Universität Bielefeld

Technische Fakultät
Center for Biotechnology (CeBiTec)

MeltDB

A software platform for the analysis and integration of metabolomics experiment data.

Zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften an der Technischen Fakultät der Universität Bielefeld vorgelegte Dissertation

von

Heiko Neuweger

9. November 2009

Heiko Neuweger
Ellerstraße 33
33615 Bielefeld
hneuwege@cebitec.uni-bielefeld.de

Supervisors: Prof. Dr. Jens Stoye
Prof. Dr. Karsten Niehaus

Printed on acid-free paper.

Summary

The recent advances in metabolomics have created the potential to measure the levels of hundreds of metabolites which are the end products of cellular regulatory processes. The automation of the sample acquisition and subsequent analysis in high throughput instruments that are capable of measuring metabolites is posing a challenge on the necessary systematic storage and computational processing of the experimental datasets. Whereas a multitude of specialized software systems for individual instruments and preprocessing methods exists, there is clearly a need for a free and platform-independent system that allows the standardized and integrated storage and analysis of data obtained from metabolomics experiments. Both support for preprocessing of raw datasets and also means to visualize and integrate the results of higher level statistical analyses within a functional genomics context are required

To facilitate the systematic storage, analysis and integration of metabolomics experiments, MeltDB was designed, implemented, and applied. MeltDB is a web-based software platform for the analysis and annotation of datasets from metabolomics experiments. The software supports open and standardized file formats (netCDF, mzXML, mzDATA) and facilitates the integration and evaluation of existing preprocessing methods. The system provides researchers with means to consistently describe and store their experimental datasets. Comprehensive analysis and visualization features of metabolomics datasets are offered to the community through a web-based user interface. The newly developed system covers the process from raw data management to the visualization of results in a knowledge-based background and is integrated into the existing software platforms of genomics, proteomics, and transcriptomics at Bielefeld University.

This work demonstrates the functionality of MeltDB by means of several application examples where e.g. the influence of three different carbon sources on the gram-negative bacterium *Xanthomonas campestris* pv. campestris or the differences between healthy and disease human blood plasma samples are dissected. Novel visualization and analysis methods based on the MeltDB API have been developed and evaluated in the context of the extensible software platform.

Contents

List of Figures

List of Tables

CHAPTER 1

Motivation and Overview

The first completely sequenced genome, namely that of the pathogenic bacterium *Haemophilus influenzae*, was published in 1995. This represented a milestone in genome research and molecular biology. Less than two decades later, more than 1100 completely sequenced organisms are available (Genomes Online, October 2009) in the public databases and currently the number of sequences entered in the public repositories doubles in less than 12 months. The genome sequence delivers the blueprint for all cellular functions and the set of encoded genes defines the functional space for an organism. Yet, the number of genes with hypothetical or even unknown function is still growing with the number of sequenced genomes. This resulted in a paradigm shift in biology towards the so called functional genomics approaches that aim at elucidating the function and interaction partners of the individual genes.

Together with advances in bioinformatics, a number of high-throughput experimental techniques have been developed that enable the analysis of a large number of compounds in living cells. Among them are DNA arrays for the analysis of the mRNAs, 2D-gel electrophoresis together with mass spectrometry for the analysis of a large fraction of the proteins and yeast-two hybrid approaches for the identification of protein-protein interactions. These techniques are often referred to as *Omics* techniques and the different analytical approaches are described by the terms transcriptomics, proteomics and interactomics. The term "-ome" is used similar to "-omics" to describe all components of a given group of compounds or interactions.

One of the more recent contributions to the *Omics* family is the field of *metabolomics*, which addresses the analysis of the *metabolome*. The term metabolome has been introduced approximately one decade ago. It refers to all

low molecular mass compounds modified and synthesized by a living cell, tissue or organism. During the last couple of years, the field emerged in the biological sciences and achieved tremendous popularity and development. Metabolomics revives the classical biochemical concepts and analytical techniques and puts them in context with genomic information and other system wide approaches. The field is unique in modern science due to its multidisciplinary requirements: knowledge from biology, chemistry, engineering, physics as well as mathematics and statistics needs to be integrated. Furthermore, metabolomics finally allows to connect the different levels of biological information on the molecular level. As a high-throughput technology, metabolomics puts high demands on the analytical and computational analysis and requires streamlined and robust processing methods. To cover and address these requirements, the MeltDB software has been developed and will be presented in this work.

