



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Semi-automated Comparison and Alignment of
LC-HRMS Data from Different Metabolomics Experiments

verfasst von / submitted by

Stefan Koch, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

A 066 910

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Computational Science UG2002

Betreut von / Supervisor:

Prof. Dr. Rainer Schuhmacher

Abstract:

Retrospective and comparative statistical analysis of LC-HRMS (liquid chromatography coupled to high-resolution mass spectrometry) chromatograms of separate biological experiments is of particular interest since they hold a considerable amount of information that can be used to investigate aspects beyond the scope of their initial acquisition and evaluation. However, shifts and drifts in retention times (R_t) and mass to charge ratio values (m/z) or the formation of different ion species caused by the limited stability of instrument parameters are big obstacles for the comparative analysis of LC-HRMS chromatograms. This is especially problematic when results obtained in multiple analytical batches using the same instrument are compared. Even more problems arise when the data are obtained from measurements on different instruments. Several algorithms and software tools have been proposed to address this problem, but many of them suffer from severe drawbacks such as incorrectly modeling drifts and shifts, complex user interaction and long run times.

In the presented thesis, the novel software tool MetMatch has been developed, which is intended to be used for the comparison of LC-HRMS chromatograms. Chromatograms are matched to a feature list, which can be created manually or by processing one or multiple datafiles that serve as the reference. Constant shifts in m/z are automatically detected and corrected. Both linear and non-linear R_t shift correction is achieved using a fast, iterative algorithm based on the observation that metabolites/compounds exhibiting similar R_t values within a measurement usually share analogous R_t shifts between measurements. Furthermore, MetMatch is able to account for the formation of different ion species during electrospray ionization (ESI). The developed software offers a convenient graphical user interface and diagnostic plots that allow for easy optimization and verification of every processing step.

The quality of results obtained by MetMatch was assessed using two different test cases. The first case involved 75 newly characterized Phe-derived metabolites that were matched to a wheat experiment obtained several years earlier. Manual validation of 23 selected features was performed, 19 of which could be confirmed in all 10 inspected target files. Additionally, 36 T-2 toxin/HT-2 toxin-derived metabolite ions were matched to another barley experiment, 32 of which could be manually verified in all three target files. In both cases, MetMatch-derived R_t shifts were highly similar to XCMS processing results (correlation > 0.9), which is a software tool frequently used for untargeted metabolomics research. Furthermore, software evaluation by means of artificially manipulated reference data confirmed that MetMatch is well suited to account for the formation of different ion species between reference and target datasets and that both m/z and R_t shifts are detected with high accuracy (on average, 96.9% of all peaks were correctly matched).

Zusammenfassung:

Die retrospektive und vergleichende Analyse von LC-HRMS (Flüssigchromatographie mit hochauflösender Massenspektrometrie-Kopplung) Chromatogrammen unterschiedlicher biologischer und/oder analytischer Experimente ist von besonderem Interesse, da diese eine beträchtliche Menge an Informationen enthalten können, die über den Umfang der ursprünglichen Auswertung hinausgehen. Die vergleichende Analyse von LC-HRMS-Chromatogrammen wird jedoch durch einige Faktoren, wie beispielsweise der Verschiebungen der Retentionszeit (R_t) und des Masse-zu-Ladungs-Verhältnisses (m/z), erschwert. Zusätzlich führt die Bildung unterschiedlicher Ionenspezies über mehrere Experimente hinweg zu einem erheblichen Mehraufwand in der Datenprozessierung, der von aktuellen Softwareprogrammen nicht genügend beachtet wird. Dies gilt sowohl für den Vergleich von Ergebnissen, die in mehreren Messungen mit demselben Gerät erzielt wurden, als auch insbesondere für Messungen, die an verschiedenen Geräten durchgeführt wurden. Viele der existierenden Algorithmen und Softwaretools für diese Problemstellung weisen gravierende Nachteile, wie beispielsweise falsch modellierte Verschiebungen auf oder haben lange Programmlaufzeiten.

In der vorliegenden Arbeit wurde das neuartige Softwaretool MetMatch entwickelt, das für den Vergleich von LC-HRMS-Chromatogrammen verwendet werden kann. Die auszuwertenden Chromatogramme werden mit einer als Referenz dienenden Liste an Messergebnissen verglichen und auf diese abgebildet. Konstante Verschiebungen in der m/z Dimension werden automatisch erkannt und korrigiert. Sowohl die lineare als auch die nichtlineare Korrektur von R_t Verschiebungen werden durch einen schnellen, iterativen Algorithmus erkannt und entsprechend korrigiert. Dieser basiert auf der Beobachtung, dass Metabolite, die ähnliche R_t -Werte innerhalb einer Messung zeigen, im Normalfall stark korrelierte R_t -Verschiebungen zwischen Messungen aufweisen. Darüber hinaus kann MetMatch die Bildung verschiedener Ionenspezies berücksichtigen. Die entwickelte Software bietet eine komfortable graphische Benutzeroberfläche, die eine einfache Optimierung und Überprüfung aller Verarbeitungsschritte ermöglicht.

Die Qualität der mit MetMatch erzielten Ergebnisse wurde anhand von zwei unterschiedlichen Experimenten bewertet. Das erste Experiment befasste sich mit 75 Phenylalanin Metaboliten, die in einem früheren Weizenversuch gefunden werden sollten. Eine manuelle Validierung von 23 ausgewählten Metaboliten zeigte, dass 19 davon in allen 10 inspizierten Datensätzen korrekt zugeordnet und korrigiert wurden. Darüber hinaus wurden für Gerstenproben vorliegende Messergebnisse auf 36 T-2-Toxin / HT-2-Toxin Metabolite, welche bereits aus einem früheren Experiment bekannt waren, überprüft. Von diesen konnten 32 Substanzen in allen drei Zieldateien manuell verifiziert werden. In beiden Fällen waren die von MetMatch detektierten R_t -Verschiebungen den durch das XCMS-Paket (eine im Bereich der untargeted Metabolomics häufig verwendete Software) errechneten Ergebnissen sehr ähnlich (Korrelation > 0.9). Darüber hinaus wurde die Software anhand künstlich manipulierter Referenzdaten evaluiert, wodurch gezeigt werden konnte, dass MetMatch gut geeignet ist, um die Bildung unterschiedlicher Ionenspezies zwischen Referenz- und Zieldatensätzen zu berücksichtigen, und dass sowohl m/z - als auch R_t -Verschiebungen mit hoher Genauigkeit detektiert werden konnten (durchschnittlich 96,9% aller Peaks wurden korrekt abgeglichen).

Contents

1. Introduction.....	7
1.1 Biological Background.....	7
1.1.1 Metabolomics: The Targeted Approach	9
1.1.2 Metabolomics: The Untargeted Approach	9
1.2 Analytical Background.....	9
1.2.1 Liquid Chromatography (LC)	10
1.2.2 Mass Spectrometry (MS)	12
1.2.3 Causes for Analytical Shifts.....	13
1.3 Computational Background	13
1.3.1 LCMS Datasets and their Alignment.....	13
1.3.2 Peak Picking.....	16
1.3.3 Existing Software Tools.....	18
1.4 Aims of this Thesis	18
1.5 Declaration of Thesis by Publication.....	19

2. Publication.....	21
----------------------------	-----------

MetMatch: A Semi-Automated Software Tool for the Comparison and Alignment of LC-HRMS Data from Different Metabolomics Experiments

1. Introduction.....	21
2. Materials and Methods.....	23
2.1 Computational Workflow.....	23
2.2 Inclusion of Alternative Ion Forms	23
2.3 <i>m/z</i> Offset Detection.....	23
2.4 Peak Picking.....	23
2.5 Rt Shift Detection.....	25
2.6 Output of Generated Results	25
2.7 Phe-Derived Biological Data.....	27
2.8 T-2 Toxin/HT-2 Toxin-Derived Biological Data	28
3. Results and Discussion.....	28
3.1 Accepted Input Files.....	30
3.2 Detection of Differently Formed Ions	31
3.3 <i>m/z</i> Shift Detection.....	31
3.4 Rt Shift Detection.....	31
3.5 Manual Corrections of Automatically Generated Results.....	32
3.6 Output Format.....	32
3.7 Runtimes	32
3.8 Evaluation of the Phe-Derived Biological Metabolites.....	33
3.9 Evaluation of the T-2 Toxin/HT-2 Toxin-Derived Biological Data	34
3.10 Evaluation of Artificial Data	35
3.10.1 Formation of Different Ion Species.....	36
3.10.2 <i>m/z</i> Shift Detection	36
3.10.3 Rt Shift Detection	37
3.10.4 Minimum Required Common Feature Groups	38
4. Conclusion.....	39
Abbreviations and Terminology.....	40
References	42
Supplementary Materials.....	44

3. MetMatch's User Interface	55
4. Conclusion.....	59
Abbreviations and Terminology.....	60
Bibliography	62
Acknowledgements	66

List of Figures

Figure T1: Omics Cascade	7
Figure T2: Exemplification of metabolomics by targeted and untargeted approaches.	8
Figure T3: Chromatography	11
Figure T4: HPLC dataset	14
Figure T5: Shifts between datasets	15
Figure T6: Peak picking	17

MetMatch: A Semi-Automated Software Tool for the Comparison and Alignment of LC-HRMS Data from Different Metabolomics Experiments

Figure 1: MetMatch computational workflow	24
Figure 2: m/z shift detection	25
Figure 3: Rt shift detection procedure.....	26
Figure 4: Exemplification of the iterative Rt shift detection algorithm	27
Figure 5: MetMatch data processing workflow	29
Figure 6: User interface	30
Figure 7: EIC correction with MetMatch (alignment before and after Rt and m/z correction)	35
Figure 8: Box-plot showing the differences between artificially imposed and detected m/z offset	37
Figure 9: Results for six different simulated Rt shifts.....	38
Figure 10: Illustration of the F-measure for reference feature lists of decreasing size	39
Figure S1: Pseudocode of the iterative Rt shift detection algorithm.....	45
Figure S2: Comparison of Rt shift in the Phe-derived biological experiment detected by MetMatch and XCMS.....	51
Figure S3: Comparison of the Rt shift detected in the Phe-derived biological experiment using chromatograms obtained in positive and negative ionization mode	52
Figure S4: Comparison of the Rt shift in the T-2 toxin/HT-2 toxin-derived biological experiment detected by MetMatch and XCMS.....	53

Figure T7: MetMatch's main window.....	55
Figure T8: EICs of an inspected ion adduct before and after Rt alignment	56
Figure T9: Animated Rt shift detection	56
Figure T10: Individual iterations of the animated Rt shift detection algorithm.....	57
Figure T11: Exemplification of manual verification within the user interface	58

List of Tables

MetMatch: A Semi-Automated Software Tool for the Comparison and Alignment of LC-HRMS Data from Different Metabolomics Experiments

Table 1: Structure of a typical reference feature list	31
Table 2: Detail of an output feature matrix	33
Table 3: Benchmark results for artificial data evaluation	36
Table 4: Results of the artificial ion species manipulation	36
Table S1: Sample runtimes with different peak picking algorithms.....	44
Table S2: Manual validation of the results obtained by matching the target dataset to the reference feature list containing Phe-derived metabolites.....	52

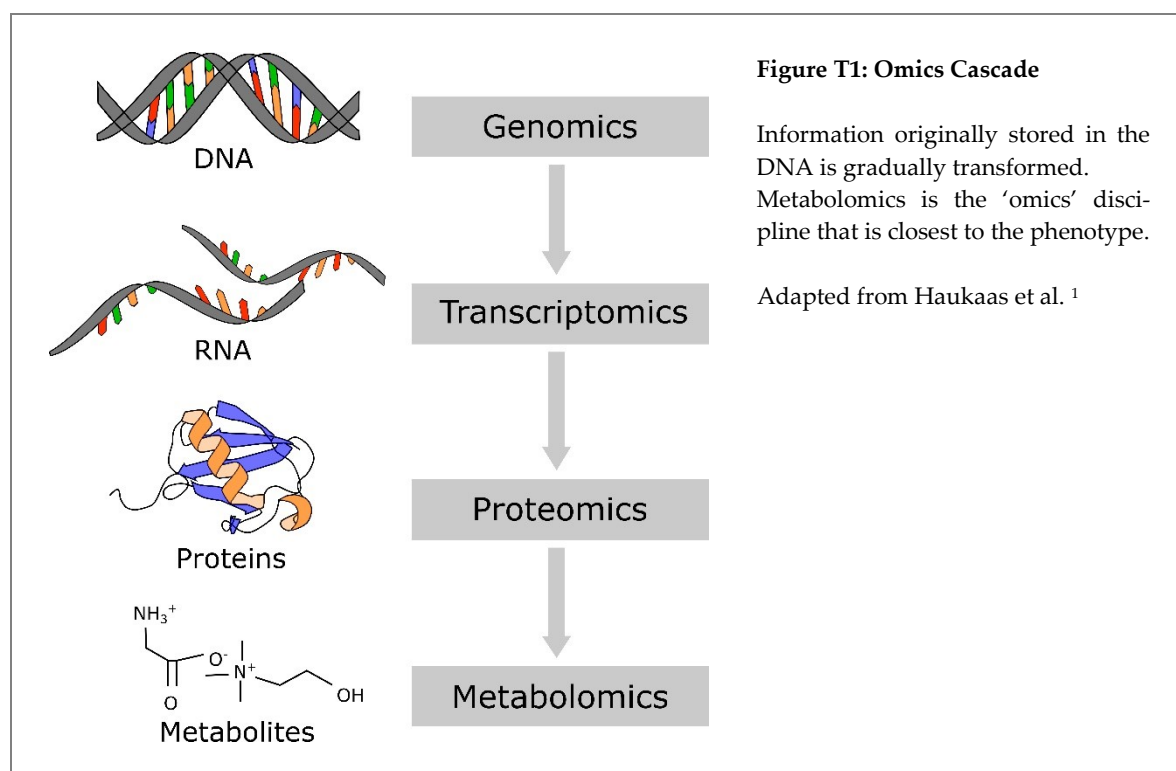
Chapter 1

Introduction

Comparative analysis of different LC-HRMS experiments often has to deal with analytical drifts and shifts between the datasets. Suitable algorithms have to be employed to correct for these shifts in order to minimize the analytical and technical variation of the datasets. During the development of such algorithms, the biological meaning, analytical peculiarities and computational limitations of this task have to be considered. This chapter is aimed at providing an overview of the most important factors relevant for the alignment of LC-HRMS metabolomics datasets.

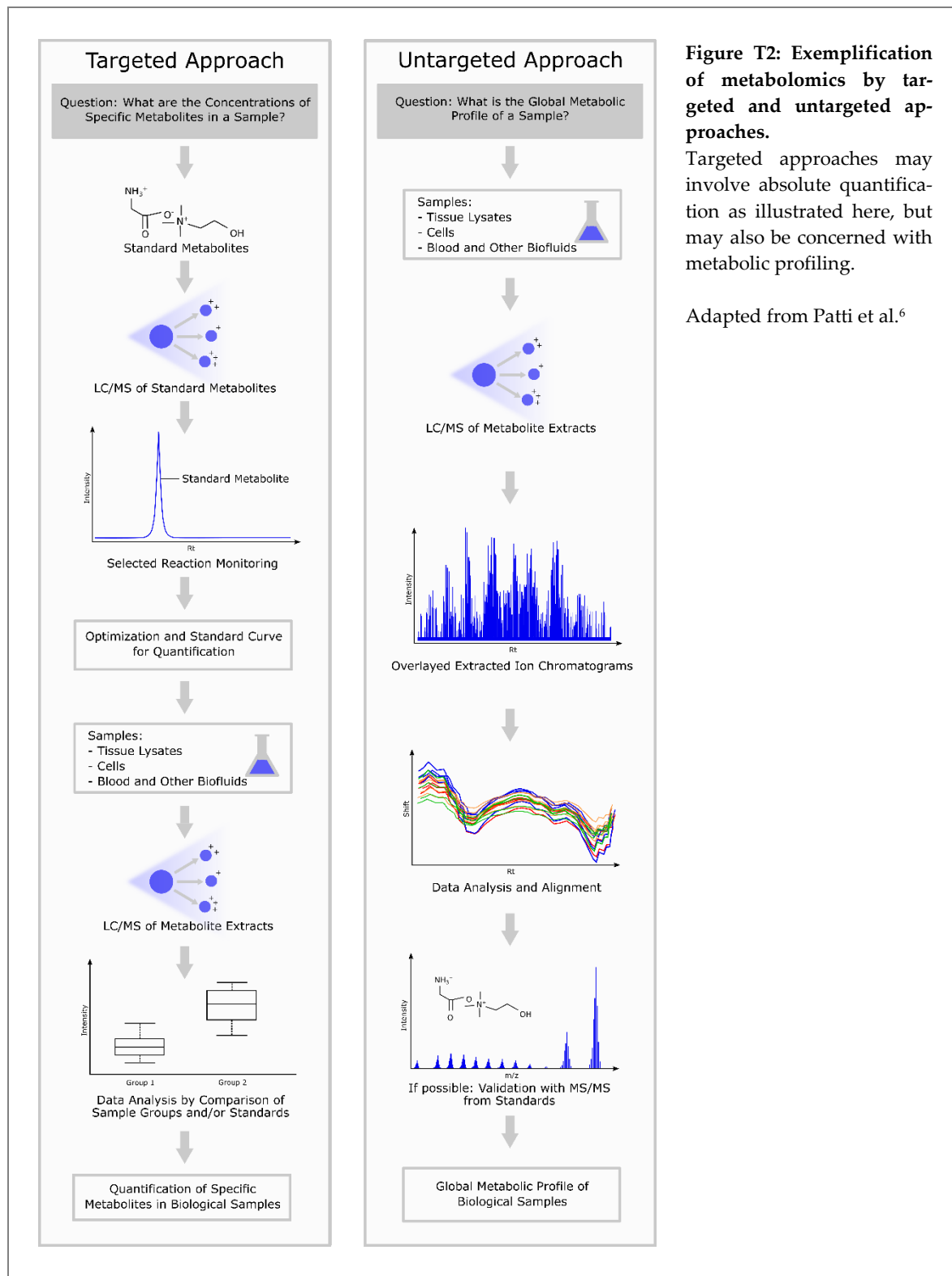
1.1 Biological Background

Comprehensive understanding of life and its biological processes at a molecular level still poses a big challenge, but fueled by the ongoing technological progress, we are getting closer to this major goal of systems biology each year. The omics cascade (outlined in **Figure T1**) is an invaluable tool in this undertaking, containing the disciplines of genomics, transcriptomics, proteomics and metabolomics; the latter of which is used to describe the metabolome, thereby encompassing all small and low-molecular-weight molecules produced and transformed in metabolism.



Metabolites span a vast number of compound classes, including amino acids, lipids, organic acids and nucleotides, just to name a few.² While the number of human metabolites was estimated to be around 2000 in the year 2003³, 18,557 metabolites have been detected and quantified in human biofluids, cells and tissues by the year 2018⁴. Not all of them are produced within the human body however, as the presence of xenobiotics (e.g., drugs or plant metabolites) is to be expected. This number is still dwarfed by the plant kingdom, with an estimated number of at least 200,000 metabolites.⁵

Metabolomics research covers a wide range of applications, organisms, analytical techniques and types of generated data. There are however two very distinct approaches: targeted and untargeted (Figure T2).



1.1.1 Metabolomics: The Targeted Approach

Targeted approaches are concerned with the analysis of a subset of metabolites rather than the whole metabolome. Usually, the subset is chosen according to a pre-existing hypothesis and/or available analytical methods. Based on the subset of interest, appropriate analytical procedures are applied, and the results are typically limited to a small number of pathways. One possible application is to measure biomarkers correlated to diseases or toxins.⁷ A targeted approach could be focused on measuring these biomarkers directly and comparing the quantities of the various biomarkers in the different treatment groups.

Thanks to technological advances, the subset of metabolites that can be analyzed simultaneously has steadily increased over the past decades and more and more substances become available as purified compounds at lower prices. In fact, the analysis of specific biomarkers can be as complex as profiling all metabolites in a specific pathway.² Therefore, the observation of the effects of perturbations on a larger number of pathways became much more feasible, paving the way for untargeted approaches.

1.1.2 Metabolomics: The Untargeted Approach

Untargeted approaches are concerned with the measurement of the whole metabolome, rather than just a small, predefined subset. Due to the large scale, the identification of all measured metabolites is often not feasible, and instead, the pattern or profile of the metabolites is used for further investigations.² These profiles are typically described by m/z (mass to charge ratio) values, intensities, and R_t (retention time) of the detected metabolites.² Suitable sample preparation and analytical methods have to be applied in order to not exclude or discriminate any metabolites in the analysis – an impossible task, yet certain protocols enable to detect more metabolites than others.

The generated datasets can be very complex and often exceed file sizes of multiple gigabytes.⁶ Therefore, inspecting and comparing every single peak manually is not possible. Comparison of different samples is further complicated by shifts in R_t , usually caused by temperature fluctuations or column degradation.⁶ This is why tailored software tools are used to analyze the data in an automated fashion.

Untargeted approaches enable researchers to examine the effects of genetic alterations or external influences such as toxins on a large number of biochemical pathways, not just on a few selected ones. This is especially valuable if pleiotropic effects cause a system-wide perturbation, not foreseen in the initial hypothesis of the experiments. Because of this, untargeted approaches sometimes generate new hypotheses for research.

Common applications of untargeted approaches include metabolic engineering, with the goal to maximize the metabolic flux into desired pathways, for instance to increase the pharmaceutical output of microbes or plants⁸, therefore reducing production costs while increasing the amount of the desired compounds. They are also used in determining whether genetically modified organisms may have unforeseen properties compared to their unaltered cultivars.⁹ Furthermore, they have proven to be of great use in the investigation of plant-pathogen interaction⁷; two such studies were used to exemplify the performance of MetMatch.

1.2 Analytical Background

Apart from the scientific questions a certain dataset is supposed to answer, it is also imperative to understand the details of the analytical methods applied in its generation in order to develop suitable analysis software tools.

Analytical methods in the field of metabolomics need to be able to analyze samples with a high degree of variability. Whereas genomics and proteomics are concerned with biopolymers built from a small set of building blocks (the 4-letter code of our genes and the 20 amino acids of our proteins),

metabolomics is dependent on analytical methods capable of distinguishing the plethora of molecules encountered in organic chemistry. Due to the complex, branched structure of molecules, it is not possible to simply sequence metabolites in order to identify them. In contrast, direct comparison to authentic standard compounds is required to unambiguously identify molecules in biological samples.¹⁰ Chromatography (for separation) and mass spectrometry (for qualification and quantification) are most commonly used today as they provide a complementary yet sensitive view into the metabolic composition of the studied samples.

1.2.1 Liquid Chromatography (LC)

Liquid Chromatography is a method of separation, which is achieved by varying levels of attraction of a sample's constituents with the mobile and the stationary phase. The mobile phase carries the sample through a column typically packed with a solid substance – the stationary phase. The more the individual components interact with the stationary phase (adsorption), the longer they need to pass through the analytical column (**Figure T3**).

