = 2. The parameters were additionally optimized using the R package IPO 1.3.3 (ref. 24), but better alignment precision was achieved with manual control and knowledge of instrument settings[25].

DOI:10.1038/sdata.2018.179

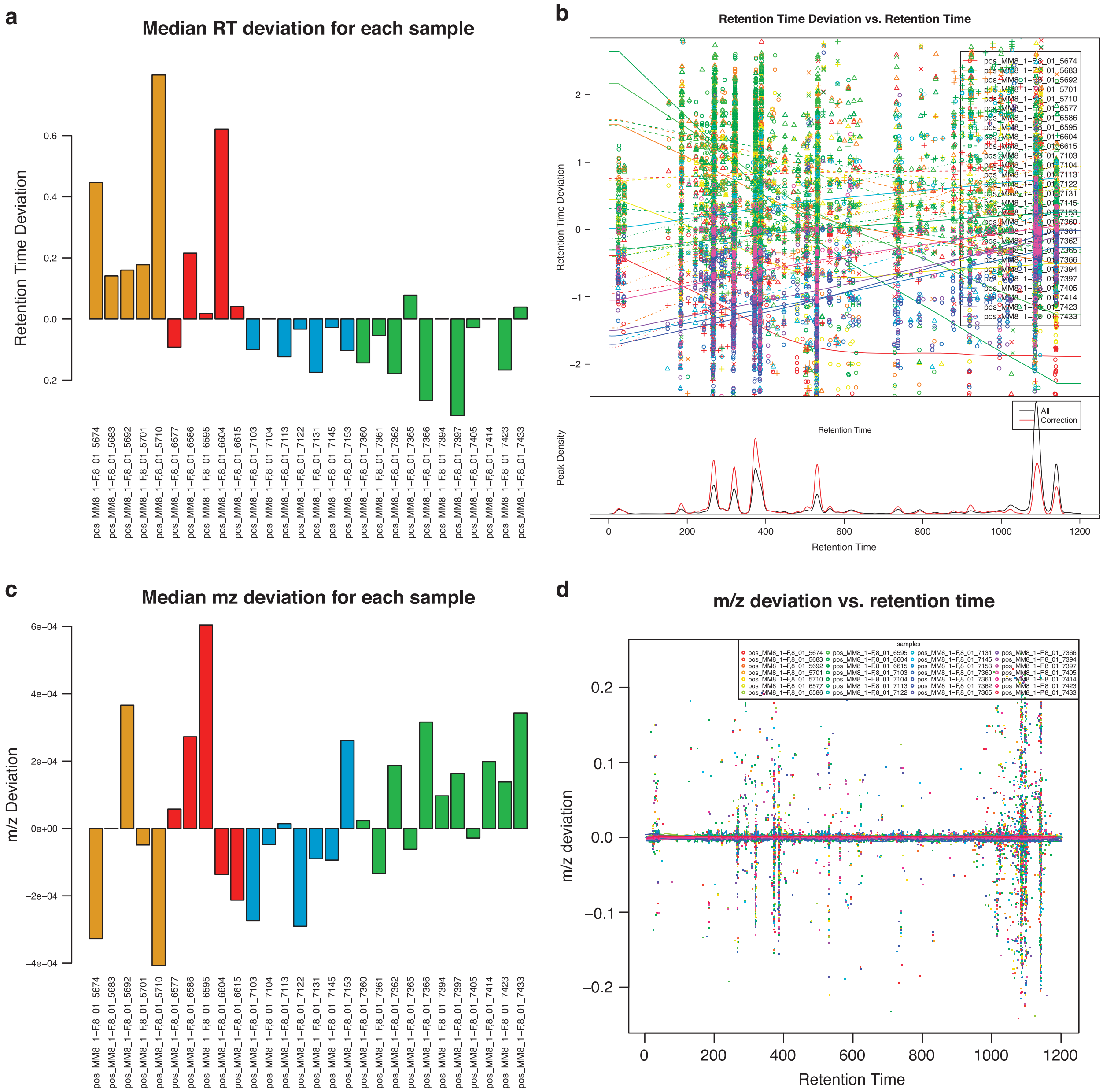


Figure 3. Quality control plots to assess shifts in retention time (RT) and mass-to-charge ratio (m/z) in the four sets of MM8 profiles. Green = spring, yellow = summer, red = autumn, blue = winter. n = 28. (**a**) Median retention time deviation for the sets of MM8 profiles. X axis: Name of MM8 profile. Y axis: Retention time deviation [s]. (**b**) Retention time deviation plotted against retention time. X axis: Retention time [s]. Y axis: Retention time deviation [s] per profile. (**c**) Median mass-to-charge deviation for each profile. X axis: MM8 profile. Y axis: m/z deviation. (**d**) Mass-to-charge deviation plotted against retention time. X axis: Retention time [s]. Y axis: m/z deviation per profile.

Peak annotation

Adduct annotation was performed with the R package CAMERA 1.33.3 (ref. 26) by using the following functions: xsAnnotate, groupFWHM, findIsotopes, groupCorr, findAdducts; with the following parameters: perfwhm = 0.6, ppm = 5, mzabs = 0.005, calcIso = TRUE, calcCiS = TRUE, calcCaS = TRUE, graphMethod = lpc, pval = 0.05, cor_eic_th = 0.75. In order to improve subsequent statistical analyses instead of the CAMERA function getPeaklist the function getReducedPeaklist was written that aggregates the adducts of putative compounds into a feature list with singular components (see pull request in GitHub: https://github.com/sneumann/CAMERA/pull/16). Since version 1.33.3 the function getReducedPeaklist is officially part of CAMERA. The parameter method = median was chosen for the study.

DOI:10.1038/sdata.2018.179

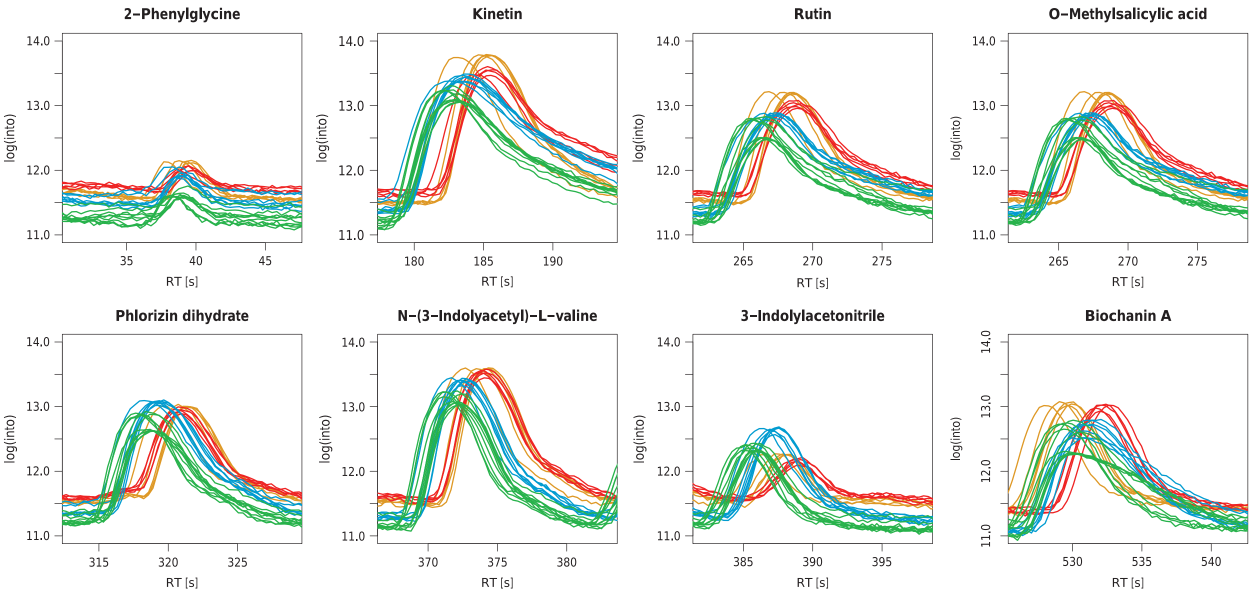


Figure 4. The eight compounds used for the internal lab standard mix (MM8) plotted next to each other. Shown are the regions of the respective compounds in the raw chromatograms before the alignment of XCMS. Green = spring, yellow = summer, red = autumn, blue = winter. X axis: Retention time [s]. Y axis: Logarithmic total ion current. n = 28.

Exemplary compound annotation

Compounds were putatively annotated for the follow-up validation and biochemical interpretation with the software Bruker Compass IsotopePattern 4.4. Annotation was performed by calculating accurate masses (mass-to-charge values) from known compounds in *M. polymorpha* and other liverworts found in PubChem, the KNApSAcK database and Asakawa *et al.*[27,28]. In the software Bruker Compass DataAnalysis 4.4 the mass-to-charge was matched to device-specific retention times in the metabolite profile. To validate whether the known compound was present in the profile, Extracted Ion Chromatograms (EIC) and area-under-curve (integrated intensities) were checked manually.

Diversity analysis

Statistical analyses were performed using the additional R packages: multtest, RColorBrewer, vegan, multcomp, multtest, nlme, ape, pvclust, dendextend, phangorn, Hmisc, gplots and VennDiagram. A presence-absence matrix was generated from the feature matrix to determine the differences in metabolite features between the experimental factors species and season. In accordance with the minfrac parameter in the alignment step in XCMS (see above), a feature was considered present when it was detected at least in two out of three replicates. The presence-absence matrix was used for measuring the metabolite richness for each species and season by calculating the Shannon diversity index (H') for each sample *i* using the function diversity in vegan with the parameter index = shannon[29]. The following equation was used for calculation:

$$H' = \sum_{i=1}^{t} pi \ \mathrm{In}(pi)$$

where t represents the number of samples in the particular group.

The total number of features and the number of unique features were calculated from the presence-absence matrix accordingly. To test factor levels for significant differences, the Tukey HSD on a one-way ANOVA was performed post-hoc using the multcomp package.

Variability was calculated with the Pearson Correlation Coefficient (PCC, Pearson's r) using the function rcorr in the package Hmisc. Venn diagrams were created for each species separately using the package VennDiagram. Each set in the Venn diagram represents one season and shows distinct and shared features in all possible combinations between the sets.

Multivariate statistical analysis

Variation partitioning was performed using the function varpart in the package vegan to analyze the influence of the factors species and seasons on the metabolite profiles. Distance-based redundancy analysis (dbRDA) using the function capscale with Bray-Curtis distance and multidimensional scaling in

DOI:10.1038/sdata.2018.179

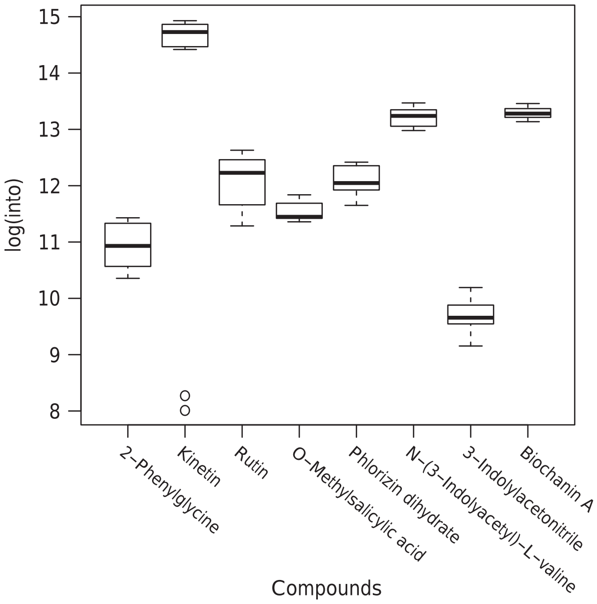


Figure 5. Boxplots of the variation in the intensities of the eight compounds used in the internal lab standard mix of all the MM8 profiles. X axis: Compound. Y axis: Logarithmic total ion current. n = 28 for each box.

the package vegan was chosen to analyse the relation of the ecological characteristics with the species metabolite profiles[30,31]. Ordinal and categorical ecological characteristics were transformed to presence-absence matrices for the ordination. The optimal model for the dbRDA was chosen with forward and backward selection using the function ordistep in the package vegan. Ecological characteristics were added to the plots as post-hoc variables using the function envfit in the package vegan.

Chemotaxonomic comparison to phylogeny

Relationships between metabolite profiles and phylogeny were analysed by calculating dissimilarities for phylogeny and the feature matrix using Bray-Curtis distance (function vegdist in vegan) followed by hierarchical clustering using the function hclust and the complete linkage method. In order to improve the visual comparison between the two trees, the chemotaxonomic plot was reordered using the function order.optimal (package cba) and leaves of Polstr and Plaund were swapped using the function reorder in vegan. The similarity of the two trees was determined with the normalized Robinson-Foulds metric (function RF.dist in package phangorn). The similarity of the distance matrices was determined with the Mantel statistics (function mantel in vegan).

Computational workflow

For the computational workflow, the required software tools, their dependencies, as well as software libraries and R packages were containerized using Docker technology[32]. The container was based on Linux and Ubuntu 16.04 and included R version 3.4.2 from the R apt repository. The commands for building the container can be found in the Dockerfile (Table 1 (available online only)). The resulting container image was made available at DockerHub (https://hub.docker.com/r/korseby/mtbls520/).

The computational workflow was constructed with the Galaxy workflow management system[33]. It consists of 20 modules and each individual module represents one or more dedicated steps in the Peters *et al.* study[9], e.g. data retrieval, feature detection, alignment or statistical analysis (Fig. 1). For the workflow, individual Galaxy modules were written in XML format. Each Galaxy module executes a shell or R script with defined inputs and outputs. Scripts are only executed inside the software container. Thus, code execution is encapsulated and all required software dependencies were resolved in the software container. In order to comply with the *Interoperability* criterion in the FAIR guidelines[34], the PhenoMeNal cloud e-infrastructure was used to test the workflow in different computational environments (https://phenomenal-h2020.eu). To ensure that the workflow generates the same results in different computational environments, continuous automatic workflow testing was implemented with wft4galaxy[35].

Data Records

The primary access site for the dataset is MetaboLights (Data Citation 1), which includes the 108 metabolite profles of the bryophytes in positive and negative mode, QC profiles, ecological data and meta-data (see Table 2 (available online only) for an overview of sample names and associated factor

DOI:10.1038/sdata.2018.179

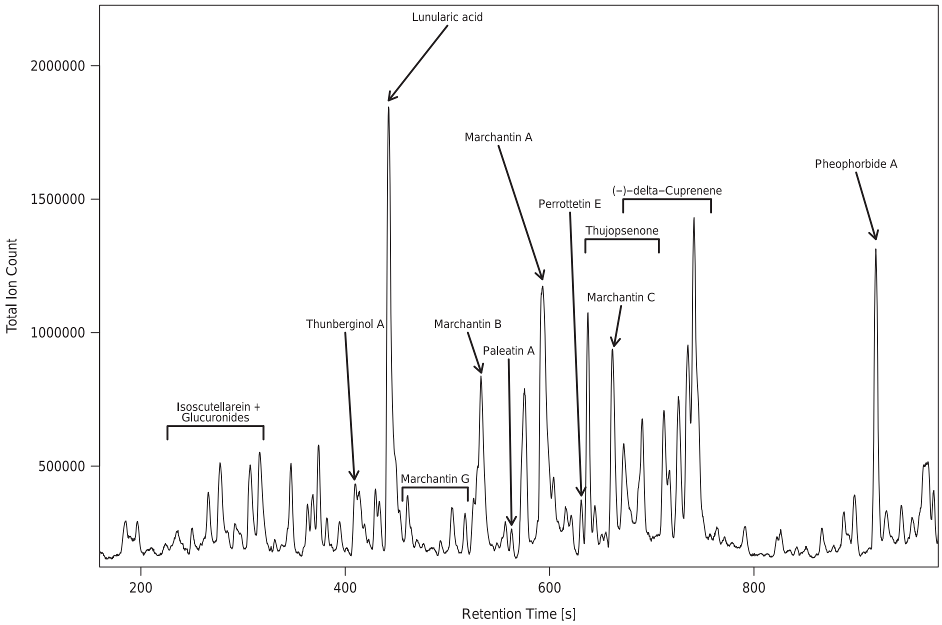


Figure 6. Total Ion Count (TIC) chromatogram obtained from the extracts of *Marchantia polymorpha*. This exemplary chromatogram was obtained from the third sample of summer. Values of the retention times (RT values), accurate masses and sum formulas are available in Table 3 (available online only).

levels). Table 1 (available online only) provides an overview of data files, formats and functions in the computational workflow.

Code Availability

The source code (also deposited at https://github.com/korseby/container-mtbls520/) was published[36] and made available under the terms of APACHE license 2.0. Please refer Table 1 (available online only) for an overview of the function of each file of the source code.

Code for building the software container image and the workflow including Galaxy modules and scripts that are executed inside the container were published under Open Access[36]. A pre-built binary software container image was made available at DockerHub (retrievable at https://hub.docker.com/r/korseby/mtbls520/).

Technical Validation

Quality control

Four sets of 27 bryophyte samples were generated in the experiment. One set for each season was analyzed with UPLC/ESI-QTOF-MS (see methods below) which resulted in a total of 108 bryophyte metabolite profiles. In order to validate the instrument performance and to detect batch effects between the four instrument runs, a quality control (QC) protocol was implemented. Sets of 27 species samples were interspersed by samples of a lab-internal standard mix (MM8) before and after 7 bryophyte samples. Peak detection in these MM8 profiles was performed with the identical parameters as for the bryophyte samples.

The four sets containing the MM8 metabolite profiles were checked visually for differences by plotting them against each other (Fig. 2a) and stacked next to each other (Fig. 2b). The density distribution of the intensities within the sets of MM8 profiles were also checked and compared to each other with a density plot (histogram) (Fig. 2c).

Mass-to-charge ratio and retention time deviation (in seconds) and correction made by XCMS were checked with diagnostic plots made by XCMS (Fig. 3). We found maximum retention time deviations within 2 s (Fig. 3a and b) which are in the expected range of the analytical setup[18]. The determined mass-to-charge deviations (Fig. 3c and d) are within instrument specification as well[18].

The variation in the intensities of the internal lab standards was also checked for each reference compound individually as shown in Figs. 4 and 5. In general, the variation for each reference compound and the deviations between MM8 profiles are both well within the typical range of 10 to 15% (ref. 18).

We conclude that there are no significant batch effects in the technical replicates to overlap with the factor *seasons* of the experiment. Thus, the automatic retention time correction made by XCMS is validated for the parameters used in the peak detection process.

Exemplary annotation of *Marchantia polymorpha* profile

With known accurate masses (m/z values) and calculated retention time values (see methods), we confirm the annotation of many known compounds which are described in literature for the model species *Marchantia polymorpha*[27,28] (Fig. 6). Many of these known compounds also constitute the most abundant features in the profile of *M. polymorpha* (Fig. 6).

Computational workflow

We have implemented the computational workflow in the Galaxy workflow management system[33] and have made the workflow and underlying code available as Open Source[36]. The Galaxy workflow represents the entire computational processing pipeline that is used in the Peters *et al.* study[9] (Fig. 1). Each of the individual modules represents a particular step in the workflow and has defined inputs (e.g. pre-processed peak table data matrix) and outputs (e.g. PDF containing the plot of a particular statistical method) (Fig. 1). We used data standards and minimum information criteria for constructing the modules of the workflow[20,22]. Continuous automatic testing of the workflow was performed with wft4galaxy[35] in the PhenoMeNal e-infrastructure (https://phenomenal-h2020.eu) to ensure that the workflow generates the same results in different computational environments.

We proceeded according to the FAIR guiding principles[34] in order to implement a reusable computational workflow. The acronym FAIR stands for Findable, Accessible, Interoperable and Reusable and encompasses several criteria to support the reuse of scholarly data. So far, the FAIR guidelines have only been aspired to make data reusable. However, as the conceptual formulation within FAIR are quite generalized[37], these principles can also be applied to computational workflows. Nonetheless, there are some computational challenges involved. For example, software runs in different software environments and software dependencies need to be resolved. We tackle this by creating software containers which can be run on multiple systems and contain the software tools, all required libraries and R packages[32,38]. As dependencies in the container have already been resolved, sharing the container image greatly facilitates to allow the software to be run in multiple environments.