The following chapters describe the design, development, implementation, and application of MeltDB. Chapter 2 begins with a brief overview of the biological foundations of metabolomics. The relation to the other *Omics* techniques is detailed and the reader is introduced to the fundamentals of hyphenated mass spectrometry based metabolomics technology. Possible problems and pit-falls when working with hyphenated mass spectrometry data are highlighted and discussed; some methods to overcome these issues, including normalization and statistical inference will be presented that allow to extract information from noisy data.

Chapter 3 is dedicated to progressive methods of standardization and software systems in the field of metabolomics. The first results of standardization efforts are presented which involve the content of communication, machine-readable formats and vocabularies, and, finally, pieces of software implementing them.

Chapter 4 deals with the more formal aspects of designing a system. It is almost infeasible to build a complex application as a monolithic piece of source-code. One should rather use a modular approach and decompose the whole system into smaller reusable and manageable portions communicating with each other via well defined interfaces.

After the initial design, the formally described components are realized in source code and combined to the working application. Chapter 5 describes the implementation of MeltDB, which software and programming languages were applied, and how their employment resulted in an extensible database system with a highly versatile interface. At the same time, novel analysis methods realized using the MeltDB API are detailed

The availability of the MeltDB system led directly to the application of the functionality in real-world projects. Recently, more than a dozen projects were initialized. Some are internal evaluation projects, dealing with assessment of new methods of statistics and visualization. Others are tutorial projects used for teaching and in laboratory courses. But the largest range of projects is dedicated to metabolomics research in a diverse range of organisms and environments. In how far MeltDB has contributed to academic research is detailed in Chapter 6 which describes several recent application examples.

A discussion of the system, its implementation, and new insights obtained is given in Chapter 7 and a conclusion together with an outlook on the perspectives of MeltDB is presented in Chapter 8.

CHAPTER 2

Background

The sequencing of complete genomes marked the starting point for a system wide understanding of living cells, as the genome represents the complete blueprint of genetic potential of an organism. However, this information is static and insufficient to describe which part of the genomic functionality is actually realized under certain environmental conditions. Methods to obtain snapshots of the actual state of the living cell on various levels have been developed in the last decade and the term *functional genomics* was coined to subsume these efforts. One important aim is the elucidation of the function of every gene. Gene expression analyses (transcriptomics) provide information about transcribed genes (Lockhart *et al.*, 1996) and modern microarray technology allows to analyze virtually all transcriptional changes of an organism or tissue. Thereby the effects of genetic and environmental manipulations can be investigated.

Protein analyses (proteomics) reflect protein abundances and their amounts in a cell (Shevchenko *et al.*, 1996). Advances in protein separation by two-dimensional gel-electrophoresis and subsequent identification using e.g. matrix assisted laser desorption coupled to one- or two-dimensional mass spectrometry allow to determine the abundance of proteins by analyzing up to several thousand proteins in parallel.

The combination of transcriptomics and proteomics already provides a large amount of functional information. Nonetheless, several recent studies revealed the limitations of transcriptomic and proteomic approaches to predict gene functions.

In accordance with the terms transcriptome and proteome, the complement of all metabolites of an organism was termed *metabolome* (Oliver *et al.*, 1998).

The qualitative and quantitative analysis of all metabolites (present in a cell, a tissue or an organism at a given time point) was termed metabolomics[1] (Sumner *et al.*, 2003).

In order to achieve a complete understanding of the biological behavior of a complex system, it is essential to monitor the response of an organism to a conditional perturbation at the transcriptome, proteome and metabolome levels (Bino *et al.*, 2004). An important step towards this goal is the integration of experimental data and results from all fields of functional genomics. For a truly integrative approach towards systems biology it is essential to include metabolic data (Fiehn, 2002; Sumner *et al.*, 2003; Weckwerth, 2003) especially because changes in metabolites are most closely connected to the observed phenotype of an organism. The backward link to the genotype of an organism can be established through knowledge about biochemical pathways and gene regulatory networks.

Nonetheless, there are several additional challenges to tackle in metabolome analysis compared to the analysis of the genome, transcriptome or proteome. In contrast to the linearly arranged sequences of four nucleotides in DNA and RNA or the 20 amino acids in proteins, metabolites belong to chemically diverse compound classes and vary in molecular composition. Thus, metabolites can not be subjected to a general sequence analysis, i.e. the determination of the order of the nucleotides or amino acids (Fiehn, 2002). The identification is complex since the elemental composition, the atomic configuration, as well as the stereochemic orientation need to be determined for each metabolite. Furthermore, as the concentration of individual compounds present in biological material can vary by as much as nine orders of magnitude (Sumner *et al.*, 2003), the dynamic range of most modern analytical systems is easily exceeded.