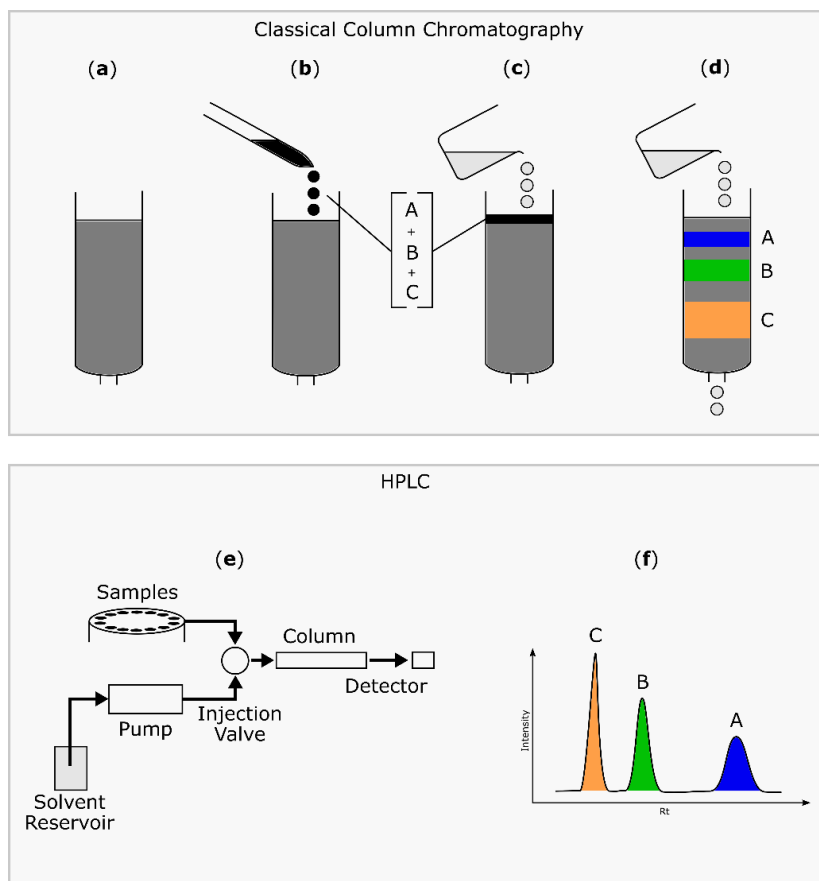


Figure T3: Chromatography

a-d) classical column chromatography outlining the principle of separation.

a) column packed with the stationary phase.

b) sample consisting of three components (A, B and C) is applied.

c) mobile phase solvent is added.

d) through gravity and interactions with the stationary and mobile phase, the components of the sample move down the column at different speeds, gradually separating themselves.

e-f) high performance liquid chromatography (HPLC) separation.

e) schematic of a modern HPLC setup. Samples are automatically injected into the column, and separated by the continuous flow of solvent. As the components leave the column, their intensities are recorded by a detector.

f) chromatogram created by plotting the output of the detector.

Adapted from Snyder et al¹¹.

Some analytes do not interact with the stationary phase at all and thus no retention and therefore separation of these analytes can be achieved with the chosen analytical conditions. Consequently, they have to be excluded from further analysis. If, on the other hand, the analytes have a very strong attraction to the stationary phase, they are not transported by the mobile phase, consequently accumulate in the column and are therefore not detected. The composition of mobile and stationary phase should be chosen in such a way, as to reduce the number of components that are either too strongly or too weakly attracted to the stationary phase.

High performance liquid chromatography has been developed to increase the efficiency of LC by reducing the diameter of the stationary phase particles while simultaneously increasing the pressure at which the mobile phase is pumped through the column. Using these two alterations, higher degrees of separation can be achieved within shorter runtimes compared to standard LC, making HPLC one of the most commonly used analytical methods.¹¹

In theory, chromatography can be used to identify compounds based on their R_t . However, R_t is highly dependent on the experimental conditions. Identification would require a R_t catalogue of reference compounds acquired under the same conditions. And even if such a catalogue would exist, most compounds couldn't be identified with sufficient certainty, as the number of compounds with similar R_t in a typical biological sample is just too high.¹² This is why in metabolomics research chromatography is often combined with additional analytical techniques such as mass spectrometry.

1.2.2 Mass Spectrometry (MS)

Mass spectrometry is an analytical method used to qualify and quantify compounds within a sample. In simplified terms, the sample is introduced into the mass spectrometer, where ions are formed from the compounds, which are then separated by their mass to charge ratio and finally routed to a detector.¹²

Analysis of complex mixtures comprised of a large number of compounds will generate mass spectra that consist of mass signals originating from all compounds contained in the mixture and are thus much harder to evaluate than spectra of pure compounds. The combination of mass spectrometry with sample separation steps such as liquid chromatography is an elegant solution for this problem.

Not all ionization methods are well suited for the complex mixture of sample fractions and solvent typically used for liquid chromatography. Electron ionization, chemical ionization, fast-atom bombardment, thermospray, electrospray and atmospheric-pressure chemical ionization have proven to generate ions and fragments from non-charged molecules under these conditions.¹²

In its essence, a mass spectrum (the result of a mass spectrometer) is a list of ions falling into certain m/z bins. The size of these bins, and therefore the resolution of the resulting mass spectrum is largely influenced by the method of ion separation. At low resolutions, it is only possible to differentiate ions with different nominal mass to charge ratios such as the most abundant ^{13}C isotopologs. If the experiment requires a differentiation at the level of isotopes, more sophisticated methods of ion separation have to be applied and instruments with a higher resolution need to be used. Additionally, each method works best under a certain set of conditions (e.g., operating pressures, composition of sample matrix).

In the following section, a short overview of commonly used ion separation methods is given:

The Quadrupole Mass Analyzer separates ions by an electric field generated between four parallel rods. Depending on the voltages applied to the rods, only ions with a specific m/z can pass through the electric field and subsequently be detected. This method is very well suited for the combination with liquid chromatography due to its tolerance for high operating pressures. In general, the resolving power of a quadrupole mass analyzer is set to unit resolution (i.e. full width of mass peak at half max height (FWHM) is set to e.g. 0.2 amu) and kept constant over the whole scan range, so that the major ^{13}C isotopologs can be resolved properly.

The Ion-Trap Mass Analyzer, as the name suggests, traps ions in a complex yet stable trajectory. Through a change in voltage, the trajectories of ions with specific m/z become unstable, directing them towards the detector. Similar to the linear quadrupole analyzer, the ion-trap is a low-resolution mass analyzer, typically operated at unit resolution with high compatibility for liquid chromatography.¹²

The Time-of-Flight (ToF) Mass Analyzer is based on the idea that propelled by the same kinetic energy, ions with higher m/z values take longer to reach the detector than ions with lower m/z values. This principle requires a pulsed sample injection, otherwise the difference between the time of injection and detection cannot be measured precisely. ToF measurements result in high-resolution data¹², typically up to resolving power of ca. 40.000.

The Orbitrap Mass Analyzer traps ions on an axially oscillating orbit around an inner electrode. This induces an electric current in the outer electrode, which is measured and converted into a mass spectrum using Fourier transformation. Commercial Orbitrap mass analyzers can be operated at very high resolution settings¹³ of up to ca. 280.000, which is even sufficient to separate ^{13}C and ^{34}S or for lower m/z values even ^{13}C and ^{15}N isotopolog peaks.

1.2.3 Causes for Analytical Shifts

Even with modern, robust LC-HRMS instruments, shifts and drifts in R_t and m/z within a dataset are inevitable. There are many factors that cause such shifts. Sometimes MS instruments show a mass deviation (i.e., offset) compared to the exact mass of an ion. Chromatographic results (e.g., retention time) can be heavily influenced by column age, the state of the pump or pH of the mobile phase. And even small temperature fluctuations can have a noticeable impact.^{14,15}

Samples that have been analyzed with the same instruments and instrument parameter settings usually exhibit only very minor shifts. These shifts become more pronounced if samples were analyzed in different batches, which is often inevitable (e.g., large number of samples in an experiment). The comparison of datasets between experiments analyzed in different instrumental setups is usually accompanied by large shifts which prevent straightforward matching of metabolites across different batches. Yet, such comparisons are very valuable, as they enable researchers to investigate novel biological aspects in already existing datasets. Therefore, suitable algorithms and software are required for efficient data processing. Comparative analysis of datasets exhibiting substantial shifts would not be feasible without them.

1.3 Computational Background

1.3.1 LCMS Datasets and their Alignment

From a strictly computational point of view, the task is to align 3-dimensional datasets. The dimensions are R_t , m/z and intensity. An example of such a dataset is shown in **Figure T4**. These datasets show shifts of varying degree, typically only in two dimensions (R_t and m/z), as visualized in **Figure T5**. Alignment and matching of peaks between datasets are complicated by the large amount of noise that is present in LC-HRMS data. Additionally, the number of peaks with identical metabolic origin (i.e., the peaks that should be matched) between two datasets can be very low, increasing the risk of overfitting and potentially leading to incorrect matches.

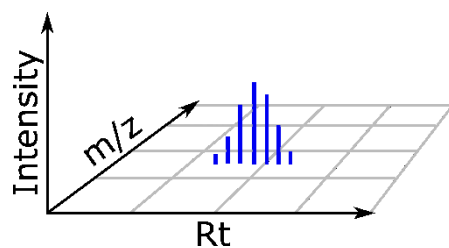


Figure T4: HPLC dataset

This plot represents a 3-dimensional HPLC dataset. Each signal is defined by three values: R_t , m/z and intensity. Ideally, signals corresponding to a detected metabolite ion are clustered, resembling a bell curve due to chromatographic elution of metabolites.

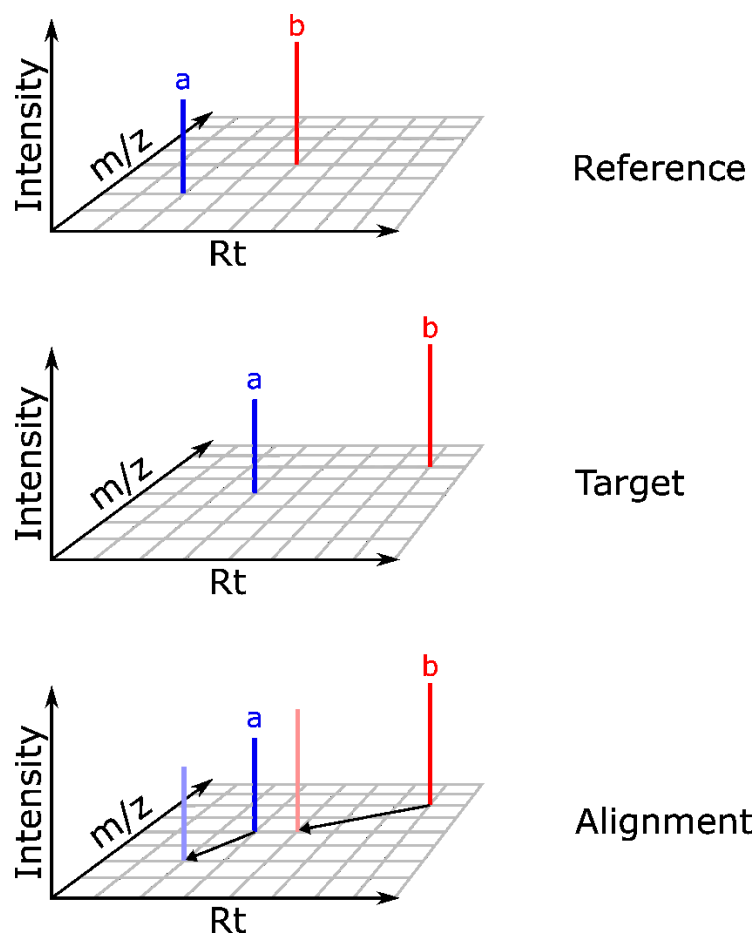


Figure T5: Shifts between datasets

In this example, reference and target dataset both contain peaks corresponding to two different metabolites (a and b). The peaks of the target dataset are shifted in m/z and Rt relative to the reference. The goal of alignment is to find a function that maps peaks of the same metabolite between datasets, correcting for shifts of the target relative to the reference. Such a function allows for the matching of peaks between datasets.

Earlier algorithms performed alignment primarily based on the TIC (total ion current) chromatogram^{16–18}, whereas more recent approaches also take the m/z dimension into account^{19–21}. The majority of the available tools achieve alignment through various warping techniques.¹⁵ Warping refers to means of manipulating one or both dimensions (Rt and m/z) by stretching and/or compressing them in order to maximize data consistency.²² Warping functions can be either constant, linear or non-linear. In many cases, constant and linear functions don't reflect the real-world nature of the observed shifts. Non-linear warping functions (e.g., polynomial functions or splines) are usually applied by more sophisticated algorithms, yielding more accurate results as they also take local shifts into consideration and try to model them. However, they often suffer from undesirably long runtimes.¹⁵

1.3.2 Peak Picking

As illustrated in **Figure T4**, signals corresponding to a certain metabolite isotopolog are clustered and usually resemble a bell curve. It is therefore beneficial to perform peak picking (**Figure T6**) prior to any alignment, as noise signals are filtered out and the amount of information that needs to be considered during the alignment step can be decreased dramatically. Apart from data reduction, peak picking is necessary for further analysis of the dataset. Researchers are interested in chromatographic peaks and their total area rather than individual signals. The goal of peak picking is to cluster signals originating from the same ion and to reduce them to a smaller set of datapoints (i.e., peaks), each with a single R_t and m/z value.

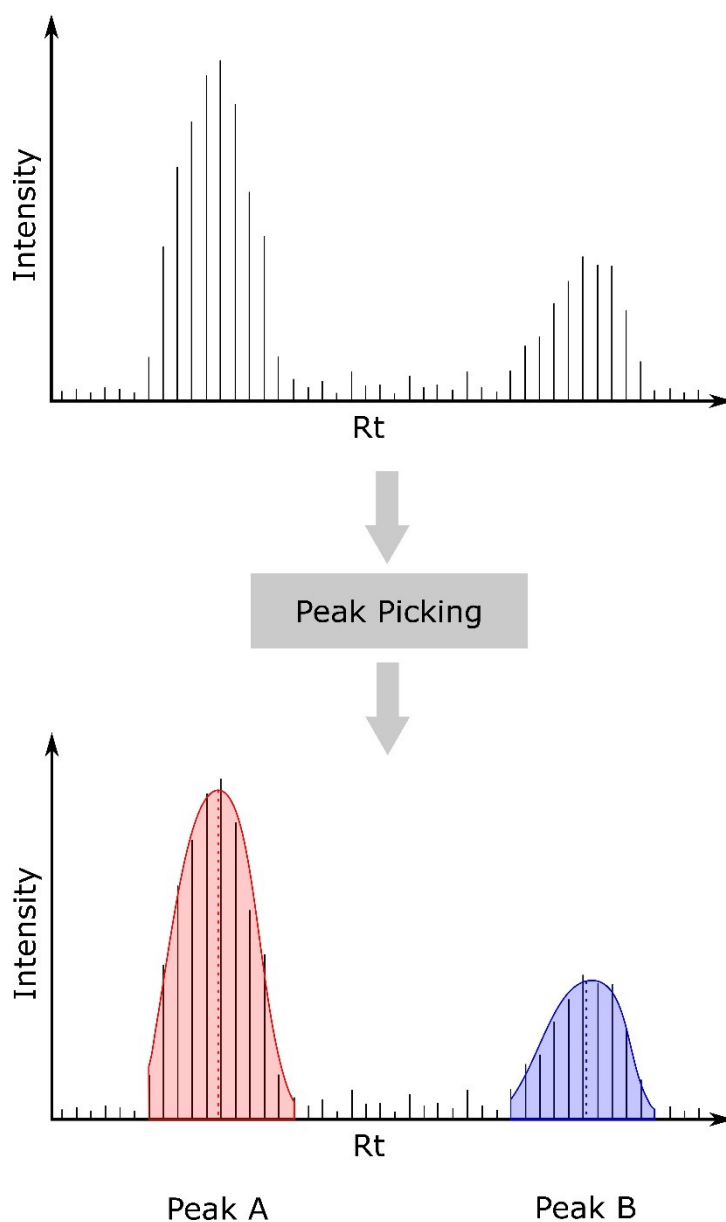


Figure T6: Peak picking

A chromatographic dataset is composed of individual signals. In order to derive the abundance of specific metabolites within the dataset and filter out noise signals, peak picking has to be performed. In this example, the applied peak picking algorithm detected two metabolic peaks (peak A and peak B). Peaks are usually characterized by their mean m/z , total area (colored) and their R_t , which is either the center of gravity (dotted line) or apex (signal with highest intensity), depending on the peak picking algorithm.

1.3.3 Existing Software Tools

The following section aims at providing a short summary of some of the most commonly used software tools for the alignment of LC-HRMS datasets. Many software tools in the field of metabolomics are proprietary or only usable in conjunction with instruments from a specific vendor. This overview will focus on free, vendor-independent tools only.

OpenMS, first released in 2005, is a feature-rich library that includes a unique alignment algorithm based on affine transformation. A first, rough alignment is achieved via a transformation that fits as many features of a reference sample to a target sample as possible. This alignment is then refined using linear regression.²³

XCMS, released in 2006, is a comprehensive tool for non-linear Rt alignment, peak detection and peak matching. It enables the automated analysis of many samples without the need for user intervention.²⁴ Ion features with similar Rt and m/z values are clustered and used to warp the datasets under investigation.¹⁵ A new web-based interface called XCMS-Online has since been added, offering a graphical interface.²⁵

MetAlign, released in 2009, first clusters all ion features within an adjustable time window by intensity and m/z value. Ion features that are found in all samples are then used as landmarks for the alignment, in which the position of the individual features is iteratively optimized. MetAlign also offers a graphical interface.²⁶ Multi-threading has been added for increased performance in 2012.²⁷

MZmine 2, released in 2010, features a rich user interface and is aimed at flexibility and extendibility. It offers a range of different algorithms for peak picking and alignment; most notably an alignment algorithm based on RANSAC (RANDOM Sample Consensus).²⁸

1.4 Aims of this Thesis

This thesis is concerned with the development of a novel software tool for the semi-automated comparison of LC-HRMS metabolomics experiments as well as the documentation of the applied workflows, computational methods, design decisions and algorithms. More precisely, the following aspects have been addressed and considered during the implementation of MetMatch:

- (a) Drifts and shifts in Rt and m/z make it necessary to align datasets prior to comparative data analysis. Thus, for proper alignment, the software tool has to account for the complex and often times non-linear nature of real-world shifts observed in experiments that have been measured with long time periods in between.
- (b) Differences in ion formation can often be observed between experiments. In order to generate a precise alignment that correctly matches as many metabolic peaks as possible, the underlying algorithm shall be capable of detecting differently formed ion species and correctly assign them between the different experiments.
- (c) As high performance and therefore short runtimes are desirable, this criterion has to be considered and addressed during software development.
- (d) A user-friendly interface shall be implemented, which facilitates the manual inspection, manipulation and fine-tuning of the alignment and ideally provides the user with comprehensive information about the achieved alignment quality.

While there already exists a plethora of software tools that are capable of LC-HRMS chromatogram alignment, most of them don't offer ideal solutions for one or multiple of these criteria, as their main focus is usually the in-depth processing of a single experiment.

1.5 Declaration of Thesis by Publication

MetMatch was developed in 2016 in collaboration with the working group Metabolomics and Bioactive Compounds at the Center for Analytical Chemistry, Department of Agrobiotechnology, IFA-Tulln, University of Natural Resources and Life Sciences, Vienna. The author of this thesis was responsible for the design and implementation of the software and the newly developed algorithms, as well as testing and analyzing the resulting tool. An article about MetMatch has previously been written and published by the author in an issue of the journal "Metabolites" ahead of this thesis.

Koch, S., Bueschl, C., Doppler, M., Simader, A., Meng-Reiterer, J., Lemmens, M. & Schuhmacher, R. MetMatch: A Semi-Automated Software Tool for the Comparison and Alignment of LC-HRMS Data from Different Metabolomics Experiments. *Metabolites* 6, 39 (2016).

The article is included in its entirety (with supplementary materials) in the following chapter of this thesis.

The Introduction gives a short overview of the scientific context, causes for shifts between different LC-HRMS chromatograms, existing software tools for LC-HRMS chromatogram alignment and their shortcomings.

The workflow of MetMatch is presented in Materials and Methods. Each step is described and illustrated in detail. Furthermore, the two biological experiments used to evaluate MetMatch are outlined.

Results and Discussion provides further information on the finished implementation. Design decisions that led to the individual workflow steps are explained and the user interface is briefly highlighted (a more extensive description of the user interface can be found on **page 55** of this thesis). Results generated by the software when applied to the previously mentioned experimental datasets, as well as artificial data are evaluated and discussed, before the publication is finally concluded.

The Supplementary Materials contain a list of available peak picking algorithms. Also, the Rt shift detection algorithm is presented in the form of pseudocode and described in detail. Processing parameters for the two biological experiments are provided, as well as a comparison with XCMS processing results.

Moreover, the software can be accessed free-of-charge for academic users at:

<https://metabolomics-ifa.boku.ac.at/MetMatch>, last accessed December 2018

Chapter 2: Publication



metabolites



Article

MetMatch: A Semi-Automated Software Tool for the Comparison and Alignment of LC-HRMS Data from Different Metabolomics Experiments

Stefan Koch ¹, Christoph Bueschl ¹, Maria Doppler ¹, Alexandra Simader ¹,
Jacqueline Meng-Reiterer ², Marc Lemmens ² and Rainer Schuhmacher ^{1,*}

¹ Center for Analytical Chemistry, Department of Agrobiotechnology (IFA-Tulln),
University of Natural Resources and Life Sciences, Vienna (BOKU), Tulln an der Donau 3430, Austria;
stefan.koch@boku.ac.at (S.K.); christoph.bueschl@boku.ac.at (C.B.); maria.doppler@boku.ac.at (M.D.);
alexandra.simader@boku.ac.at (A.S.)

² Institute for Biotechnology in Plant Production, Department of Agrobiotechnology (IFA-Tulln),
University of Natural Resources and Life Sciences, Vienna (BOKU), Tulln an der Donau 3430, Austria;
jacqueline.reiterer@boku.ac.at (J.M.-R.); marc.lemmens@boku.ac.at (M.L.)