We have chosen the Galaxy Workflow Management system[33,39] to implement the whole data processing pipeline (Fig. 1) as it is already known to facilitate reproducible results[40]. Several processing modules were constructed that represent the individual steps of the Peters *et al.* study[9]. Software tools are invoked from the Galaxy modules and are executed inside the container, thus, adding a level of encapsulation and eliminating the need for the user to install additional software[41]. Galaxy has a graphical user interface that hides the technical complexity from the end user and does not need intensive bioinformatic background knowledge to run the particular modules and workflows. This greatly contributes to the adoption by the end users (biochemists and ecologists) and facilitates future studies in the research field of Eco-Metabolomics.

Statistical analyses

With untargeted metabolomics analysis in ecology, diversity analysis is typically used to characterize the richness and the abundance of biochemical features in the metabolite profiles of biological species[42]. Metabolite richness is a simple measure that counts the individual biochemical features in the metabolite profiles of the species[43]. The abundance of features in the metabolite profiles is usually calculated by diversity indices such as the Shannon diversity index (H') in order to characterize simple relationships with regard to the study factors[44].

Ordination methods such as Redundancy Analysis (RDA) and distance-based Redundancy Analysis (dbRDA) are frequently used in Ecology[30]. They allow to derive correlations of specific variables between the matrix of predictors containing the measurements (X matrix) and the response matrix with the ecological traits (Y matrix)[30,45]. These methods are also suitable for Eco-Metabolomics data as they allow the use of multiple (non-categorial) variables in a single model and allow to calculate the amount of explained variance of the model. We have chosen the dbRDA, which can also be regarded as a constrained version of metric scaling (MDS)[46,47]. We have implemented dedicated modules for these statistical operations in our computational workflow (see Methods section and Fig. 1).

Usage Notes

Building the container image

Following are instructions to manually build the container image. The file Dockerfile in Table 1 (available online only) contains the ruleset. The container has been built using Docker version 17.05-ce under Linux Ubuntu 16.04. The following commands were run to generate the image:

```
sudo apt-get install apt-transport-https ca-certificates git
sudo echo deb http://apt.dockerproject.org/repo ubuntu-xenial main
>>/etc/apt/sources.list
sudo apt-key adv --keyserver hkp://ha.pool.sks-keyservers.net:80
--recv-keys 58118E89F3A912897C070ADBF76221572C52609D
sudo apt-get update && sudo apt-get install docker
git clone https://github.com/korseby/container-mtbls520
cd container-mtbls520
docker build -t korseby/mtbls520.
```

DOI:10.1038/sdata.2018.179

Installing and using Galaxy to run the workflow

The workflow was tested with Galaxy version 17.09. Instructions how to install Galaxy can be found in the training material of the Galaxy project (accessible at https://galaxyproject.github.io/training-material/). However, it is recommended that an official Galaxy server is used, such as those from the PhenoMeNal infrastructure (available at https://public.phenomenal-h2020.eu/).

After being logged into Galaxy, a click on "Workflow" in the menu bar on the top and then a click on the "Upload" button opens up a new page. In the field "Galaxy workflow URL:" enter the following address "https://raw.githubusercontent.com/korseby/container-mtbls520/develop/galaxy/mtbls520_workflow.ga" or upload the .ga file from the GitHub repository (Table 1 (available online only)) and then clicking on the button "Import". This will import the workflow of the study into Galaxy. The workflow will now be available in Galaxy under Workflows as "Metabolights 520 Eco-Metabolomics Workflow". From there, clicking on the drop-down menu there are options to "Edit" (visually view the complete workflow in the Galaxy workflow editor) or to "Run" the workflow. Required data can be downloaded from MetaboLights with the Galaxy module "mtbls520_01_mtbls_download" (Table 1 (available online only)). Once the download has been completed, data can be extracted with the Galaxy module "mtbls520_02_extract" (Table 1 (available online only)). The workflow can be directly run once the inputs have been assigned to the extracted data files. Processing will take approx. 40 min depending on the work load of the computational infrastructure.

References

1. Sardans, J., Peñuelas, J. & Rivas-Ubach, A. Ecological metabolomics: overview of current developments and future challenges. *Chemoecology* **21,** 191–225 (2011).
2. Jones, O. A. H. *et al.* Metabolomics and its use in ecology: Metabolomics in Ecology. *Austral Ecol.* **38,** 713–720 (2013).
3. Peters, K. *et al.* Current Challenges in Plant Eco-Metabolomics. *Int. J. Mol. Sci* **19,** 1385 (2018).
4. van Dam, N. M. & van der Meijden, E. A Role for Metabolomics in Plant Ecologyin *Annual Plant Reviews Volume 43* (ed. Hall, R. D.) 87–107 (Wiley-Blackwell, 2011).
5. Rivas-Ubach, A. *et al.* Are the metabolomic responses to folivory of closely related plant species linked to macroevolutionary and plant-folivore coevolutionary processes? *Ecol. Evol* **6,** 4372–4386 (2016).
6. Haug, K. *et al.* MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41,** D781–D786 (2013).
7. Gil, Y. *et al.* Examining the Challenges of Scientific Workflows. *Computer* **40,** 24–32 (2007).
8. Peng, R. D. Reproducible Research in Computational Science. *Science* **334,** 1226–1227 (2011).
9. Peters, K., Gorzolka, K., Bruelheide, H. & Neumann, S. Seasonal variation of secondary metabolites in nine different bryophytes. *Ecol. Evol.* https://doi.org/10.1002/ece3.4361 (2018).
10. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput. Biol.* **9,** e1003285 (2013).
11. Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.***18,** 530–536 (2017).
12. Cornelissen, J. H. C., Lang, S. I., Soudzilovskaia, N. A. & During, H. J. Comparative Cryptogam Ecology: A Review of Bryophyte and Lichen Traits that Drive Biogeochemistry. *Ann. Bot.* **99,** 987–1001 (2007).
13. Urmi, E. Bryophyta (Moose) in Flora indicativa*Ecological Indicator Values and Biological Attributes of the Flora of Switzerland and the Alps* 283–310 (Haupt, 2010).
14. During, H. J. Ecological classification of bryophytes and lichens in *Bryophytes and lichens in a changing environment* 1–31 (Clarendon Press, 1992).
15. Frisvoll, A. A. Bryophytes of Spruce Forest Stands in Central Norway. *Lindbergia* **22,** 83–97 (1997).
16. Smith, A. J. E. *The liverworts of Britain and Ireland* (Cambridge University Press, 1990).
17. Smith, A. J. E. *The Moss Flora of Britain and Ireland* (Cambridge University Press, 2004).
18. Böttcher, C. *et al.* The Multifunctional Enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) Converts Cysteine-Indole-3-Acetonitrile to Camalexin in the Indole-3-Acetonitrile Metabolic Network of Arabidopsis thaliana. *Plant Cell.* **21,** 1830–1845 (2009).
19. Martens, L. *et al.* mzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* **10**(R110): 000133 (2011).
20. Spicer, R. A., Salek, R. & Steinbeck, C. Compliance with minimum information guidelines in public metabolomics repositories. *Sci. Data* **4,** 170137 (2017).
21. Sansone, S.-A. *et al.* Toward interoperable bioscience data. *Nat. Genet.* **44,** 121–126 (2012).
22. Rocca-Serra, P. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26,** 2354–2356 (2010).
23. Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9,** 504 (2008).
24. Libiseller, G. *et al.* IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16,** 118 (2015).
25. Mönchgesang, S. *et al.* Natural variation of root exudates in Arabidopsis thaliana-linking metabolomic and genomic data. *Sci. Rep* **6,** 29033 (2016).
26. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84,** 283–289 (2012).
27. Nakamura, Y. *et al.* KNApSAcK Metabolite Activity Database for Retrieving the Relationships Between Metabolites and Biological Activities. *Plant Cell Physiol.* **55,** e7 (2014).
28. Asakawa, Y. *et al. Chemical constituents of bryophytes: bio- and chemical diversity, biological activity, and chemosystematics* (Springer Verlag, 2013).
29. Li, D., Heiling, S., Baldwin, I. T. & Gaquerel, E. Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. *Proc. Natl. Acad. Sci* **113,** E7610–E7618 (2016).
30. Legendre, P. & Legendre, L. *Numerical ecology* Volume 243rd edn, (Elsevier, 2012).
31. Legendre, P. & Anderson, M. J. Distance-based Redundancy Analysis: Testing Multispecies Responses In Multifactorial Ecological Experiments. *Ecol. Monogr.* **69,** 24 (1999).
32. Miksa, T., Rauber, A. & Mina, E. Identifying impact of software dependencies on replicability of biomedical workflows. *J. Biomed. Inform.* **64,** 232–254 (2016).
33. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44,** W3–W10 (2016).

DOI:10.1038/sdata.2018.179

34. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3,** 160018 (2016).
35. Piras, M. E., Pireddu, L. & Zanetti, G. wft4galaxy: a workflow testing tool for galaxy. *Bioinformatics* **33,** 3805–3807 (2017).
36. Peters, K., Gorzolka, K., Bruelheide, H. & Neumann, S. Code for the computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes using the MTBLS520 dataset (Version v1.1). Zenodo https://doi.org/10.5281/zenodo.1284246 (2018).
37. Dunning, A. C., De Smaele, M. M. E. & Böhmer, J. K. Evaluation of data repositories based on the FAIR Principles for IDCC 2017 practice paper TU Delft https://doi.org/10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f (2017).
38. Merkel, D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux Journal* https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment (2014).
39. Goecks, J., Nekrutenko, A. & Taylor, J. & Galaxy Team, T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).
40. Giacomoni, F. *et al.* Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **31,** 1493–1495 (2015).
41. Piccolo, S. R. & Frampton, M. B. Tools and techniques for computational reproducibility. *GigaScience* **5,** 30 (2016).
42. Richards, L. A. *et al.* Phytochemical diversity drives plant–insect community diversity. *Proc. Natl. Acad. Sci* **112,** 10973–10978 (2015).
43. Ristok, C. *et al.* Leaf litter diversity positively affects the decomposition of plant polyphenols. *Plant Soil* **419,** 305–317 (2017).
44. Tewes, L. J., Michling, F., Koch, M. A. & Müller, C. Intracontinental plant invader shows matching genetic and chemical profiles and might benefit from high defence variation within populations. *J. Ecol.* **106,** 714–726 (2018).
45. von Wehrden, H., Hanspach, J., Bruelheide, H. & Wesche, K. Pluralism and diversity: trends in the use and application of ordination methods 1990-2007. *J. Veg. Sci.* **20,** 695–705 (2009).
46. Field, K. J. & Lake, J. A. Environmental metabolomics links genotype to phenotype and predicts genotype abundance in wild plant populations. *Physiol. Plant.* **142,** 352–360 (2011).
47. Zuppinger-Dingley, D. *et al.* Selection for niche differentiation in plant communities increases biodiversity effects. *Nature* **515,** 108–111 (2014).

Data Citation

1. Peters, K., Gorzolka, K., Neumann, S. & Bruelheide, H. *MetaboLights* MTBLS520 (2018).

Acknowledgements

K.P. acknowledges funding from the European Commission PhenoMeNal Grant EC654241. Further, we like to thank the Leibniz Foundation for supporting this study, the PhenoMeNal e-infrastructure for supplying the cloud environment for testing the Galaxy workflow, Stefanie Döll for helping with annotation, Sylvia Krüger and Julia Taubert for technical assistance and Dierk Scheel for advice and corrections to the manuscript.

Author Contributions

K.P.: Design of the experiment, Field sampling, Statistics, Quality Control, Code, Galaxy workflow, Docker container image, Writing the first draft of the manuscript K.G.: Extraction protocol and LC/MS data acquisition H.B.: Advice on multivariate statistics S.N.: Advice on the Quality Control with XCMS All authors contributed to the final version of the manuscript.

Additional Information

Tables 1, 2 and 3 are available only in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Peters, K. *et al.* Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes. *Sci. Data* 5:180179 doi: 10.1038/sdata.2018.179 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files made available in this article.

© The Author(s) 2018

DOI:10.1038/sdata.2018.179

Received: 21 February 2018 | Revised: 12 June 2018 | Accepted: 22 June 2018
DOI: 10.1002/ece3.4361

ORIGINAL RESEARCH

WILEY Ecology and Evolution Open Access

Seasonal variation of secondary metabolites in nine different bryophytes

Kristian Peters[1] | **Karin Gorzolka**[1] | **Helge Bruelheide**[2,3] | **Steffen Neumann**[1,3]

[1]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Halle, Germany

[2]Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle Wittenberg, Halle, Germany

[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

Correspondence
Kristian Peters, Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Halle, Germany.
Email: kpeters@ipb-halle.de

Funding information
H2020 Research Infrastructures, Grant/Award Number: 654241; Leibniz-Gemeinschaft; European Commission PhenoMeNal, Grant/Award Number: EC654241

Abstract

Bryophytes occur in almost all land ecosystems and contribute to global biogeochemical cycles, ecosystem functioning, and influence vegetation dynamics. As growth and biochemistry of bryophytes are strongly dependent on the season, we analyzed metabolic variation across seasons with regard to ecological characteristics and phylogeny. Using bioinformatics methods, we present an integrative and reproducible approach to connect ecology with biochemistry. Nine different bryophyte species were collected in three composite samples in four seasons. Untargeted liquid chromatography coupled with mass spectrometry (LC/MS) was performed to obtain metabolite profiles. Redundancy analysis, Pearson's correlation, Shannon diversity, and hierarchical clustering were used to determine relationships among species, seasons, ecological characteristics, and hierarchical clustering. Metabolite profiles of *Marchantia polymorpha* and *Fissidens taxifolius* which are species with ruderal life strategy (R-selected) showed low seasonal variability, while the profiles of the pleurocarpous mosses and *Grimmia pulvinata* which have characteristics of a competitive strategy (C-selected) were more variable. *Polytrichum strictum* and *Plagiomnium undulatum* had intermediary life strategies. Our study revealed strong species-specific differences in metabolite profiles between the seasons. Life strategies, growth forms, and indicator values for light and soil were among the most important ecological predictors. We demonstrate that untargeted Eco-Metabolomics provide useful biochemical insight that improves our understanding of fundamental ecological strategies.

KEYWORDS
biochemistry, bryophytes, chemotaxonomy, ecology, ecometabolomics, environment, liverworts, mosses, phylogeny

1 | INTRODUCTION

There are approx. 20,000 bryophyte species known to science. Bryophytes are classified into three major groups: liverworts ("hepatics", *Marchantiophyta*), mosses *s. str.* ("musci", *Bryophyta*), and hornworts (*Anthocerophyta*) (Bowman et al., 2017; Goffinet & Shaw, 2009; Qiu et al., 2006; Shaw, Szovenyi, & Shaw, 2011). They occur in nearly every land ecosystem (Vanderpoorten & Goffinet, 2009).

Bryophytes contain many unique chemical compounds with high biological and ecological relevance (Asakawa, Ludwiczuk, & Nagashima, 2013a). Due to unique oil bodies, liverworts are biochemically very distinctive from other mosses. Secondary metabolites in

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2018 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

DOI:10.1002/ece3.4361

oil bodies are mostly composed of lipophilic terpenoids, abundant (bis-)bibenzyls, and small aromatic compounds (Asakawa et al., 2013a). Liverworts represent a phylogenetic group of plants that were the first colonizers of land; thus, they share many biochemical features of both algae and land plants (Bowman et al., 2017). It has been acknowledged that there must have been many biochemical innovations involved during evolution from water to land (He, Sun, & Zhu, 2013; Suire et al., 2000). Even though oil bodies in *M. polymorpha* are usually restricted to only few vegetative cells of the thallus, relative number of oil bodies has been correlated to growth conditions, availability of nutrients, level of plant-herbivory, and biodiversity (Tanaka et al., 2016). The compounds unique to liverworts are involved in many biotic interactions and act as defense to herbivory (Asakawa, Ludwiczuk, & Nagashima, 2013b).

Despite the fact that the majority of bryophytes (approx. 14,000 species) belong to the group of mosses (Bryophyta), fewer compounds have been characterized in mosses than in liverworts. Mosses contain terpenoids; benzoic, cinnamic, and phthalic acid derivatives; coumarins; and some nitrogen-containing aromatic compounds, which sometimes are structurally similar to those found in vascular plants (Asakawa et al., 2013a).

As secondary metabolite profiles are similar among phylogenetically closely related species (Maksimova, Klavina, Bikovens, Zicmanis, & Purmalis, 2013; Wink, 2003; Wu, 1992), metabolomics can also be used to support phylogenies based on genetic markers, for example, to find marker compounds to assist current phylogenetic classifications, to discriminate several ecotypes of bryophyte species, or even to propose new chemical taxonomic markers (Heinrichs, Anton, Gradstein, & Mues, 2000; Pejin et al., 2010; Rycroft, Heinrichs, Cole, & Anton, 2001).

Several hundred new compounds have been isolated from bryophytes in recent years. Species produce secondary metabolites as a defense against mechanical damage, environmental stress, herbivores, and pathogens, as well as to capture and conserve resources (Cornelissen, Lang, Soudzilovskaia, & During, 2007). However, there is still a knowledge gap with regard to the ecological relevance of compounds (Asakawa et al., 2013b).

Bryophytes exhibit allelopathic interactions with other organisms by releasing allelochemicals. For example, as some slugs feed on bryophytes, mosses such as *Dicranum scoparium* have evolved acetylic oxylipins that act as a defense against herbivorous slugs (Boch, Prati, & Fischer, 2016; Rempt & Pohnert, 2010). Other oxylipins or related compound classes have also been found to induce defense reactions in vascular plants. In this context, several studies found both inhibition and facilitation effects of bryophytes on seed germination and seedling growth of vascular plants (Donath & Eckstein, 2010; Michel, Burritt, & Lee, 2011; Zamfir, 2000). In addition, positive and negative effects of bryophytes on species diversity have been described. As a result, the effect of bryophytes on diversity cannot be generalized as it has been found to depend on the type of habitat and environmental conditions (Ehlers, Damgaard, & Laroche, 2016; Gornall, Woodin, Jónsdóttir, & van der Wal, 2011; Hüllbusch, Brandt, Ende, & Dengler, 2016; Jeschke & Kiehl, 2008; Müller et al., 2012).

Despite their small size, bryophytes show remarkable biochemical adjustments to environmental changes (During, 1992; Klavina, 2015). For example, bryophyte species that occur as colonizers in early successional stages collect debris, store water, and deposit and solidify soil. Thus, bryophytes can reduce erosion and often act as prerequisite for establishing vascular plants by creating microhabitats (Streitberger, Schmidt, & Fartmann, 2017; Zamfir, 2000). In late successional stages in grasslands, even low bryophyte abundances can facilitate the regeneration of vascular plants by influencing nutrient retention and water cycling (Virtanen, Eskelinen, & Harrison, 2017). However, the net outcome is often depending on environmental conditions (Doxford, Ooi, & Freckleton, 2013).

There are many studies that link the abundance and the distribution of bryophytes with the environment (Aranda et al., 2014; Smith, 1982). Altitudinal gradients were often used to study the effects of seasons and environments in combination (Mateo et al., 2016; Sun et al., 2013; Wagner, Zotz, Salazar Allen, & Bader, 2013). However, there are only few studies that analyzed the biochemical responses of bryophytes to different environments or seasons. For example, studies with the liverwort *Conocephalum conicum* revealed largely different metabolite profiles of morphologically mostly indistinguishable specimen that were collected in contrasting environments (Ghani, Ludwiczuk, Ismail, & Asakawa, 2016; Ludwiczuk, Odrzykoski, & Asakawa, 2013). A different study analyzed three leafy liverwort species and found seasonal variation in antioxidant and polyphenol oxidase enzymes, as well as in the flavonoid and phenolic content (Thakur & Kapila, 2017).

Bryophytes have adopted different types of ecological strategies (During, 1992; Frisvoll, 1997) (Table 1). Grime (1977) described three basic types of life strategies for plants (the so-called CSR triangle). Competitive species (C-selected) show high nutrient turnover, large relative growth rates, morphological plasticity, a long life span, and usually low reproduction (During, 1992). They are typically found in late successional habitats. The S-selected group consists of stress-tolerant species that are slowly growing, have a conservative nutrient uptake, and are usually found in habitats that have abiotic constraints, for example, limited resource availability. Many ruderal species are R-selected and have traits related to fast growth, rapid nutrient uptake, high reproduction, and a short life span (Ayres, van der Wal, Sommerkorn, & Bardgett, 2006). They are usually found in early successional habitats and are quickly overgrown by competitors. There are also many species with intermediary strategies, especially epiphytic and epilithic bryophytes (During, 1992; Frisvoll, 1997).