Despite these limitations and pitfalls, comprehensive analysis of metabolic responses has been made possible and useful with the advent of modern computational and analytical tools. The analytical procedure for metabolomics can be subdivided into three experimental sections. First, the biological samples have to be harvested without disturbance of the metabolome; the metabolites must be extracted and eventually derivatized. In a second step, the metabolites are separated by chromatographic techniques such as gas chromatography (GC) or high performance liquid chromatography (HPLC). In the third step, the identification and quantification of the metabolites is achieved. Hyphenated mass spectrometry is currently one of the most widely applied technologies in metabolomics, as it provides rapid, sensitive, and selective qualitative and quantitative analyses (Dunn and Ellis, 2005).

Estimations for the number of metabolites present in living organisms range from hundreds to many thousands; for the model organism *Escherichia coli* ≈ 750 metabolites have been ascertained (Nobeli *et al.*, 2003). With the application of modern hyphenated technologies, a significant part of the metabolome (Hall, 2006)

[1] The term metabonomics was proposed in parallel, but is used less often due to the inconsistency in terminology with transcriptomics and proteomics

can be analyzed. Yet, the extraction and analysis of entire metabolomes in a single step is impossible due to the chemical complexity and heterogeneity of compounds (Goodacre *et al.*, 2004). The number of estimated compounds for eukaryotic cells ranges from 4,000 to 20,000 (Fernie *et al.*, 2004) and in the entire plant kingdom up to 200,000 metabolites are expected to exist. Therefore, truly metabolomic studies will rely on new technical inventions (Hall, 2006).

At present, comprehensive metabolite profiles of plants and bacteria need to be acquired by using combinations of gas chromatography - mass spectrometry (GC-MS) (Fiehn *et al.*, 2000; Roessner *et al.*, 2000; Weckwerth *et al.*, 2004), liquid chromatography - mass spectrometry (LC-MS) (Buchholz *et al.*, 2002; Huhman and Sumner, 2002; Tolstikov and Fiehn, 2002), capillary electrophoresis - mass spectrometry (CE-MS) (Soga *et al.*, 2002, 2003; Nesbitt *et al.*, 2008) and also liquid chromatography - nuclear magnetic resonance (LC-NMR) (Wolfender *et al.*, 2003). The experimental and analytical effort for such extensive analyses is immense. Considering the tradeoff between effort and the number of detectable and identifiable compounds in biological samples, GC-MS is currently deemed to be the most popular method for global metabolite profiling (Kopka, 2006) which can be complemented by LC-MS analysis for incompatible compounds (Broeckling *et al.*, 2005).

Apart from the comprehensive analysis of all metabolites which requires an integrated analysis, several other approaches are employed in metabolomic research. These approaches can be classified in the following main categories according to Goodacre *et al.* (2004):

In *metabolite target analyses*, the number of metabolites is restricted to e.g. a particular enzyme system that would be directly affected by abiotic or biotic perturbation (Fiehn, 2001). The number of metabolites that need to be detected and quantified is low and analytical procedures can be optimized towards these.

The more general *metabolite profiling* approach focuses on a group of identifiable metabolites, for example, a class of compounds such as carbohydrates, amino acids or those associated with a specific pathway. Compared to target analyses, the quantification of the individual compounds may be less accurate but a broader analysis reduces the over-interpretation of data (Fiehn, 2001).

If the goal of the analysis is the classification of samples on the basis of provenance of either their biological relevance or origin, *metabolic fingerprinting* is employed (Kell *et al.*, 2005). Quantitative data for all biochemical pathways is not strictly necessary; it is often sufficient to obtain enough information to unravel metabolic alterations (Fiehn, 2001).

The analytical methodology and instrumentation necessary for these and the comprehensive metabolomics analyses will be detailed in the following.