* Correspondence: rainer.schuhmacher@boku.ac.at; Tel.: +43-1-47654-97307

Academic Editor: Peter D. Karp

Received: 30 August 2016; Accepted: 28 October 2016; Published: 2 November 2016

Abstract: Due to its unsurpassed sensitivity and selectivity, LC-HRMS is one of the major analytical techniques in metabolomics research. However, limited stability of experimental and instrument parameters may cause shifts and drifts of retention time and mass accuracy or the formation of different ion species, thus complicating conclusive interpretation of the raw data, especially when generated in different analytical batches. Here, a novel software tool for the semi-automated alignment of different measurement sequences is presented. The tool is implemented in the Java programming language, it features an intuitive user interface and its main goal is to facilitate the comparison of data obtained from different metabolomics experiments. Based on a feature list (i.e., processed LC-HRMS chromatograms with mass-to-charge ratio (m/z) values and retention times) that serves as a reference, the tool recognizes both m/z and retention time shifts of single or multiple analytical datafiles/batches of interest. MetMatch is also designed to account for differently formed ion species of detected metabolites. Corresponding ions and metabolites are matched and chromatographic peak areas, m/z values and retention times are combined into a single data matrix. The convenient user interface allows for easy manipulation of processing results and graphical illustration of the raw data as well as the automatically matched ions and metabolites. The software tool is exemplified with LC-HRMS data from untargeted metabolomics experiments investigating phenylalanine-derived metabolites in wheat and T-2 toxin/HT-2 toxin detoxification products in barley.

Keywords: feature detection; retention time correction; metabolite data matching

1. Introduction

Technological advances in analytical techniques have helped metabolomics to emerge into a valuable discipline for studying metabolic phenotypes caused by various perturbations such as environmental influences, biotic/abiotic factors, or changes in genotype [1]. The analysis of metabolites can provide unique insights into biological processes since the metabolome represents the final readout of any biological system [2]. Especially untargeted metabolomics experiments generate thousands of features [3], which may be unknown at the time of measurement and may turn interesting in later studies. Unlike targeted liquid chromatography coupled to high resolution mass spectrometry

(LC-HRMS (liquid chromatography coupled to high resolution mass spectrometry); a list of all abbreviations used throughout this manuscript is provided in “Abbreviations and Terminology”) approaches, an untargeted metabolomics experiment based on full-scan LC-HRMS also allows for retrospective data investigation. In order to get as much information as possible from a single experiment, it can be beneficial to compare novel biological information to former experiments, thus addressing novel biological aspects (e.g., newly identified metabolites) that have not been investigated in former experiments.

LC-HRMS is one of the most commonly employed analytical techniques in metabolomics research. Chemical compounds are separated by two orthogonal properties, i.e., retention time (R_t) and mass-to-charge ratio (m/z). Although modern LC-HRMS instruments are well established and robust, the generated raw data tend to show shifts and drifts in both the chromatographic as well as mass spectrometric dimension. The extent of these shifts is more distinct for samples which have been analyzed in different batches or when the time span of data acquisition was rather long (e.g., several weeks or months). Instabilities can be caused by: (i) the instrument (e.g., systematic mass shifts, column age and new hardware such as pumps); (ii) the chemistry of the sample and separation system; and (iii) random variations (e.g., temperature fluctuations) [4,5]. Furthermore, the ion species formed by the analytes can vary depending on the instrument's condition (e.g., contaminations in the ion source) and the composition of the mobile phase or sample matrix. Apart from the sheer size of the data, which makes manual analysis impractical, the comparison of different experiments is aggravated by the aforementioned drifts and shifts.

Several algorithms and software tools have been developed to address this problem. Some of these algorithms are restricted to the use of the total ion current (TIC) chromatogram [6–8], whereas most of the more recent approaches do not limit themselves to the chromatographic dimension and also include information from the m/z dimension [9–11]. The process of reducing the aforementioned m/z and R_t shifts between different LC-HRMS chromatograms is called alignment. The majority of the available tools, such as XCMS [12], MZMine [13]/MZMine2 [14] or the MapAligner tool of OpenMS [15], use warping techniques [4]. Warping shifts, stretches and/or compresses one or both dimensions (i.e., R_t and m/z) simultaneously to improve data consistency across different datafiles under investigation [16]. The applied warping function can be constant (i.e., a constant shift value across the whole chromatogram), linear (i.e., a constant change in shift over time) or non-linear (i.e., the change in shift is not constant). While many advanced warping algorithms exist [9,17,18], most of them are refined versions of either correlation optimized warping (COW) or dynamic time warping (DTW). Existing simple solutions suffer from various shortcomings such as model assumptions that fail to capture the behavior of real-world data, e.g., when applying a linear R_t shift function. More sophisticated and complex algorithms can yield increasingly accurate results using the non-linear retention time correction function (e.g., by the use of polynomial functions or splines). However, they often have undesirably long runtimes [4], which, combined with the fact that many of them are solely command line-based and lack a convenient user interface, hinders them from being widely used in the broad field of metabolomics.

Here we present a software tool called MetMatch for the user-guided alignment of different LC-HRMS chromatograms. Chromatograms of interest can be aligned to a reference feature list consisting of reference metabolites and their ions (i.e., ID, m/z , retention time, etc.). The reference feature list can be generated manually or by processing single or multiple LC-HRMS chromatograms with one of the currently available software tools for untargeted metabolomics such as XCMS [12], MZMine2 [14], or MetExtract [19]. The alignment is performed using novel, efficient algorithms that correct offsets in the m/z as well as non-linear shifts in the R_t dimension. Apart from m/z and R_t shift correction, the formation of different ion species is also accounted for, which is a distinctive feature of MetMatch. Users can tweak all processing parameters through a simple, convenient and interactive user interface.

The software is available free of charge for academic, non-commercial use and can be downloaded for Windows from [20]. Here, the software is exemplified with LC-HRMS data from untargeted metabolomics experiments investigating phenylalanine-derived metabolites in wheat and T-2 toxin/HT-2 toxin detoxification products in barley as well as artificial test data.

2. Materials and Methods

2.1. Computational Workflow

The general data processing steps performed by MetMatch are outlined in Figure 1. Apart from generation and import of the reference feature list and output of the resulting data matrix, all processing steps are successively performed on a single datafile and then repeated for every file of a processed dataset. Details regarding the individual processing steps are provided in the following subsections.

2.2. Inclusion of Alternative Ion Forms (Step 1)

For each feature specified in the reference feature list, MetMatch generates an extended list of putative other ion adducts to account for differences in ion adduct formation between reference and target. These are typically caused by changes in eluent or other instrumental settings. Possible alternative ion species are provided by the user. If the ion species of a reference feature is known, MetMatch generates new ions based on the mass of the non-charged compound. In case the ion species of the reference feature is unknown, MetMatch first assumes this feature to be any ion species from the user-defined list, and then calculates the feature's neutral mass and the m/z values of all putative ion species that can potentially occur for the particular compound.

2.3. m/z Offset Detection (Steps 2 and 3)

All signals of a target datafile that are within a certain R_t and m/z tolerance window (so-called search frame) around each reference feature are parsed from the respective datafile. For each search frame, the m/z shift of all included signals relative to the reference feature is calculated and weighted by the intensity of the signals. A histogram containing the weighted average m/z shift of all search frames of the datafile is then constructed. The m/z shift value of the bin with the highest frequency marks the constant systematic m/z offset of the respective datafile relative to the reference, as exemplified in Figure 2. While averaging m/z shifts between individual search frames helps to properly estimate the overall m/z offset, multiple chromatographic peaks with similar retention time and m/z values within single search frames or peaks partly split between neighboring search frames may cause conflicts which are not resolved by MetMatch or any other metabolomics data processing software.

After the detection of the m/z offset, m/z values of all signals inside the chromatogram are corrected by the calculated m/z shift. Corrected signals with an m/z deviation higher than a user defined ppm value are rejected, resulting in narrowed search frames.

2.4. Peak Picking (Step 4)

Following m/z offset detection, peak picking is performed on the remaining signals inside each narrowed search frame. Currently three different chromatographic peak picking algorithms have been implemented and can be selected in MetMatch. The MassSpecWavelet algorithm, while being very reliable, has significantly longer runtimes than the other two peak picking algorithms. These are based on Gaussian peak correlation and the Savitzky-Golay filter. For further details, please refer to Supplementary Materials Section 1.1.

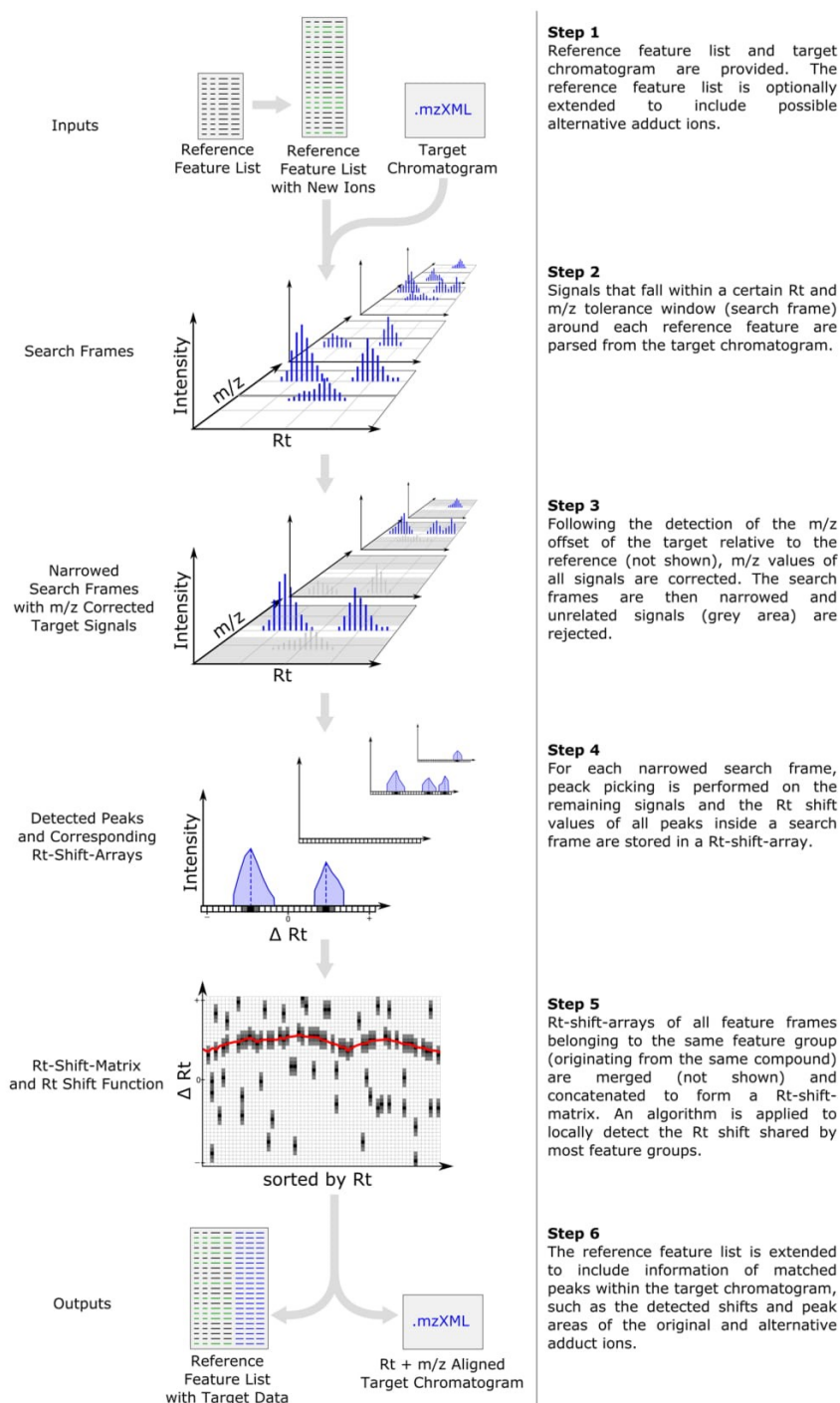


Figure 1. Workflow depicting the main data processing steps of MetMatch for a single target chromatogram. If more than one target chromatogram is going to be processed, each of those is successively carried through the steps 2–5 and a single data table containing the (extended) reference list as well as the matched features of every target chromatogram is generated. After processing of the target chromatograms, all new information (chromatographic peak area, corrected m/z value and retention time, calculated m/z and Rt shift) for each reference feature and processed target chromatogram are saved as an extended data matrix. Moreover, the target chromatograms can optionally be saved as m/z - and Rt -shift-corrected $mzXML$ files.

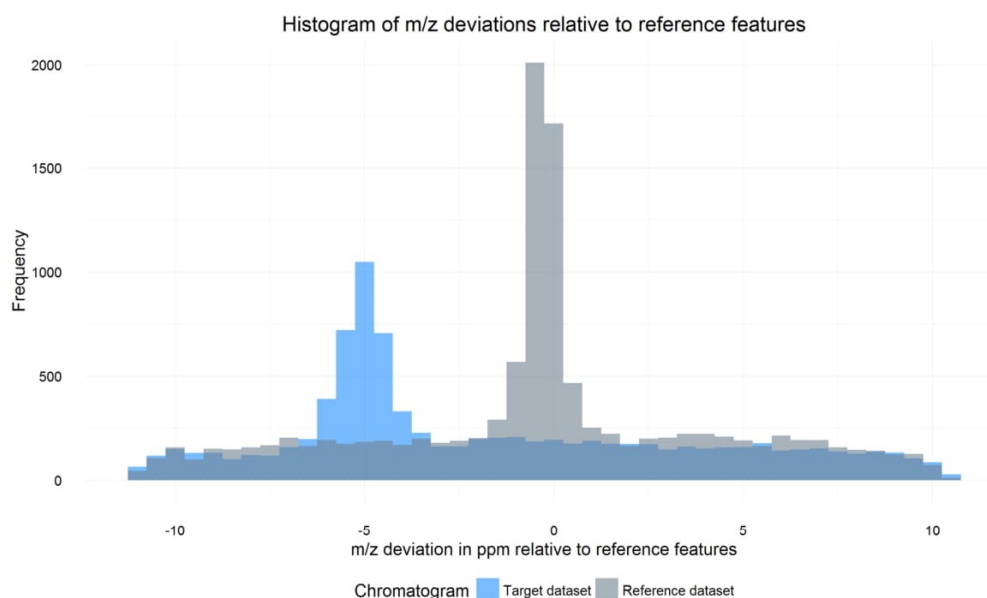


Figure 2. Overlay of two histograms showing the distribution of the weighted average mass to charge ratio (m/z) shifts of all search frames of a target file (blue) relative to a reference file (grey). The baseline frequency of about 200 is caused by unrelated signals, which have a random m/z shift relative to their respective reference feature. The m/z shift of signals corresponding to reference features is, however, not random and shows a significant increase around the systematic m/z offset of the respective file relative to the reference. This causes a maximum around the aforementioned offset. For the illustrated target datafile, this m/z offset was determined to be -5 ppm.

2.5. Rt Shift Detection (Step 5)

Next, the Rt shift function is determined using all entries from the reference list. Figure 3a–c outlines the creation of the initial Rt shift matrix, containing Rt shift information of all peaks of a datafile detected during the peak picking step (step 4).

An iterative algorithm is then applied to deduce an Rt shift function through the Rt shift matrix, as exemplified in Figure 4. The target peak that is closest to the Rt shift function but still within a user defined maximum tolerance is matched to the reference feature. For further details regarding the Rt shift detection algorithm, please refer to Supplementary Materials Section 1.2.

After several iterations of the chromatographic alignment algorithm, a sufficient agreement of the correction function with the region of high density of putative matching peaks has been achieved, which can also be visually inspected and accepted/rejected by the user (Figure 3d).

2.6. Output of Generated Results (Step 6)

After the m/z and chromatographic alignment steps have been determined, MetMatch applies these corrections to all chromatographic peaks from the reference feature list detected in the respective target chromatogram. This additional, new information (chromatographic peak area, corrected m/z value and retention time, calculated m/z and Rt shift) for each reference feature and processed target chromatogram are saved as an extended data matrix. Additionally, MetMatch also offers the option to save the processed target chromatograms as new mzXML files, in which all MS signals have been corrected by the determined m/z and Rt shifts.

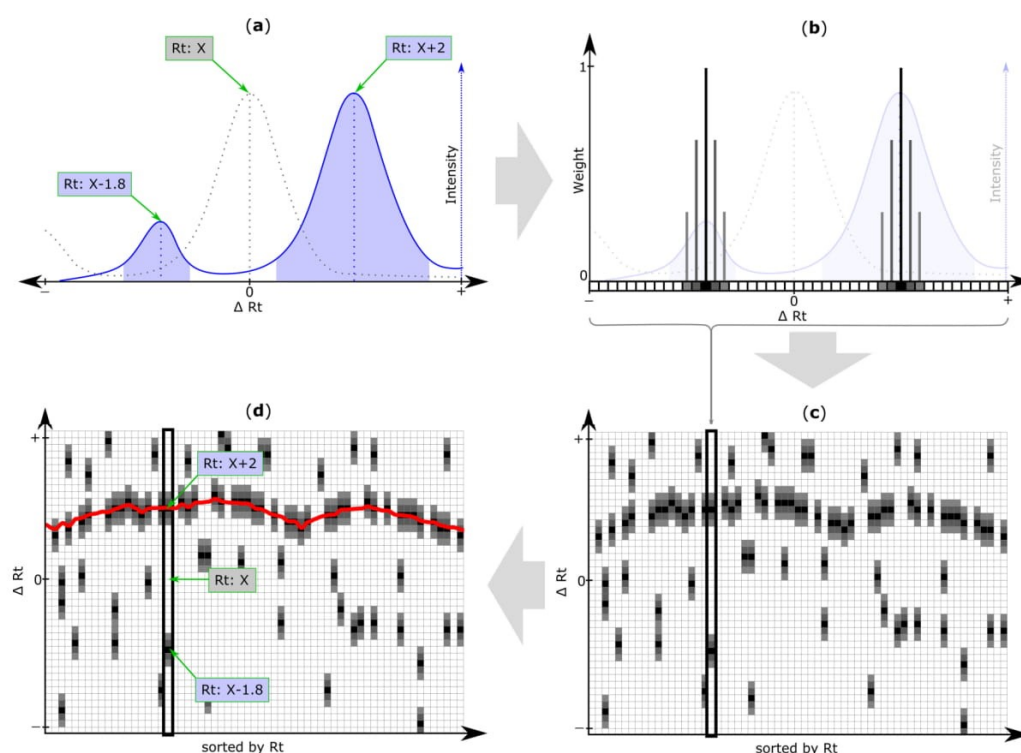


Figure 3. Retention time (Rt) shift detection procedure: (a) Extracted ion chromatogram (EIC) of a target chromatogram (blue solid line) with two chromatographic peaks (blue areas) and their peak apexes or centers (blue dotted line), depending on the selected peak picking algorithm. Determination of the chromatographic peak that correctly matches the reference peak (grey) is non-trivial, as the actual Rt shift is unknown; (b) All chromatographic peaks detected within a search frame are considered to be putative matching peaks and their Rt deviation relative to the original reference feature is discretized through binning. The resulting array of Rt shift bins is referred to as Rt shift array and is non-zero for all Rt shift bins at which—depending on the selected peak picking algorithm—the apex or center of the respective chromatographic peak has been detected. By default, the retention times of peaks are assigned a weight of 1 (black cells) and bins within a certain tolerance around the peak apex (or center respectively), are assigned linearly decreasing weight values (grey squares). All Rt shift arrays of search frames belonging to the same feature group (and therefore the same metabolite) are merged into a single array (step not shown); (c) All feature group Rt shift arrays are then sorted according to the Rt of their respective reference feature group and are concatenated column-wise to form the Rt shift matrix. Any non-zero entry in this matrix corresponds to the Rt shift (ΔRt) of a peak detected within a specific feature group (i.e., column) of the inspected target file. A region of higher density of putatively matching peaks marking the most probable Rt shift across the inspected chromatogram is clearly visible; (d) The Rt shift detection algorithm of MetMatch identifies this region of higher peak density by iteratively optimizing a function (red line) to be as close to as many peaks as possible, as outlined in Figure 4. This function is then used to determine the correctly aligned chromatographic peaks relative to their reference list entry. In the depicted case the aligned peak from the target chromatogram is $X + 2$.

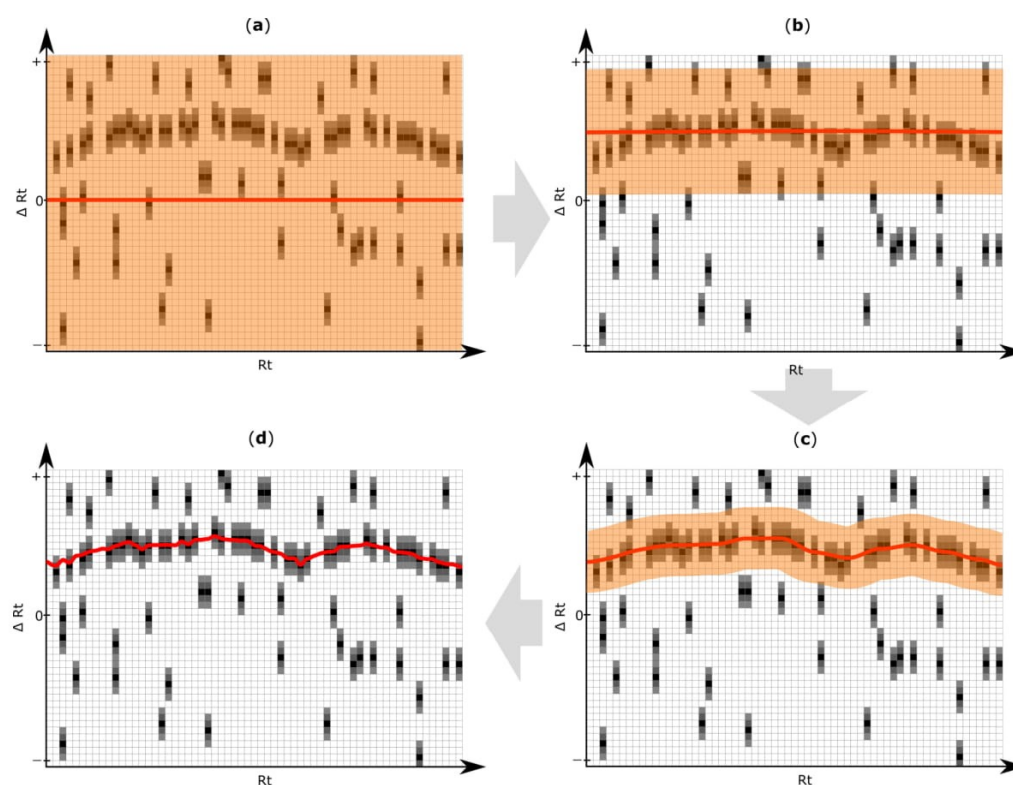


Figure 4. Exemplification of results obtained with the iterative retention time (Rt) shift detection algorithm. For improved illustration, only a certain part of the Rt shift matrix is depicted. (a) After construction of the Rt-shift matrix of the inspected target chromatogram (Figure 3a,b), the Rt shift function (red line) is initially constructed assuming no apparent shift ($\Delta R_t = 0$). The orange area highlights the maximum allowed Rt deviation which may be used in the next iteration step; (b–d) Three successive iteration steps were applied to model the most likely Rt shift of the target chromatogram relative to the reference feature list. The Rt shift matrix used in this exemplification consists of 45 rows and 63 columns. The three iterations of the Rt shift detection algorithm were using a row tolerance (orange area) of 23, 9 and 4 rows, and a column tolerance of 41, 4 and one column, respectively. Each iteration gradually adjusts the retention time correction function more to the region of higher peak density, which finally results in an optimal alignment maximizing the number of correctly aligned chromatographic peaks. The pseudocode of the algorithm for elucidation of the Rt shift correction function is listed in Supplementary Materials Figure S1.