Many morphological and physiological relationships have been described to be correlated with these plant strategy types (e.g., leaf area, growth, and photosynthesis), including the capabilities of bryophytes that drive biogeochemical processes (Caccianiga, Luzzaro, Pierce, Ceriani, & Cerabolini, 2006; Cornelissen et al., 2007; Grime, Rincon, & Wickerson, 1990). Linking metabolites to plant strategy theory contributes to a mechanistic understanding of how bryophytes are able to, for example, tolerate desiccation biochemically and are still able to grow under dry and cool conditions (Grime et al., 1990).

DOI:10.1002/ece3.4361

TABLE 1 Life history characteristics of the bryophytes used in the study were collected from the literature

Code	Species	Family	Type	Growth form	Habitat type	Substrate	Life strategy	Gametangia distribution	Mean spore size [μm]	Sexual reproduction frequency	Light index	Temperature index	Continentality index	Moisture index	Reaction index	Nitrogen index	Life-form index
Brarut	*Brachythecium rutabulum*	Brachytheciaceae	Pleurocarpous	Mat	Woods, Shrubs	Soil, Firm rocks	Perennial stayer competitive	Autoicous	20	Common	5	5	5	4	5	9	C,(E)
Calcus	*Calliergonella cuspidata*	Amblystegiaceae	Pleurocarpous	Mat	Meadows, Herbaceous	Soil, Turf	Perennial stayer competitive	Dioicous	20	Occasional	8	3	5	7	7	8	C
Fistax	*Fissidens taxifolius*	Fissidentaceae	Acrocarpous	Turf	Woods, Shrubs	Soil	Colonist	Autoicous	15	Occasional	5	4	5	6	7	5	H
Gripul	*Grimmia pulvinata*	Grimmiaceae	Acrocarpous	Cushion	Exposed Rocks	Firm rocks	Pioneer	Autoicous	10	Very common	8	5	5	1	7	7	C
Hypcup	*Hypnum cupressiforme* s. l.	Hypnaceae	Pleurocarpous	Mat	Woods, Shrubs	Dead wood, Bark	Perennial stayer stress-tolerant	Dioicous	14	Common	5	5	5	4	4	8	C, E
Marpol	*Marchantia polymorpha* s. l.	Marchantiaceae	Liverwort	Thalloid	Ruderal, Banks	Soil, Loose rocks	Colonist	Dioicous	14	Common	8	5	5	6	5	8	T
Plaund	*Plagiomnium undulatum*	Mniaceae	Acrocarpous	Dendroid	Woods, Shrubs	Soil	Long-lived shuttle	Synoicous	28	Rare	4	3	5	6	6	7	H, C
Polstr	*Polytrichum strictum*	Polytrichaceae	Acrocarpous	Turf	Woods, Shrubs	Turf, Soil	Perennial stayer competitive	Dioicous	16	Common	8	2	6	6	1	4	H
Rhysqu	*Rhytidiadelphus squarrosus*	Hypnaceae	Pleurocarpous	Mat	Meadows, Herbaceous	Soil	Perennial stayer competitive	Dioicous	19	Rare	7	3	6	6	5	7	C

Note. Family and type are based on the taxonomic classification found in Smith (1990, 2004); The characteristics “growth form,” “habitat type,” and “substrate” were added from the tables in Urmi (2010); “life strategy” is based on the classification of During (1992) and was added from tables in Frisvoll (1997); “spore size,” “gametangia distribution,” and “sexual reproduction frequency” were collected from Smith (1990, 2004); Ellenberg indicator values (light, temperature, continentality, moisture, reaction, nitrogen, and life-form indices) were added from Urmi (2010).

DOI:10.1002/ece3.4361

Recent advances in analytical methods (e.g., liquid chromatography coupled with mass spectrometry—LC/MS) allow to simultaneously measure most semipolar metabolites of an organism at once in an untargeted way (without specifically targeting some known compounds). In an ecological context, this is known as Eco-Metabolomics (Hall, 2006; Sardans, Peñuelas, & Rivas-Ubach, 2011). When compared to typical biochemical experiments, where plants are usually grown under controlled conditions in glasshouses or growth chambers, in Eco-Metabolomics, metabolite profiles are typically acquired from wild plant species in their natural environment (van Dam & van der Meijden, 2011; Rivas-Ubach et al., 2016; Sardans et al., 2011). As a result, experiment designs are more complex and metabolite profiles are expected to be highly variable.

Discovering patterns in the metabolite profiles can reveal new ecological and biogeochemical relationships as the biochemistry of bryophytes is related to the environment, climate, and biotic interactions (Sardans et al., 2011). For example, metabolite profiling of higher plants grown in field plots showed that resource limitation results in decreased performance of small-statured herbs with increasing species diversity (Scherling, Roscher, Giavalisco, Schulze, & Weckwerth, 2010). Multivariate statistical methods such as principal components analysis (PCA) allow to discriminate species based on their metabolite profiles. Furthermore, profiles can also be used to discriminate species that were grown in different environments or had a history of different ecological interactions (van Dam & van der Meijden, 2011; Hall, 2006; Jones et al., 2013).

Studying the biochemistry of bryophytes is often targeting the discovery of novel potentially active compounds and natural product chemistry (Asakawa et al., 2013a). We have found only a few studies in the literature that performed untargeted metabolomics analyses (LC/MS, GC/MS, NMR) with bryophytes, and none that were performed in an ecological context (e.g., Erxleben, Gessler, Vervliet-Scheebaum, & Reski, 2012; Klavina, 2015; Pejin et al., 2010; Rycroft et al., 2001).

In this study, we introduce an integrative Eco-Metabolomics approach to connect biochemistry with ecology using bioinformatics methods (Hall, 2006; Sardans et al., 2011). The aims of this study are as follows: (a) to investigate metabolic differences between species as explained by ecological characteristics, in particular, with regard to the CSR life strategy types; (b) to determine biochemical differences in species profiles with regard to the seasons; (c) to find out how the metabolomes of the bryophytes reflect their phylogeny; and (d) to present a reproducible bioinformatic workflow that can be reused by other subsequent Eco-Metabolomics studies.

2 | MATERIALS AND METHODS

2.1 | Field campaign and sampling

Samples of the nine moss species, *Brachythecium rutabulum* (Hedw.) Schimp., *Calliergonella cuspidata* (Hedw.) Loeske, *Fissidens taxifolius* Hedw., *Grimmia pulvinata* (Hedw.) Sm., *Hypnum cupressiforme* Hedw. *s.l.*, *Marchantia polymorpha* L., *Plagiomnium undulatum* (Hedw.) T.J. Kop., *Polytrichum strictum* Menzies ex Brid., and *Rhytidiadelphus squarrosus* (Hedw.) Warnst., were collected in the Botanical Garden of Martin Luther University Halle-Wittenberg, Germany (see Supporting Information Figure S4 for photographs of the species). Sampling was performed in summer (2016/08/08), autumn (2016/11/09), winter (2017/01/27), and spring (2017/05/11) under stable weather conditions with sunshine at least 2 days prior to sampling and during sampling. Sampling was conducted between 13:00 and 15:00.

Three composite samples of different individuals of each species were taken in each season, leading to a total of 3 × 9 × 4 = 108 samples. Only aboveground parts of the moss gametophytes were taken for sampling. Visible archegonia or antheridia, sporophytes, and any belowground parts were removed with a sterile tweezer before sampling. The gametophytic moss parts were put in Eppendorf tubes and were frozen instantly on dry ice. Life strategies and other life characteristics were collected from the literature (Table 1).

2.2 | Biochemical protocol

Frozen moss samples were extracted according to Böttcher et al. (2009) with the following modifications: After adding 200 mg of ceramic beads (0.5 mm diameter, Roth), samples were homogenized with a tissue homogenizer (2 × 20 at 6,500 rpm; Precellys® 24, Bertin Technologies, Montigny-le-Bretonneux, France). 1 ml ice-cold 80/20 (v/v) methanol/water was added. Metabolites were extracted by shaking/ultrasonification/shaking for 15 min at 1000 rpm. After 15 min centrifugation at 15,000 g (rcf), 500 μl of supernatant was dried in a vacuum centrifuge at 40°C and reconstituted in 80/20 (v/v) methanol/water with the volume adjusted to the initial fresh weight of the sample to a final concentration of 10 mg fresh weight per 100 μl extract.

Ultra-performance liquid chromatography (Waters Acquity UPLC equipped with a HSS T3 column (100 × 1.0 mm)) coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (UPLC/ESI-QToF-MS) was performed using a high-resolution MicrOTOF-Q II hybrid quadrupole time-of-flight mass spectrometer (Bruker Daltonics), as described in Böttcher et al. (2009). Data were acquired in centroid mode with the following MS instrument settings for positive mode: nebulizer gas: nitrogen, 1.4 bar; dry gas: nitrogen, 6 L/min, 190°C; capillary:, 5,000 V; end plate offset: −500 V; funnel 1 radio frequency (RF): 200 Volts peak-to-peak (Vpp); funnel 2 RF: 200 Vpp; in-source collision-induced dissociation (CID) energy: 10 eV; hexapole RF: 100 Vpp; quadrupole ion energy: 3 eV; collision gas: nitrogen; collision energy: 7 eV; collision cell RF: 250 Vpp; transfer time: 70 μs; prepulse storage: 5 μs; pulser frequency: 10 kHz; and spectra rate: 3 Hz.

2.3 | Data analyses

Raw LC/MS data were converted to the open data format mzML with the software Bruker CompassXPort 3.0.9. Raw data and metadata

DOI:10.1002/ece3.4361

were published in the metabolomics repository MetaboLights as MTBLS520 (Haug et al., 2013; Peters, Gorzolka, Bruelheide, & Neumann, 2018). A computational workflow was constructed in the Galaxy workflow management system for the entire data processing pipeline of this study (Supporting information Figure S3). Required software tools, their dependencies, as well as software libraries and R packages were containerized using Docker technology to facilitate reusability on different computational environments. Source code was made publicly available on GitHub (Peters et al., 2018).

Profiles of positive mode were used for the data analyses as many important and known secondary metabolites classes in bryophytes such as flavonoids, phenylpropanoids, anthocyanins, glycosides, and previously characterized compounds such as marchantins, communins, and ohioensins ionize well in positive mode with our instrumental setup.

Detection of chromatographic peaks was performed in R with the package XCMS 1.52.0 (Tautenhahn, Bottcher, & Neumann, 2008) with two grouping factors in "phenoData": seasons (summer, autumn, winter, spring) and species (Brarut, Calcus, Fistax, Gripul, Hypcup, Marpol, Plaund, Polstr, Rhysqu). Quality control was performed with a laboratory internal standard mix (Peters et al., 2018). As the quality control revealed no significant differences between batches, no additional corrections on the peak detection with XCMS were performed. Intensities in the peak table were log transformed before grouping. For further analysis, only features between the retention times 20 and 1,020 were kept.

Adduct annotation was performed with the package CAMERA 1.33.3 (Kuhl, Tautenhahn, Böttcher, Larson, & Neumann, 2012). A specific function getReducedPeaklist was written (method = median) that aggregates the adducts of putative compounds into a feature matrix with singular components in order to improve subsequent statistical analyses (Peters et al., 2018).

Statistical analyses were performed in R 3.4.2 using the additional packages: multtest, RColorBrewer, vegan, multcomp, multtest, nlme, ape, pvclust, dendextend, phangorn, Hmisc, gplots, and VennDiagram. A presence–absence matrix was generated from the feature matrix to determine the differences in metabolite features between the experimental factors species and season. In concordance with the "minfrac" parameter in the alignment step in XCMS, a feature was considered present if it was detected in two out of three replicates. The presence–absence matrix was used for measuring the biochemical diversity by calculating the Shannon index for each sample using the function "diversity" in vegan (Li, Heiling, Baldwin, & Gaquerel, 2016). The total number of features and the number of unique features were calculated from the presence–absence matrix accordingly.

To test factor levels for significant differences, the Tukey HSD on a one-way ANOVA was performed post hoc using the multcomp package. Intraspecific variability of species profiles in response to the seasons was calculated with the Pearson correlation coefficient (Pearson's r) on the presence–absence matrix using the function "rcorr" in the package Hmisc. Venn diagrams were created for each species separately using the package VennDiagram.

Variation partitioning was performed using the function "varpart" in the package vegan to analyze the influence of the factors species and seasons on the metabolite profiles. Distance-based redundancy analysis (dbRDA) using the function "capscale" with Bray–Curtis distance and multidimensional scaling in the package vegan was chosen to analyze the relation of the ecological characteristics with the species metabolite profiles (Legendre & Anderson, 1999). Ordinal and categorical ecological characteristics were transformed to the presence–absence matrices for the ordination. The model for the dbRDA was chosen with forward and backward selection using the function "ordistep" in the package vegan. Ecological characteristics were added to the plots as post hoc variables using the function "envfit" in the package vegan.

Relationships between metabolite profiles and phylogeny were analyzed by calculating Bray–Curtis distances for phylogeny and the feature matrix (function "vegdist" in vegan) followed by hierarchical clustering (function "hclust) with the complete linkage method. The chemotaxonomic plot was reordered using the function "order.optimal" (package cba), and branches of *P. strictum* and *P. undulatum* were swapped using the function "reorder" in vegan. The similarity of the two trees was determined with the normalized Robinson–Foulds metric (function "RF.dist" in package phangorn). The similarity of the distance matrices was determined with the Mantel statistic (function "mantel" in vegan).

More detailed methods and further information on the computational workflow are described in Peters et al. (2018).

3 | RESULTS

Preprocessing of the LC/MS raw data with XCMS and CAMERA (see Materials and Methods) resulted in a feature matrix with 108 samples and 4,032 features. The corresponding data table is available in MetaboLights and was also used for biostatistics and for the components of the entire computational workflow (Peters et al., 2018).

3.1 | Diversity of metabolite features between the species

Marchantia polymorpha had significantly more biochemical features than the other species with our analytical setup (Supporting Information Table S1). In general, we observed fewer features in pleurocarpous than in acrocarpous species (Figure 1a and b, Supporting information Table S1). The relationships were also reflected in the Shannon index for the species (Figure 1a). Further, *M. polymorpha* was the species in which significantly more unique features were detected (131 ± 18) (Figure 1b). The pleurocarpous species had fewer unique features (25 ± 14) than the acrocarpous species (59 ± 17) (indicated green vs. red colors in Figure 1b; Supporting information Table S1). *M. polymorpha* and *P. undulatum* had significantly higher metabolic content per extracted gram fresh weight than the other species (Figure 1c).

DOI:10.1002/ece3.4361

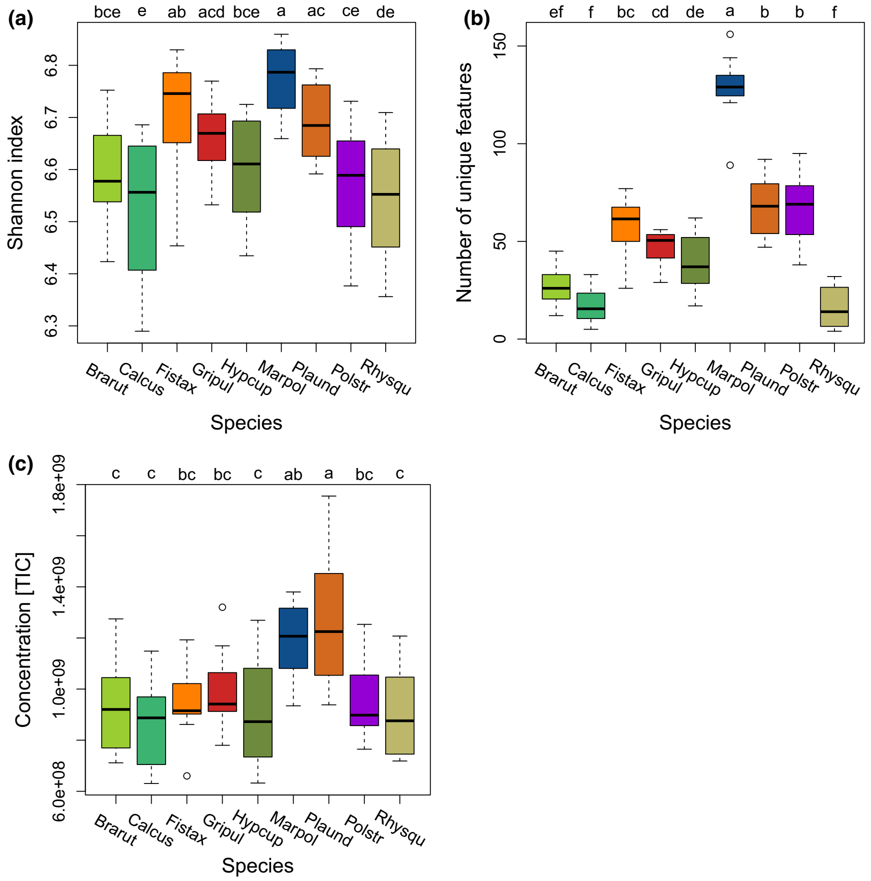


FIGURE 1 The diversity of biochemical features of the metabolite profiles of the nine bryophyte species. (a) Shannon diversity indices (H') for the total number of features present in the species profiles. (b) Number of unique features that were exclusively present in one of the nine species. (c) Total intensities of features (= sum of total ion current) for the species. Groups for each species were calculated with performing post hoc Tukey HSD on a one-way ANOVA. *n* = 12 for each species

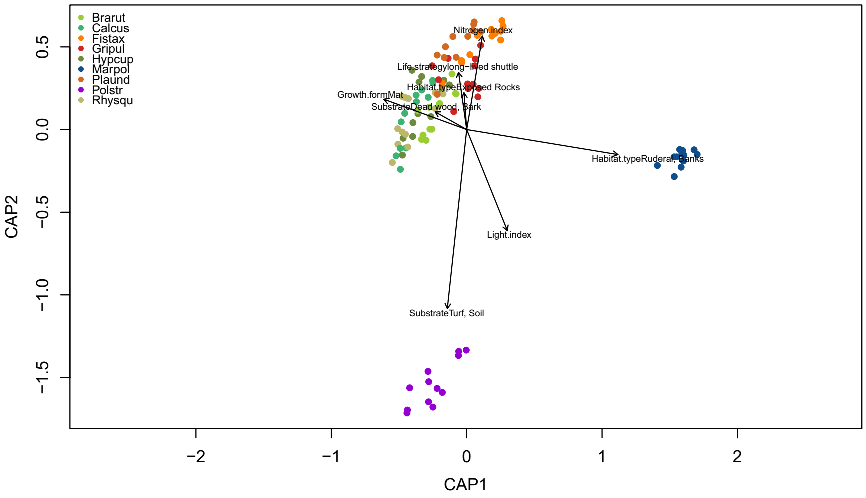


FIGURE 2 dbRDA plot of species samples (colored scores) and ecological characteristics (arrows). The length of the arrows represents the explanation power of the characteristics for the features in the matrix of metabolite profiles. The relative position of the samples to the direction of the axis describes the relationship of the sample with the characteristic. The two axes of the plot explain a total variation of 48.7% in the feature matrix. *n* = 108 samples

3.2 | Metabolic differences between species related to ecological characteristics

Variation partitioning revealed that species identity accounted for 33% of the variation in the feature matrix and seasonal effects for 9% (Supporting Information Figure S1). Distance-based redundancy analysis (dbRDA) was performed to assess the relation between ecological characteristics (Table 1) and the metabolite features of the species (Figure 2). Model selection resulted in a model of eight characteristics which explained 48.7% of the variation in the species metabolite profiles (Figure 2).

Habitat type "ruderal, banks" was responsible for the separation of *M. polymorpha* in the plot. The substrate "turf" (turfs and soils characterized by low pH) was the most powerful predictor for *P. strictum* (Figure 2). The dbRDA suggested nonlinear relationships of several indicator values with the metabolite profiles of the species. Model selection included light and nitrogen index in the model (Table 1). Profiles of *F. taxifolius* and *G. pulvinata* were correlated to the "nitrogen" indicator value. Habitat type "exposed rocks" was a powerful predictor for the epilithic *G. pulvinata*, whereas profiles of *P. undulatum* were correlated to the life strategy "long-lived shuttle". Growth form "mat" was the main predictor for the pleurocarpous mosses (green colored scores in Figure 2).