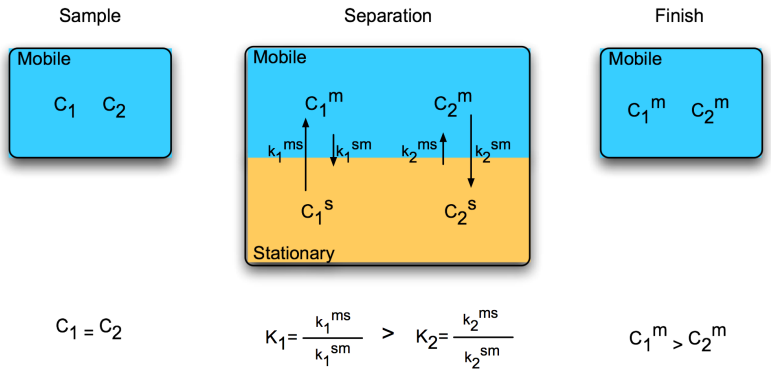

Figure 2.1.: The chromatographic separation used in metabolomics usually relies on the distribution between two phases. One phase is stationary while the second phase (gas or liquid) is mobile and can be exchanged. Differences in the distribution coefficients K_1 and K_2 allow to separate the compounds C_1 and C_2 (Figure adapted from Villas Bôas (2007)).

2.1. Separation and chromatography

Chromatography is a physical method of separation in which components that need to be separated are distributed between a stationary and a mobile phase that moves into a defined direction (Ettre, 1994). The first description of chromatographic separation dates back to the turn of the 20th century when M. S. Tswett used columns of powdered calcium carbonate to separate green leaf pigments into a series of colored bands (Poole, 2002). Tswett also coined the name chromatography (color writing) to describe this process. Since then, chromatography has evolved into a large number of applied methods. The main advances of the technique occurred between the 1960s and 1990s. Improvements of columns, detectors and electronics resulted in the possibility to separate nearly all types of chemical compounds even when they are found in complex mixtures (Villas Bôas, 2007).

2.1.1. Basics of chromatography

Small differences in the distribution coefficients and their temperature dependence are used to separate compounds in two-phase systems (e.g. liquid-liquid or gas-liquid). In these systems, one phase is termed the *stationary* phase which is chemically bound to a surface and fixed in a column. The second phase is termed *mobile* and can either be gaseous or liquid.

The separation process is dynamic, the differences in the distribution coefficients as described in Figure 2.1 determine how long a compound remains in the stationary phase. Since the mobile phase is continuously fed to the chromatographic system, the compounds are dynamically separated until the end of the column is reached. A plot of the concentration of compounds eluting at the end of the column vs. the time is called *chromatogram*. A simple chromatogram is shown in Figure 2.2

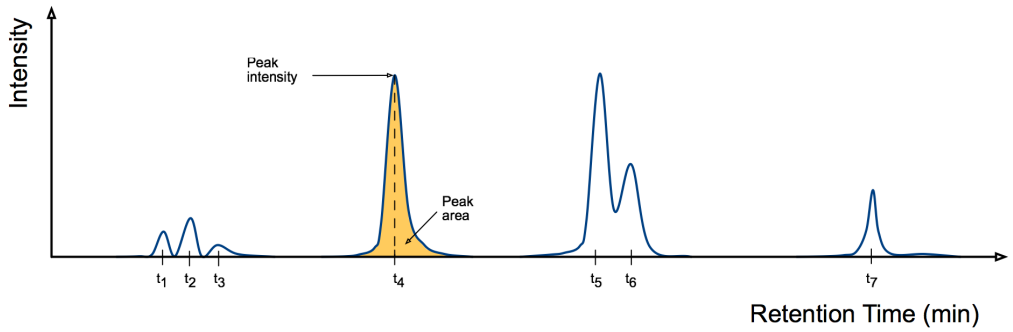

Figure 2.2.: The figure displays a typical chromatogram. The retention time of the analytes found is presented on the x-Axis of the diagram and the intensity measured by a detector is presented on the y-Axis. In the sample image, the retention time of seven peaks is marked. For each peak, the maximal intensity and the area can be determined. The peaks occurring at timepoints t_5 and t_6 are not completely resolved by the chromatographic separation.

illustrating the most important parameters *retention time*, *peak height*, and *peak area* by which the peak of an analyte can be described. The process of detecting and integrating peaks is addressed by computational signal analysis methods.