2.7. Phe-Derived Biological Data

Two different test cases were chosen to demonstrate the MetMatch software tool. The first involves a dataset from a study investigating the metabolism of the aromatic amino acid phenylalanine (Phe) in wheat ears.

For generation of the reference feature list, wheat plants of the cultivar “Remus” were grown in a glass house under controlled conditions and inoculated with $^{13}\text{C}_9$ phenylalanine solution (~99.5% ^{13}C). Subsequent LC-HRMS measurement and TracExtract version 2 [21] data processing resulted in 151 Phe-derived features, which were detected in the ^{13}C -Phe-treated sample. In order to increase the number of features on the reference list and thereby enhance m/z and Rt correction accuracy, plants of the same cultivar were also raised in a tailor made growth chamber with ^{13}C enrichment of 98.8% as well as under native $^{12}\text{CO}_2$ atmosphere in parallel. LC-HRMS analysis and

data processing by AllExtract version 2 [22] resulted in 1482 features, which exclusively served to enlarge the reference list for matching Phe-derived metabolites in the samples of biological interest. The two resulting feature lists were merged (total of 1549 features) to create the reference feature list (for more details refer to Supplementary Materials Section 2.1) used in Section 3.8 as well as for artificial data evaluation (Section 3.10). The generation of the reference feature list is outlined in Supplementary Materials Section 2.2. It shall also be mentioned that only the information related to features of the native metabolite ions has been used by MetMatch.

For the target dataset, a study involving the two wheat cultivars “Remus” and “CM-82036” was chosen. Wheat plants (native, no ^{13}C -labeling) were grown in a glass house, treated with water at the flowering stage and harvested 96 h after inoculation. These samples represent a part of an untargeted metabolomics experiment consisting of about 450 samples, which was carried out in the year 2012. Twenty of these samples have been used in the presented manuscript to demonstrate MetMatch. For more details on the experiment, refer to Warth et al. [23]. Samples were prepared and measured with LC-HRMS as described by Bueschl et al. [22] on an LTQ Orbitrap XL instrument in August and September 2012. Measurement of the two samples serving as the reference was carried out in December 2015. Between the measurement of the reference and target datasets, the pump of the respective HPLC had been replaced (Accela 600 pump to Accela 1250 pump) and the chromatographic column had been renewed (same type and dimensions, different production charge). All other method parameters and instrumental setup remained constant for all measurements.

2.8. T-2 Toxin/HT-2 Toxin-Derived Biological Data

The second test case chosen to demonstrate the developed tool is a study investigating plant-derived metabolites of the mycotoxins T-2 toxin and HT-2 toxin in barley reported by Meng-Reiterer et al. [24]. For comparison, matching of results and generation of the reference feature list, two barley samples were treated with native and ^{13}C -labeled T-2 toxin and HT-2 toxin respectively. Measurements were performed using an LC-Q-TOF-system. The two chromatograms were processed using XCMS version 1.40.0, resulting in 20,682 features. T-2 toxin/HT-2 toxin-derived detoxification products reported by Meng-Reiterer et al. were matched to the list of barley metabolites. The time interval between generating the reference datafiles and measurement of the target files amounted to two years. Degradation of T-2 toxin/HT-2 toxin metabolites could therefore not be excluded. For further details regarding the generation of the reference feature list, please refer to Supplementary Materials Section 2.3.

3. Results and Discussion

We present MetMatch, a novel software tool for the comparison and alignment of LC-HRMS data from different metabolomics experiments. MetMatch is implemented in Java 8, as its performance is comparable to C/C++ while offering flexibility and abstraction from low-level operations [25]. Additionally, a plethora of high quality Java libraries exists, which helps to reduce development time. Java packages used by MetMatch are listed in Supplementary Materials Section 1.3.

Figure 5 shows a schematic representation of the MetMatch data processing workflow and how it can be integrated into already existing analysis pipelines. Refer to Section 2 for more details on the internal computational processing steps performed by MetMatch.

The purpose of the developed tool is to provide a fast, easy and accurate method to automatically detect m/z and R_t shifts between multiple measurements or LC-HRMS datasets, align all features between reference and individual target files and generate a table of matched metabolic features for further investigation (e.g., statistical processing, metabolite annotation). Due to the successive pairwise matching of target datafiles against a single reference list, MetMatch can also be applied to correct for drifts across a measurement sequence. This can be used to investigate if a dataset exhibited systematic changes in R_t and m/z shift over time. Usability is facilitated by a comprehensive and interactive user interface (Figure 6).

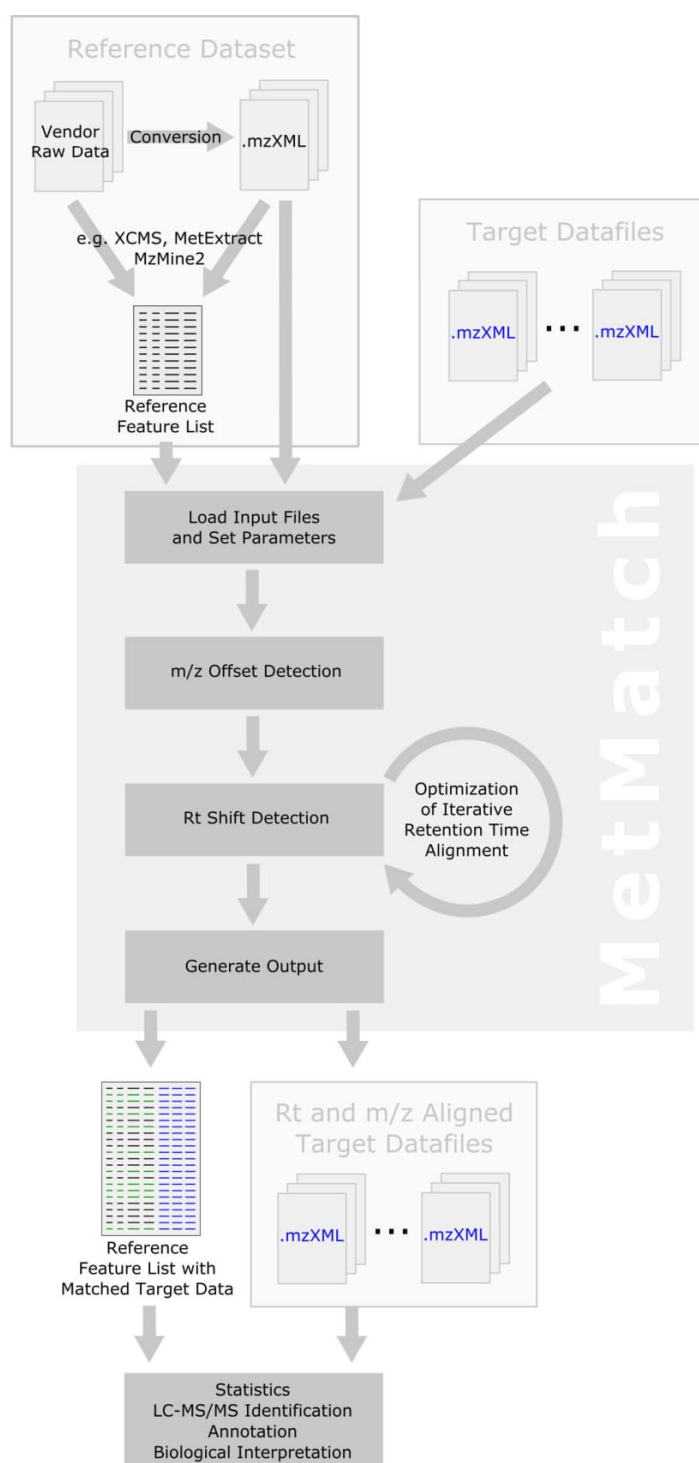


Figure 5. Schematic data processing workflow illustrating how MetMatch can be integrated into existing analysis pipelines.

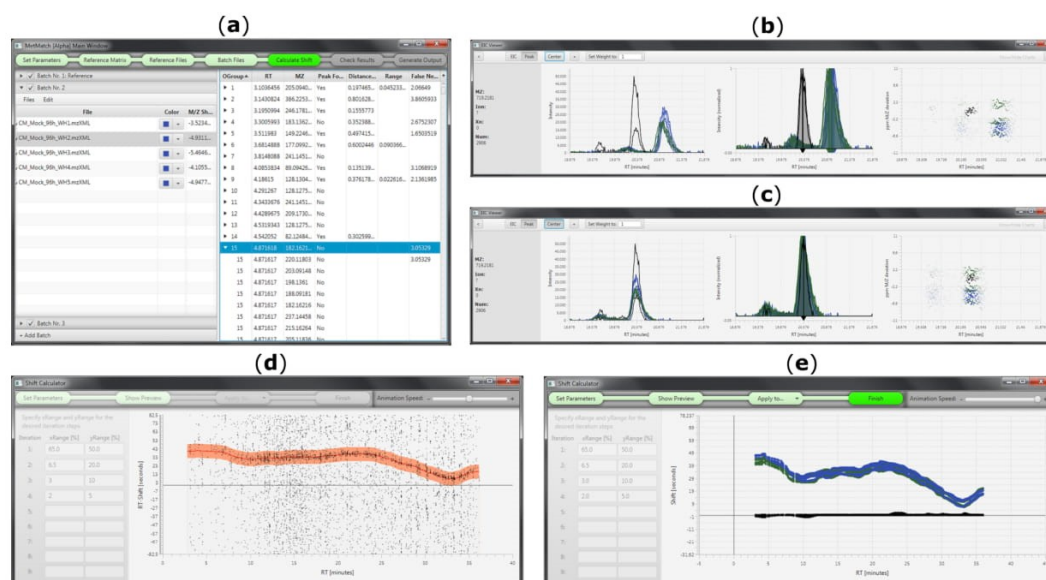


Figure 6. MetMatch's user interface during various processing steps. (a) Main window showing the reference feature list and all loaded datafiles. The step indication bar on top helps the user to keep track of the data processing and also serves as a menu bar; (b) EICs of reference (black) and target datafiles (blue, green) before Rt alignment; (c) Overlay of EICs after Rt alignment; (d) Animated Rt shift detection algorithm (animated version available at [26]); (e) Overlay of Rt shift functions calculated for multiple target datafiles.

The main motivation behind MetMatch is to support the comparison of results between various metabolomics experiments, for instance the investigation of newly generated biological knowledge (e.g., novel or identified/annotated metabolites) in previous datasets or the comparison of LC-HRMS data from experiments with slightly modified biological conditions or sampling times. Sections 3.8 and 3.9 demonstrate the flexibility of MetMatch with newly characterized, phenylalanine-derived metabolites as well as T-2 toxin/HT-2 toxin detoxification products in barley.

3.1. Accepted Input Files

MetMatch requires a list of reference features in the form of a tab-separated values (tsv) file. This list is typically obtained by processing one or more chromatograms with software for data processing of untargeted metabolomics approaches and serves as a reference for further metabolite matching in target datafiles. Each feature must be specified by the retention time (Rt), the mass-to-charge ratio (m/z), a unique identification number and the identification number of the associated feature group (Table 1). Because of the generic nature of the tsv file format, output from various applications such as XCMS [12], MzMine2 [14] or MetExtract [19] can be used as a reference either directly or after minor modification. Such a reference list can also be created from scratch using any text editor and should at least contain the information specified in Table 1.

The target chromatograms that are to be matched to the reference list have to be provided in the mzXML file format [27]. Most common vendor file formats can be easily converted to mzXML using freely available conversion tools such as ProteoWizard [28]. MetMatch also offers the option to load the unprocessed reference LC-HRMS chromatograms in the mzXML format for visualization. These are treated similar to target chromatograms by MetMatch and facilitate the visual assessment of EIC peaks underlying reference features as well as the success of the alignment process.

Table 1. Structure of a typical reference feature list. Each feature is specified by four required values and four optional values relating to its ion species. “Ion ID” is an integer that uniquely identifies a feature. Each feature belongs to a feature group (i.e., metabolite), which is specified in the column “Group ID”. All features with the same “Group ID” are different ions of the same metabolite. “Rt” is the retention time of the feature in minutes and “ m/z ” is its mass-to-charge ratio. The remaining columns are optional and used for the generation of differently formed ions if specified.

Ion ID ^a	Group ID ^a	Rt ^a	m/z ^a	Ion Form ^b	Uncharged Ion Mass ^b	Ion Charge ^b	Ionisation Mode ^b
1	1	2.78	160.075	[M + H] ⁺	159.068	1	+
2	1	2.78	177.102	[M + NH ₄] ⁺	159.068	1	+
3	2	3.19	482.296	[M + NH ₄] ⁺	464.262	1	+
4	3	3.27	330.458	[M + NH ₄] ⁺	312.424	1	+

^a Mandatory information required by MetMatch; ^b optional information (used if provided)

3.2. Detection of Differently Formed Ions

A chemical substance can form different ion species in the ion source of the mass spectrometer, such as protonated molecules ([M + H]⁺) or sodium ([M + Na]⁺) or ammonium ([M + NH₄]⁺) adducts, each of which is detected at a unique m/z value. Especially when the ionization conditions vary between measurements (e.g., alteration of mobile phase composition or contamination of MS ion source), the relative fraction of the individual ion species may also change. Because of this, different ion species of the same metabolite might have been formed in the reference and target experiments, which poses a problem during comparison.

MetMatch addresses this problem by expanding the reference feature list with new alternative ion species. As a result, all ion forms specified on a user-created list are automatically searched for in the target files and matched against the extended reference feature list.

3.3. m/z Shift Detection

LC-HRMS data often exhibit small shifts in mass accuracy within analytical batches, even if the MS instrument has been calibrated immediately before the measurements. Therefore, m/z shifts have to be accounted for, especially when comparing different batches or longer sequences that do not use some method of online mass calibration. Shifts in the m/z dimension typically show negligible variation within a single file [4], which is why MetMatch assumes a constant m/z offset per file. This constant m/z offset is estimated using a histogram containing the weighted average m/z offset of all search frames of a chromatogram file. Search frames are defined as the set of all signals of a chromatogram that fall within a certain user-defined Rt and m/z tolerance around a reference feature. Using reference histograms, which are generated as outlined in Section 2.3, the m/z shift of the respective target file is determined, which is represented by the bin with the highest number of assigned ion signals.

3.4. Rt Shift Detection

While chromatograms within an analytical batch commonly only exhibit retention time shifts of small magnitude (usually less than the width of typical chromatographic peaks), this Rt shift can be more distinctive when comparing chromatograms from different analytical batches. In order to detect these Rt shifts, suitable alignment algorithms have to be employed [29]. MetMatch’s Rt shift correction method has been specifically developed for the detection of various forms of Rt shift, ranging from constant offsets to non-linear shifts in retention time.

It is assumed that most metabolites eluting within a narrow Rt range share a similar Rt shift between different chromatograms. If this assumption is wrong, every alignment attempt solely based on a single Rt shift function yields erroneous results [30]. Therefore, most of a chromatogram’s metabolite peaks should locally share an Rt shift, which can be described by a smooth function across

a chromatogram. The Rt-alignment algorithm implemented in MetMatch uses an iterative approach to detect this Rt shift function. Since MetMatch provides an animated preview of the algorithm, the resulting function to be used for alignment of chromatograms can immediately be assessed visually and parameter settings can therefore be easily and quickly optimized by the user. When an appropriate balance between global and local alignment has been achieved, the user-guided settings may be applied to all samples in a dataset. The resulting alignment of each chromatogram can then be reviewed and adapted if necessary.

The applied Rt shift detection method has two limitations: (i) Depending on differences in their chemical structures, some metabolites may change their elution order due to variations in experimental conditions between the reference and target datafiles [30]. Based on Rt and m/z information alone, it is impossible to detect such changes, and therefore these features cannot be correctly matched; (ii) Reference list and target datafiles have to be compatible with regard to their features. If too few compounds/feature groups are shared between both of them, the quality of the resulting estimated Rt shift can decrease significantly. The minimum number of common feature groups has been exemplified with artificial data in Section 3.10.4. It is worth mentioning that not all features contained within the reference feature list have to be of biological interest. Additional features may be included only for the purpose of increasing the accuracy of the m/z and Rt shift detection. We have also chosen this approach in the two experiments described below to exemplify MetMatch.

3.5. Manual Corrections of Automatically Generated Results

All processing results can be verified and corrected manually using MetMatch's interactive user interface. EICs are visualized and can be used to assess the initial signal extraction parameters as well as the quality of the detected peaks. The user is able to easily delete or manually pick peaks if necessary. Furthermore, the weight of individual EIC peaks can be increased according to their likelihood of representing the correctly matched metabolite ions. This way, the user can guide the automated Rt alignment algorithm with useful information in the form of user-verified "anchor" metabolites. Matched peaks are highlighted directly in the corresponding EIC's and can be visually aligned to assess the automated processing result.

3.6. Output Format

The final result of MetMatch's alignment routine is a tab-separated values (tsv) file containing the matched chromatographic peaks of all processed target chromatograms and datasets. It is essentially an extended reference feature list, containing the additional information derived from the matched chromatograms as exemplified in Table 2. If new ions were detected, they are included as additional rows in the data matrix. For each processed datafile, every matched chromatographic peak is specified by its EIC peak area, Rt and m/z value as well as their respective offsets by default. Additionally, m/z - and Rt-shift-corrected mzXML files can also be generated, which will contain all signals of the respective target chromatograms, however, their m/z and Rt values will be corrected using the MetMatch-derived shifts. Features which were not part of the reference list and were therefore not considered for the m/z and Rt correction by MetMatch are still corrected and linear interpolation is used whenever needed. These corrected mzXML files can then be viewed or further processed with other software tools capable of reading the mzXML format.

3.7. Runtimes

The runtime of MetMatch is linear with respect to the number of files and the number of entries in the reference feature list. Actual runtimes heavily depend on the applied peak-picking algorithm. For further details on the individual peak-picking algorithms, please refer to Supplementary Materials Section 1.1. Typical runtimes of the different algorithms for chromatographic peak picking on an Intel Core 2 Duo (2.8 GHz) are listed in Supplementary Materials Table S1.

Table 2. Detail of an output feature matrix generated by matching a target chromatogram to the reference feature list from Table 1. The columns “Ion Charge” and “Ionization Mode” have been omitted for simplicity. Features 901–903 are newly generated ion forms that were not contained in the reference list. Trying to align the target chromatogram to the reference without accounting for the formation of different ion species would have resulted in no matches for the displayed features.

Ion ID ^a	Group ID ^a	Rt ^a	<i>m/z</i> ^a	Ion Form ^a	Uncharged Ion Mass ^a	Reference Peak Area ^a	File Peak Area ^b	File Peak Rt ^b	File Peak <i>m/z</i> ^b
1 ^a	1	2.78	160.075	[M + H] ⁺	159.068	103,038	-	-	-
2 ^a	1	2.78	177.102	[M + NH ₄] ⁺	159.068	6,234,458	-	-	-
901 ^b	1	2.78	182.057	[M + Na] ⁺	159.068	-	3,423,472	2.83	182.057
3 ^a	2	3.19	482.296	[M + NH ₄] ⁺	464.262	523,238	-	-	-
902 ^b	2	3.19	487.251	[M + Na] ⁺	464.262	-	2,137,423	3.26	487.251
4 ^a	3	3.27	330.458	[M + NH ₄] ⁺	312.424	13,459,147	-	-	-
903 ^b	3	3.27	487.251	[M + Na] ⁺	312.424	-	15,358,291	3.38	487.251

^a Part of the reference feature list; ^b New column/row added by MetMatch.

3.8. Evaluation of the Phe-Derived Biological Metabolites

A metabolomics experiment studying phenylalanine (Phe)-derived metabolites in wheat was chosen as a reference. Seventy-five Phe-derived metabolites were detected and subsequently matched to a previous wheat experiment using MetMatch. Before matching, the reference list was extended to include metabolites that are not of interest (i.e., non-Phe-derived wheat metabolites), but help improve the accuracy of the *m/z* and Rt shift calculations. Refer to Supplementary Materials Section 2.2. for details on the creation of the reference feature list. While all features contained in the reference feature list have been used by MetMatch to correct the *m/z* and Rt shifts, data evaluation has been limited to the Phe-derived features since only those have been of scientific interest in the respective biological study. Rt shifts were to be expected since the two datasets were generated in the years 2012 and 2015. Furthermore, the HPLC pump and the chromatographic column had been replaced between the two experiments.

Processing all 22 files using MetMatch took 24 min on an Intel Core 2 Duo (2.8 GHz) with the parameters listed in Supplementary Materials Section 2.1. Matching chromatographic peaks for 92 of the 151 Phe-derived metabolite features contained in the reference feature list were detected in at least five target datafiles using MetMatch. Only 42 of these features were matched if differently formed ion adducts were not accounted for. On average, the software determined an *m/z* shift of −3.4 ppm as well as a non-linear retention time shift ranging from 7 to 52 s as illustrated in Figure S2 in the Supplementary Materials.

To compare the detected Rt shift of MetMatch with another commonly used software tool in the field of metabolomics, the same chromatograms have also been processed and aligned with XCMS. XCMS processing took 73 min on a similar computer (Intel Core 2 Duo, 2.9 GHz, no multithreading). However, XCMS uses an untargeted approach and therefore does not use the reference feature list. The resulting Rt shift functions obtained by MetMatch and XCMS agree across the majority of the Rt range of interest (3 to 36 min). Rt values smaller than 7 min, however, show reduced consensus. The underlying data shows only few signals in this Rt region, and it was not possible to definitively confirm or reject any of the two proposed Rt shifts by means of manual investigation. The calculated retention time shift functions of MetMatch and XCMS were compared using the Pearson correlation coefficient. For this, the calculated retention time shift functions of MetMatch and XCMS were compared pairwise. Between the retention time interval of 3 to 38 min, a mean correlation of 0.858 (standard deviation of 0.042) was calculated. However, for the retention time interval of 7 to 38 min, the average correlation increased to 0.926 (standard deviation of 0.011), which confirms that MetMatch better detected the retention time shifts in areas where more reference features are present and that the detected retention time shift functions of MetMatch and XCMS are very similar. A comparative illustration of the retention time shifts detected by MetMatch and XCMS is shown in Figure S2 in Supplementary Materials.