3.3 | Biochemical differences in species profiles with regard to the seasons

The total number of features present in summer (856 ± 48) was significantly higher in all species than in the seasons autumn (748 ± 108), winter (738 ± 98), and spring (762 ± 42). This was reflected by the Shannon index (Figure 3a), but not by the number of unique features in the seasons (Figure 3b). The Venn diagrams break down the proportions for each species separately (Supporting Information Figure S2). Total metabolic extracts (TIC) were also significantly higher in summer than in the other seasons (Figure 3c).

The dbRDA using seasons as constrained variables explained 14.8% of the variation present in the feature matrix. Seasons were clearly distinct from each other (Figure 4). The dbRDA shows that metabolite profiles from autumn and winter were more similar than those from spring and summer (Figure 4). The pleurocarpous species (filled symbols in Figure 4) were less separated than the acrocarpous species. These results are in line with the number of unique features in the different species per season (Venn diagrams in Supporting Information Figure S2).

The metabolite profiles of *M. polymorpha*, *F. taxifolius*, and *P. strictum* had significantly larger Pearson Correlation Coefficients. This means that the profiles with regard to the number of features were less variable among seasons than those of the other species (Figure 3d). This lower variation among seasons is also seen in the Venn diagrams, which show the number of features that are distinct and shared between all possible combinations of the seasons and for each species separately (Supporting Information Figure S2). In contrast to the acrocarpous species, the pleurocarpous species had more distinct features between the seasons, but less shared features across the seasons.

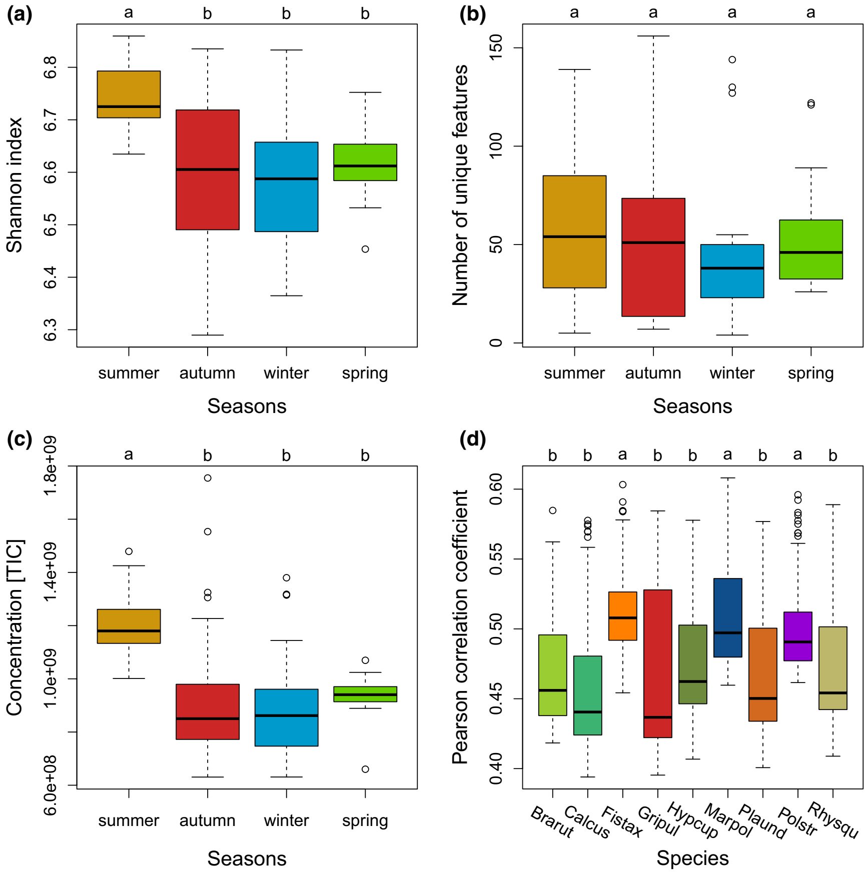


FIGURE 3 The diversity of biochemical features in the four seasons. (a) Shannon diversity indices (H') for the total number of features present in the seasons. (b) Number of unique features that were exclusively present in one of the four seasons. (c) Total intensities of features (= sum of total ion current, TIC) per season. (d) Pearson's correlation coefficients (PCC) that show the intraspecific variability of the profiles of the species in response to the seasons. The lower the PCC values are, the more dissimilar they are, meaning higher difference in the number of features between the seasons. Groups were calculated with performing the Tukey HSD post hoc on a one-way ANOVA. *n* = 12 for each species

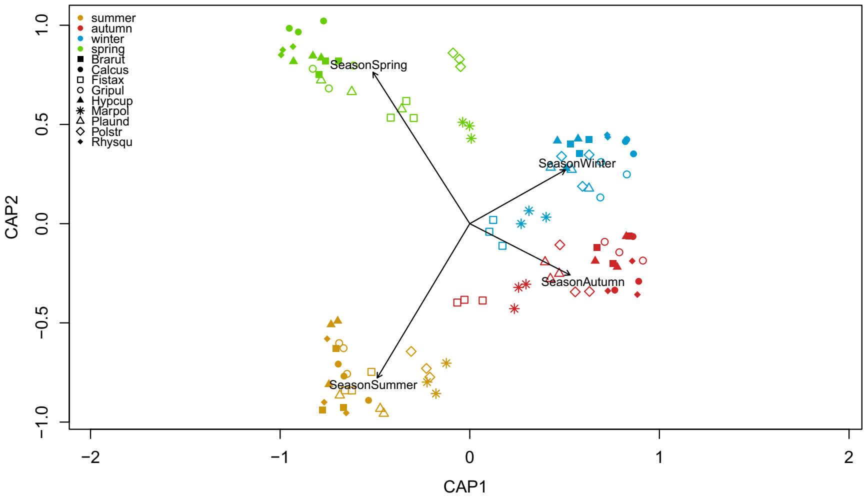


FIGURE 4 Constrained dbRDA plot of samples (colored scores) to the seasons (arrows). The length of the arrows represents the explanatory power of the season for the metabolite features. The position of the samples relative to the direction of the arrow represents the relationship of the sample with the season. The first two axes of the plot explain a total variation of 14.8% in the feature matrix. *n* = 108 samples

3.4 | Relationships of metabolite profiles and phylogeny

In accordance with the phylogenetic tree (Figure 5a), *M. polymorpha* and *P. strictum* were identified by clustering based on metabolite features as the two most basal species with largest distances (Figure 5b). In contrast to the phylogeny, where *P. undulatum* was closer related to the group of pleurocarps than to *G. pulvinata* and *F. taxifolius*, *P. undulatum* was more dissimilar with regard to metabolite features than the other species in this clade (Figure 5b). This resulted in a higher intergroup dissimilarity of the clade.

The pleurocarpous species also formed a clade in the chemotaxonomic tree, but with different distances as in the phylogenetic tree. Comparing the two trees showed a normalized Robinson–Foulds similarity of 0.57 (where a value of 0 means total similarity and 1 means no similarity) and comparing the distance matrices of the two trees resulted in a Mantel statistics of 0.39 (Figure 5a and b).

4 | DISCUSSION

A bioinformatic workflow was created that can be run to reproduce the results from this study (Supporting Information Figure S3). It can be reused by Eco-Metabolomics studies with a comparable approach and with different data. Overall, our analyses revealed strong species-specific differences in the metabolite profiles between the seasons, which could be related to the ecology of the bryophytes.

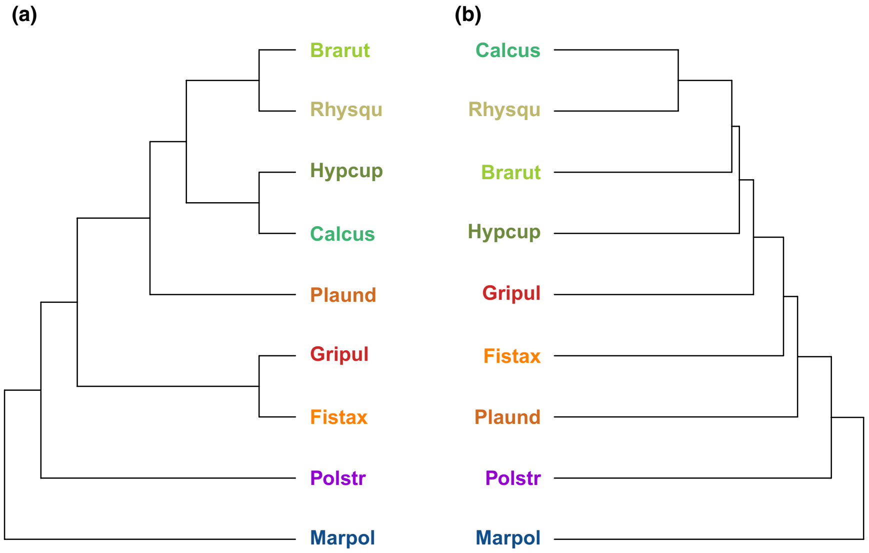


FIGURE 5 Hierarchical clustering of the bryophyte species. (a) Phylogenetic tree constructed from the phylogenetic distances of the species showing the taxonomic relationships of the bryophytes. (b) Chemotaxonomic tree resulting from hierarchical clustering of the species metabolite profiles. Height specifies the distances between the nodes

4.1 | Bioinformatic workflow

The Galaxy workflow management provides an easy to use graphical user interface which runs in different software environments and can be operated via a web browser (Afgan et al., 2016). Our computational workflow implements the entire data processing pipeline ranging from preprocessing the metabolite profile data to multivariate statistics (Figure S3) (Peters et al., 2018). Each analysis is represented by a dedicated module in Galaxy and can be run independently to give identical results in different software environments. More importantly, modules can be adapted to other use-cases and reused with other metabolomics data by utilizing the code which has been made available as open source (Peters et al., 2018).

Most Eco-Metabolomics studies relate metabolite profiles to growth, stress, environment, diversity, interactions, and even geographical regions (e.g., van Dam & van der Meijden, 2011; Fester, 2015; Sardans et al., 2011; Scherling et al., 2010; Szakiel, Pączkowski, & Henry, 2011). However, comparative studies that link ecological characteristics with metabolomics are still widely missing. A comparable methodological approach was made by Frisvad, Andersen, and Thrane (2008) who related diversity in secondary metabolite profiles of filamentous fungi to life strategies. Ivanišević, Thomas, Lejeusne, Chevaldonné, and Pérez (2011) analyzed metabolic fingerprints of sponges and linked them to metabolite diversity.

With our computational workflow, we address typical challenges in Eco-Metabolomics by analyzing data tables (one for the metabolite feature matrix and one data matrix for the ecological characteristics) conjointly with suitable statistical methods commonly used in ecology (Legendre & Legendre, 2012). As our approach follows the FAIR guiding principles for data management and stewardship (Wilkinson et al., 2016), we facilitate the reuse of our workflow by other subsequent Eco-Metabolomics studies.

4.2 | Relationships of metabolite diversity and phylogeny

The liverwort *Marchantia polymorpha* had significantly higher diversity of metabolite features than the other mosses with our analytical setup. This can be explained by oil bodies which are unique to liverworts and are known to contain many specialized secondary metabolites such as flavonoids, phenylpropanoids, anthocyanins, and glycosides that deter pathogens and herbivores (Bowman et al., 2017; Suire et al., 2000; Tanaka et al., 2016). In the metabolite profiles of *M. polymorpha*, we annotated many known compounds which are described as unique to liverworts in the literature (Asakawa et al., 2013a; Peters et al., 2018). The distant metabolite profiles explain also the most basal position and the largest distance of *M. polymorpha* in chemotaxonomic clustering.

The chemotaxonomic distance of *P. strictum* may be related to recent evolutionary developments such as secondary cell structures (Ligrone, Carafa, Duckett, Renzaglia, & Ruel, 2008). For example, although lignin is already present in *M. polymorpha*, its function as desiccation protective substance is less effective than in mosses where it is embedded in secondary cell structures (Ligrone et al., 2008).

In general, the dissimilarities between the phylogenetic and the chemotaxonomic tree were likely the result of different life strategies and biochemical responses of the bryophytes to the specific conditions prevalent in the habitat and may ultimately result from the differential expression of corresponding genes (Wink, 2003). This was especially evident for *P. undulatum* and could further be explained by the large separation in the dbRDA. The branch with pleurocarpous mosses represents a relatively young phylogenetic clade which can, in part, explain the weak biochemical separation of the pleurocarpous species from the others (Shaw, Cox, Goffinet, Buck, & Boles, 2003).

4.3 | Metabolic differences between species as explained by ecological characteristics

We identified two groups of bryophytes whose metabolite profiles were either R- or C-selected (During, 1992; Grime, 1977).

The R-selected group was composed of *M. polymorpha* and *F. taxifolius*. These species had significantly more features and were significantly less variable across seasons than the other bryophyte species. These results suggest that these species rely on only a few metabolic adjustments with regard to the seasons. The two species also have ruderal characteristics such as being adaptive to the conditions in disturbed areas, fast growth and loosely growth forms, high reproduction, and being quickly overgrown by other plants with progressing succession (Frisvoll, 1997; Grime, 1977; Hedwall, Skoglund, & Linder, 2015).

Furthermore, in ruderal habitats, there could be fewer mycorrhizal associations of bryophytes and fungi as in late successional habitats (Chapin, Walker, Fastie, & Sharman, 1994). Accordingly, for the genome of *M. polymorpha* it was found that some gene families were missing that were described to be required for successful mycorrhizal associations (Bowman et al., 2017). These findings could partly explain the relatively large inventory of different metabolites that is expressed consistently throughout the whole year.

The C-selected group included all tested pleurocarpous species *B. rutabulum*, *C. cuspidata*, *H. cupresiforme*, *R. squarrosus*, and the epilithic species *G. pulvinata*. They had low metabolite diversity, but—more significantly—showed a high seasonal variability of metabolites and, thus, produced many different features only seasonally. Except the epilithic *G. pulvinata*, species in this group were categorized as competitive (C-selected) in the literature (Frisvoll, 1997).

Our results suggest that species in this group are specialized to the conditions in late successional stages with regard to their biochemistry, as well as to grow in mats or cushions and to have high relative growth rates in order to withstand the competition from

DOI:10.1002/ece3.4361

vascular plants (During, 1992; Hedwall et al., 2015; Virtanen et al., 2017). Producing metabolites only on demand seems to be favorable for bryophyte species in late successional stages.

Grimmia pulvinata was categorized as pioneer by Frisvoll (1997), and as such, it should be R-selected. However, our metabolomic data suggest that it realizes a C-selected strategy. When only considering rocks or stones as immediate habitat, the species is very competitive to other species as it usually grows solitary.

The metabolite profiles of *Polytrichum strictum* showed an intermediary R- and S-selected strategy, whereas the profiles of *Plagiomnium undulatum* showed evidence for C- and S-selection. Profiles of *P. strictum* had relatively low total number of metabolite features but a high number of unique features and made little metabolic adaptations across the seasons. By contrast, profiles of *P. undulatum* had many unique and relatively high numbers of metabolites that did change considerably between the seasons. This is in accordance with the plant strategy theory which explicitly describes transitions between the different life strategies (During, 1992; Grime, 1977). According to results of Wang, Bader, Liu, Zhu, and Bao (2017), the intermediary life strategies of *Polytrichum* and *Plagiomnium* may be explained by specialized traits related to photosynthesis and growth forms.

4.4 | Biochemical differences in species profiles with regard to the seasons

The total number of features present in summer was significantly higher than in the other seasons in any species. This can generally be explained by biological activities that are more intense during summer (Doxford et al., 2013; Lambers, Chapin, & Pons, 2008; Rousk, Pedersen, Dyrnum, & Michelsen, 2017; Thakur & Kapila, 2017). With our experimental setup, we could not measure interactions with other organisms. However, in the literature, it is also described that ecological interactions are also more manifold in the summer season in temperate regions (Grime, 1977; Lambers et al., 2008).

Bryophytes often respond sensitively to sudden climatic changes. Hence, they are considered good indicators for environmental changes (Gignac, 2001; Gilbert, 1968). It is likely that the profiles of the bryophytes we measured during summer contained also many protective substances such as sugars or polyphenols to tolerate desiccation (Erxleben et al., 2012; Garcia, Rosenstiel, Graves, Shortlidge, & Eppley, 2016; He et al., 2013; Proctor et al., 2007). However, we suggest to use additional LC/MS-MS or NMR to identify significant metabolite features in order to make conclusions at the mechanistic level (Sardans et al., 2011).

Our results suggest that bryophytes respond species-specifically to different seasonal conditions. The responses of bryophytes to seasons are not only depending on their ecology and the type of life strategy (see above). They are also seemed to be determined by their phylogenetic history, as metabolite profiles of pleurocarpous species were less well distinguished from those of phylogenetically more distant acrocarpous species.

5 | CONCLUSION

We found that seasonal changes have great impact on the biochemistry of bryophytes and that the tested bryophytes realize common as well as species-specific biochemical adjustments to the different conditions prevalent in the seasons. We further found that metabolite profiles were driven by the particular ecological characteristics and life strategies such as growth form, light availability, nutrient supply, and pH soil value. With regard to seasonal changes, the biochemistry of bryophytes is still largely unexplored. Our results warrant further biochemical investigation of bryophytes and to study relationships with ecological characteristics, life strategies, and phylogeny. With this study, we present first evidence that bryophytes realize life strategies that follow plant strategy theory by Grime (1977) at the biochemical scale. Our results demonstrate that untargeted Eco-Metabolomics are useful to answer fundamental questions in ecology and that the ecological strategy concepts also apply to biochemical scales.

ACKNOWLEDGMENTS

KP acknowledges funding from the European Commission PhenoMeNal Grant EC654241. Further, we like to thank the Leibniz Foundation for supporting this study, Stefanie Döll for helping with annotation, Sylvia Krüger and Julia Taubert for technical assistance, and Dierk Scheel for advice and corrections to the manuscript.

CONFLICT OF INTEREST

None.

AUTHOR CONTRIBUTIONS

Kristian Peters designed the experiment, participated in field sampling and collection, performed data analysis, and wrote the first draft of the manuscript. Karin Gorzolka contributed to extraction protocol and LC/MS data acquisition. Helge Bruelheide provided advice on multivariate statistics. Steffen Neumann provided advice on the bioinformatics pipeline. All authors contributed to the final version of the manuscript.

DATA ACCESSIBILITY

Raw Metabolite profiles, metabolite feature matrices, and metadata: MetaboLights MTBLS520 (https://www.ebi.ac.uk/metabolights/MTBLS520). Computational workflow code version 1.1: Zenodo https://doi.org/10.5281/zenodo.1284246

ORCID

Kristian Peters http://orcid.org/0000-0002-4321-0257
Helge Bruelheide http://orcid.org/0000-0003-3135-0356
Steffen Neumann http://orcid.org/0000-0002-7899-7192

DOI:10.1002/ece3.4361

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., … Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, *44*, W3–W10. https://doi.org/10.1093/nar/gkw343

Aranda, S. C., Gabriel, R., Borges, P. A. V., Santos, A. M. C., de Azevedo, E. B., Patiño, J., … Lobo, J. M. (2014). Geographical, temporal and environmental determinants of bryophyte species richness in the Macaronesian Islands. *PLoS One*, *9*, e101786. https://doi.org/10.1371/journal.pone.0101786

Asakawa, Y., Ludwiczuk, A., & Nagashima, F. (Eds.) (2013a). *Chemical constituents of bryophytes: Bio- and chemical diversity, biological activity, and chemosystematics*. New York, NY: Springer Verlag.