2.1.2. Chromatographic systems

The analysis of complex mixtures is a recurring requirement in biological experiments, and chromatography can achieve the important step of physical separation. Most important for the field of metabolomics which requires the separation of complex metabolite mixtures are high performance liquid chromatography (HPLC), gas chromatography (GC) and also capillary electrophoresis (CE). The general principles and theories of gas and liquid chromatography are quite similar as described in the previous section. Both analytical systems consist of a supply for the mobile phase, an injection system, the column, and a detector. The systems are controlled and data is recorded though an integrated computer system. The exemplary setup of a GC instrument is presented in Figure 2.3.

An important and critical part of a gas chromatograph is the injector unit which transfers the typically liquid sample into the gaseous phase and focuses it at the beginning of the column. The injector can easily evaporate small molecules with boiling points below 300°C but non-volatiles need special chemical modifications to be made amenable for GC analysis. In metabolomics, amino acids, sugars, small organic acids and other polar metabolites are of great interest. By protecting their polar groups through chemical modifications e.g. methylation or silylation, the volatility can be increased. A comprehensive overview of modern derivatization methods is given by Toyo'oka (2000)

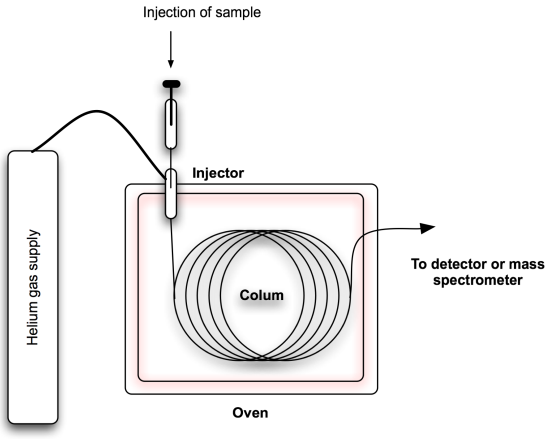

Figure 2.3.: The main components of a gas chromatograph are a gas supply coupled to an injector which transfers the sample into the mobile gas phase and a column located in an oven. The temperature program of the oven is controlled and the column is connected to a detector, e.g. a mass spectrometer (Figure adapted from Villas Bôas (2007)).

In order to unambiguously identify and quantify the compounds after their separation, both sensitive and selective detectors need to be employed. A further requirement for the detector is the compatibility with the separation technique which results in three main technologies. At first, there are so-called optical detectors that use ultra violet light (UV), photo diode arrays, or laser induced fluorescence (Fraser *et al.*, 2000; Britz-Mckibbin *et al.*, 2003). Alternatives are nuclear magnetic resonance (NMR) which allows for structural elucidation of the compounds but requires high sample volumes (Wolfender *et al.*, 2003) and finally mass spectrometry detectors which are deemed to be the most generic and comprehensive technology for metabolomics approaches (Bedair and Sumner, 2008).

2.2. Mass spectrometry

The development of *mass spectrometry* (MS) dates back to the year 1910 when J.J. Thomson showed that the noble gas Neon consists of a mixture of two isotopes of nominal mass 20 and 22. The basic principle of mass spectrometry is the generation of ions from organic or anorganic compounds followed by the separation of these ions by their mass-to-charge (m/z) ratio. The m/z ratio is dimensionless by definition, it calculates from the dimensionless *mass number*, m, of a given ion and the number of its elemental charges, z. If the number of elemental charges is one, the m/z scale directly matches the m scale. The associated value that is measured in mass spectrometry is the abundance of ions of a specific m/z ratio.

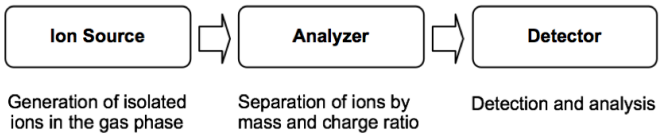

Figure 2.4.: The figure displays a block diagram of a mass spectrometer together with the function of the individual modules. Analytes are ionized in the ion source and are afterwards transferred to the high vacuum of the mass spectrometer.

2.2.1. Mass spectrometer

A mass spectrometer always consists of a common basic setup. The first module is called *ion source* followed by a *mass analyzer* and a *detector* operated under vacuum conditions as presented in Figure 2.4. The ion source realizes several key processes which are the transfer of the samples into the gas phase, ionization, and finally the transfer of the analyte to the vacuum. Compounds may be ionized thermally, by electromagnetic fields, or the impact of accelerated electrons or ions with the analyte. The commonly positive ions are single ionized atoms, clusters, molecules or their fragments or associates. The separation of the ions can be achieved by static or dynamic electromagnetic fields or by time-of-flight analyzers (Kienitz and Aulinger, 1968).