MetMatch's capability of matching different ion species has been tested by means of matching the same target samples as before and additionally using chromatograms measured in negative ionization mode. The used reference feature list only contained ions measured in positive ionization mode. Matching them to the samples obtained in negative ionization mode requires the inclusion of alternative negative adduct ions in the reference feature list. This was automatically done by MetMatch as described in Section 2.2. Possible positive (e.g., $[M + H]^+$, $[M + Na]^+$, $[M + NH_4]^+$) as well as negative (e.g., $[M - H]^-$, $[M + Na - 2H]^-$, $[M + Cl]^-$) ion species had to be provided by the user. Figure S3 in Supplementary Materials shows a comparison of the Rt shift functions obtained by matching positive and negative files to the positive reference feature list. The resulting Rt shifts for positive and negative chromatograms agree well. The Pearson correlation coefficient for chromatograms obtained in the positive and negative ionization modes for the same sample were calculated. On average, a correlation of 0.826 (standard deviation of 0.038) was calculated for the retention time interval of 3 to 38 min. The correlation improved to an average of 0.977 (standard deviation of 0.008) when the retention time interval of 7 to 38 min was used, in which most of the features of the reference feature list were present. This demonstrates that MetMatch was able to match target files and a reference feature list that share no common ion species by extending the reference feature list with alternative putative ion species. Without accounting for differently formed ion species, matching positive and negative ionization data would not be possible in that way.

To validate the results generated by MetMatch, selected Phe-derived features have been manually investigated in the raw data of reference and target files (Thermo Xcalibur Software, version 2.2). To this end, the correctness and plausibility of the proposed m/z and Rt shifts was examined for 23 features in 10 files, ranging from low to high m/z - as well as Rt values. The calculated shift of 19 chromatographic peaks could be confirmed in all files. One peak could not be correctly matched in any file, because it was detected at a very early Rt. Very few other peaks were detected around this Rt, leading to an inaccurate Rt shift function for this Rt region. Two peaks were not matched in some files since their retention time shifts were slightly above the maximum allowed Rt shift tolerance around the Rt shift function. This could have been solved by increasing the tolerance, at the risk of including false positives. One peak could not be reliably matched in all files because of the peak shape (shoulder peak). For a table containing the manually investigated features, please refer to Supplementary Materials Table S2.

EICs of one of the manually validated chromatographic peaks are shown in Figure 7. Here, the usefulness of MetMatch becomes apparent, since manual comparison of the two experiments could have easily resulted in an incorrect match.

3.9. Evaluation of the T-2 Toxin/HT-2 Toxin-Derived Biological Data

In the second test case, 36 T-2 toxin/HT-2 toxin-derived metabolite ions were matched to a previous barley experiment (Supplementary Materials Section 2.3). Since two years have passed between the two experiments, possible Rt shifts were expected. The reference list was again extended to improve the accuracy of Rt correction. This time, all features detected with an untargeted approach facilitated by XCMS were used as the reference feature list. Processing all five files using MetMatch took 5 min on an Intel Core 2 Duo (2.8 GHz) using the parameters listed in Supplementary Materials Section 2.3. MetMatch determined a non-linear retention time shift ranging from -9 to -24 s as illustrated in Figure S4 in Supplementary Materials.

Again, the retention time alignment calculated by MetMatch was compared to that detected by XCMS for the same chromatograms. XCMS processing took 40 min on a similar computer (Intel Core 2 Duo, 2.9 GHz, no multithreading). However, it shall be mentioned that XCMS uses an untargeted approach and therefore does not make use of the reference feature list. The resulting Rt shift functions obtained by MetMatch and XCMS agree across the majority of the Rt range of interest (3 to 22 min). Rt values below 5 min, however, show no consensus. The underlying data show only few signals in this Rt region, and thus it was not possible to definitively confirm or reject any of the two proposed

Rt shifts by means of manual investigation. Again, the Pearson correlation coefficient was used to calculate the similarity of the retention time shift functions calculated by MetMatch and XCMS. The average correlation was 0.966 (standard deviation of 0.013) in the retention time interval of 4 to 22.5 min, demonstrating that both MetMatch and XCMS calculated very similar retention time shift functions for the chromatograms. An illustration of the retention time shifts detected by MetMatch and XCMS is shown in Figure S4 in Supplementary Material.

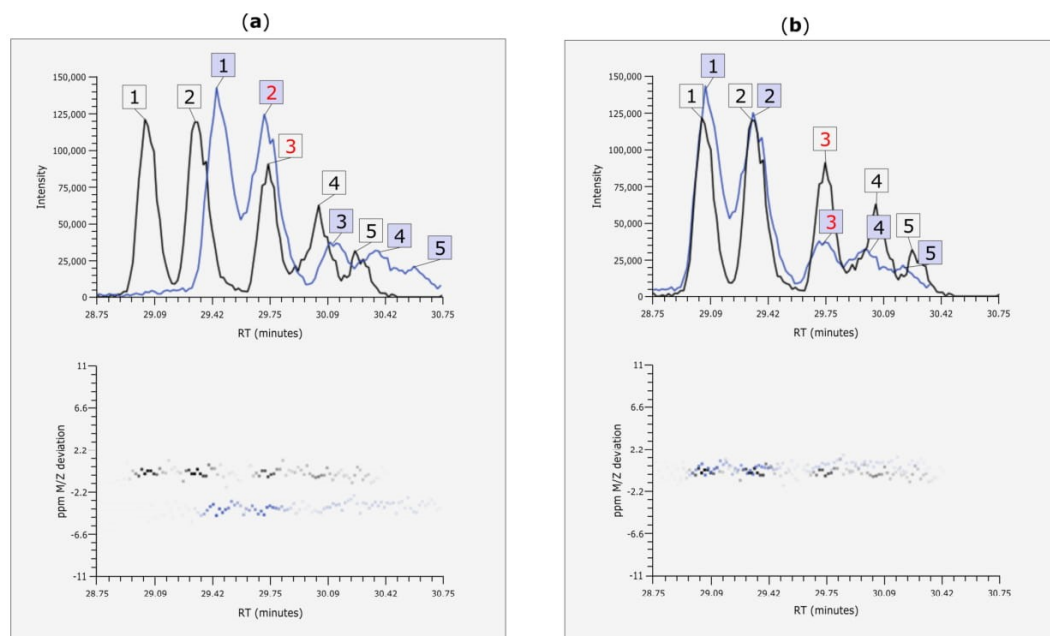


Figure 7. EIC (alignment before and after retention time (Rt) alignment and mass-to-charge ratio (m/z) correction with MetMatch. (a) A phenylalanine-derived chromatographic peak was annotated at position 3 in the reference file (black) and thus included in the reference feature list. The target peak (blue) at position 2 would probably be matched to it by means of manual matching or any other alignment approach that does not account for Rt shift; (b) Extracted ion chromatograms (EICs) after Rt and m/z shift correction using MetMatch. When the two corrected EICs are viewed simultaneously, it is apparent that each target peak should be matched to the reference peak denoted by the same ID number and that a systematic m/z shift is present between the two chromatograms.

Thirty-two out of 36 T-2 toxin/HT-2 toxin-derived metabolite ions were successfully matched in all target chromatograms. Three chromatographic peaks could not be correctly matched in any file because they were detected at an early Rt. Too few other peaks were detected around this Rt, leading to an inaccurate Rt shift function for this Rt region. One chromatographic peak was missed in two of the files during peak picking, because of its low abundance (peak height lower than 3000 counts) and its poor peak shape.

3.10. Evaluation of Artificial Data

While the evaluation of real-world data can provide important insight into the general quality of the resulting matches, knowledge of the true shifts is required in order to properly assess the results generated by MetMatch. To this end, several artificial datasets have been generated to test MetMatch's performance. For each of these artificial tests, the reference list used in the Phe experiment (Section 2.7) has been modified by applying various m/z and Rt shifts. The original, unmodified reference chromatogram served as the target. These two datasets were then used as input for MetMatch

and the generated results were compared to the artificially applied modifications. This approach was used to assess the number of true positives (Tp), false positives (Fp), false negatives (Fn) as well as precision, recall and F-measure. When matching the chromatogram directly to the unmodified reference feature list, results for a match without the need to correct for Rt or m/z shifts were obtained. This also served as a benchmark for further tests. The results of this match are displayed in Table 3.

Table 3. Benchmark results obtained by matching LC-full scan HRMS data of a wheat sample (serving as a target file) to a list of MetExtract-derived features from the same datafile (serving as the reference list).

Total Number of Features	Matched Features	Tp	Fp	Fn	Precision	Recall	F-Measure
1482	1446	1446	0	36	100%	97.6%	0.988

From the 1482 features in the reference list (originating from two separate datafiles, for more details see Materials and Methods), 1446 were correctly matched, while 36 could not be matched. The reference feature list is based on two chromatograms (Section 2.7), including some metabolites that were only detected in the chromatogram that was not used as a target here. Therefore, the corresponding EIC peaks of the underlying metabolites were not detectable during peak picking. Moreover, although both MetExtract and MetMatch applied the MassSpecWavelet peak-picking algorithm, slightly different results were obtained. Since the algorithm is applied in the context of two different software tools, which have very similar EIC extraction approaches but do not share a common implementation, the output of the peak-picking algorithm can vary slightly. Even though these chromatographic peaks could not directly be matched to features on the reference list, their expected Rts were correctly interpolated.

3.10.1. Formation of Different Ion Species

The automated search for differently formed ion species of the same metabolite is a distinct feature of MetMatch. To estimate the performance of the applied technique, the m/z values of all features in the reference list have been modified. The goal was to simulate a situation in which all metabolites contained in the reference list constituted the same type of ion species (A1), while for all of these metabolites a single different ion species (A2) was assumed for the target datafiles. Thus, the extended lists contained the m/z values of hypothetical adducts, which are not part of the original chromatogram. The inclusion of A1 and A2 in the target ion list, which is used to detect alternatively formed ion species, should allow MetMatch to detect all features regardless of their (modified) ion species. The results of this performance test are displayed in Table 4.

Table 4. Results obtained by matching LC full scan HRMS data of a wheat sample (serving as a target file) to a list of MetExtract-derived features from the same datafile (serving as the reference list) after modifying the m/z values to simulate the formation of completely different ion species in the reference and target datafile.

Total Number of Features	Matched Features	Tp	Fp	Fn	Precision	Recall	F-Measure
1482	1446	1446	0	36	100%	97.6%	0.988

MetMatch successfully matched 1446 chromatographic peaks to features although they were simulated to be present as different ion species. The results are identical to the benchmark evaluation (see above), suggesting that MetMatch is able to match features regardless of their differing ion adducts.

3.10.2. m/z Shift Detection

To test the accuracy and reliability of m/z shift correction, various modified reference feature lists have been created which display m/z offsets of -10 to $+10$ ppm in 1 ppm incremental steps. In total, 21 different lists have been created to assess the difference between the artificially imposed

m/z offset and the value calculated by MetMatch (Figure 8). The bins used in the underlying m/z offset calculation were set to a size of 0.01 ppm. Further decreasing the bin size did not increase the accuracy.

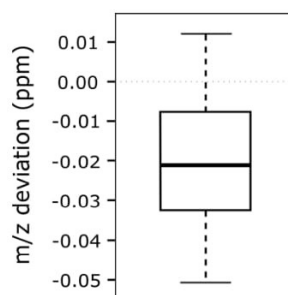


Figure 8. Box-plot showing the differences between imposed and detected mass-to-charge ratio (m/z) offset. A chromatogram was matched to 21 modified feature lists obtained by processing the same datafile with MetExtract and artificially imposing m/z offset values of -10 to $+10$ ppm, with 1 ppm increments per list.

No correlation between the magnitude of artificially imposed m/z offset and the accuracy of MetMatch with respect to the correctly matched reference features in the target chromatogram has been observed. Differences between imposed and detected m/z offset were within the range of $+0.01$ to -0.05 ppm for all simulated cases. This is significantly less than the mass variation across a typical chromatographic peak recorded on the used LTQ Orbitrap instrument. In this test, the observed m/z discrepancy is explained by the fact that the m/z values of the reference features were calculated based on two chromatograms (Section 2.7). Thus, it is not surprising that the chromatogram, which was matched to this reference list, shows a marginal m/z deviation.

3.10.3. Rt Shift Detection

Six modified reference feature lists were created displaying varying degrees of Rt shift relative to the original chromatogram. The applied retention time transformations are illustrated in Figure 9, column “Simulated Rt Shift”. On top of the respective systematic shifts, all feature groups were shifted randomly by ± 6 s, simulating small Rt shifts observed in real-world data. Additionally, differently formed ion adducts were simulated analogous to Section 3.10.1.

Irrespective of the tested simulation, MetMatch achieved a high degree of correctly matched chromatographic peaks that are almost identical to the benchmark test. Occasionally, values exceeding the benchmark could also be observed due to slight variations in the generated EICs, which sometimes led to different peak-picking results. Although some features were incorrectly matched, the number of Fp (5 to 16) was quite low compared to Tp (1389 to 1448). Some of the Fp originated from the alternative ion species that were added to the reference feature list. In case of the modified reference feature list simulating a discontinuous Rt shift (simulation 3), the lowest number of correctly matched chromatographic peaks to reference features was observed. This is not surprising, since the algorithm tries to detect a continuous Rt shift (see Section 2.5) and only approximates the discontinuous shift applied in simulation 3. Especially around the erratic increase, MetMatch failed to fit the theoretical shift correctly. However, the number of 1389 correctly matched chromatographic peaks in contrast to 16 incorrectly matched peaks still corresponds to an F-measure value of 0.962. All other artificial Rt shifts resulted in F-measure values ranging from 0.983 to 0.987, which, compared to the benchmark F-measure value of 0.988, shows that MetMatch reliably detected the applied Rt shifts and produced high quality results.

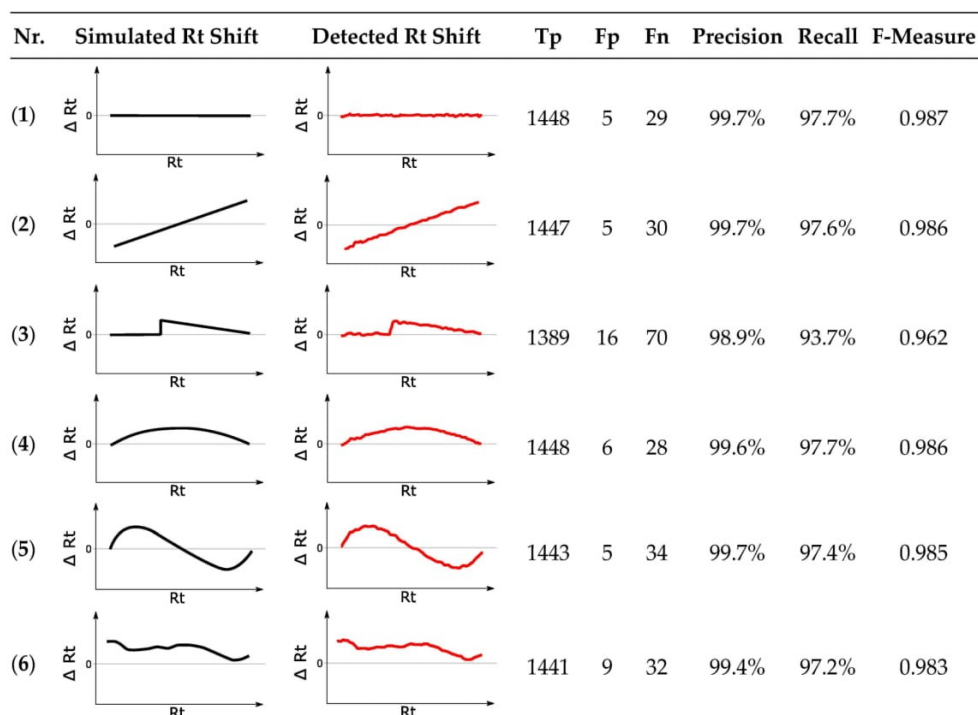


Figure 9. Results for six different simulated retention time (Rt) shifts. The reference chromatogram was matched to six modified feature lists, displaying artificially modified Rt shifts visualized by the black curves. The Rt shifts obtained by MetMatch are shown in red. (1) No Rt shift apart from random ± 6 s; (2) Monotonically increasing linear Rt shift; (3) Discontinuous Rt shift, as may, e.g., happen by a defective LC pump; (4) Non-linear Rt shift without inflection; (5) Non-linear Rt shift with inflection; (6) Complex non-linear Rt shift based on the type of Rt shift observed in the real-world wheat experiment.

3.10.4. Minimum Required Common Feature Groups

As outlined in Section 3.4, a minimum number of feature groups has to be part of the reference feature list as well as the target chromatograms; otherwise, the Rt shift function cannot be reliably calculated. To test how large this number has to be in order to obtain reliable results, the reference feature list of Section 3.10.3, simulation 6, has been repeatedly matched to the original chromatogram, as described in Section 3.10. The number of common feature groups has been varied from 505 (complete list) to one. Starting from the complete list, subsets of decreasing size (each list being 25% smaller than its predecessor) have been created. This way, 20 reference feature lists with a decreasing number of common feature groups were simulated. Feature groups were deleted at random. This approach was repeated 10 times, cumulating in 200 MetMatch processing runs. The resulting F-measures for the various number of matched feature groups are shown in Figure 10.

On average, 159 common feature groups were required to maintain the same accuracy as achieved by 505 common feature groups. All test runs resulted in F-measures greater than 0.9 if at least 89 common feature groups were present. Decreasing the number to 49 common features groups still resulted in F-measures greater than 0.8 for all test runs. Below this number, the accuracy dropped more quickly. However, these numbers might vary for other datasets, as they are highly dependent on the number of isomers in each narrowed search frame (i.e., after m/z shift correction) as well as the number of narrowed search frames solely containing unrelated peaks. The higher these numbers, the higher the minimum required number of common feature groups.

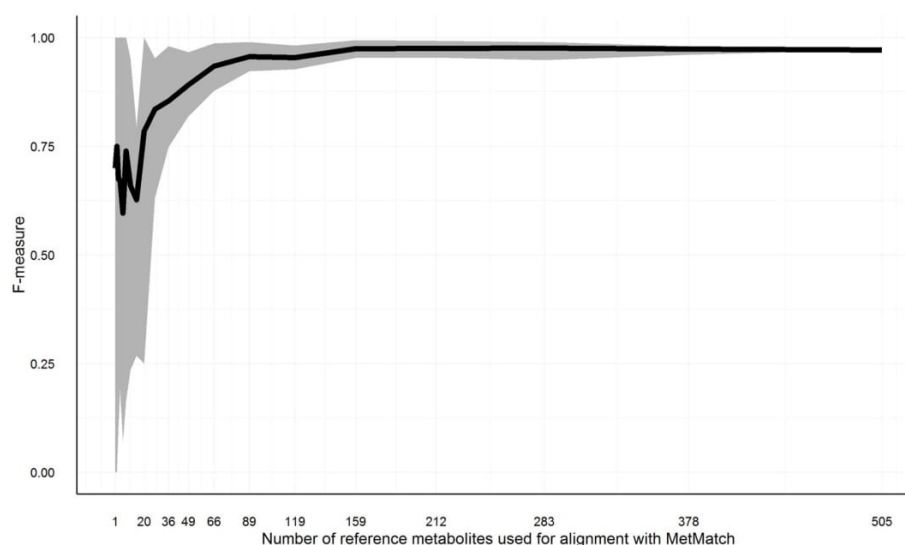


Figure 10. Illustration of the F-measure range (grey area) and average (black line) for reference feature lists of decreasing size. To estimate the minimum number of entries in the reference feature list, a simulation was carried out. In total, 200 subsets of decreasing size were generated. For each of the generated subsets, the m/z and R_t alignment step was calculated with MetMatch. The minimum (lower border of grey area), average (black line) and maximum (upper border of grey area) values are illustrated for each of the iterations.

4. Conclusions

Metabolomics is a valuable tool for studying phenotype changes [1] and is therefore widely used. Retrospective and comparative analysis of LC-HRMS chromatograms of separate biological experiments is of particular interest since they hold a considerable amount of information that can be used to investigate aspects beyond the scope of their initial acquisition and evaluation [2]. However, the reliable, comparative analysis is still an issue, due to shifts and drifts of retention time (R_t) and mass-to-charge ratio (m/z) or the formation of different ion species caused by the limited stability of experimental and instrument parameters. This holds true when comparing results obtained in multiple analytical batches using the same instrument, and especially when obtained by different instruments [5]. Several algorithms and software tools have been proposed to address this problem, but many of them suffer from severe drawbacks such as incorrect model assumptions and long run times [4].

In this respect, MetMatch is a novel software tool for the comparison of LC-HRMS data utilized through a convenient user interface. Chromatograms are matched to a feature list, which can be created manually or by processing one or multiple datafile(s), which serves as a reference. Constant shifts in m/z are automatically detected and corrected for. Both linear and non-linear R_t shift correction is achieved using a fast, iterative algorithm based on the observation that metabolites/compounds exhibiting similar R_t values within a measurement usually share analogous R_t shifts between measurements. Furthermore, MetMatch is able to account for the formation of different ion species that may be formed under slightly differing experimental conditions. The developed software offers a convenient graphical user interface that allows for easy optimization and verification of every processing step.

The quality of results obtained by MetMatch was assessed using two different test cases. The first case involved 75 newly characterized Phe-derived metabolites that were matched to a previous wheat experiment. Manual validation of 23 selected features was performed, 19 of which could be confirmed in all 10 inspected target files. Additionally, 36 T-2 toxin/HT-2 toxin-derived

metabolite ions were matched to another barley experiment, 32 of which could be manually verified in all three target files. In both cases, MetMatch-derived Rt shifts were very similar to XCMS processing results. Furthermore, software evaluation by means of artificially manipulated reference data showed that MetMatch is well suited to account for the formation of different ion species between reference and target datasets and that both m/z and Rt shifts are detected with high accuracy (on average, 96.9% of all peaks were correctly matched).

In conclusion, we think that MetMatch will be a useful tool for the comparative analysis of metabolomics LC-HRMS experiments and we plan to add additional functionalities such as detection of non-linear m/z shifts and automatic assignment/correction of questionable matched features to further improve accuracy.

Supplementary Materials: The Supplementary materials are available online at <http://www.mdpi.com/2218-1989/6/4/39/s1>: Figure S1. Pseudocode of the iterative retention time (Rt) shift detection algorithm; Figure S2. Comparison of the Rt shift detected in the Phe-derived biological experiment using MetMatch and XCMS relative to the selected reference chromatogram; Figure S3. Comparison of the Rt shift detected in the Phe-derived biological experiment using chromatograms obtained in positive and negative ionization mode for the same samples; Figure S4. Comparison of the Rt shift detected in the T-2 toxin/HT-2 toxin-derived biological experiment using MetMatch and XCMS relative to the selected reference chromatograms; Table S1. Sample runtimes measured on an Intel Core 2 Duo (2.8 GHz); Table S2. Manual validation of the results obtained by matching the target dataset to the reference feature list containing Phe-derived metabolites.