Asakawa, Y., Ludwiczuk, A., & Nagashima, F. (2013b). Phytochemical and biological studies of bryophytes. *Phytochemistry*, *91*, 52–80. https://doi.org/10.1016/j.phytochem.2012.04.012

Ayres, E., van der Wal, R., Sommerkorn, M., & Bardgett, R. D. (2006). Direct uptake of soil nitrogen by mosses. *Biology Letters*, *2*, 286–288. https://doi.org/10.1098/rsbl.2006.0455

Boch, S., Prati, D., & Fischer, M. (2016). Gastropods slow down succession and maintain diversity in cryptogam communities. *Ecology*, *97*, 2184–2191. https://doi.org/10.1002/ecy.1498

Böttcher, C., Westphal, L., Schmotz, C., Prade, E., Scheel, D., & Glawischnig, E. (2009). The Multifunctional Enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) Converts Cysteine-Indole-3-Acetonitrile to Camalexin in the Indole-3-Acetonitrile Metabolic Network of Arabidopsis thaliana. *The Plant Cell Online*, *21*, 1830–1845. https://doi.org/10.1105/tpc.109.066670

Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., … Schmutz, J. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, *171*, 287–304 e15. https://doi.org/10.1016/j.cell.2017.09.030

Caccianiga, M., Luzzaro, A., Pierce, S., Ceriani, R. M., & Cerabolini, B. (2006). The functional basis of a primary succession resolved by CSR classification. *Oikos*, *112*, 10–20. https://doi.org/10.1111/j.0030-1299.2006.14107.x

Chapin, F. S., Walker, L. R., Fastie, C. L., & Sharman, L. C. (1994). Mechanisms of primary succession following deglaciation at Glacier Bay, Alaska. *Ecological Monographs*, *64*, 149–175. https://doi.org/10.2307/2937039

Cornelissen, J. H. C., Lang, S. I., Soudzilovskaia, N. A., & During, H. J. (2007). Comparative cryptogam ecology: A review of bryophyte and lichen traits that drive biogeochemistry. *Annals of Botany*, *99*, 987–1001. https://doi.org/10.1093/aob/mcm030

Donath, T. W., & Eckstein, R. L. (2010). Effects of bryophytes and grass litter on seedling emergence vary by vertical seed position and seed size. *Plant Ecology*, *207*, 257–268. https://doi.org/10.1007/s11258-009-9670-8

Doxford, S. W., Ooi, M. K. J., & Freckleton, R. P. (2013). Spatial and temporal variability in positive and negative plant-bryophyte interactions along a latitudinal gradient. *Journal of Ecology*, *101*, 465–474. https://doi.org/10.1111/1365-2745.12036

During, H. J. (1992). Ecological classification of bryophytes and lichens. *Bryophytes and lichens in a changing environment* (pp. 1–31). Oxford: Clarendon Press.

Ehlers, B. K., Damgaard, C. F., & Laroche, F. (2016). Intraspecific genetic variation and species coexistence in plant communities. *Biology Letters*, *12*, 20150853. https://doi.org/10.1098/rsbl.2015.0853

Erxleben, A., Gessler, A., Vervliet-Scheebaum, M., & Reski, R. (2012). Metabolite profiling of the moss Physcomitrella patens reveals evolutionary conservation of osmoprotective substances. *Plant Cell Reports*, *31*, 427–436. https://doi.org/10.1007/s00299-011-1177-9

Fester, T. (2015). Plant metabolite profiles and the buffering capacities of ecosystems. *Phytochemistry*, *110*, 6–12. https://doi.org/10.1016/j.phytochem.2014.12.015

Frisvad, J. C., Andersen, B., & Thrane, U. (2008). The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. *Mycological Research*, *112*, 231–240. https://doi.org/10.1016/j.mycres.2007.08.018

Frisvoll, A. A. (1997). Bryophytes of spruce forest stands in Central Norway. *Lindbergia*, *22*, 83–97.

Garcia, E. L., Rosenstiel, T. N., Graves, C., Shortlidge, E. E., & Eppley, S. M. (2016). Distribution drivers and physiological responses in geothermal bryophyte communities. *American Journal of Botany*, *103*, 625–634. https://doi.org/10.3732/ajb.1500422

Ghani, N. A., Ludwiczuk, A., Ismail, N. H., & Asakawa, Y. (2016). Volatile components of the stressed liverwort *Conocephalum conicum*. *Natural Product Communications*, *11*, 103–104.

Gignac, L. D. (2001). Bryophytes as indicators of climate change. *The Bryologist*, *104*, 410–420. https://doi.org/10.1639/0007-2745(2001)104[0410:BAIOCC]2.0.CO;2

Gilbert, O. L. (1968). Bryophytes as indicators of air pollution in the tyne valley. *New Phytologist*, *67*, 15–30. https://doi.org/10.1111/j.1469-8137.1968.tb05450.x

Goffinet, B., & Shaw, A. J. (2009). *Bryophyte biology*. Cambridge, NY: Cambridge University Press.

Gornall, J. L., Woodin, S. J., Jónsdóttir, I. S., & van der Wal, R. (2011). Balancing positive and negative plant interactions: How mosses structure vascular plant communities. *Oecologia*, *166*, 769–782. https://doi.org/10.1007/s00442-011-1911-6

Grime, J. P. (1977). Evidence for the existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. *The American Naturalist*, *111*, 1169–1194. https://doi.org/10.1086/283244

Grime, J. P., Rincon, E. R., & Wickerson, B. E. (1990). Bryophytes and plant strategy theory. *Botanical Journal of the Linnean Society*, *104*, 175–186. https://doi.org/10.1111/j.1095-8339.1990.tb02217.x

Hall, R. D. (2006). Plant metabolomics: From holistic hope, to hype, to hot topic: *Tansley review*. *New Phytologist*, *169*, 453–468. https://doi.org/10.1111/j.1469-8137.2005.01632.x

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., … Steinbeck, C. (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, *41*, D781–D786. https://doi.org/10.1093/nar/gks1004

He, X., Sun, Y., & Zhu, R.-L. (2013). The oil bodies of liverworts: Unique and important organelles in land plants. *Critical Reviews in Plant Sciences*, *32*, 293–302. https://doi.org/10.1080/07352689.2013.765765

Hedwall, P.-O., Skoglund, J., & Linder, S. (2015). Interactions with successional stage and nutrient status determines the life-form-specific effects of increased soil temperature on boreal forest floor vegetation. *Ecology and Evolution*, *5*, 948–960. https://doi.org/10.1002/ece3.1412

Heinrichs, J., Anton, H., Gradstein, S. R., & Mues, R. (2000). Systematics ofPlagiochila t.Glaucescentes Carl (Hepaticae) from tropical America: A morphological and chemotaxonomical approach. *Plant Systematics and Evolution*, *220*, 115–138. https://doi.org/10.1007/BF00985374

Hüllbusch, E., Brandt, L. M., Ende, P., & Dengler, J. (2016). Little vegetation change during two decades in a dry grassland complex in the Biosphere Reserve Schorfheide-Chorin (NE Germany). *Tuexenia*, *36*, 395–412.

Ivanišević, J., Thomas, O. P., Lejeusne, C., Chevaldonné, P., & Pérez, T. (2011). Metabolic fingerprinting as an indicator of biodiversity: Towards understanding inter-specific relationships among Homoscleromorpha sponges. *Metabolomics*, *7*, 289–304.

Jeschke, M., & Kiehl, K. (2008). Effects of a dense moss layer on germination and establishment of vascular plants in newly created calcareous grasslands. *Flora - Morphology, Distribution, Functional*

DOI:10.1002/ece3.4361

Ecology of Plants, *203*, 557–566. https://doi.org/10.1016/j.flora.2007.09.006

Jones, O. A. H., Maguire, M. L., Griffin, J. L., Dias, D. A., Spurgeon, D. J., & Svendsen, C. (2013). Metabolomics and its use in ecology: Metabolomics in Ecology. *Austral Ecology*, *38*, 713–720. https://doi.org/10.1111/aec.12019

Klavina, L. (2015). A study on bryophyte chemical composition–search for new applications. *Agronomy Research*, *13*, 969–978.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., & Neumann, S. (2012). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, *84*, 283–289. https://doi.org/10.1021/ac202450g

Lambers, H., Chapin, F. S., & Pons, T. L. (2008). *Plant physiological ecology* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-0-387-78341-3

Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, *69*, 24.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd ed.). Amsterdam: Elsevier.

Li, D., Heiling, S., Baldwin, I. T., & Gaquerel, E. (2016). Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, E7610–E7618. https://doi.org/10.1073/pnas.1610218113

Ligrone, R., Carafa, A., Duckett, J. G., Renzaglia, K. S., & Ruel, K. (2008). Immunocytochemical detection of lignin-related epitopes in cell walls in bryophytes and the charalean alga Nitella. *Plant Systematics and Evolution*, *270*, 257–272. https://doi.org/10.1007/s00606-007-0617-z

Ludwiczuk, A., Odrzykoski, I. J., & Asakawa, Y. (2013). Identification of cryptic species within liverwort Conocephalum conicum based on the volatile components. *Phytochemistry*, *95*, 234–241. https://doi.org/10.1016/j.phytochem.2013.06.011

Maksimova, V., Klavina, L., Bikovens, O., Zicmanis, A., & Purmalis, O. (2013). Structural Characterization and Chemical Classification of Some Bryophytes Found in Latvia. *Chemistry & Biodiversity*, *10*, 1284–1294. https://doi.org/10.1002/cbdv.201300014

Mateo, R. G., Broennimann, O., Normand, S., Petitpierre, B., Araújo, M. B., Svenning, J.-C., … Vanderpoorten, A. (2016). The mossy north: An inverse latitudinal diversity gradient in European bryophytes. *Scientific Reports*, *6*, 1–9.

Michel, P., Burritt, D. J., & Lee, W. G. (2011). Bryophytes display allelopathic interactions with tree species in native forest ecosystems. *Oikos*, *120*, 1272–1280. https://doi.org/10.1111/j.1600-0706.2010.19148.x

Müller, J., Klaus, V. H., Kleinebecker, T., Prati, D., Hölzel, N., & Fischer, M. (2012). Impact of land-use intensity and productivity on bryophyte diversity in agricultural grasslands. *PLoS One*, *7*, e51520. https://doi.org/10.1371/journal.pone.0051520

Pejin, B., Vujisic, L., Sabovljevic, M., Sabovljevic, A., Tesevic, V., & Vajs, V. (2010). Preliminary analysis of fatty acid chemistry of *Kindbergia praelonga* and *Kindbergia stokesii* (Brachytheciaceae). *Journal of the Serbian Chemical Society*, *75*, 1637–1640. https://doi.org/10.2298/JSC100209129P

Peters, K., Gorzolka, K., Bruelheide, H., & Neumann, S. (2018). Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes. *Scientific Data*, 5, 180179. https://doi.org/10.1038/sdata.2018.179

Proctor, M. C. F., Oliver, M. J., Wood, A. J., Alpert, P., Stark, L. R., Cleavitt, N. L., & Mishler, B. D. (2007). Desiccation-tolerance in bryophytes: A review. *The Bryologist*, *110*, 595–621. https://doi.org/10.1639/0007-2745(2007)110[595:DIBAR]2.0.CO;2

Qiu, Y.-L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., … Davis, C.C. (2006). The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 15511–15516. https://doi.org/10.1073/pnas.0603335103

Rempt, M., & Pohnert, G. (2010). Novel acetylenic oxylipins from the moss *Dicranum scoparium* with antifeeding activity against herbivorous slugs. *Angewandte Chemie International Edition*, *49*, 4755–4758. https://doi.org/10.1002/anie.201000825

Rivas-Ubach, A., Hódar, J. A., Sardans, J., Kyle, J. E., Kim, Y.-M., Oravec, M., … Peñuelas, J. (2016). Are the metabolomic responses to folivory of closely related plant species linked to macroevolutionary and plant-folivore coevolutionary processes? *Ecology and Evolution*, *6*, 4372–4386. https://doi.org/10.1002/ece3.2206

Rousk, K., Pedersen, P. A., Dyrnum, K., & Michelsen, A. (2017). The interactive effects of temperature and moisture on nitrogen fixation in two temperate-arctic mosses. *Theoretical and Experimental Plant Physiology*, *29*, 25–36. https://doi.org/10.1007/s40626-016-0079-1

Rycroft, D. S., Heinrichs, J., Cole, W. J., & Anton, H. (2001). A phytochemical and morphological study of the liverwort *Plagiochila retrorsa* Gottsche new to Europe. *Journal of Bryology*, *23*, 23–34. https://doi.org/10.1179/jbr.2001.23.1.23

Sardans, J., Peñuelas, J., & Rivas-Ubach, A. (2011). Ecological metabolomics: Overview of current developments and future challenges. *Chemoecology*, *21*, 191–225. https://doi.org/10.1007/s00049-011-0083-5

Scherling, C., Roscher, C., Giavalisco, P., Schulze, E.-D., & Weckwerth, W. (2010). Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. *PLoS One*, *5*, e12569. https://doi.org/10.1371/journal.pone.0012569

Shaw, A. J., Cox, C. J., Goffinet, B., Buck, W. R., & Boles, S. B. (2003). Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (Bryophyta). *Evolution*, *57*, 2226–2241. https://doi.org/10.1111/j.0014-3820.2003.tb00235.x

Shaw, A. J., Szovenyi, P., & Shaw, B. (2011). Bryophyte diversity and evolution: Windows into the early evolution of land plants. *American Journal of Botany*, *98*, 352–369. https://doi.org/10.3732/ajb.1000316

Smith, A. J. E. (1982). *Bryophyte ecology*. London, NY: Chapman and Hall. https://doi.org/10.1007/978-94-009-5891-3

Smith, A. J. E. (1990). *The liverworts of Britain and Ireland*, Digital repr. Cambridge: Cambridge University Press.

Smith, A. J. E. (2004). *The moss flora of Britain and Ireland*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511541858

Streitberger, M., Schmidt, C., & Fartmann, T. (2017). Contrasting response of vascular plant and bryophyte species assemblages to a soil-disturbing ecosystem engineer in calcareous grasslands. *Ecological Engineering*, *99*, 391–399. https://doi.org/10.1016/j.ecoleng.2016.11.037

Suire, C., Bouvier, F., Backhaus, R. A., Bégu, D., Bonneu, M., & Camara, B. (2000). Cellular Localization of Isoprenoid Biosynthetic Enzymes in *Marchantia polymorpha*. Uncovering a New Role of Oil Bodies. *Plant Physiology*, *124*, 971–978. https://doi.org/10.1104/pp.124.3.971

Sun, S.-Q., Wu, Y.-H., Wang, G.-X., Zhou, J., Yu, D., Bing, H.-J., & Luo, J. (2013). Bryophyte species richness and composition along an altitudinal gradient in Gongga Mountain, China. *PLoS One*, *8*, e58131. https://doi.org/10.1371/journal.pone.0058131

Szakiel, A., Pączkowski, C., & Henry, M. (2011). Influence of environmental abiotic factors on the content of saponins in plants. *Phytochemistry Reviews*, *10*, 471–491. https://doi.org/10.1007/s11101-010-9177-x

Tanaka, M., Esaki, T., Kenmoku, H., Koeduka, T., Kiyoyama, Y., Masujima, T., … Matsui, K. (2016). Direct evidence of specific localization of sesquiterpenes and marchantin A in oil body cells of *Marchantia polymorpha* L. *Phytochemistry*, *130*, 77–84. https://doi.org/10.1016/j.phytochem.2016.06.008

DOI:10.1002/ece3.4361

Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, *9*, 504. https://doi.org/10.1186/1471-2105-9-504

Thakur, S., & Kapila, S. (2017). Seasonal changes in antioxidant enzymes, polyphenol oxidase enzyme, flavonoids and phenolic content in three leafy liverworts. *Lindbergia*, *5*, 39–44. https://doi.org/10.25227/linbg.01076

Urmi, E. (2010). Bryophyta (Moose). In E. Landolt (Ed.), *Flora indicativa, ecological indicator values and biological attributes of the flora of Switzerland and the alps* (pp. 283–310). Bern: Haupt.

van Dam, N. M., & van der Meijden, E. (2011). A role for metabolomics in plant ecology. In R. D. Hall (Ed.), *Annual plant reviews*, Vol. 43 (pp. 87–107). Oxford, UK: Wiley-Blackwell.

Vanderpoorten, A., & Goffinet, B. (2009). *Introduction to bryophytes*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511626838

Virtanen, R., Eskelinen, A., & Harrison, S. (2017). Comparing the responses of bryophytes and short-statured vascular plants to climate shifts and eutrophication. *Functional Ecology*, *31*, 946–954. https://doi.org/10.1111/1365-2435.12788

Wagner, S., Zotz, G., Salazar Allen, N., & Bader, M. Y. (2013). Altitudinal changes in temperature responses of net photosynthesis and dark respiration in tropical bryophytes. *Annals of Botany*, *111*, 455–465. https://doi.org/10.1093/aob/mcs267

Wang, Z., Bader, M. Y., Liu, X., Zhu, Z., & Bao, W. (2017). Comparisons of photosynthesis-related traits of 27 abundant or subordinate bryophyte species in a subalpine old-growth fir forest. *Ecology and Evolution*, *7*, 7454–7461. https://doi.org/10.1002/ece3.3277

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry*, *64*, 3–19. https://doi.org/10.1016/S0031-9422(03)00300-5

Wu, C.-L. (1992). Chemosystematic correlations of Taiwanese Hepaticae. *Journal of the Chinese Chemical Society*, *39*, 655–667. https://doi.org/10.1002/jccs.199200101

Zamfir, M. (2000). Effects of bryophytes and lichens on seedling emergence of alvar plants: Evidence from greenhouse experiments. *Oikos*, *88*, 603–611. https://doi.org/10.1034/j.1600-0706.2000.880317.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Peters K, Gorzolka K, Bruelheide H, Neumann S. Seasonal variation of secondary metabolites in nine different bryophytes. *Ecol Evol*. 2018;8:9105–9117. https://doi.org/10.1002/ece3.4361

DOI:10.1002/ece3.4361

GigaScience, 8, 2019, 1–12

doi: 10.1093/gigascience/giy149
Advance Access Publication Date: 7 December 2018
Technical Note

TECHNICAL NOTE

PhenoMeNal: processing and analysis of metabolomics data in the cloud

Kristian Peters[1,*], James Bradbury[2,*], Sven Bergmann[3,4], Marco Capuccini[5,6], Marta Cascante[7], Pedro de Atauri[7], Timothy M. D. Ebbels[9], Carles Foguet[7], Robert Glen[9,10], Alejandra Gonzalez-Beltran[11], Ulrich L. Günther[12], Evangelos Handakas[9], Thomas Hankemeier[14], Kenneth Haug[15], Stephanie Herman[6,16], Petr Holub[17], Massimiliano Izzo[11], Daniel Jacob[18], David Johnson[11,19], Fabien Jourdan[20], Namrata Kale[15], Ibrahim Karaman[21], Bita Khalili[3,4], Payam Emami Khonsari[16], Kim Kultima[16], Samuel Lampa[6], Anders Larsson[6,22], Christian Ludwig[23], Pablo Moreno[15], Steffen Neumann[1,24], Jon Ander Novella[6,22], Claire O'Donovan[15], Jake T.M. Pearce[9], Alina Peluso[9], Marco Enrico Piras[25], Luca Pireddu[25], Michelle A.C. Reed[12], Philippe Rocca-Serra[11], Pierrick Roger[26], Antonio Rosato[27], Rico Rueedi[3,4], Christoph Ruttkies[1], Noureddin Sadawi[8,9], Reza M. Salek[15], Susanna-Assunta Sansone[11], Vitaly Selivanov[7], Ola Spjuth[6], Daniel Schober[1], Etienne A. Thévenot[26], Mattia Tomasoni[3,4], Merlijn van Rijswijk[13,28], Michael van Vliet[14], Mark R. Viant[2,29], Ralf J. M. Weber[2,29], Gianluigi Zanetti[25] and Christoph Steinbeck[30,*]

[1]Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany, [2]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, [4]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [5]Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden, [6]Department of Pharmaceutical Biosciences, Uppsala University, Box 591, 751 24 Uppsala, Sweden, [7]Department of Biochemistry and Molecular Biomedicine,

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

Received: 6 September 2018; **Revised:** 19 October 2018; **Accepted:** 20 November 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI:10.1093/gigascience/giy149