2.2.2. Ionization methods

Ionization is crucial to mass spectrometry since only ionized molecules can be transported and separated by electromagnetic fields in the vacuum that is required for mass detection. Different ionization methods will be briefly sketched according to (Kopka *et al.*, 2004) in the following subsections.

Electron ionization (EI)

In electron ionization (EI), the analyte of interest is bombarded in the vapor phase with high-energy electrons (usually 70 eV). Some of this energy is absorbed by the analyte molecules ($\approx$ 20 eV) causing a number of processes. The analyte may simply lose a single electron and is thereby ionized. This results in the so called molecular ion ($M^{+\bullet}$), the m/z of which corresponds to the molecular weight of the analyte. Typically, 10 eV of the energy is required for this process. Since bond energies in organic molecules are typically around 4-5 eV, the remaining excess energy may lead to fragmentation. Complex spatial rearrangements of the atoms in the molecule or simple scission of bonds may occur. This process may help to elucidate the nature of the analyte since the generated fragment ions yield direct information on the structure of the analyte. An example of the scissions that can

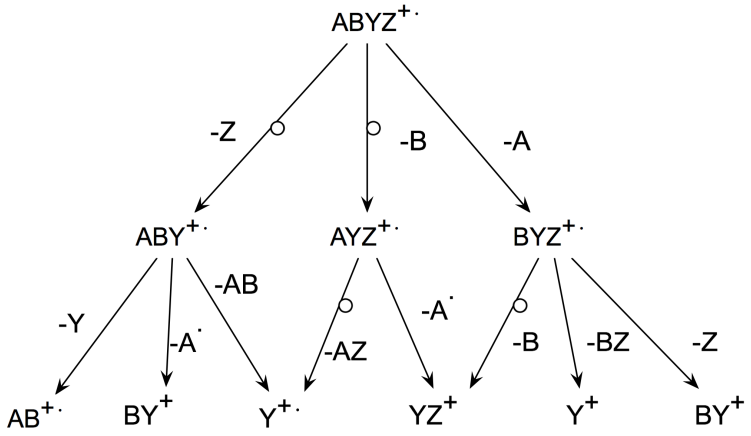

Figure 2.5.: A possible fragmentation pathway of an ionized molecule $ABYZ^{+\bullet}$. The excess energy of electron ionization typically leads to a number of rearrangement and fragmentation processes. In an EI mass spectrum of the molecule ABYZ, all the formed ionic species may be detected. Nonetheless, the intensity of the respective peaks can not be predicted by simple rules due to numerous competing and consecutive reactions. (Figure adapted from Gross (2004))

occur are presented in Figure 2.5 and for further consideration of the interpretation of EI spectra the textbook by McLafferty (1993) is recommended.

Electrospray ionization (ESI)

Electrospray ionization (ESI) is typically used in combination with liquid chromatography. The ionization process of the molecules that are dissolved in large amounts of solvent starts with the formation of small charged liquid droplets. These are sprayed into a vacuum. While the solvent evaporates, the charge is concentrated at the surface of the shrinking droplets. At the end of the process, the charge is transferred to the solutes. The electrostatic repulsion inside the small droplets leads to a discharge of the ions which can subsequently be analyzed by mass-spectrometric detectors. Molecules with a low ionization potential are preferentially ionized and if molecules compete for ionization, those molecules that do not readily form ions may be suppressed.

Chemical ionization (CI)

The *hard ionization* techniques such as EI lead to a fast fragmentation of the molecular ion. This is disadvantageous as the molecular ion is one of the most important pieces of analytical information as it can be used to derive potential sum formulas of the analyte. In order to reduce the fragmentation associated with ionization and thereby increase the production of molecular species, chemical ionization (CI)

has been developed. This approach is often termed as 'soft ionization' where the analyte molecules are subjected into a mass spectrometer which contains a reagent gas. Although the mixture is bombarded with electrons as described in the previous section, the ionization of gas molecules is predominant since the reagent gas is present in vast excess ($>$ 1,000:1). After ionization of the gas atoms, ion-molecule reactions take place between neutral analyte molecules and the reactant gas atoms in the high-pressure regime of the mass spectrometer source. Typically, adducts of reagent ions with the analyte molecules occur in low energy processes which result in little fragmentation.