Acknowledgments: The authors thank Elisabeth Varga and Franz Berthiller for providing technical assistance with the QTOF-LC-HRMS measurements of the T-2/HT-2 example. We also thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. The Austrian Science Fund (projects: Metabolism of T-2 and HT-2 P26213 and SFB-Fusarium-F37, subproject 3715) and the Government of Lower Austria (projects: NovAlgo and NoBiTUM) are acknowledged for financial support.

Author Contributions: Stefan Koch designed, implemented and tested the MetMatch software and wrote large parts of the manuscript. Christoph Bueschl co-supervised Stefan Koch and performed data processing of the reference datasets. Maria Doppler analyzed the wheat samples with LC-Orbitrap-HRMS. Alexandra Simader verified the MetMatch software. Jacqueline Meng-Reiterer analyzed the T-2/HT-2 samples with LC-QTOF-HRMS. Marc Lemmens was responsible for the biological experiments. Rainer Schuhmacher contributed to the design of MetMatch and supervised the work. All authors contributed to the manuscript and read and approved the final version of the article.

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

Abbreviations and Terminology

The following abbreviations and terminologies are used in this manuscript:

Binning	A range (e.g., m/z or Rt shift values) is divided into a series of equally long intervals (bins).
Dataset	A list of LC-HRMS datafiles belonging to one analytical batch, originating from the same experiment.
EIC	Extracted ion chromatogram. An EIC represents the abundance/intensity of a certain m/z value in each mass spectrum of a chromatogram. The mass spectra are sorted according to their acquisition time relative to injecting the sample into the HPLC instrument. Usually a small deviation from the selected m/z value ($\pm$ ppm; depending on the used HRMS instrument) is allowed to account for minor scan-to-scan m/z inconsistencies observed during acquisition of a chromatogram.
Feature	A feature represents a particular metabolite ion of interest. It is characterized by its m/z value (experimental and theoretical) and retention time (as calculated by the used peak-picking algorithm). Features are usually the result of an untargeted metabolomics data processing approach and typically comprise multiple chromatographic peaks, one per LC-HRMS chromatogram.

Feature group	All features originating from the same chemical substance and hence sharing the same Rt and peak shape.
F-measure	Harmonic mean of precision and recall.
Fn	False negatives. Features that were not matched.
Fp	False positives. If a matched feature does not meet the criteria to be considered as Tp, it is an Fp.
LC-HRMS	Liquid chromatography coupled to high resolution mass spectrometry.
<i>m/z</i>	Mass-to-charge ratio.
Peak/Chromatographic peak	A chromatographic peak representing a particular feature in a chromatogram. Other than a feature, a peak has a defined <i>m/z</i> value, a retention time (as defined by the used peak-picking algorithm) as well as a peak area. Peaks are detected by the means of a peak-picking algorithm.
Peak apex	The time at which the most abundant signal of a respective chromatographic peak is detected after injection of the sample into the HPLC instrument.
Peak center	The time at which a chromatographic peak has its center of gravity. (Time) measurement is started with injecting the sample into the HPLC instrument.
Precision	Percentage of matched reference features that were correctly matched. Calculation: Number of Tp divided by the number of Tp plus Fp.
Recall	Percentage of all reference features that were correctly matched. Calculation: Number of Tp divided by the number of all reference features.
Reference dataset	A single or multiple LC-HRMS datafiles, which are used to generate the reference feature list.
Reference feature list	List of features to which target datasets are matched.
Rt	Retention time.
Rt shift function	A function that holds the calculated retention time shift between a target chromatogram relative to the reference feature list. It is used to map the retention times of detected chromatographic peaks in the target chromatogram to the retention time of the features in the reference list.
Search frame	Signals of a chromatogram (single datafile) that fall within a certain Rt and <i>m/z</i> tolerance window around a reference feature. These values have to be slightly larger than the maximum expected shift. (e.g., 1.5 min Rt shift and 10 parts per million (ppm) <i>m/z</i> shift for a typical LC-HRMS chromatogram).
Rt shift array	An array containing the Rt shift of all peaks detected within a search frame relative to the retention time of the respective reference feature. First, a search frame is discretized into smaller bins (number of bins is set by the user of MetMatch) each representing a certain Rt offset relative to the retention time of the respective reference feature. Then, those Rt bins for which chromatographic peaks of the respective reference feature were detected in the target chromatogram, are assigned weight values greater than zero (either automatically or user-guided). Any Rt offset bin for which no chromatographic peak was detected, is assigned zero value and thus do not contribute to the Rt shift detection algorithm. The number of entries in each Rt shift array is set by the user during the iterative Rt alignment step. Each entry in the Rt shift array corresponds to a certain Rt deviation between the retention times of the reference features and their corresponding chromatographic peaks in the target chromatograms.

Rt shift matrix	The individual Rt shift arrays of all reference feature groups are concatenated column-wise and sorted by the retention time of their reference feature group. Each column represents a feature group, while rows represent discretized Rt shift values.
Target dataset	A list of LC-HRMS datafiles to be matched to a reference feature list.
Tp	True positives. Matched features are considered Tp if their m/z deviation is ≤ 5 ppm and their Rt deviation is ≤ 8.4 s (manually validated threshold value for the distinction of Tp and Fp in this datafile) relative to the reference feature after shift correction. The formation of different ion species is taken into account when comparing the m/z .

References

1. Dettmer, K.; Aronov, P.A.; Hammock, B.D. Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.* **2007**, *26*, 51–78. [[CrossRef](#)] [[PubMed](#)]
2. Fiehn, O. Metabolomics—The link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171. [[CrossRef](#)] [[PubMed](#)]
3. Patti, G.J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: The apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269. [[CrossRef](#)] [[PubMed](#)]
4. Smith, R.; Ventura, D.; Prince, J.T. LC-MS alignment in theory and practice: A comprehensive algorithmic review. *Brief. Bioinform.* **2015**, *16*, 104–117. [[CrossRef](#)] [[PubMed](#)]
5. Aberg, K.M.; Alm, E.; Torgrip, R.J. The correspondence problem for metabonomics datasets. *Anal. Bioanal. Chem.* **2009**, *394*, 151–162. [[CrossRef](#)] [[PubMed](#)]
6. Kassidas, A.; MacGregor, J.F.; Taylor, P.A. Synchronization of batch trajectories using dynamic time warping. *AIChE J.* **1998**, *44*, 864–875. [[CrossRef](#)]
7. Nielsen, N.-P.V.; Carstensen, J.M.; Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* **1998**, *805*, 17–35. [[CrossRef](#)]
8. Eilers, P.H. Parametric time warping. *Anal. Chem.* **2004**, *76*, 404–411. [[CrossRef](#)] [[PubMed](#)]
9. Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal. Chem.* **2008**, *80*, 3095–3104. [[CrossRef](#)] [[PubMed](#)]
10. Prince, J.T.; Marcotte, E.M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **2006**, *78*, 6140–6152. [[CrossRef](#)] [[PubMed](#)]
11. Zhang, X.; Asara, J.M.; Adamec, J.; Ouzzani, M.; Elmagarmid, A.K. Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics* **2005**, *21*, 4054–4059. [[CrossRef](#)] [[PubMed](#)]
12. 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. [[CrossRef](#)] [[PubMed](#)]
13. Katajamaa, M.; Miettinen, J.; Oresic, M. Mzmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006**, *22*, 634–636. [[CrossRef](#)] [[PubMed](#)]
14. 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*, 395–404. [[CrossRef](#)] [[PubMed](#)]
15. Bertsch, A.; Gropl, C.; Reinert, K.; Kohlbacher, O. Openms and TOPP: Open source software for LC-MS data analysis. *Methods Mol. Biol.* **2011**, *696*, 353–367. [[PubMed](#)]
16. Bloemberg, T.G.; Gerretzen, J.; Lunshof, A.; Wehrens, R.; Buydens, L.M. Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. *Anal. Chim. Acta* **2013**, *781*, 14–32. [[CrossRef](#)] [[PubMed](#)]
17. Christin, C.; Hoefsloot, H.C.; Smilde, A.K.; Suits, F.; Bischoff, R.; Horvatovich, P.L. Time alignment algorithms based on selected mass traces for complex LC-MS data. *J. Proteome Res.* **2010**, *9*, 1483–1495. [[CrossRef](#)] [[PubMed](#)]

18. Christin, C.; Smilde, A.K.; Hoefsloot, H.C.; Suits, F.; Bischoff, R.; Horvatovich, P.L. Optimized time alignment algorithm for LC-MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.* **2008**, *80*, 7012–7021. [CrossRef] [PubMed]
19. Bueschl, C.; Kluger, B.; Berthiller, F.; Lirk, G.; Winkler, S.; Krska, R.; Schuhmacher, R. Metextract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. *Bioinformatics* **2012**, *28*, 736–738. [CrossRef] [PubMed]
20. MetMatch. Available online: <https://metabolomics-ifa.boku.ac.at/MetMatch/index.html> (accessed on 31 October 2016).
21. Kluger, B.; Bueschl, C.; Neumann, N.; Stuckler, R.; Doppler, M.; Chassy, A.W.; Waterhouse, A.L.; Rechthaler, J.; Kamplaitner, N.; Thallinger, G.G.; et al. Untargeted profiling of tracer-derived metabolites using stable isotopic labeling and fast polarity-switching lc-esi-hrms. *Anal. Chem.* **2014**, *86*, 11533–11537. [CrossRef] [PubMed]
22. Bueschl, C.; Kluger, B.; Lemmens, M.; Adam, G.; Wiesenberger, G.; Maschietto, V.; Marocco, A.; Strauss, J.; Bodi, S.; Thallinger, G.G.; et al. A novel stable isotope labelling assisted workflow for improved untargeted LC-HRMS based metabolomics research. *Metabolomics* **2014**, *10*, 754–769. [CrossRef] [PubMed]
23. Warth, B.; Parich, A.; Bueschl, C.; Schoefbeck, D.; Neumann, N.K.; Kluger, B.; Schuster, K.; Krska, R.; Adam, G.; Lemmens, M.; et al. GC-MS based targeted metabolic profiling identifies changes in the wheat metabolome following deoxynivalenol treatment. *Metabolomics* **2015**, *11*, 722–738. [CrossRef] [PubMed]
24. Meng-Reiterer, J.; Varga, E.; Nathanail, A.V.; Bueschl, C.; Rechthaler, J.; McCormick, S.P.; Michlmayr, H.; Malachova, A.; Fruhmman, P.; Adam, G.; et al. Tracing the metabolism of ht-2 toxin and t-2 toxin in barley by isotope-assisted untargeted screening and quantitative LC-HRMS analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 8019–8033. [CrossRef] [PubMed]
25. Fourment, M.; Gillings, M.R. A comparison of common programming languages used in bioinformatics. *BMC Bioinform.* **2008**, *9*. [CrossRef] [PubMed]
26. Algorithm1. Available online: <https://metabolomics-ifa.boku.ac.at/metmatch/figures/Algorithm1.gif> (accessed on 31 October 2016).
27. Pedrioli, P.G.; Eng, J.K.; Hubley, R.; Vogelzang, M.; Deutsch, E.W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R.H.; Apweiler, R.; et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **2004**, *22*, 1459–1466. [CrossRef] [PubMed]
28. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. Proteowizard: Open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534–2536. [CrossRef] [PubMed]
29. Podwojski, K.; Fritsch, A.; Chamrad, D.C.; Paul, W.; Sitek, B.; Stuhler, K.; Mutzel, P.; Stephan, C.; Meyer, H.E.; Urfer, W.; et al. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics* **2009**, *25*, 758–764. [CrossRef] [PubMed]
30. Mitra, V.; Smilde, A.; Hoefsloot, H.; Suits, F.; Bischoff, R.; Horvatovich, P. Inversion of peak elution order prevents uniform time alignment of complex liquid-chromatography coupled to mass spectrometry datasets. *J. Chromatogr. A* **2014**, *1373*, 61–72. [CrossRef] [PubMed]



© 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/>).

Supplementary Materials: MetMatch: A Semi-Automated Software Tool for the Comparison and Alignment of LC-HRMS Data from Different Metabolomics Experiments

Stefan Koch, Christoph Bueschl, Maria Doppler, Alexandra Simader, Jacqueline Meng-Reiterer, Marc Lemmens and Rainer Schuhmacher

1. Implementation Details of MetMatch

1.1. Peak-Picking Algorithms

Currently, three different chromatographic peak-picking algorithms have been implemented and can be selected in MetMatch. Table S1 shows an overview of the different algorithm's runtimes.

- MassSpecWavelet

In order to provide a reliable peak-picking algorithm, Du et al. [1] implemented this continuous wavelet transform (CWT)-based algorithm that includes the consideration of the shape of the peaks. The retention time of a detected chromatographic peak is the peak's apex.

- Gauss Peak Correlation

The extracted ion chromatogram (EIC) of a search frame is correlated to Gaussian peaks of different widths, after which the resulting array of Pearson correlations is smoothed using a moving average calculation. Minima and maxima of the correlation array are calculated using the first derivative and are then utilized to determine peak retention time and borders. The retention time of a detected chromatographic peak is the peak's center. This algorithm has been developed in-house.

- Savitzky-Golay Filter

To further improve the performance of the above algorithm while keeping the general peak-picking principle, EICs are directly smoothed using a Savitzky-Golay filter [2]. Minima and maxima of the smoothed EIC are calculated using the first derivative and are then utilized to determine peak retention time and borders. The retention time of a detected chromatographic peak is the peak's center.

Table S1. Sample runtimes measured on an Intel Core 2 Duo (2.8 GHz). Reference Features: number of features in the reference feature list. Additional Adduct Ions: number of alternative adduct ions specified by the user. Total Features: number of features after alternative adduct ion generation. Note that the theoretical number would be higher, but some features have been rejected as their retention time and mass-to-charge ratio values are too similar to an already included feature. Runtimes for the individual peak-picking algorithms are displayed in hours:minutes:seconds.

Files	Reference Features	Additional Adduct Ions	Total Features	Runtime		
				MassSpecWavelet	Peak Correlation	Savitzky-Golay Filter
1	500	0	500	00:00:15	00:00:01	00:00:01
5	500	0	500	00:01:13	00:00:03	00:00:02
10	500	0	500	00:02:25	00:00:04	00:00:03
1	1000	0	1000	00:00:31	00:00:02	00:00:01
1	1500	0	1500	00:00:59	00:00:02	00:00:01
1	1500	3	9439	00:01:25	00:00:03	00:00:02
1	1500	5	28,660	00:02:38	00:00:04	00:00:02
1	20,000	5	394,110	01:56:13	00:08:10	00:01:28

1.2. Rt Shift Detection Algorithm

A novel, iteratively applied algorithm for the efficient calculation of the Rt shift present in a target chromatogram relative to the reference feature list has been developed. A matrix with peak shift information (**M**), an array containing the current Rt shift function (**c**) and two tolerance parameters (**x**, **y**) are required as inputs. The output is an array containing the updated Rt shift function (**c'**). Figure S1 shows the pseudocode of the algorithm.

Input: (**M**, **c**, **x**, **y**)

matrix **M** Rt-shift matrix. Columns represent the feature groups from the reference feature list and rows represent discretized Rt shifts relative to the retention time of the respective reference features. Each cell in the matrix with a value greater than 0 indicates a chromatographic peak in the respective target chromatogram that has a certain Rt-offset (=row-index) relative to its reference feature (=column-index). These values are weights, which have been set to 1 by MetMatch or higher values by the user.

array **c** current Rt shift values, corresponding to row indices in the matrix **M**, for each feature group of matrix **M**.

integer **x** maximum allowed feature group tolerance, corresponding to column indices in the matrix **M**

integer **y** maximum allowed Rt shift tolerance, corresponding to row indices in the matrix **M**

```

1  start
2      Initialize:    array weights, where  $weights_f = 1$  for each  $f \in \{1, \dots, length[c]\}$ 
3                      array c' = copy of c
4
5      for each  $f \in \{1, \dots, length[c]\}$  do
6          maxproximity = 0
7          newshift = -1
8          for each  $r \in \{c_f - y, \dots, c_f + y\}$  and  $r$  within borders of M do
9              (a)      proximity =  $\sum_{i=f-x}^{f+x} \sum_{j=c_f-y}^{c_f+y} \frac{M_{i,j}}{(|j-r|+1)^2}$  only for  $M_{i,j}$  within borders of M
10                 if proximity > maxproximity then
11                     maxproximity = proximity
12                     newshift = r
13                 end if
14             end for
15             if newshift > -1 then
16                 for each  $i \in \{f - x, \dots, f + x\}$  and  $i$  within borders of M do
17                     xdistance =  $\frac{x+1-|i-f|}{x+1}$ 
18                 (b)       $c'_i = \frac{c'_i * weights_i + newshift * maxproximity * xdistance}{weights_i + maxproximity * xdistance}$ 
19                     weightsi = weightsi + maxproximity * xdistance
20                 end for
21             end if
22         end for
23     end

```

Figure S1. Pseudocode of the iterative retention time (Rt) shift detection algorithm. An Rt shift matrix **M**, an array **c** containing the current Rt shift for each column of the matrix and two range parameters **x** and **y** are required as inputs. The algorithm optimizes the current Rt shift **c** to be as close to as many chromatographic peaks as possible in the respective calculation window (a box drawn around each reference feature). The range parameters determine the size of this calculation window. The maximum Rt deviation from the current solution is set using the row range (**y**). The column range (**x**) allows optimizing for local maxima (low values) or more global maxima (high values). (a) Detection of the Rt shift value shared by most of the peaks within the calculation window is achieved by minimizing the squared distance to all peaks in the current window. (b) The final Rt shift of a feature group corresponds to the weighted average of the Rt shift of all feature groups

within the column range. The weight linearly decreases with distance, and increases proportional to the weight of all peaks in close proximity to the respective R_t shift.

1.2.1. Description of Input

Each column of the matrix $\mathbf{M}$ represents one feature group from the reference feature list, while each row of matrix $\mathbf{M}$ corresponds to a specific R_t shift value. A cell in matrix $\mathbf{M}$ holds the maximum weight of all peaks of a feature group (specified by its column) that have a certain R_t shift (specified by its row) relative to the reference. A value greater than 0 in a cell of the matrix $\mathbf{M}$ therefore indicates that a chromatographic peak for the respective metabolite has been detected in a certain target chromatogram at a certain retention time offset corresponding to the row index of the cell. Columns are sorted by R_t of their respective feature group.

For example: The maximum allowed R_t shift tolerance is equal to 50 s, and the number of R_t shift bins is equal to 100. The first row would now hold the weights of peaks that have an R_t shift of >49 to 50 s. The R_t shift range for the second row is then equal to >48 to 49 s and so on. The middle row of the matrix, in this example the 50th row, corresponds to no retention time shift and covers the R_t shift of >0 to 1 s. A chromatographic peak detected at an R_t -shift of 20 s relative to the retention time of its reference feature would therefore be saved in the 30th row (corresponding to an R_t shift of >20 to 21 s) of the matrix $\mathbf{M}$. The cell of the 30th row and the respective reference feature (column in $\mathbf{M}$) would have a value of 1 if the user has not change the weight for that reference feature.

An entry or “column” of array $\mathbf{c}$ corresponds to the same feature group as the respective column of matrix $\mathbf{M}$. Each entry contains the current value of the R_t shift function of the respective feature group. The value is specified by the index of the R_t shift bin (row of matrix $\mathbf{M}$), but does not need to be an integer (interpolation between rows of the matrix $\mathbf{M}$ is performed for $\mathbf{c}$).

For example: For the previously depicted example of a chromatographic peak detected at an R_t offset of 20 s relative to its reference feature, the value in $\mathbf{c}$ for this reference feature is likely close to 30, since the 30th row in the matrix $\mathbf{M}$ corresponds to a relative R_t shift of >20 to 21 s. However, the actual value depends on the search window and on the neighboring chromatographic peaks in that region of the target chromatogram.

The tolerance parameter x specifies the maximum distance (in column indices of matrix $\mathbf{M}$) two feature groups can have for them to still contribute to the updated R_t shift value (stored in $\mathbf{c}'$) of one another. It therefore determines whether to optimize for a global fit (bigger values) or a more local fit (smaller values).

The tolerance parameter y specifies the maximum distance from the current R_t shift value (in row indices of matrix $\mathbf{M}$ = R_t shift bins) a peak can have to still contribute to the updated R_t shift value of a feature group

1.2.2. Algorithm

Here, the algorithm presented in the pseudocode in Figure S1 is described in detail, referencing the current line in brackets. Generally, the algorithm updates the R_t shift function value c'_f of each feature group (for loop from lines 5 to 22). The array $\mathbf{c}'$ is a copy of $\mathbf{c}$, and only written to but never read from an iteration of the algorithm. This way, intermediate results and the direction in which the matrix is processed do not influence the alignment iteration. Every c'_f is updated in the following way:

A box (i.e., calculation window) is considered around the current c_f whose dimensions are specified by the tolerance parameters x (first sum in line 9) and y (second sum in line 9). The R_t shift value *newshift* that is in the center of the biggest cluster of peaks is detected (for the loop in lines 8 to 14). This is done by iterating over all row indices r that are within the box. Next, the weight of all peaks within the box, divided by the squared row distance from the current r , is summed up and stored as *proximity* (line 9). Then, *newshift* is assigned the row index r for which *proximity* was maximized (line 12). The distance is squared to promote the detection of clusters.

The *newshift* of a feature group is not only used to update its own Rt shift function value c'_f , but also the value of neighboring feature groups (similar retention time shifts are assumed for local neighbors, for loop from lines 16 to 22). Having found the Rt shift value *newshift* that is within the biggest cluster of the box, the values of c' that correspond to any of the feature groups within the box are updated. The final value of c'_f is a weighted average of the *newshift* values of all feature groups within the box. All contributing *newshift* values are weighted by the column distance of their feature group to the current feature group (*xdistance*, line 17) and the corresponding *proximity*. Rather than calculating the weighted average once the whole matrix *M* has been processed, calculation is performed using a running weighted average (line 19). The current sum of weights for all c'_f is stored in the **weights** array (line 19). The weight of each *newshift* that contributes to the weighted moving average is calculated by multiplying the corresponding *proximity* with *xdistance* (lines 18, 19).

To summarize, a box (defined by the row and column tolerance) around the current Rt shift function value of each feature group is drawn. The biggest cluster of peaks inside this box is detected, and the Rt shift function values of all feature groups inside the box are updated.