Universitat de Barcelona; Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Instituto de Salud Carlos III (ISCIII), Spain, [8]Department of Computer Science, College of Engineering, Design and Physical Sciences, Brunel University, London, UK, [9]Department of Surgery & Cancer, Imperial College London, South Kensington, London, SW7 2AZ, United Kingdom, [10]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB21EW, United Kingdom, [11]Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, OX1 3QG, Oxford, United Kingdom., [12]Institute of Cancer and Genomic Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [13]Netherlands Metabolomics Center, Leiden, 2333 CC, Netherlands, [14]Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden, 2333 CC, The Netherlands, [15]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom, [16]Department of Medical Sciences, Clinical Chemistry, Uppsala University, 751 85 Uppsala, Sweden, [17]BBMRI-ERIC, Graz, Austria, [18]INRA, University of Bordeaux, Plateforme Métabolome Bordeaux-MetaboHUB, 33140 Villenave d'Ornon, France, [19]Department of Informatics and Media, Uppsala University, Box 513, 751 20 Uppsala, Sweden, [20]INRA - French National Institute for Agricultural Research, UMR1331, Toxalim, Research Centre in Food Toxicology, Toulouse, France, [21]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St. Mary's Campus, Norfolk Place, W2 1PG, London, United Kingdom, [22]National Bioinformatics Infrastructure Sweden, Uppsala University, Uppsala, Sweden, [23]Institute of Metabolism and Systems Research (IMSR), University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [24]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany, [25]Distributed Computing Group, CRS4, Pula, Italy, [26]CEA, LIST, Laboratory for Data Analysis and Systems' Intelligence, MetaboHUB, Gif-Sur-Yvette F-91191, France, [27]Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence and CIRMMP, 50019 Sesto Fiorentino, Florence, Italy, [28]ELIXIR-NL, Dutch Techcentre for Life Sciences, Utrecht, 3503 RM, Netherlands, [29]Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, [30]Cheminformatics and Computational Metabolomics, Institute for Analytical Chemistry, Lessingstr. 8, 07743 Jena, Germany, [31] and [32]

*Correspondence address. Kristian Peters, Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany. E-mail: kpeters@ipb-halle.de http://orcid.org/0000-0002-4321-0257; James Bradbury, E-mail: j.bradbury@bham.ac.uk; Christoph Steinbeck, E-mail: christoph.steinbeck@uni-jena.de http://orcid.org/0000-0001-6966-0814

Abstract

Background: Metabolomics is the comprehensive study of a multitude of small molecules to gain insight into an organism's metabolism. The research field is dynamic and expanding with applications across biomedical, biotechnological, and many other applied biological domains. Its computationally intensive nature has driven requirements for open data formats, data repositories, and data analysis tools. However, the rapid progress has resulted in a mosaic of independent, and sometimes incompatible, analysis methods that are difficult to connect into a useful and complete data analysis solution. **Findings:** PhenoMeNal (Phenome and Metabolome aNalysis) is an advanced and complete solution to set up Infrastructure-as-a-Service (IaaS) that brings workflow-oriented, interoperable metabolomics data analysis platforms into the cloud. PhenoMeNal seamlessly integrates a wide array of existing open-source tools that are tested and packaged as Docker containers through the project's continuous integration process and deployed based on a kubernetes orchestration framework. It also provides a number of standardized, automated, and published analysis workflows in the user interfaces Galaxy, Jupyter, Luigi, and Pachyderm. **Conclusions:** PhenoMeNal constitutes a keystone solution in cloud e-infrastructures available for metabolomics. PhenoMeNal is a unique and complete solution for setting up cloud e-infrastructures through easy-to-use web interfaces that can be scaled to any custom public and private cloud environment. By harmonizing and automating software installation and configuration and through ready-to-use scientific workflow user interfaces, PhenoMeNal has succeeded in providing scientists with workflow-driven, reproducible, and shareable metabolomics data analysis platforms that are interfaced through standard data formats, representative datasets, versioned, and have been tested for reproducibility and interoperability. The elastic implementation of PhenoMeNal further allows easy adaptation of the infrastructure to other application areas and 'omics research domains.

Keywords: metabolomics; data analysis; e-infrastructures; NMR; mass spectrometry; computational workflows; galaxy; cloud computing; standardization; statistics

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

Findings

Background

The field of metabolomics has seen remarkable progress over the last decade and has enabled fascinating discoveries in many different research areas. Metabolomics is the study of small molecules in organisms that can reveal detailed insights into metabolic biochemistry, e.g., changes in concentrations of specific molecules, metabolic fluxes between cells or compartments, identification of molecules that are involved in the pathogenesis of a disease, and the study of the biochemical phenotype of animals, plants, and even soil microorganisms [1–3].

The principal metabolomics technologies of mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) typically generate large datasets that require computationally intensive analyses [4]. Biomedical investigations can involve large cohorts with many thousands of metabolite profiles and can produce hundreds of gigabytes of data [5–8]. With such large datasets, processing becomes impracticable and unmanageable on commodity hardware. Cloud computing can offer a solution by enabling the outsourcing of calculations from local workstations to scalable cloud data centers, with the possibility to allocate thousands of central processing unit (CPU) cores simultaneously. Furthermore, cloud computing allows for resources to be instantiated on-demand (CPUs, random access memory, network, storage) and allows access to computational tools in the form of microservices that can dynamically grow or shrink.

MS and NMR data processing usually involves selection of parameters (that are often specific to the analytical instrumentation), algorithmic peak detection, peak alignment and grouping, annotation of putative compounds, and extensive statistical analyses [9, 10]. Many open-source tools have been developed that address these different steps in data processing and analysis. These tools, however, usually come with their own software dependencies, resource requirements, and scripting languages. As a consequence, configuring and running them is often complicated, especially for researchers who are untrained in computer science [4]. Furthermore, many tools require users to input parameters that can significantly affect results and performance, and reporting of these parameters is not always clear [11].

A number of infrastructures and integration efforts have been initiated in the past five years, including metabolomics data repositories with a global scope [6, 12], platforms for reproducible workflow analysis [13, 14], as well as initiatives to integrate and coordinate data standards [15]. Simultaneously, multiple networks of service centers such as the international Phenome Centers [16] and MetaboHub [17] have formed with the goal to facilitate the acquisition, processing, and analysis of metabolomics data [6–8] at ever increasing scales.

Currently, several web-based metabolomics data processing platforms are available. XCMSOnline provides a platform based on XCMS for downstream data analysis, visualization, data sharing, and access to Metlin to facilitate metabolite identification and pathway analysis [18]. MetaboAnalyst presents a wide variety of data processing and analysis tools including statistical analysis, time-series analysis, functional analysis, and pathway analysis [19]. Workflow4Metabolomics is based on Galaxy and provides various metabolomics processing workflows, including NMR [13, 20]. These common tools for analyzing metabolomics data provide web-based graphical user interfaces (GUIs) with different functionality.

Here, we present PhenoMeNal (Phenome and Metabolome aNalysis), a unique, easy-to-use, complete, robust, and performant cloud e-infrastructure that provides a large suite of standardized and interoperable metabolomics data processing tools as a complete data analysis solution. In contrast to current metabolomics processing platforms, PhenoMeNal provides Infrastructure-as-a-Service (IaaS) and seamlessly integrates a wide array of existing open-source tools.

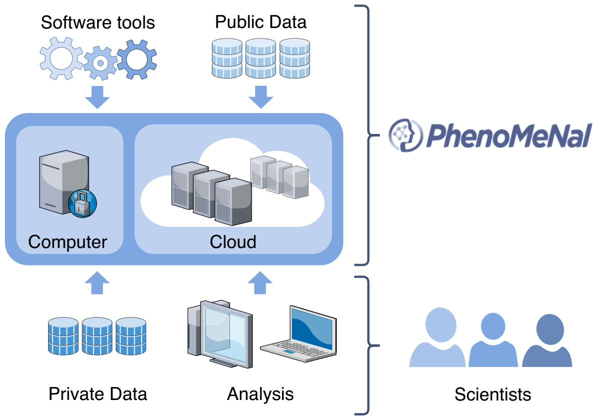


Figure 1: Conceptual design of the PhenoMeNal cloud e-infrastructure, which brings compute to the data for any large number of data scientists.

A major advantage over other platforms is that PhenoMeNal make it possible to instantiate many different services in the cloud and provides a number of standardized, automated, and published analysis workflows in the user interfaces Galaxy, Jupyter, Luigi, and Pachyderm (Fig. 1). Moreover, the PhenoMeNal e-infrastructure can be easily deployed onto public and private cloud environments and can be configured elastically to fit into any cloud-based environment, thus enabling scalable and cost-effective high-performance metabolomics data analysis in a way that hides the technical complexity from the user. PhenoMeNal further facilitates reproducible analyses through automated, sharable, and citable workflows.

Overview

The features of the PhenoMeNal e-infrastructure are encapsulated as a cloud research environment (CRE). The PhenoMeNal CRE can be instantiated on major commercial public cloud providers, including Amazon web services (AWS) and Google cloud platform (GCP), as well as OpenStack-based private clouds and in custom environments. Technical complexity is hidden from the users, simplifying setting up the cloud infrastructure for administrators (Fig. 2).

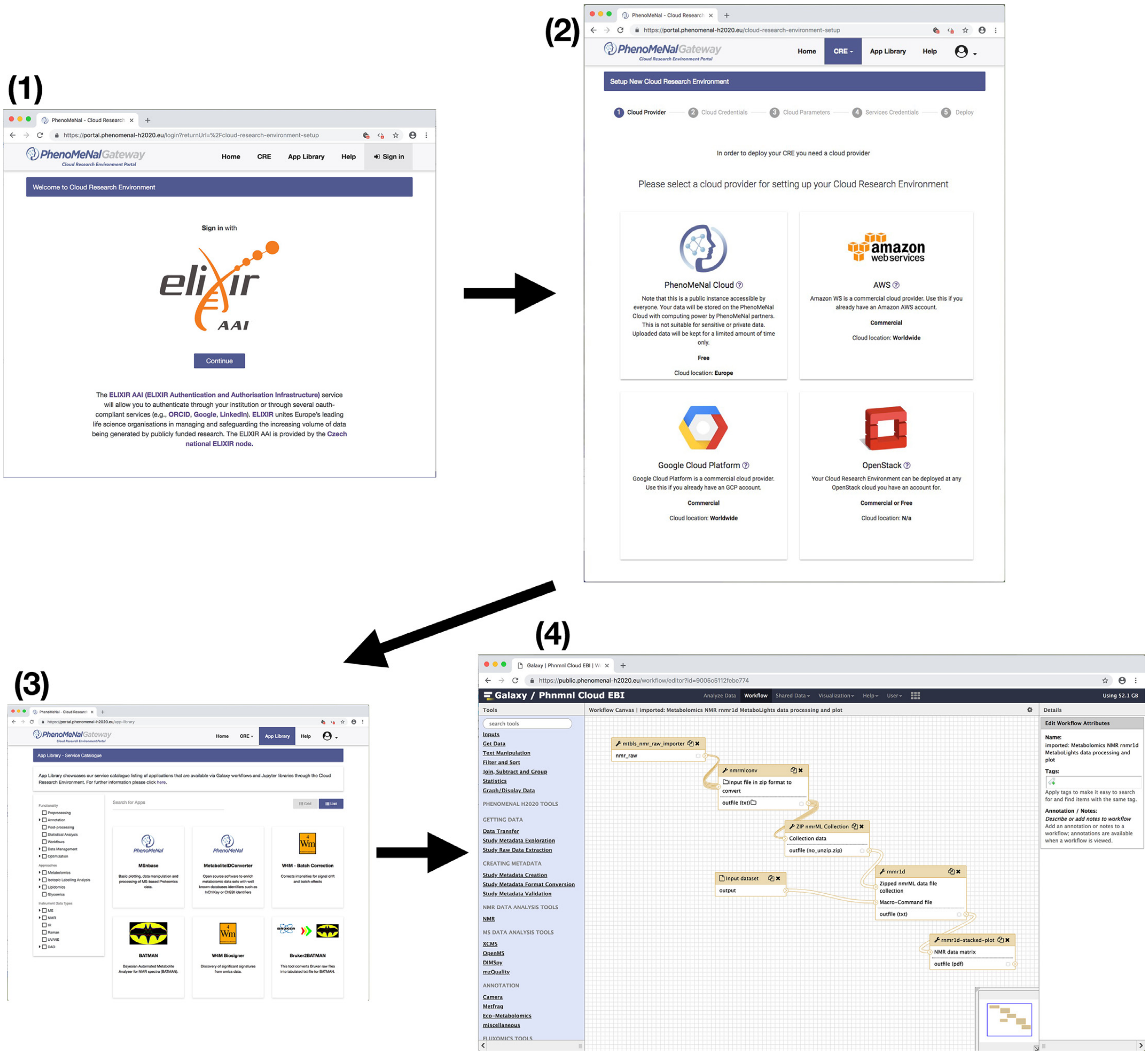
From a web-based portal, users can deploy the CRE, which includes several web services and software tools (Fig. 2). Data can be processed directly in the e-infrastructure without the need to install additional software. Scientific workflows can be executed via user-friendly web-based platforms such as Galaxy, as well as programmatic interfaces and notebooks. Each service has been supplied with a rich source of documentation and training material to assist researchers.

The PhenoMeNal Portal

The PhenoMeNal Portal [21] allows users to deploy, manage, and delete PhenoMeNal CREs simply through a web interface. Deployments to major commercial cloud platforms (AWS and GCP) as well as OpenStack, an open-source cloud platform, can be made using an easy-to-follow wizard (Fig. 2). OpenStack deploy-

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149


(1)
PhenoMeNal - Cloud Research
https://portal.phenomenal-h2020.eu/login?returnUrl=%2Fcloud-research-environment-setup
PhenoMeNal Gateway
Cloud Research Environment Portal
Home
CRE
App Library
Help
Sign in
Welcome to Cloud Research Environment
Sign in with
elixir AAI
Continue
(2)
https://portal.phenomenal-h2020.eu/cloud-research-environment-setup
Setup New Cloud Research Environment
1 Cloud Provider
2 Cloud Credentials
3 Cloud Parameters
4 Services Credentials
5 Deploy
amazon web services
(3)
App Library - Service Catalogue
(4)
Galaxy / Phnmnl Cloud EBI

The ELIXIR AAI (ELIXIR Authentication and Authorisation Infrastructure) service will allow you to authenticate through your institution or through several oauth-compliant services (e.g., ORCID, Google, LinkedIn). ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research. The ELIXIR AAI is provided by the Czech national ELIXIR node.

In order to deploy your CRE you need a cloud provider

Please select a cloud provider for setting up your Cloud Research Environment

PhenoMeNal Cloud
Note that this is a public instance accessible by everyone. Your data will be stored on the PhenoMeNal Cloud with computing power by PhenoMeNal partners. This is not suitable for sensitive or private data. Uploaded data will be kept for a limited amount of time only.
Free
Cloud location: Europe

AWS
Amazon WS is a commercial cloud provider. Use this if you already have an Amazon AWS account.
Commercial
Cloud location: Worldwide

Google Cloud Platform
Google Cloud Platform is a commercial cloud provider. Use this if you already have an GCP account.
Commercial
Cloud location: Worldwide

OpenStack
Your Cloud Research Environment can be deployed at any OpenStack cloud you have an account for.
Commercial or Free
Cloud location: N/a

Figure 2: Screenshots of creating and using the PhenoMeNal cloud e-infrastructure. First, log in with ELIXIR to the cloud research environment (CRE) portal. Second, select a public or private cloud provider. After entering cloud credentials and setting up parameters in the dedicated portal, the deployment of the PhenoMeNal e-infrastructure into the cloud environment can be made. Third, in the PhenoMeNal Portal app library there are several services ready to be deployed and used in the set-up infrastructure. Fourth, dedicated web services such as Galaxy are readily available in the cloud e-infrastructure. All steps can be operated from an easy-to-use web interface that is accessible from any standard web browser.

ments can be deployed behind clinical firewalls, which is especially pertinent when dealing with sensitive (i.e., patient) data.

The PhenoMeNal public instance allows users to test-run a CRE without the need to deploy on a cloud platform. It can be deployed and accessed through the portal. Once credentials for users have been generated, analyses can be run through a Galaxy instance containing the tools and workflows present in any deployed CRE. The portal also includes user and developer documentation, workflow tutorials, and links to training videos.

Scientific workflows

A scientific workflow is a set of computational steps that are carried out to process and analyze data [22]. Usually, a workflow is comprised of several linked software tools that are each executed during a particular step of the workflow. In order to manage and automate scientific workflows, in PhenoMeNal the well-established dedicated workflow management system Galaxy can be deployed, which presents the user with an easy-to-use graphical user interface as well as providing a programmatic interface [20, 23]. Galaxy facilitates collaborative exchange, reproducibility, and traceability of data analysis by enabling users to share entire workflows and analysis histories [24]. In addition to Galaxy, programmatic executable notebooks (Jupyter) and the workflow tools exposed as programmatic interfaces Luigi and Pachyderm are also supported [25].

In order to cover typical use cases in metabolomics and to illustrate the usage and applicability of given analytical pipelines and software tools, five representative scientific workflows are available in the PhenoMeNal Galaxy (Table 1), each having different computational demands and purposes. More than 250 individual modules have been integrated in Galaxy (see the subsection Scientific Workflows in the Methods section).

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

Table 1: List of workflows that are representative for their respective metabolomics domains (identification in NMR, Fluxomics, Annotation, and identification in MS and eco-metabolomics)

Workflow name	Description	Reference
1D NMR	Processes 1D NMR experiments from raw data to a data matrix required for visualization and statistical analysis, building on nmrML and NMRProcFlow. The automatic workflow is based on the MTBLS1 dataset, describing urinary changes in type 2 diabetes in humans.	[26, 27, 28]
Fluxomics	Quantifies steady-state fluxes following ^{13}C metabolic flux analysis. The workflow was first based on the analysis of the MTBLS412 dataset with 13C tracer data of human umbilical vein endothelial cells under hypoxia.	[29, 30]
LC-MS/MS	Processes, quantifies, and annotates/identifies features in mass spectra using MetFrag — a tool that annotates molecules from compound databases of tandem mass spectrometry (MS/MS) spectra. The workflow is based on MTBLS558.	[31, 32, 33]
Univariate and Multivariate Statistics	Applies univariate and multivariate statistical analysis and illustrates how datasets may be explored, enabling the identification of variables of interest and the construction of predictive models. The workflow is based on MTBLS404.	[13, 34]
Eco-Metabolomics	Implementation of a resource demanding metabolomics use case in ecology, used in large field experiments to describe interactions between different species of organisms in remarkable detail. The workflow is based on MTBLS520.	[35]
ISA-Create-Validate-Upload	A workflow to create Investigation, Study, and Assay data model framework-compliant metadata files based on study design information, augmented with semantic markup as source, implementing UK Phenome center naming conventions. Following validation, the workflow also allows visualization of overall study design and deposition to EMBL-EBI.	

Software tools

The Portal App Library [36] shows the software tools packaged in PhenoMeNal that are available through the CRE deployment (Fig. 2). The range of software tools available covers several metabolomics domains, making PhenoMeNal relevant for use in a wide range of data analysis scenarios. The domains covered include clinical metabolomics, plant metabolomics, fluxomics, and eco-metabolomics. Data from both targeted and untargeted analysis can be analyzed for metabolite profiling and fingerprinting approaches [1, 2]. NMR and MS (liquid chromatography coupled with mass spectrometry, gas chromatography coupled with mass spectrometry, direct infusion mass spectrometry) data can be processed.

PhenoMeNal also provides tools for data management (e.g., via the Investigation, Study, and Assay data model framework [ISA] format and application programming interface [API]), metabolite feature detection (e.g., XCMS, CAMERA, nmrProcFlow), metabolite identification (MetFrag, BATMAN, MetaboMatching), and (bio)statistics (e.g., univariate, multivariate, and power analyses) (Supplementary Table S1). Tools can be filtered for functionality, approaches, and instrument (data) types to readily find the most appropriate software tools. Some tools that implement specific functionality (e.g., Rnmr1D, which performs baseline correction of NMR spectra as part of nmrProcFlow) are available through dedicated Galaxy modules or through software containers (Supplementary Table S1).

Study design

PhenoMeNal was designed to use standardized protocols and software tools and to comply with state-of-the-art dedicated specifications and data formats across the entire project. Development was geared toward implementation of open standards for tracking provenance of both data and metadata generated by clinical phenotyping projects. In PhenoMeNal, the ISA model and specifications were implemented using the ISA format to generate, annotate, validate, and deposit experimental metadata information of datasets and studies to public repositories such as MetaboLights [37, 38]. ISA-based metadata tracking is used for the different analysis pipelines that are specific to the distinct metabolomics domains. PhenoMeNal reached native support for the ISA format by developing a dedicated Galaxy composite data type. Such component affords direct recognition of the ISA format by the Galaxy environment, thus ensuring seamless integration with the downstream workflow component.