The m/z of the ions observed does not give the molecular weight directly, it arises from the combination of the analyte with an adduct. The mass of the adduct (1 for methane gas or 18 in the case of ammonia) must in this case be subtracted from the observed m/z value.

To summarize, hard ionization techniques transfer more energy than is required for the first ionization step. All molecules are ionized but the remaining excess energy often causes strong molecular fragmentation. Soft ionization technologies, such as CI and ESI, transfer a smaller amount of energy. Molecules are therefore less likely to form fragments but not all compounds can be ionized using CI or ESI.

2.2.3. Mass-detection approaches

The determination of the m/z ratio can be achieved through a combination of electric and/or magnetic fields in the mass analyzer module of the mass spectrometer. All mass analyzers operate under high vacuum conditions to ensure that the generated ions do not collide with uncharged molecules (e.g. the air) or with each other. The main quantitative difference between mass analyzers is the mass resolution they achieve. Nominal mass analyzers obtain integer mass accuracy and feature a resolution of $\approx$ 1:1,000-2,000. High resolution mass analyzers can achieve resolutions of up to 1:100,000 which corresponds to a mass accuracy lower than 1 part-per-million (ppm).

Quadrupole detectors (QUAD)

A quadrupole detector (QUAD) is a mass-spectrometric device for the detection and quantification of ions. These are generated from molecules of interest as described above. In a quadrupole instrument, a set of four electronically operated metal rods acts as a mass-selective ion filter. The ions that pass this filter at each of the successively monitored masses are recorded and counted by a detector which finally generates a mass spectra. Triple quadrupole detectors are a variation of the basic single quadrupole setup. For these detectors, three quadrupole devices are coupled in a linear array which can realize two selective filtering steps. Thereby all unwanted components can be removed from a complex mixture and the target metabolite of interest can be extracted.

Ion-trap technology (TRAP)

Ion-traps (TRAP) can also be employed for the mass-spectrometric detection and quantification of ions. An ion-trap device collects and stores ions by forcing them into stable orbits. Afterwards, the ions are released from the device and analyzed. By trapping and release of ions of successive masses, a mass spectrum can be generated through this two-step process. An interesting feature of ion-trap instruments is the possibility to analyze secondary fragments originating from the collected and stored 'primary ions'. Researchers can employ this approach for determining the structure of molecules, or to monitor known metabolites in complex samples.

Time-of-flight technology (TOF)

TOF was initially developed for the analysis of macromolecules, such as proteins, peptides and polysaccharides, by matrix-assisted laser desorption (MALDI-TOF). Improvements to this technology allow a coupling to gas chromatographic separation instruments. TOF-MS is ideally suited for fast-scanning analysis of small volatile molecules (Kopka *et al.*, 2004) since a processing cycle that consists of bundling, accelerating and detection of the ions in an evacuated flight tube does only take a few milliseconds. The flight tube has a fixed distance and as ions of low mass travel faster than those with a high mass, they can be distinguished by their time of flight which is recorded at the end of the tube.

As mass spectrometry is a destructive method, the analytes are consumed during the examination. This is a drawback in comparison to non destructive methods such as infrared (IR) or nuclear magnetic resonance (NMR) spectroscopy but because of the sensitivity of MS this is rarely a problem. MS is even more considered as the method of choice when other analytical techniques fail to obtain clear analytical information from amounts of sample in the nanogram scale (Gross, 2004).

2.2.4. Mass spectrum

A *mass spectrum* (often abbreviated as MS) represents the two dimensional information on ion abundance in relation to m/z ratio. The intensity is derived from the area or simply the height of the measured signals (so called *peaks*). It is common to normalize all measured intensities relative to the most intense peak in the spectrum the so called *base peak* (rel. % Int.). Mass spectra can either be represented as discrete spectra or as profile spectra which resemble the peak shape. A sample mass spectrum of the compound valine is presented in Figure 2.6.