1.3. Libraries and Tools Used in MetMatch

The following libraries and tools have been used in the presented implementation of MetMatch

- Apache Commons Math 3
The normal distribution and Pearson correlation coefficient modules are utilized to calculate peak correlation with an idealized Gauss peak. Furthermore, linear interpolator is used to generate corrected mzXML files.
- JRI (Java/R Interface)
The JRI package is used to interface MetMatch with an R-instance, as one algorithm for peak picking (MassSpecWavelet) is implemented in R.
- JavaFX
MetMatch's user interface is set up with JavaFX, which is integrated into the Java Runtime Environment (JRE).
- ControlsFX
This library is an add-on to JavaFX and includes many useful user interface control elements, resulting in a user-friendly and responsive interface.
- Java NIO
The usage of memory-mapped files promises the possibility to store information on a disk (HDD/SSD) without discernible performance loss, as long as it is only processed occasionally.
- Michael Thomas Flanagan's Java Scientific Library
Peak smoothing using either a moving average algorithm or a Savitzky-Golay filter is achieved through this library.
- uniVocity
Input files are read and output files are formatted as tab-separated files (.tsv), which is performed with this efficient library.
- Maven
The maven build system is used for managing project dependencies as well as building MetMatch.

2. Biological Experiments and MetMatch Data Processing

2.1. Phenylalanine-Derived Metabolites

2.1.1. Phenylalanine as Tracer

To provide insights into the set of Phe-derived metabolites of wheat, plants of the cultivar "Remus" were grown in a glass house under controlled conditions. At flowering stage which was reached after nine weeks, eight spikelets were inoculated with a 5g/l $^{13}\text{C}_9$ phenylalanine solution

(Cambridge Isotope Laboratories, ^{13}C enrichment of 99.5%). Seventy-two hours after inoculation, samples were harvested and immediately frozen in liquid nitrogen. For mock/control samples wheat ears were treated with water only.

2.1.2. Full Metabolome-Labeled Wheat Samples

^{13}C wheat plant ("Remus") was grown in a tailor made plant growing chamber (ECH-Halle, Halle, Germany) under controlled conditions (regulated temperature, light regime, nutrient solution, $^{13}\text{CO}_2$ concentration) for 62 days until the flowering stage (^{13}C enrichment of 98.8%). ^{12}C plants were grown under the same conditions.

The following *MetMatch* processing parameters were used to match the target dataset to the reference feature list:

Input Parameters

- Retention Time: RT
- Mass/Charge: MZ
- Ion ID: Num
- Metabolite ID: Group_ID
- Rest: (they were left blank)

General Parameters

- Start: 0
- End: 40
- RT Tolerance (Parsing Parameters): 1.5
- M/Z Tolerance (Parsing Parameters): 11
- M/Z Tolerance (Slice Parameters): 3.5
- Minimum Signals: 5
- Minimum Consecutive Signals: 3
- Intensity Cutoff: 5000
- Peak Picking: MassSpecWavelet
- S/N Threshold: 3
- Scales: 3, 19
- RT Tolerance (Rt Shift): 0.12
- Number of Bins: 100

Adducts

- Ion Name: $1\text{M} + \text{H}$, Charge: 1+, Ion Mass: $\text{M} + 1.007$
- Ion Name: $1\text{M} + \text{NH}_4$, Charge: 1+, Ion Mass: $\text{M} + 18.034$
- Ion Name: $1\text{M} + \text{Na}$, Charge: 1+, Ion Mass: $\text{M} + 22.989$
- Ion Name: $1\text{M} + \text{CH}_3\text{OH} + \text{H}$, Charge: 1+, Ion Mass: $\text{M} + 33.033$
- Ion Name: $1\text{M} + \text{K}$, Charge: 1+, Ion Mass: $\text{M} + 38.963$
- Rest: (they were left blank)

Rt Shift Calculator Parameters

- 1st iteration: xRange: 65, yRange: 50
- 2nd iteration: xRange: 6.5, yRange: 20
- 3rd iteration: xRange: 3, yRange: 10
- 4th iteration: xRange: 2, yRange: 5

2.2. Creation of the Reference List (Phe Experiment)

2.2.1. Data Processing with TracExtract

Samples of wheat ears were searched for (^{13}C -) Phe-derived metabolites using an updated version of the in-house-developed TracExtract algorithm (unpublished software) [3,4].

- Search for Isotope Patterns from Native and ^{13}C -Labeled Metabolites

Signals of native and ^{13}C -labeled metabolite ions (M and M') had to show a minimum abundance of 5000 counts, and a maximum m/z deviation of 1.5 ppm (intra-scan) between M and M' was allowed with respect to the detected number of ^{13}C atoms. The isotopologs of both the native and labeled forms (M+1 and M'-1) were allowed to deviate up to 25% from their theoretical ratios for the detected number of ^{13}C atoms and no intensity threshold check was applied for these signals. TracExtract searched for six, seven, eight, and nine ^{13}C -atoms. Additionally, it also searched for 12, 14, 16, and 18 ^{13}C -atoms to be able to detect metabolites consisting of two ^{13}C -Phe units. Each such detected pair of signals (M and M') was reported as MS signal pairs.

- Binning of MS Signal Pairs

Detected MS signal pairs were subsequently binned using hierarchical clustering (HC, Euclidean distance) using the m/z values of the native, monoisotopic isotopolog MS signals. After calculating the dendrogram, only those subclusters remained, which had less than 8 ppm m/z deviation between the MS signal pair with the highest and lowest m/z value.

- Detection of Chromatographic Peaks

Next, each subcluster from the previous binning step was used to calculate the EIC (± 5 ppm) of the native, monoisotopic isotopologs as well as the EIC of the ^{13}C -labeled ions. The two EICs were then inspected separately for chromatographic peaks. Only those chromatographic peaks that were present in both EICs at approximately the same retention time (± 10 scans) were considered to be derived from native and ^{13}C -labeled ions of the same metabolite. Next, the chromatographic peak shapes of such chromatographic peak pairs were compared using the Pearson correlation coefficient. Any putative peak pair not having a sufficient correlation (min. correlation: 0.85) was discarded, while those with a high correlation were reported as feature pairs of native and ^{13}C -labeled ion species and thus as ion species of Phe-derived metabolite ions.

- Convolution of Feature Pairs into Metabolite Groups

Detected feature pairs were then convoluted into feature groups again using the Pearson correlation coefficient. However, for this step not the chromatographic peak shapes of the native and the ^{13}C -labeled ions were compared, but rather the peak shapes of the chromatographic peaks of the native isotopologs of different feature pairs were compared. If their correlation was higher than 0.85, these feature pairs were said to be derived from the same metabolite and hence put into one feature group, each representing a single Phe-derived metabolite in the sample.

- Mock Samples Processing

To check for false positives the mock sample (water instead of ^{13}C -Phe) was processed in the same manner as the ^{13}C -Phe sample was processed.

- Generated Results

After data processing with TracExtract, a total of 151 feature pairs corresponding to 75 metabolites were detected in the ^{13}C -Phe sample as Phe-derived metabolites. In the mock sample, which was spiked with water instead of ^{13}C -Phe, eight feature pairs were detected. These eight feature pairs are false positives.

2.2.2. Data Processing with AllExtract

The AllExtract module of MetExtract was utilized to search for wheat metabolites using native and uniformly ^{13}C -labeled metabolite isotopologs [4,5]. Thus, any compound present in the LC-HRMS sample of the wheat ear was reported with this data processing strategy.

The search was performed as described earlier for the ^{13}C -labeled Phe-derived metabolites. However, as in these samples all metabolites were present as native and uniformly ^{13}C -labeled isotopologs, the number of carbon atoms was not restricted and AllExtract searched for a minimum of three and a maximum of 60 carbon atoms.

2.2.3. Combination of Phe-Derived Metabolites and Wheat-Metabolite Results

The 151 feature pairs detected with TracExtract (Phe-derived metabolites) and the 1482 feature pairs detected with AllExtract (all wheat metabolites) were merged into one reference feature list. The comparison of the two lists was done with a custom Python script, which compared the m/z values of the monoisotopic isotopologs as well as the retention time of putatively matched feature pairs. A maximum m/z deviation of ± 5 ppm and a maximum retention time deviation of ± 15 s were allowed for this matching.

2.3. T-2 toxin/HT-2 Toxin-Derived Metabolites

Reference and target datasets are based on an experiment performed by Meng-Reiterer et al. [6].

2.3.1. Reference Dataset:

The following reference files were used: one barley sample treated with native and ^{13}C -labeled T-2 toxin and one treated with native and ^{13}C -labeled HT-2 toxin. Both were measured on an LC-Q-TOF-system in positive full-scan mode with gradient method 2 and mass spectrometric settings according to Meng-Reiterer et al. [6]. All features present in the chromatogram were searched for with XCMS version 1.40.0 [7] and annotated with CAMERA version 1.20.0 [8]. Processing parameters were: peakwidth 10–45 s, ppm 5, prefilter c (5, 10,000), mzdifff 0.025, noise 2500, snthr 10, maxcharge 2. Using this approach, 20,682 features were detected and subsequently convoluted into 6764 metabolites. However, other than the TracExtract or AllExtract data evaluation strategies used for the Phe-derived metabolites, this list also includes different isotopologs and is therefore longer.

T-2 toxin/HT-2 toxin-derived detoxification products reported by Meng-Reiterer et al. [6] were matched to the list of barley metabolites using a maximum ppm deviation of five and a maximum retention time shift of 20 s. To match the peaks, the adducts $[\text{M} + \text{H}]^+$, $[\text{M} + \text{NH}_4]^+$, $[\text{M} + \text{Na}]^+$ and $[\text{M} + \text{K}]^+$ were calculated.

2.3.2. Target Dataset:

The target dataset consisted of three measurements of one sample from the same dataset, which were performed on the same instrument two years later. Degradation of certain metabolites was therefore expected.

2.3.3. MetMatch Processing Parameters:

The following parameters were used to match the target dataset to the reference feature list:

Input Parameters

- Retention Time: rt_min
- Mass/Charge: mz
- Ion ID: Num
- Metabolite ID: pcgroup
- Rest: (they were left blank)

General Parameters

- Start: 0
- End: 25
- RT Tolerance (Parsing Parameters): 0.75

- M/Z Tolerance (Parsing Parameters): 70
- M/Z Tolerance (Slice Parameters): 40
- Minimum Signals: 5
- Minimum Consecutive Signals: 3
- Intensity Cutoff: 0
- Peak Picking: Savitzky-Golay Filter
- NU Threshold: 3
- RT Tolerance (Rt Shift): 0.14
- Number of Bins: 100

Adducts

- No additional adducts were generated

Rt Shift Calculator Parameters

- First iteration: xRange: 65, yRange: 50
- Second iteration: xRange: 6.5, yRange: 20
- Third iteration (only for reference files): xRange: 3, yRange: 10

3. Evaluation of MetMatch

3.1. Comparison of MetMatch with XCMS (Phe Dataset)

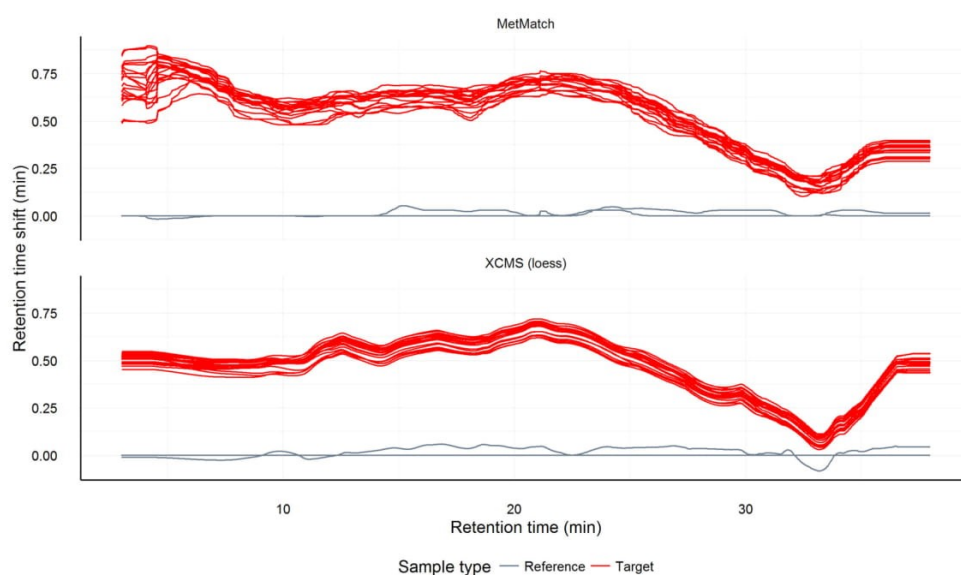


Figure S2. Comparison of the Rt shift detected in the Phe-derived biological experiment using MetMatch and XCMS relative to the selected reference chromatogram. Both tools calculated a similar Rt shift and thus a comparable chromatographic alignment for the two datasets relative to the selected reference chromatograms. The Pearson correlation coefficient was calculated pairwise for corresponding target chromatograms to compare the retention time shift functions calculated by MetMatch and XCMS. The correlation is on average 0.858 with a standard deviation of 0.042 for the retention time interval of 3 to 38 min. Between the interval of 7 to 38 min, which corresponds to the time interval in that most features of the reference feature list were detected, the average correlation for the target chromatograms was 0.926 with a standard deviation of 0.011. For the processing parameters of MetMatch see Supplementary Materials Section 2.1. Processing parameters for the XCMS peak picking were: findPeaks centwave, peakwidth 10–45 s, ppm 5, prefilter $5 \times 10,000$, noise 2500. Peaks before 3 and after 37.5 min were removed. XCMS group and Rt alignment parameters were: group density, bw 30, retcor loess, group density, bw 10.

3.2. Evaluation of Generation and Matching of Extended Reference Feature List

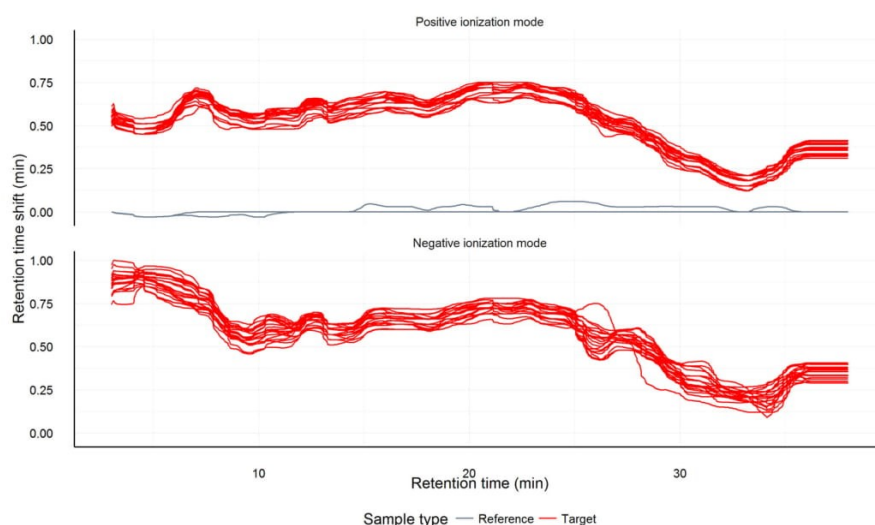


Figure S3. Comparison of the Rt shift detected in the Phe-derived biological experiment using chromatograms obtained in positive and negative ionization mode for the same samples. Reference files (grey) were obtained in positive ionization mode only. All target samples (red) of both ionization modes have been matched to the feature list exclusively containing ions detected in the positive ionization mode. The Rt shift functions of the negative and positive ionization mode chromatograms agree, except for retention times earlier than 7 min. This can be explained by the low number of reference features and chromatographic peaks in this region. One negative ionization mode chromatogram showed irregular behavior between Rt 15 to 35 min, which is explained by a reduced number of signals in this region of that particular chromatogram. The Pearson correlation coefficient was calculated pairwise for corresponding samples measured in the positive and negative ionization modes to compare the calculated retention time shifts. The correlation is on average 0.826 with a standard deviation of 0.038 for the retention time interval of 3 to 38 min. Between the interval of 7 to 38 min, which corresponds to that time interval in that most features of the reference feature list were detected, the average correlation for the target chromatograms was 0.977 with a standard deviation of 0.008.

3.3. Results of Manual Curation of Matched Phe-Derived Metabolites

Table S2. Manual validation of the results obtained by matching the target dataset to the reference feature list containing Phe-derived metabolites. Twenty-three matched features ranging from early to late Rt as well as low and high m/z values have been checked in 10 files (Remus, CM, 96h) using Thermo Xcalibur Software, version 2.2. Each feature is characterized by its unique identification number (Ion ID), the identification number its feature group (Group ID), its Rt and m/z values. Increasing the maximum allowed Rt shift tolerance would have resulted in successful matches of peaks 116 and 125 in all files. However, this could have caused an increased number of false positives.

Ion ID	Group ID	Rt	m/z	Correctly matched	Manual Validation Comment
23	19	5.48	105.0698	No	Peaks not found (too little features for such low m/z or Rt for correction function)
106	107	12.72	373.1129	Yes	-
108	107	12.72	390.1395	Yes	-
116	107	12.75	179.0701	Some	Not all peaks found because slightly out of maximum allowed Rt shift tolerance

125	107	12.74	197.0807	Some	Not all peaks found because slightly out of maximum allowed Rt shift tolerance
602	107	12.72	373.1128	Yes	-
1904	260	21.08	427.1384	Yes	-
1909	260	21.08	445.1492	Yes	-
1913	260	21.04	467.1309	Yes	-
1986	493	35.19	360.3258	Yes	-
2283	275	21.93	497.1441	Yes	-
2292	236	19.67	515.1544	Yes	-
3041	236	19.68	843.2318	Yes	-
3131	267	21.41	880.2867	Yes	-
3132	267	21.44	885.2420	Yes	-
10002	10002	4.15	120.0806	Some	Not all peaks found because it is a shoulder peak
10010	10010	7.78	207.1035	Yes	-
10011	10011	7.84	519.1971	Yes	-
10025	10025	13.23	274.0684	Yes	-
10030	10030	13.54	596.1689	Yes	-
10035	10035	16.50	640.1955	Yes	-
10059	10059	29.75	393.2609	Yes	-
10064	10064	34.88	834.5245	Yes	-

3.4. Comparison of MetMatch with XCMS (T-2 Toxin/HT-2 Toxin Dataset)

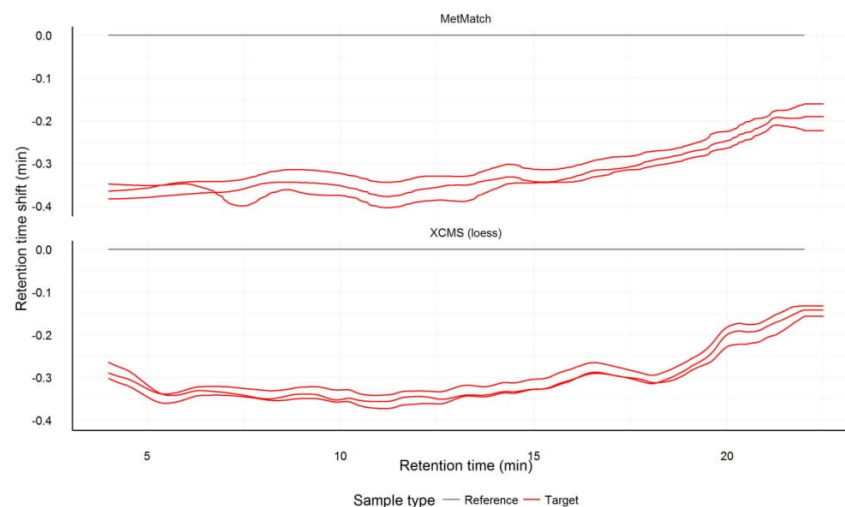


Figure S4. Comparison of the Rt shift detected in the T-2 toxin/HT-2 toxin-derived biological experiment using MetMatch and XCMS relative to the selected reference chromatograms. Both tools calculated a similar Rt shift and thus a comparable chromatographic alignment for the two datasets relative to the selected reference chromatograms. The Pearson correlation coefficient was calculated pairwise for corresponding target chromatograms to compare the retention time shift functions calculated by MetMatch and XCMS. The correlation is on average 0.966 with a standard deviation of 0.013 for the retention time interval of 5 to 22.5 min. For the processing parameters of MetMatch see Supplementary Materials Section 2.3. Processing parameters for the XCMS peak picking were: findPeaks centwave, peakwidth 10–45 s, ppm 5, prefilter $5 \times 10,000$, mzdiff 0.025, noise 2500. Peaks before 2 and after 22 min were removed. XCMS group and Rt alignment parameters were: group density, bw 30, retcor loess, group density, bw 10.

References

1. Du, P.; Kibbe, W.A.; Lin, S.M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **2006**, *22*, 2059–2065.
2. Savitzky, A.; Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639.
3. Kluger, B.; Bueschl, C.; Neumann, N.; Stuckler, R.; Doppler, M.; Chassy, A.W.; Waterhouse, A.L.; Rechthaler, J.; Kamplleitner, N.; Thallinger, G.G.; et al. Untargeted profiling of tracer-derived metabolites using stable isotopic labeling and fast polarity-switching LC-ESI-HRMS. *Anal. Chem.* **2014**, *86*, 11533–11537.
4. Bueschl, C.; Kluger, B.; Berthiller, F.; Lirk, G.; Winkler, S.; Krska, R.; Schuhmacher, R. Metextract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. *Bioinformatics* **2012**, *28*, 736–738.
5. Bueschl, C.; Kluger, B.; Lemmens, M.; Adam, G.; Wiesenberger, G.; Maschietto, V.; Marocco, A.; Strauss, J.; Bodi, S.; Thallinger, G.G.; et al. A novel stable isotope labelling assisted workflow for improved untargeted lc-hrms based metabolomics research. *Metabolomics* **2014**, *10*, 754–769.
6. Meng-Reiterer, J.; Varga, E.; Nathanail, A.V.; Bueschl, C.; Rechthaler, J.; McCormick, S.P.; Michlmayr, H.; Malachova, A.; Fruhmann, P.; Adam, G.; et al. Tracing the metabolism of ht-2 toxin and t-2 toxin in barley by isotope-assisted untargeted screening and quantitative lc-hrms analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 8019–8033.
7. 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.
8. Kuhl, C.; Tautenhahn, R.; Bottcher, 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.



© 2016 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Chapter 3

MetMatch's User Interface

The original publication only gave a brief overview of the user interface, which is featured in more detail in this section. The intuitive user interface enables researchers to match datasets without the use of console commands. **Figure T7** shows the main window. The step indication bar on top helps the user to keep track of the state of data processing and also serves as a menu bar. The current step is highlighted in bright green and any previous step can be repeated or adapted (i.e., through the change of parameters) by clicking on it. The left pane lists all loaded data files, clustered into batches by the user. The reference feature list (optionally including alternative ion species) can be found on the right.