Data deposition

PhenoMeNal encourages the metabolomics data repository MetaboLights as a primary source of data deposition [39]. Private and public datasets are supported, as are download and upload to MetaboLights. If the storage in a data repository such as MetaboLights is not possible, data can be stored locally or in the cloud e-infrastructure. Access to the data is strictly controlled and secured. To support data deposition, ISA-based Galaxy modules are available making it possible to publish and disseminate scientific results in standard compliant ways.

Reproducibility

One of the challenges of cloud computing is that analyses need to be run continuously and successfully in different environments [40]. Specifically, it has to be ensured that, given the same input, workflows and tools produce identical results regardless of the underlying environment [4, 40]. When these requirements are fulfilled, end users can be confident that their data will be analyzed correctly. PhenoMeNal has implemented three major testing strategies to ensure technical reproducibility using a continuous integration framework [41]. Tests were implemented for the infrastructure components, individual software containers, and data involved in computational workflows.

Sustainability

PhenoMeNal is part of a number of initiatives (BioMedBridges, COSMOS, and ELIXIR) to foster the role of metabolomics and to harmonize experimental data and metadata usage [15, 42]. Col-

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

laborations were established with EGI [43] and Indigo Datacloud [44] infrastructure providers and initiatives [45, 46] to ensure that PhenoMeNal uses technologies that are well supported and ensure their widespread usage, continuity, and further development. For example, the development of KubeNow and contributions to the Galaxy and Workflow4Metabolomics community are essential for PhenoMeNal [47]. Core development will continue on GitHub and is fostered by collaborations with tool developers.

Dependencies on specific technologies and frameworks were avoided by focusing on open standards such as ISA-Tab/ISA-JSON, mzML and nmrML, and widely accepted software [48]. By being able to deploy PhenoMeNal on multiple types of cloud environments, lock-ins to specific computing resource providers are avoided. PhenoMeNal implemented continuous integration and delivery, validated by extensive testing and with clear maintenance responsibilities (see Methods section).

Privacy and security

With human or animal material, the collection, storage, and analysis of metabolomics data introduce a number of constraints due to ethical, legal, and social implications (ELSI) [49]. In particular, data initially derived from human clinical studies may be identifiable and will require consent for use, usually for a defined objective, such as diagnosis, or be related to a particular disease study. Where data is identifiable or pseudonymized, users can deploy PhenoMeNal on local secure resources, thus avoiding the export of data. In this scenario, access to the e-infrastructure should be strictly controlled through local access and authorization. It is recommended that clinical data be fully anonymized before analysis in PhenoMeNal [49, 50].

The PhenoMeNal portal provides substantial guidance to enable users to comply with ELSI and general data protection regulation (GDPR) requirements. Users must register in order to use the individual parts of the e-infrastructure. PhenoMeNal was implemented to use secured and encrypted transport and network communications.

Documentation and training materials

Extensive user documentation and tutorials are provided via the PhenoMeNal Wiki page [51]. The Wiki includes detailed developer resources including information about the PhenoMeNal release schedule; guidelines for tool, workflow, and portal developers; continuous integration; and testing. Further documentation is also provided detailing, creating, and managing PhenoMeNal CREs and tutorials for the Galaxy modules and preconfigured workflows, as well as Galaxy tours that provide step-by-step guidance for inexperienced users.

Community engagement

The PhenoMeNal project is open source and is hosted on GitHub [52]. Developers can contribute tools to PhenoMeNal and are encouraged to do so. To add a tool to PhenoMeNal, it must be containerized using Docker and then integrated into the build process. Detailed documentation is available in the project's Wiki for developers who wish to add their tools to PhenoMeNal.

Collaborations with other projects have been actively encouraged during the development of PhenoMeNal, including Workflow4Metabolomics [13] and the developers of both nmrML and nmrProcFlow [26]. These collaborations are essential to fostering greater standardization within PhenoMeNal and to increasing compatibility with other metabolomics data processing infrastructures.

Availability

Information on how to access PhenoMeNal can be found at the project's website [53]. The GitHub repository hosts the source code of all development projects [52]. The project container-galaxy-k8s-runtime contains all of the developments regarding Galaxy. The Wiki containing documentation is also hosted on GitHub [51]. The PhenoMeNal Portal can be reached at [21]. The public instance of Galaxy is accessible at [54]. Source code and documentation are available under the terms of the Apache 2.0 license. Integrated open-source projects are available under the respective licensing terms.

Conclusions

PhenoMeNal has succeeded in increasing the robustness and coverage of representative metabolomics data processing in scientific cloud e-infrastructures. The presented cloud e-infrastructure covers a wide range of analysis pipelines including data generation and download, data pre- and post-processing, (bio)statistics, and result deposition in data repositories. A large effort has been made to introduce lower-level changes to cloud e-infrastructures (e.g., the cloud deployment software KubeNow) to meet the demands of the biomedical domain. Furthermore, Galaxy has been enriched with metabolomics data standards, in particular, the ISA format for study metadata and mzML and nmrML for acquired data files, as well as support for Kubernetes. PhenoMeNal has fostered the visibility of new metabolomics tools and has enabled the development of more sophisticated data analysis workflows. Our efforts were also guided by feedback from real-life test scenarios collected at workshops with users from the clinical domain.

PhenoMeNal constitutes a keystone solution in cloud platforms available for metabolomics data analysis. The platform was designed to deliver optimal performance and functionality for typical use cases in the metabolomics domain. While the needs of clinicians and researchers in the biomedical and biochemical domains have been targeted, PhenoMeNal is not limited to a specific domain as the cloud infrastructure, tools, and workflows can be adapted to other use cases as demonstrated with the inclusion of the eco-metabolomics workflow. The technological advancements can be reused in other scientific cloud environments and could be integrated with solutions from other ‘omics domains in the future.

Methods

Cloud e-infrastructure

The PhenoMeNal CRE is designed as a microservice architecture, with services being implemented as virtual machine images and software containers. Containers are used to provide microservices for metabolomics data analysis tools and also long-running services such as workflow management systems. A container orchestrator runs containers on top of the scalable infrastructure. The orchestrator takes a group of machines that act as a distributed cluster and receives requests for tools as well as service executions. PhenoMeNal implements various layers to providea container orchestrator on top of either bare metal hardware or IaaS given by a cloud provider [55] (Supplementary Fig. S1).

During the setup process and while PhenoMeNal is deployed, data storage and CPU limits can be configured and dynamically scaled to fit any cloud environment. Deployments can be made

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

to GCE, AWS, and OpenStack-based private clouds from the PhenoMeNal portal. Deployments are also supported from the command line to Microsoft Azure [56], the European Science Cloud [57], and local servers (bare metal) [58]; we provide step-by-step instructions for these solutions.

PhenoMeNal provides IaaS for three different cloud environments:

"local cloud": local workstations or bare metal clusters where data are not allowed to leave the facility.
"public cloud": the flexible use of commercial cloud providers such as GCE and AWS.
"shared cloud": using OpenStack—a free and open-source software platform for cloud computing, ideal for custom environments and research networks.

Software tools

The PhenoMeNal portal has an application library that allows users to deploy tools as microservices into the cloud infrastructure (Fig. 2, Supplementary Table S1). The portal is packaged into frontend and backend engines on top of Kubernetes.

Most software tools in PhenoMeNal are compiled from source code and use a variety of programming languages. Linux versions of software tools and user interfaces such as Galaxy are supported in dedicated encapsulated Docker containers that are implemented as minimum-sized microservices. PhenoMeNal currently hosts 100 such projects in its GitHub repository [59] (Supplementary Table S1). Projects are indicated by the trailing Àcontainer-À name and include a ruleset to build and run the containerized tools, as well as datasets for testing and other necessary files.

PhenoMeNal provides tutorials for developers who want to integrate their tools into our e-infrastructure [60].

Scientific workflows

In PhenoMeNal, a number of options are available for running reproducible and standardized workflows (Table 1).

Galaxy

The Galaxy workflow management system is widely regarded as one of the most popular scientific workflow platforms [20, 61]. It provides a user-friendly web-based GUI to make it easy for the end user to configure and run individual modules and entire workflows without programming experience. Command-line tools and scripts are encapsulated into modules that are launched via the web interface. Galaxy also supports more powerful features such as programmatic access through a REST API and helper libraries to access the running instance of Galaxy [62].

PhenoMeNal has been able to adapt Galaxy for use with a microservices-based architecture [31]. To this end, modules are encapsulated into Docker containers that can be flexibly launched within the cloud e-infrastructure. Galaxy is available in all deployed PhenoMeNal CREs and contains more than 250 modules that have been implemented as part of PhenoMeNal.

Six representative metabolomics Galaxy workflows have been fully integrated into PhenoMeNal (Table 1), and more workflows (mzQuality, NMR-BATMAN) are available for testing.

Jupyter

Jupyter, which started its history as the IPython notebook, is the most popular among the tools commonly referred to as executable notebooks or computational notebooks [63]. Jupyter lets users combine executable code with results from code executions such as text, tables, and figures. Usually, Jupyter notebooks are enriched with extended information that explains what the code does. As a result, they are often used for training material and for tutorials. Also, computational notebooks can, to some extent, be used as a way to document code executions and to make executions more reproducible [64].

Luigi and pachyderm

Luigi is a Python workflow programming library that was originally developed by the company Spotify. It manages pipelines of computations primarily on ”big data” systems such as Hadoop and Apache Spark but also supports local execution [63, 64]. Luigi is a very flexible library that facilitates building complex pipelines of batch jobs handling dependency resolution, workflow management, and visualization.

Similarly, Pachyderm makes it possible to process distributed data and to keep track of the data from every stage of the analysis pipeline [25]. With Pachyderm, it is possible to track the provenance of results and to accurately reproduce scientific workflows. Luigi and Pachyderm are well suited for complex scientific tasks and are easy to use from the python environment in Jupyter notebooks without additional integration tooling needed.

In PhenoMeNal, we have extended Galaxy, Jupyter, Luigi, and Pachyderm in such a way that they can be orchestrated throughout the cloud infrastructure together with the data analysis tools themselves [31]. Six important metabolomics workflows have been fully integrated into PhenoMeNal (Table 1), and more (mzQuality, NMR-BATMAN) are available for testing.

Reproducibility

Three strategies are realized to ensure technical reproducibility. They are implemented in the continuous integration (CI) software development framework Jenkins [41] which is accessible at [65]. These strategies are implemented as tests in our Jenkins and a tutorial guide is available at [66].

- Infrastructure testing: Procedures were implemented to ensure that each individual component (e.g., the deployment process of software containers, resource management, APIs/application binary interfaces [ABIs]) within the infrastructure is interacting correctly with the other components.
- Container testing: Verification that tools, which are packaged into software containers, build and run correctly in the infrastructure. Dependencies within one container and across several interdependent containers are tested.
- Data testing: The output of tools, which process demonstration data, is checked against a data set that is known to contain the expected result. This is being done for both individual tools and for several tools running in a workflow using the workflow testing tool for Galaxy called wft4galaxy [67].

Standardization

PhenoMeNal has implemented several dedicated Galaxy modules that directly retrieve and store ISA-Tab data set descriptors from and to MetaboLights, and can convert between other formats. Native Galaxy composite data types to support ISA-Tab and ISA-JSON have also been integrated, building upon the ISA API [38, 48]. The ISA data type allows for the upload of an ISA-Tab archive (a zip file containing the ISA set of files and raw

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

Table 2: Overview of the most important FAIR criteria and implementations suggested for PhenoMeNal data, tools and workflows

	Data	Tools	Workflows
(F)indability	Indexing in domain relevant databases (e.g., MetaboLights)	Indexing in domain relevant software repositories (e.g., the PhenoMeNal App Library, GitHub)	Indexing in workflow management systems such as Galaxy (e.g., PhenoMeNal, W4M), or libraries such as [69]
	Rich descriptions of metadata (e.g., ISA-Tab)	Tool descriptions follow the EDAM ontology	Persistent identifier (e.g., W4M ID, DOI) and intuitive naming patterns
(A)ccessibility	Data access and rights management based on e.g., data use ontology (DUO)	Accessible open-source licenses	Access to workflow systems can be configured to be shared or restricted
(I)nteroperability	Standard formats for experimental metadata (ISA-Tab/ISA-JSON)	Standardized tool descriptions	Standardized workflow format (e.g., Galaxy GA format, Common Workflow Language CWL)
	Domain specific standards for raw data (e.g., mzML, nmrML)	Containerization of software tools	Execution in various software environments (e.g., through the use of containers)
	OboFoundry vocabularies and established domain ontologies to annotate data	EDAM ontology to annotate tools	Workflow annotation ontologies (e.g., Ontology of workflow motifs for annotating workflow specifications [70])
(R)eusability	Deposition in data repositories (e.g., MetaboLights) and data indexing sites (e.g., OmicsDI)	Rich documentation and usage guides	Rich documentation and tutorials (e.g., Galaxy tours)

data when available), which is displayed to the users as a single Galaxy history data set. The integrated Galaxy modules include a MetaboLights downloader and uploader (for ingestion and submission), an ISAcreate module for the creation of ISA compliant archives, modules to explore study metadata through queries on study factors, ISA-Tab "slicing" where queries are used to select subsets of data files of interest, as well as format conversion (export to ISA-JSON and Workflow4Metabolomics [W4M]) and study metadata validation (Supplemental Table S1).

PhenoMeNal also advanced the specification of the nmrML standard data format [27] and contributed a dedicated composite data type for nmrML to Galaxy. nmrML is used extensively throughout the NMR 1D workflow and conversion from raw format into nmrML is supported via dedicated Galaxy modules (Table 1).

Throughout the entire analysis pipeline, modules of computational workflows were designed to accept standard formats such as mzML, XML or CSV whenever possible.

Standardized APIs/ABIs are being used for the programmatic interfaces as well as for deploying services. To this end, modern and standardized programming, scripting and meta languages were selected such as Go, HCL, Python, Shell, XML and YAML that are widely used in cloud computing.

Reusability

In an ongoing effort, PhenoMeNal is actively advancing the criteria for good data management and stewardship based on findability, accessibility, interoperability and reusability (FAIR) for good data management and stewardship [68] to be applied not only to data, but also to software tools and computational workflows (Table 2).

Privacy

PhenoMeNal supports fully anonymized data, which cannot be traced back to individuals in any way [50] and treats pseudonymized data as identifiable. As pseudonymized data are anonymous to the investigator, third parties may be able to link pseudonymized data back to identifiable individuals through mappings such as a hash or code [49]. In these cases, e.g., in a hospital environment, users must deploy PhenoMeNal within a private cloud or bare metal cluster behind their institution's firewall.

PhenoMeNal provides guidance on ethical and technical frameworks to regulate and secure the use of private or sensitive data [49, 50]. It is possible to combine data and metadata within an ELSI compliant framework [50] and in such cases users can follow the example of the European Genome Phenome Archive (EGA) [71]. In public installations of PhenoMeNal, the ELIXIR policy on privacy has been implemented within a technically secure environment to process data [42].

Security

Open-source tools are used throughout the entire e-infrastructure. This promotes community efforts to discover and resolve bugs and security issues. The container build process is steered by the continuous integration (CI) service Jenkins, which continuously builds the containers and generates reports. On success and through authentication, container images are pushed to the PhenoMeNal container registry, which is publicly available but read-only. Cloud provider credentials are not stored in the cloud but only on the deployer host. The Kubernetes cluster running the Jenkins-CI and the container registry, as well as the portal, runs on a CoreOS container, which is a self-updatable, cluster-aware system with most portions being read-only. It reboots nodes sequentially to avoid lack of availability.

KubeNow is a key component that initializes the cloud infrastructure and configures access to it via Cloudflare [72], providing dynamic Domain Name Services (DNS) and encryption for all network communication. The flexible implementation of PhenoMeNal allows the user to decide to not use Cloudflare, in

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

which case encryption is disabled. KubeAdm, which manages the setup of Kubernetes, is not reachable at runtime by default. The only way to access it is by having access to the private key stored on the computer on which it was launched. PhenoMeNal only allows access to standard ports (ssh, http, https, and port 44 for the Galaxy Downloader) and implements a cloud-specific firewall for all supported cloud providers.

Microservices are designed to be launched on-demand and terminated after completed analysis. If security issues are reported for the microservices, tool, or dependencies or if incremental security patches are available, new builds are automatically triggered in the CI system and developers and the release manager are notified to take additional actions if required. Images are built on a daily basis and tested for deployment to avoid security patches from introducing any abnormality in the deployment process.

User resources

There are many user resources for both PhenoMeNal users and developers in the form of documentation, tutorials, and training videos. The PhenoMeNal Wiki [51] contains detailed documentation on all aspects of PhenoMeNal, including general user guides, workflow and tool tutorials, developer documentation, and general information on topics such as security and the e-infrastructure landscape. The PhenoMeNal portal contains help pages generated from the Wiki [73], which are categorized as User Documentation, Developer Documentation, and Workflow Tutorials. Interactive Galaxy tours are directly integrated in Galaxy [74]. Training videos are available at the project's YouTube page [75].

Availability of source code and requirements

Project name: PhenoMeNal,
Project home page: http://phenomenal-h2020.eu
Operating system(s): Platform independent
Programming language: Go, HCL, Java, JavaScript, Python, R, Shell, XML, YAML
Other requirements: Linux, Docker, Kubernetes, Terraform, Ansible, Helm
License: MIT license for all code written by the PhenoMeNal project. Individual, Open Source Foundation approved licenses for all containerized tools.
RRID:SCR_016605

Availability of supporting data

The following MetaboLights datasets are integrated into PhenoMeNal and are used to demonstrate the cloud integration and reproducibility of Galaxy workflows: MTBLS1 (NMR1D), MTBLS404 (Uni- and multivariate statistics), MTBLS412 (Fluxomics), MTBLS520 (Eco-Metabolomics), MTBLS558 (MetFrag). Datasets are available at https://www.ebi.ac.uk/metabolights. Snapshots of the code and additional supporting data are available in the *GigaScience* repository, GigaDB [76].

Additional files

Supplemental Figure 1: PhenoMeNal implements various layers to provision containers on top of the e-infrastructure.

Supplemental Table 1: List of external software tools that were incorporated into PhenoMeNal.

Abbreviations

ABI: application binary interface; API: application programming interface; AWS: Amazon web services; CI: continuous integration; CPU: central processing unit; CRE: cloud research environment; ELSI: ethical, legal, and social implications; FAIR: criteria for good data management and stewardship based on findability, accessibility, interoperability, and reusability; GCP: Google cloud platform; GUI: graphical user interface; IaaS: Infrastructure-as-a-Service; ISA: Investigation, Study, and Assay data model framework; MS: mass spectrometry; NMR: nuclear magnetic resonance (spectroscopy); PhenoMeNal: Phenome and Metabolome aNalysis; W4M: Workflow4Metabolomics.

Competing interests

The authors declare that they have no competing interests.

Declarations

Human-derived samples in the datasets MTBLS404 and MTBLS412 were processed according to ELSI guidelines.

Author contributions

Writing original draft: K.P. and J.B.
Conceptualization: C.S.
Supervision: R.G., U.L.G., K.H., S.N., A.R., M.vR., C.S., O.S., P.R.-S., R.W.
Project administration: N.K.
Technical lead: P.M.
Review and editing: J.B., M.C., M.Cap., M.Cas., P.dA., T.M.D.E., R.G., A.G.-B., K.H., S.H., D.Ja., D.Jo., F.J., K.K., N.K., P.E.K., A.L., S.L., P.M., S.N., C.O.D., K.P., L.P., M.E.P., M.A.C.R., P.R.-S., P.R.-M., A.R., R.R., C.R., M.vR., N.S., R.M.S., S.-A.S., D.S., O.S., V.S., E.A.T., M.T., T.H., M.vV., M.R.V., R.J.M.W., G.Z., C.S.
Software: J.B., M.Cap., M.Cas., P.dA., A.G.-B., U.L.G., K.H., S.H., D.Jo., F.J., P.E.K., A.L., C.L., P.M., S.N., C.O.D., K.P., L.P., M.E.P., M.A.C.R., P.R.-S., P.R.-M., A.R., R.R., C.R., M.vR., N.S., R.M.S., S.-A.S., O.S., V.S., E.A.T., M.T., T.H., M.vV., R.J.M.W., G.Z.
External software: S.B., C.F., E.H., S.H., M.I., D.Ja., B.K., I.K., K.K., P.E.K., S.L., J.A.N., J.T.M.P., A.P., L.P., R.R.
Data curation: K.H., S.-A.S., P.R.-S.
Funding acquisition: R.G., U.L.G., K.H., S.N., A.R., M.vR., C.S., O.S., P.R.-S., R.W.
Support was provided by the Nordice-Infrastructure Collaboration (NeIC) via the Glenna2 and Tryggve2 projects. All authors contributed to, read, and approved the final manuscript.