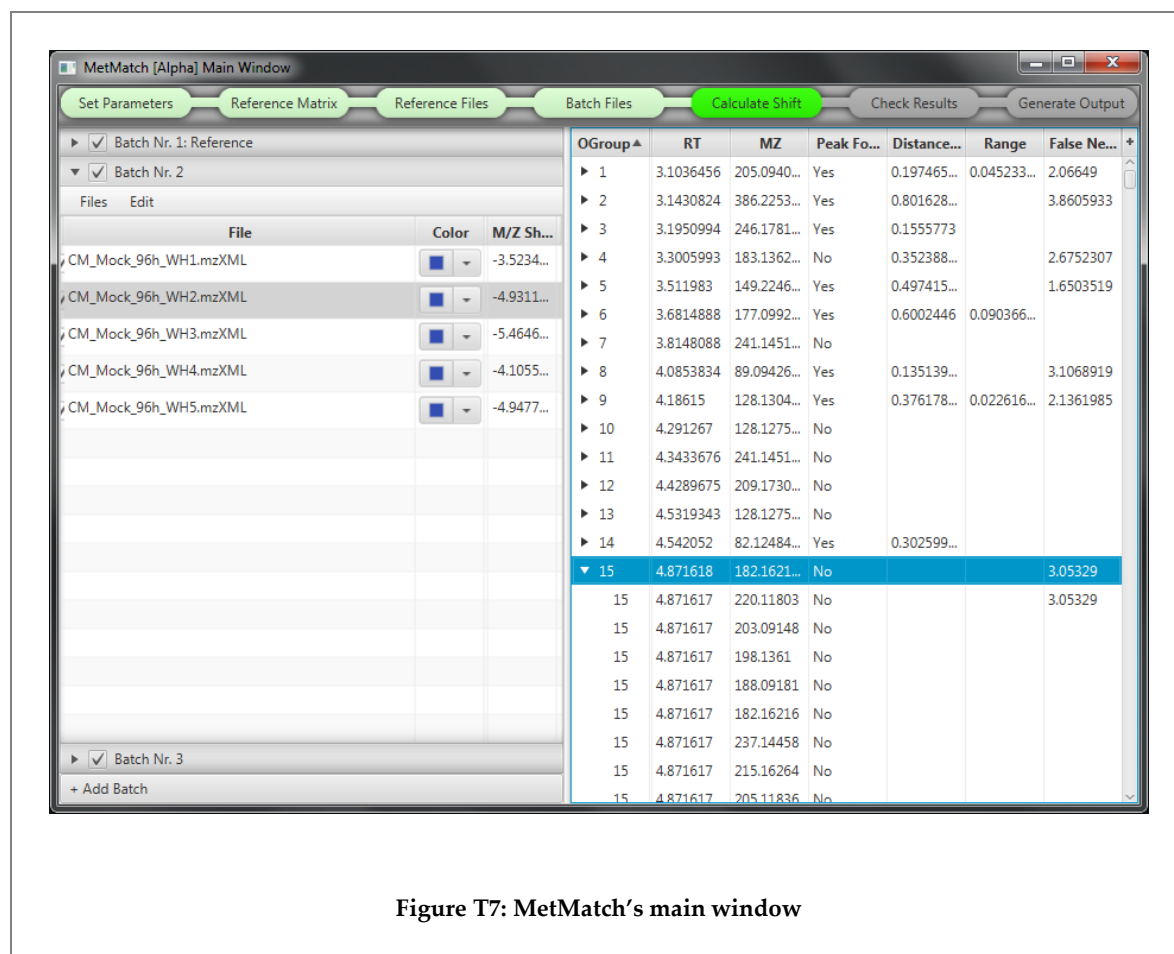
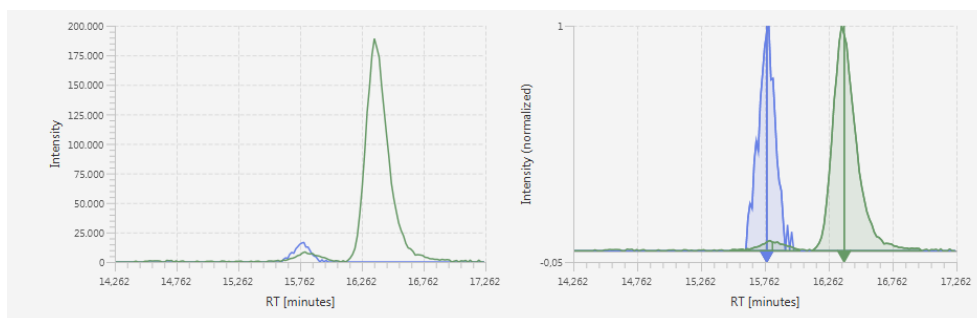


Figure T7: MetMatch's main window

Double clicking any adduct opens up a visualization of the EIC, with different chart types such as an intensity normalized EIC (**Figure T8 a**, right chart). The EICs can be visually aligned after Rt shift detection (**Figure T8 b**), allowing for easy comparison of peak shapes.

(a)



(b)

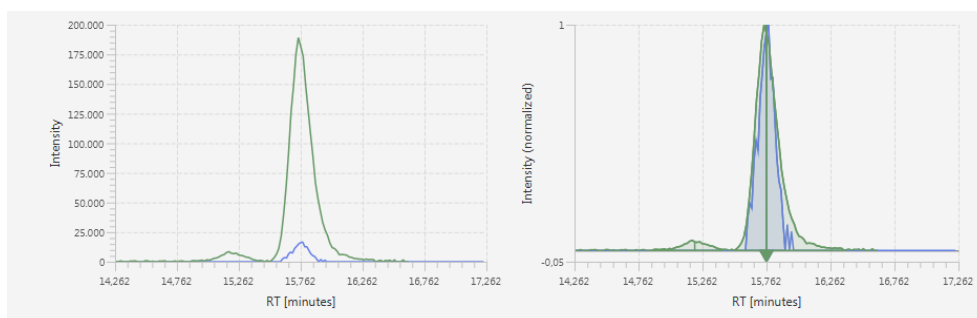


Figure T8: EICs of an inspected ion adduct before (a) and after Rt alignment (b)

The Rt shift detection step itself is completely visualized (**Figure T9**), which helps the user to understand the underlying algorithm. All parameters can easily be adjusted and tweaked and their impact is immediately illustrated by the animation of the algorithm (**Figure T10**). This allows even novice users to see the effects of the different parameters upon the achieved Rt-alignment.

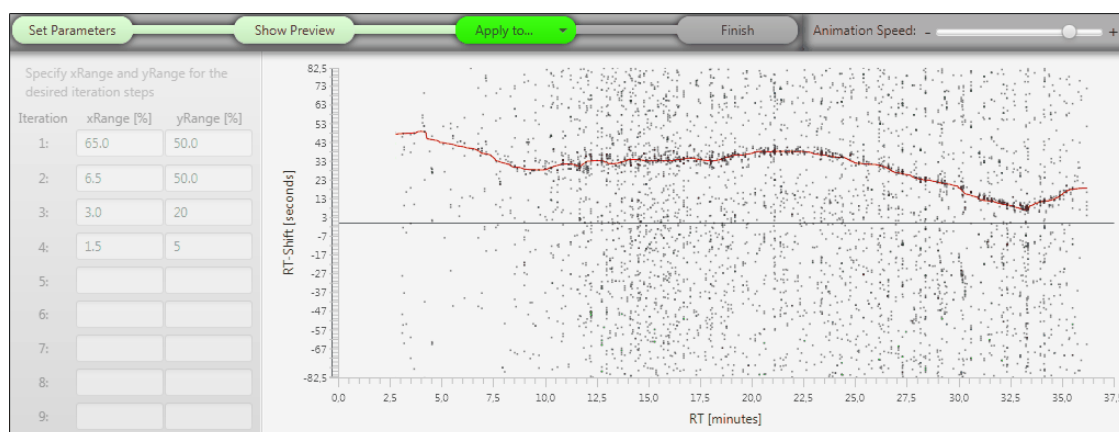


Figure T9: Animated Rt shift detection

An animated version is available at the homepage of the metabolomics department of the IFA Tulln at <https://metabolomics-ifa.boku.ac.at/MetMatch> (Accessed: December 2018).

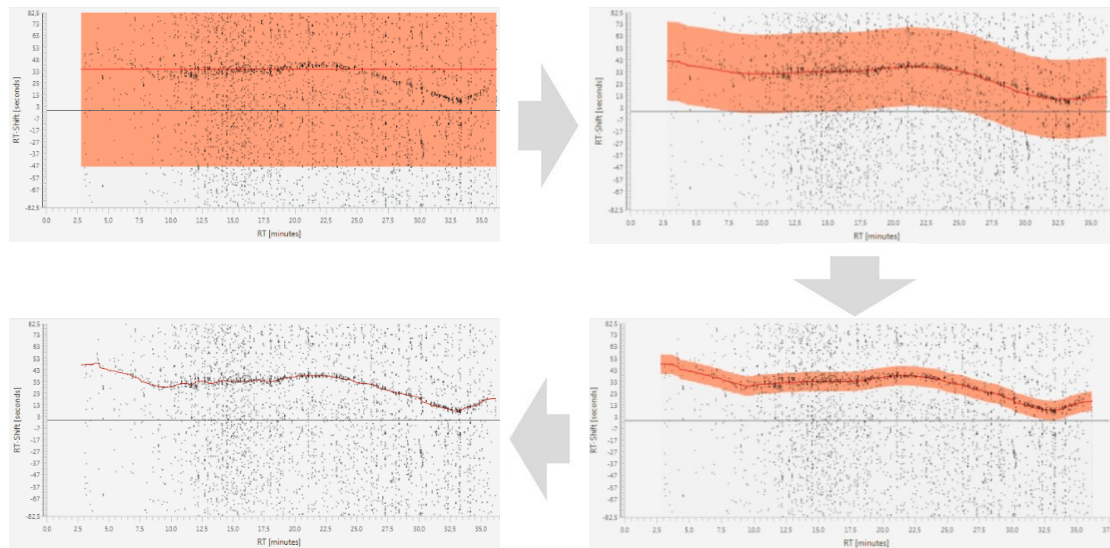


Figure T10: Individual iterations of the animated Rt shift detection algorithm

Apart from visual feedback, the user interface facilitates the manual verification and correction of the processing results. This is exemplified in **Figure T11**. In this case, the user matched two datasets consisting of multiple target (green) and reference files (blue). The Rt detection step uncovered a relatively large Rt shift of one of the reference files (**Figure T11 a**), which is unusual since the reference files are expected to show little to no shift. The user is able to inspect an adduct with a specific Rt value simply by double clicking the calculated Rt shift value of the reference file in question. The unexpected deviation in Rt shift is then easily confirmed by the EIC. The peak of the selected reference file is automatically highlighted in red and is indeed shifted relative to the other reference files (**Figure T11 b**).

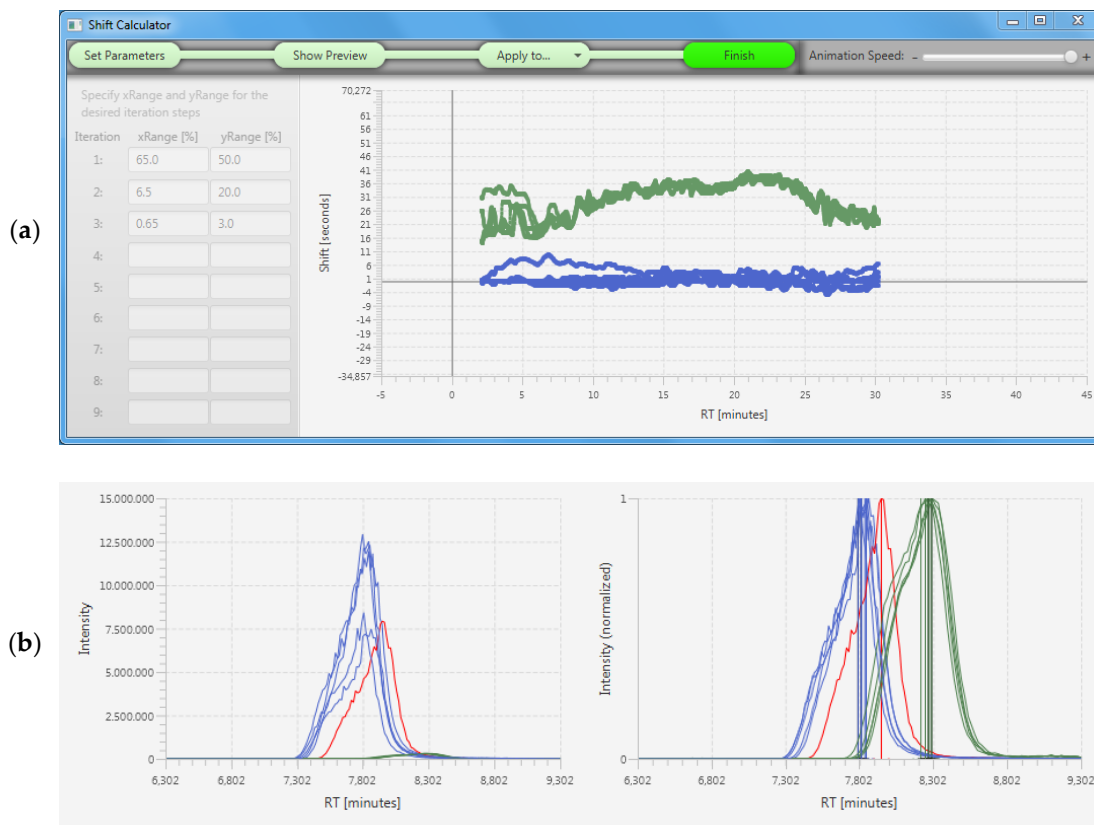


Figure T11: Exemplification of manual verification within the user interface

(a) MetMatch detected an unexpectedly large Rt shift for one of the reference files (blue) in the range of two to ten minutes.

(b) this shift is easily verified when inspecting one of the adducts. One reference file (red) is indeed shifted relative to the other reference files (blue).

Chapter 4

Conclusion

Retrospective alignment of data generated in different LC-HRMS experiments can enable researchers to investigate aspects outside the scope of their original evaluation, maximizing the scientific value of these types of experiments. For example, previously unknown metabolites that have only recently been identified as biologically relevant can easily be mapped from new to old experimental data thereby enabling their investigation in the old experiments with less effort.

MetMatch facilitates this task by correcting for linear and non-linear R_t shifts as well as constant shifts in m/z during the alignment of LC-HRMS data. Compared to other existing tools, MetMatch has three main advantages: the convenient user interface, the inclusion of alternate ion species and the fast processing speed. Apart from that, MetMatch's ability to work with pre-calculated lists of features/metabolites is a great asset when trying to identify them in other experiments.

MetMatch has already been used in the following biological studies/publications:

- Simader et al. (2019) *Fusarium graminearum* metabolites associated with the infection of wheat plants (manuscript in preparation).
- Doppler et al. (2019) Phenylalanine-derived wheat metabolites as defense compounds against *Fusarium graminearum* and its main virulence factor deoxynivalenol (manuscript in preparation).

Both studies / publications have reference datasets that were measured on an Orbitrap QExactive HF instrument and evaluated in order to find either *Fusarium graminearum*-related metabolites or polyphenol-related metabolites of wheat. These reference metabolites were then mapped back onto large metabolomics experiments consisting of a fine-grained time-course as well as different pathogen-treatments measured years earlier on an older instrument (Orbitrap LTQ XL). Although the analytical setup was almost the same, slight differences of measurement parameters such as the use of a new chromatographic column, different solvent batch and different HPLC system, led to changes in ionization efficiency, formed metabolite adducts and in-source fragments, as well as a slight shift in retention time. These changes were automatically detected and corrected using the developed software tool MetMatch. The author of this thesis hopes the aforementioned publications mark only the beginning of comparative LC-HRMS metabolomics research being enabled and facilitated by MetMatch.

Abbreviations and Terminology

Binning	A range (e.g., m/z or R_t shift values) is divided into a series of equally long intervals (bins).
Dataset	A list of LC-HRMS datafiles belonging to one analytical batch, originating from the same experiment.
EIC	Extracted ion chromatogram. An EIC represents the abundance/intensity of a certain m/z value in each mass spectrum of a chromatogram. The mass spectra are sorted according to their acquisition time relative to injecting the sample into the HPLC instrument. Usually a small deviation from the selected m/z value ($\pm$ ppm; depending on the used HRMS instrument) is allowed to account for minor scan-to-scan m/z inconsistencies observed during acquisition of a chromatogram.
Feature	A feature represents a particular metabolite ion of interest. It is characterized by its m/z value (experimental and theoretical) and retention time (as calculated by the applied peak-picking algorithm). Features are usually the result of an untargeted metabolomics data processing approach and typically comprise multiple chromatographic peaks, one per LC-HRMS chromatogram.
Feature group	All features originating from the same chemical substance and hence sharing the same R_t and peak shape.
F-measure	Harmonic mean of precision and recall.
Fn	False negatives. Features that were not matched.
Fp	False positives. If a matched feature does not meet the criteria to be considered as Tp, it is an Fp.
Isotopolog	Molecules of a metabolite with uniform isotopic composition.
LC-HRMS	Liquid chromatography coupled to high-resolution mass spectrometry.
m/z	Mass-to-charge ratio
Peak/Chromatographic peak	A chromatographic peak representing a particular feature in a chromatogram. Other than a feature, a peak has a defined m/z value, a retention time (as defined by the applied peak-picking algorithm) as well as a peak area. Peaks are detected by a peak-picking algorithm.
Peak apex	The time at which the most abundant signal of a respective chromatographic peak is detected after injection of the sample into the HPLC instrument.
Peak center	The time at which a chromatographic peak has its center of gravity. (Time) measurement is started with injecting the sample into the HPLC instrument.
Precision	Percentage of matched reference features that were correctly matched. Calculation: Number of Tp divided by the number of Tp plus Fp.
Recall	Percentage of all reference features that were correctly matched. Calculation: Number of Tp divided by the number of all reference features.
Reference dataset	A single or multiple LC-HRMS datafiles, which are used to generate the reference feature list.
Reference feature list	List of features to which target datasets are matched.
R_t	Retention time

Rt shift function	A function that holds the calculated retention time shift between a target chromatogram relative to the reference feature list. It is used to map the retention times of detected chromatographic peaks in the target chromatogram to the retention time of the features in the reference list.
Search frame	Signals of a chromatogram (single datafile) that fall within a certain Rt and m/z tolerance window around a reference feature. These values have to be slightly larger than the maximum expected shift. (e.g., 1.5 min Rt shift and 10 parts per million (ppm) m/z shift for a typical LC-HRMS chromatogram).
Rt shift array	An array containing the Rt shift of all peaks detected within a search frame relative to the retention time of the respective reference feature.
Rt shift matrix	The individual Rt shift arrays of all reference feature groups are concatenated column-wise and sorted by the retention time of their reference feature group. Each column represents a feature group, while rows represent discretized Rt shift values.
Target dataset	A list of LC-HRMS datafiles to be matched to a reference feature list.
Tp	True positives. Matched features are considered Tp if their m/z deviation is ≤ 5 ppm and their Rt deviation is ≤ 8.4 s (manually validated threshold value for the distinction of Tp and Fp in this datafile) relative to the reference feature after shift correction. The formation of different ion species is considered when comparing the m/z .

Bibliography

1. Haukaas, T., Euceda, L., Giskeødegård, G. & Bathen, T. Metabolic Portraits of Breast Cancer by HR MAS MR Spectroscopy of Intact Tissue Samples. *Metabolites* **7**, 18 (2017).
2. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* **26**, 51–78 (2007).
3. Beecher, C. W. W. The Human Metabolome. in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis* (eds. Harrigan, G. G. & Goodacre, R.) 311–319 (Springer US, 2003).
doi:10.1007/978-1-4615-0333-0_17
4. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research* **46**, D608–D617 (2018).
5. Weckwerth, W. METABOLOMICS IN SYSTEMS BIOLOGY. *Annual Review of Plant Biology* **54**, 669–689 (2003).
6. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy: Innovation. *Nature Reviews Molecular Cell Biology* **13**, 263–269 (2012).
7. Fiehn, O. Metabolomics – the link between genotypes and phenotypes. *An International Journal on Molecular Biology, Molecular Genetics and Biochemistry* 155
8. Giddings, G., Allison, G., Brooks, D. & Carter, A. Transgenic plants as factories for biopharmaceuticals. *Nature Biotechnology* **18**, 1151–1155 (2000).
9. Kuiper, H. A. *et al.* Safety assessment, detection and traceability, and societal aspects of genetically modified foods. European Network on Safety Assessment of Genetically Modified Food Crops (ENTRANS-FOOD). Concluding remarks. *Food Chem. Toxicol.* **42**, 1195–1202 (2004).
10. Creek, D. J. *et al.* Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* **10**, 350–353 (2014).
11. Snyder, L. R., Kirkland, J. J. & Dolan, J. W. *Introduction to modern liquid chromatography*. (Wiley, 2010).
12. Ardrey, R. E. *Liquid chromatography-mass spectrometry: an introduction*. (J. Wiley, 2003).
13. Zubarev, R. A. & Makarov, A. Orbitrap Mass Spectrometry. *Analytical Chemistry* **85**, 5288–5296 (2013).
14. Åberg, K. M., Alm, E. & Torgrip, R. J. O. The correspondence problem for metabonomics datasets. *Analytical and Bioanalytical Chemistry* **394**, 151–162 (2009).

15. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in Bioinformatics* **16**, 104–117 (2015).
16. Kassidas, A., MacGregor, J. F. & Taylor, P. A. Synchronization of batch trajectories using dynamic time warping. *AIChE Journal* **44**, 864–875 (1998).
17. Nielsen, N.-P. V., Carstensen, J. M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **805**, 17–35 (1998).
18. Eilers, P. H. C. Parametric Time Warping. *Analytical Chemistry* **76**, 404–411 (2004).
19. Suits, F., Lepre, J., Du, P., Bischoff, R. & Horvatovich, P. Two-Dimensional Method for Time Aligning Liquid Chromatography–Mass Spectrometry Data. *Analytical Chemistry* **80**, 3095–3104 (2008).
20. Prince, J. T. & Marcotte, E. M. Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping. *Analytical Chemistry* **78**, 6140–6152 (2006).
21. Zhang, X., Asara, J. M., Adamec, J., Ouzzani, M. & Elmagarmid, A. K. Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics* **21**, 4054–4059 (2005).
22. Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R. & Buydens, L. M. C. Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. *Analytica Chimica Acta* **781**, 14–32 (2013).
23. Lange, E. *et al.* A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics* **23**, i273–i281 (2007).
24. 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. *Analytical Chemistry* **78**, 779–787 (2006).
25. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry* **84**, 5035–5039 (2012).
26. Lommen, A. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* **81**, 3079–3086 (2009).
27. Lommen, A. & Kools, H. J. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics* **8**, 719–726 (2012).

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

Acknowledgements

Rainer and Christoph, I owe both of you more than you can imagine. Working alongside you was an invaluable inspiration and I will always think of you as role models.

I would also like to thank Maria and Jaqueline for the experimental data used to test MetMatch and a big thanks to Alex for the manual validation.

Last but not least, I have to thank Toni and everyone at Catalysts for all the great opportunities and the awesome work environment. Without you however, this thesis would have been done over two years ago. Just saying...