Funding

The project was funded by the European Commission PhenoMeNal (grant EC654241). The consortium members include J.B., M.Cap., M.Cas., P.dA., T.M.D.E., R.G., A.G.-B., K.H., M.I., D.Jo., F.J., N.K., P.E.K., A.L., P.M., S.N., C.O.D., K.P., L.P., M.A.C.R., P.R.-S., P.R.-M., A.R., R.R., C.R., T.H., M.vR., M.vV., N.S., R.M.S., S.-A.S., D.S., O.S., V.S., E.A.T., M.T., M.R.V., and R.J.M.W. C.S. received funding from the European Commission PhenoMeNal (grant EC654241).

References

1. Gowda GN, Zhang S, Gu H, et al. Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn 2008;**8**:617–33.

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

2. Bundy JG, Davey MP, Viant MR. Environmental metabolomics: a critical review and future perspectives. Metabolomics 2009;**5**:3–21.
3. Peters K, Worrich A, Weinhold A, et al. Current challenges in plant eco-metabolomics. Int J Mol Sci 2018;**19**:1385.
4. Weber RJM, Lawson TN, Salek RM, et al. Computational tools and workflows in metabolomics: an international survey highlights the opportunity for harmonisation through Galaxy. Metabolomics 2017;**13**:12.
5. Joyce AR, Palsson BØ. The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol 2006;**7**:198–210.
6. Haug K, Salek RM, Conesa P, et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res 2013;**41**:D781–6.
7. Lindon JC, Nicholson JK. The emergent role of metabolic phenotyping in dynamic patient stratification. Expert Opin Drug Metab Toxicol 2014;**10**:915–9.
8. Sumner LW, Hall RD. Metabolomics across the globe. Metabolomics 2013;**9**:258–64.
9. Rosato A, Tenori L, Cascante M, et al. From correlation to causation: analysis of metabolomics data using systems biology approaches. Metabolomics Off J Metabolomic Soc 2018;**14**:37.
10. Vignoli A, Ghini V, Meoni G, et al. High-throughput metabolomics by 1D NMR. Angew. Chem. Int. Ed., 2018, **57**, 2–29, doi:10.1002/anie.201804736.
11. Goodacre R, Broadhurst D, Smilde AK, et al. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 2007;**3**:231–41.
12. Sud M, Fahy E, Cotter D, et al. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res 2016;**44**:D463–70.
13. Giacomoni F, Le Corguille G, Monsoor M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics 2015;**31**:1493–5.
14. Haug K, Salek RM, Steinbeck C. Global open data management in metabolomics. Curr Opin Chem Biol 2017;**36**:58–63.
15. Salek RM, Neumann S, Schober D, et al. COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics 2015;**11**:1587–97.
16. IPCN. International Phenome Centre Network. http://phenomenetwork.org. 2018. Accessed 25 Oct 2018.
17. French Ministry of Research, Higher Education and the National Agency for Science. MetaboHUB. http://www.metabohub.fr/metabohub.html. 2018. Accessed 25 Oct 2018.
18. Tautenhahn R, Patti GJ, Rinehart D, et al. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 2012;**84**:5035–9.
19. Chong J, Soufan O, Li C, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 2018;**46**:W486–94.
20. Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 2016;**44**:W3–10.
21. PhenoMeNal: The PhenoMeNal Portal. https://portal.phenomenal-h2020.eu. 2018. Accessed 25 Oct 2018.
22. Hoffa C, Mehta G, Freeman T, et al. On the Use of Cloud Computing for Scientific Workflows. 2008 IEEE Fourth Int Conf EScience. Indianapolis, IN, USA: IEEE; 2008 [cited 2018 Sep 3]. p. 640–5. Available from: http://ieeexplore.ieee.org/document/4736878/.
23. Digan W, Countouris H, Barritault M, et al. An architecture for genomics analysis in a clinical setting using Galaxy and Docker. GigaScience 2017;**6**:1–9.
24. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 2010;**11**:R86.
25. Novella JA, Khoonsari PE, Herman S, et al. Container-based bioinformatics with Pachyderm, Wren J . editor. Bioinformatics 2018, 1–8; doi:10.1093/bioinformatics/bty699/5068160.
26. Jacob D, Deborde C, Lefebvre M, et al. NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics. Metabolomics 2017;**13**:36.
27. Schober D, Jacob D, Wilson M, et al. nmrML: a community supported open data standard for the description, storage, and exchange of NMR Ddta. Anal Chem 2018;**90**:649–56.
28. Salek RM, Maguire ML, Bentley E, et al. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. Physiol Genomics 2007;**29**:99–108.
29. Buescher JM, Antoniewicz MR, Boros LG, et al. A roadmap for interpreting 13 C metabolite labeling patterns from cells. Curr Opin Biotechnol 2015;**34**:189–201.
30. Niedenführ S, Wiechert W, Nöh K. How to measure metabolic fluxes: a taxonomic guide for 13 C fluxomics. Curr Opin Biotechnol 2015;**34**:82–90.
31. Emami Khoonsari P, Moreno P, Bergmann S, et al. Interoperable and scalable data analysis with microservices: Applications in Metabolomics, Journal: bioRxiv. 2018, **bioRxiv:213603**, 1–29 bioRxiv doi:10.1101/213603.
32. Ruttkies C, Schymanski EL, Wolf S, et al. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminformatics 2016;**8**:3. http://www.jcheminf.com/content/8/1/3.
33. Herman S, Khoonsari PE, Tolf A. et al. Integration of magnetic resonance imaging and protein and metabolite CSF measurements to enable early diagnosis of secondary progressive multiple sclerosis. Theranostics 2018;**8**:4477–90.
34. Thévenot EA, Roux A, Xu Y. et al. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. J Proteome Res 2015;**14**:3322–35.
35. Peters K, Gorzolka K, Bruelheide H, et al. Computational workflow to study the seasonal variation of secondary metabolites in nine different bryophytes. Sci Data 2018;**5**:180179.
36. PhenoMeNal. The Portal App Library. https://portal.phenomenal-h2020.eu/app-library. 2018. Accessed 25 Oct 2018.
37. Rocca-Serra P, Salek RM, Arita M, et al. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics. 2016;**12**:14.
38. Smith B, Ashburner M, Rosse CThe OBI Consortium,, et al., The OBI Consortium, The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;**25**:1251–5.
39. Steinbeck C, Conesa P, Haug K, et al. MetaboLights: towards a new COSMOS of metabolomics data management. Metabolomics 2012;**8**:757–60.
40. Gil Y, Deelman E, Ellisman M, et al. Examining the challenges of scientific workflows. Computer 2007;**40**:24–32.
41. Moutsatsos IK, Hossain I, Agarinis C, et al. Jenkins-CI, an

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

open-source continuous integration system, as a scientific data and image-processing platform. SLAS Discov Adv Life Sci RD 2017;**22**:238–49.
42. van Rijswijk M, Beirnaert C, Caron C, et al. The future of metabolomics in ELIXIR. F1000Research 2017;**6**:1649.
43. EGI Foundation. EGI: Advanced Computing for Research. https://www.egi.eu. 2018. Accessed 25 Oct 2018.
44. INIGO Datacloud. INtegrating Distributed data Infrastructures for Global ExplOitation. https://www.indigo-datacloud.eu. 2018. Accessed 25 Oct 2018.
45. Viljoen M, Dutka L, Kryza B, et al. Towards European Open Science Commons: the EGI Open Data Platform and the EGI DataHub. Procedia Comput Sci 2016;**97**:148–52.
46. Salomoni D, Campos I, Gaido L, et al. INDIGO-DataCloud: a Platform to Facilitate Seamless Access to E-Infrastructures, J Grid Computing, 2018, **16**, 381–408. ArXiv160309536 Cs. doi:10.1007/s10723-018-9453-3.
47. Capuccini M, Larsson A, Carone M, et al. On-demand virtual research environments using microservices, 10.1093/bioinformatics/bty699/5068160, **arXiv:1805.06180**, 1–31. ArXiv180506180 Cs. 2018; http://arxiv.org/abs/1805.06180.
48. Rocca-Serra P, Brandizi M, Maguire E, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 2010;**26**:2354–6.
49. Sariyar M, Schluender I, Smee C, et al. Sharing and reuse of sensitive data and samples: supporting researchers in identifying ethical and legal requirements. Biopreservation Biobanking 2015;**13**:263–70.
50. Heatherly R, Rasmussen LV, Peissig PL, et al. A multi-institution evaluation of clinical profile anonymization. J Am Med Inform Assoc 2016;**23**:e131–7.
51. PhenoMeNal. Wiki. https://github.com/phnmnl/phenomenal-h2020/wiki. 2018. Accessed 25 Oct 2018.
52. PhenoMeNal. GitHub Project Repository. https://github.com/phnmnl/. 2018. Accessed 25 Oct 2018.
53. PhenoMeNal. Phenome and Metabolome aNalysis. https://phenomenal-h2020.eu. 2018. Accessed 25 Oct 2018.
54. PhenoMeNal. Public Galaxy Instance. https://public.phenomenal-h2020.eu. 2018. Accessed 25 Oct 2018.
55. Mell PM, Grance T. The NIST definition of cloud computing. In: Gaithersburg MD . National Institute of Standards and Technology; 2011. Report No.: NIST SP 800-145. Available from: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.
56. PhenoMeNal. Deploy on Microsoft Azure. https://github.com/phnmnl/phenomenal-h2020/wiki/Deploy-on-Microsoft-Azure. 2018. Accessed 25 Oct 2018.
57. PhenoMeNal. Deploy on European Open Science Cloud (EOSC). https://github.com/phnmnl/phenomenal-h2020/wiki/Deploy-on-European-Open-Science-Cloud-(EOSC). 2018. Accessed 25 Oct 2018.
58. PhenoMeNal. Deploy on a local server (bare metal). https://github.com/phnmnl/phenomenal-h2020/wiki/Deploy-on-a-local-server-(bare-metal). 2018. Accessed 25 Oct 2018.
59. Phnmnl GitHub https://github.com/phnmnl/?q=container.
60. PhenoMeNal. How to make your software tool available through PhenoMeNal. https://github.com/phnmnl/phenomenal-h2020/wiki/How-to-make-your-software-tool-available-through-PhenoMeNal. 2018. Accessed 25 Oct 2018.
61. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 2012;**13**:667–72.
62. Sloggett C, Goonasekera N, Afgan E. BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 2013;**29**:1685–6.
63. Thomas K, Benjamin R-K, Fernando P, et al. Jupyter Notebooks - a publishing format for reproducible computational workflows. Stand Alone. 2016;87–90.
64. Lampa S, Alvarsson J, Spjuth O. Towards agile large-scale predictive modelling in drug discovery with flow-based programming design principles. J Cheminformatics 2016;**8**:67.
65. PhenoMeNal. Jenkins-CI Instance. http://phenomenal-h2020.eu/jenkins/. 2018. Accessed 25 Oct 2018.
66. PhenoMeNal. Jenkins Guide. https://github.com/phnmnl/phenomenal-h2020/wiki/Jenkins-Guide. 2018. Accessed 25 Oct 2018.
67. Piras ME, Pireddu L, Zanetti G. wft4galaxy: a workflow testing tool for galaxy. Bioinformatics 2017;**33**:3805–7.
68. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;**3**:160018.
69. myExperiment www.myexperiment.org. Accessed 25 Oct 2018
70. Cohen-Boulakia S, Belhajjame K, Collin O, et al. Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. Future Gener Comput Syst 2017;**75**:284–98.
71. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European genome-phenome archive of human data consented for biomedical research. Nat Genet 2015;**47**:692–5.
72. Cloudflare Inc. Cloudflare. https://www.cloudflare.com/. 2018. Accessed 25 Oct 2018.
73. PhenoMeNal. Portal Help. https://portal.phenomenal-h2020.eu/help. 2018. Accessed 25 Oct 2018.
74. PhenoMeNal. Interactive Galaxy Tours. https://public.phenomenal-h2020.eu/tours. 2018. Accessed 25 Oct 2018.
75. PhenoMeNal. The PhenoMeNal YouTube page. https://www.youtube.com/channel/UCXGAvsVNQk-aUpckjRC8Ang. 2018. Accessed 25 Oct 2018.
76. Peters K, Bradbury J, Bergmann S, et al. Supporting data for “PhenoMeNal: Processing and analysis of Metabolomics data in the Cloud.” GigaScience Database 2018. http://dx.doi.org/10.5524/100528.
77 Brikman Y. Terraform: Writing Infrastructure as Code. Sebastopol: O’Reilly Media; 2017. Available from: http://public.eblib.com/choice/publicfullrecord.aspx?p=4822376.
78 Hanwell MD, de Jong WA, Harris CJ. Open chemistry: RESTful web APIs, JSON, NWChem and the modern web application. J Cheminformatics. 2017;**9**:55.
79 Newman S. Building microservices: designing fine-grained systems. First Edition. Beijing Sebastopol, CA: O’Reilly Media; 2015.
80 Erl T (Ed.). SOA with REST: principles, patterns & constraints for building enterprise solutions with REST. Upper Saddle River, NJ: Prentice Hall; 2012.
81 Bandrowski A, Brinkman R, Brochhausen M, et al. The Ontology for Biomedical Investigations. PLoS One 2016;**11**:e0154556.
82 Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nat Genet 2012;**44**:121–6.
83 Sansone S-A, Schober D, Atherton HJ, et al. Metabolomics standards initiative: ontology working group work in progress. Metabolomics 2007;**3**:249–56.
84 Dyke SOM, Philippakis AA, Rambla De Argila, J et al. Consent Codes: upholding standard data use conditions. PLoS Genet 2016;**12**:e1005772.

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

85 Selivanov VA, Benito A, Miranda, A et al. MIDcor, an R-program for deciphering mass interferences in mass spectra of metabolites enriched in stable isotopes. BMC Bioinformatics 2017;**18**:88.

86 Hao J, Liebeke M, Astle W, et al. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. Nat Protoc 2014;**9**:1416–27.

87 Rinaudo P, Boudah S, Junot C, et al. biosigner: a new method for the discovery of significant molecular signatures from omics data. Front Mol Biosci 2016;**3**:26.

88 Kuhl C, Tautenhahn R, Böttcher C, et al. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 2012;**84**:283–9.

89 Dührkop K, Shen H, Meusel M, et al. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci 2015;**112**:12580–5.

90 Southam AD, Weber RJM, Engel J, et al. A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. Nat Protoc 2017;**12**:255–73.

91 King ZA, Dräger A, Ebrahim A, et al. Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. PLOS Comput Biol 2015;**11**:e1004321.

92 Cottret L, Frainay C, Chazalviel M, et al. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res 2018;**46**:W495–502.

93 Libiseller G, Dvorzak M, Kleb U, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics 2015;**16**:118.

94 González-Beltrán A, Neumann S, Maguire E, et al. The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again. BMC Bioinformatics 2014;**15**:S11.

95 Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nat Genet 2012;**44**:121–6.

96 Selivanov VA, Vizán P, Mollinedo F, et al. Edelfosine-induced metabolic changes in cancer cells that precede the overproduction of reactive oxygen species and apoptosis. BMC Syst Biol 2010;**4**:135.

97 Perez F, Granger BE. IPython: a system for interactive scientific computing. Comput Sci Eng 2007;**9**:21–9.

98 Ludwig C, Günther UL. MetaboLab - advanced NMR data processing and analysis for metabolomics. BMC Bioinformatics 2011;**12**:366.

99 Wohlgemuth G, Haldiya PK, Willighagen E, et al. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. Bioinformatics 2010;**26**:2647–8.

100 Rueedi R, Mallol R, Raffler J, et al. Metabomatching: using genetic association to identify metabolites in proton NMR spectroscopy. PLOS Comput Biol 2017;**13**:e1005839.

101 Helmus JJ, Jaroniec CP. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. J Biomol NMR 2013;**55**:355–67.

102 Mohamed A, Nguyen CH, Mamitsuka H. NMRPro: an integrated web component for interactive processing and visualization of NMR spectra. Bioinformatics 2016;**32**:2067–8.

103 Sturm M, Bertsch A, Gröpl C, et al. OpenMS – an open-source software framework for mass spectrometry. BMC Bioinformatics 2008;**9**:163.

104 Blaise BJ, Correia G, Tin A, et al. Power analysis and sample size determination in metabolic phenotyping. Anal Chem 2016;**88**:5179–88.

105 Scheubert K, Hufsky F, Petras D, et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 2017;**8**, 1–24. doi:10.1038/s41467-017-01318-5.

106 Chambers MC, Maclean B, Burke R, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 2012;**30**:918–20.

107 Lewis IA, Schommer SC, Markley JL. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. Magn Reson Chem 2009;**47**:S123–6.

108 Rodriguez N, Thomas A, Watanabe L, et al. JSBML 1.0: providing a smorgasbord of options to encode systems biology models: Table 1. Bioinformatics 2015;**31**:3383–6.

109 Benton HP, Wong DM, Trauger SA, et al. XCMS ^{2} : processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem 2008;**80**:6382–9.

Downloaded from https://academic.oup.com/gigascience/article-abstract/8/2/giy149/5232984 by guest on 22 March 2020

DOI:10.1093/gigascience/giy149

Curriculum Vitae

Steffen Neumann

04. August 1972, Düsseldorf (Germany)

Address

Leibniz Institute of Plant Biochemistry
Weinberg 3
Germany

Phone: +49 (0) 345 5582 1470
Fax: +49 (0) 345 5582 1409
Email: `sneumann@IPB-Halle.DE`

Academic Career and Occupation

since 2019	Independent group leader “Bioinformatics and Research Data”
2005-2019	Group leader “Bioinformatik und Massenspektrometrie” at the IPB Halle. Our projects cover the various stages in a bioinformatics- and metabolomics pipeline.
2004	Postdoc at IPK Gatersleben in the “Plant Data Warehouse” group of Dr. Grosse.
2004	Postdoc at Bielefeld University, in the group of Prof. Sagerer, applied computer science.
1999-2003	Ph.D. at Bielefeld University, in the group of Prof. Sagerer. Title: “Soft volume models for protein-protein docking”.
1994/95	Erasmus studies in “Computer Science” and “Biotechnology” at Dublin City University (Ireland).
1992-1999	Study of “Computer- and natural science” at Bielefeld University. Specialisation in applied computer science, neurobiology and -psychology.

Publications: citation summary (November 2021)

	All	Since 2016
Citations	13060	9846
h-index	40	38
i10-index	73	69

Based on `http://scholar.google.com/citations?user=EcQVenkAAAAJ`

Memberships and Board positions

since 2019	Elected Member Strategy Science Board of the German Centre for Integrative Biodiversity Research (iDiv)
2018	Clarivate/WebOfScience *Highly cited Researcher* (Category Cross-Field)
since 2017	Member German Centre for Integrative Biodiversity Research (iDiv)
2014-18	Elected director of the international Metabolomics Society
since 2014	Member of MetaboHUB International Scientific Committee
since 2013	Editorial Board “Scientific Data”, Nature Publishing Group
since 2012	Editorial Board “Frontiers in Plant Systems Biology”, Frontiers Media S.A
since 2011	Editorial Board “Metabolites”, MDPI Publishing, and guest editor of special issue “Small Molecule Identification beyond the Crystal Ball - Selected Papers from CASMI”

Halle, den 13. November 2021

Erklärung an Eides statt

Ich erkläre hiermit, dass ich die vorliegende Habilitationsschrift

“Computational Mass Spectrometry in Metabolomics”

selbständig und ohne fremde Hilfe verfasst, und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die wörtlich oder inhaltlich den benutzten Werken entnommenen Stellen sind als solche kenntlich gemacht.

Halle (Saale), 22.06.2020

Dr. Steffen Neumann