2.3. Hyphenated mass spectrometry

Most often the compounds of interest are found as part of a complex mixture. The chromatographic technique provides separation of the components of that mixture and therefore allows for their identification or quantitative determination. The

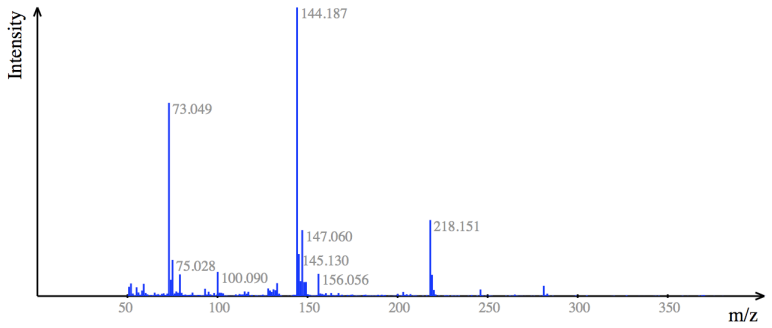

Figure 2.6.: A mass spectrum obtained through electron impact ionization (EI) of the compound valine. The mass-to-charge (m/z) ratio of the measured ions is presented on the x-axis whereas the y-axis represents the recorded number or intensity of the individual ions. Typically, mass spectra are normalized to the peak with the highest intensity (base peak). In this example, the base peak is found at 144.187 m/z.

information on the retention time of an unknown compound alone does not suffice to unequivocally identify the compound even if it matches the retention time of known standards within the limits of the experimental error. There exist simply so many different compounds that an analyst cannot say with absolute certainty that two compounds are the same based on retention time information alone. Additional information is needed and the connection to MS provides further insight.

The first approach was the coupling of *gas-chromatography to mass spectrometry* (GC-MS) which was first reported in 1958 and rapidly established as a routine method. The coupling of liquid chromatography to mass spectrometry devices was realized shortly after and was followed by further liquid phase separation methods. In any case, an additional dimension is added to the analytical measurement through the separation by one of the aforementioned techniques. For chromatographic techniques that are coupled to an MS instrument, the term *hyphenated methods* has been established. Chromatographic detectors deliver a chromatogram that represents the mass flow eluting from the chromatographic column.

If a mass spectrometer is employed as detector, repetitive scans over the m/z range of interest are measured during the chromatographic run and thereby the relation between chromatogram and mass spectra of the eluting components is established. The mass spectra of many compounds are sufficiently specific to allow their identification with high confidence. Problems may arise if the separation is incomplete and the analyte of interest is found as part of a mixture. The obtained mass spectrum will contain ions from all the compounds present in this case. Nonetheless, many compounds with similar or even identical retention times have quite different mass spectra and can therefore be differentiated.

2.3.1. Gas chromatography - mass spectrometry (GC-MS)

Gas chromatography-mass spectrometry (GC-MS) combines the features of gas chromatography and mass spectrometry to identify substances within a complex sample. A fused silica capillary is contained in the oven of a GC instrument. The inner surface of the long and fine capillaries is coated by a film of stationary phase. The selective binding of the analytes is facilitated by this stationary phase due to different physiochemical interactions (dispersion, dipole-dipole interactions and hydrogen bonding). According to the application, capillary columns are available in different dimensions with various stationary phases. A typical column is 30m long with an internal diameter of 0.25mm and a stationary phase 0.25µm thick.

Volatility and derivatization

Gas Chromatography requires a certain level of volatility and thermal robustness of the analyte since both the injection block and interface regions are at high temperatures even while the column oven is not. Compounds can be *derivatized* through silylation or acetylation (Halket and Zaikin, 2003). Thereby, the polarity of molecules decreases as e.g. XH groups are transformed in $XSiR_3$ groups. A reduced polarity leads to improved volatility even if the molecular weight of the analyte is increased. The thermal robustness is additionally increased, alcohols are for example protected from thermal dehydration.

Column bleed

The high temperatures of the GC column lead to the slow release of the stationary phase of the inner wall of the capillary. This effect is termed thermal degradation or *column bleed*. One of the characteristics is that it is proportional to the temperature in the GC oven. In case of the frequently employed methyl-phenyl-siloxane liquid phases, abundant ions are 73, 147, 207, 281 m/z. Within one series the peaks are 74 U ($OSiMe_2$) distant.

In the repetitive scanning mode in GC-MS each point of the TIC corresponds to a full mass spectrum. It is important for the analytical process that the time to acquire a mass spectrum is shorter than the time to elute a component from the column. When scan cycle times are in the order of one second the resulting chromatograms are usually represented by thousands of mass spectra. With recent GC/GC instruments the frequencies of hundreds of scans per second can easily be achieved. This leads to a proportional increase in the number of mass spectra achieved for individual experiments. Due to the high sensitivity, GC-MS can detect trace compounds and is therefore well suited for metabolomics approaches.

2.3.2. Liquid chromatography - mass spectrometry (LC-MS)

In GC-MS virtually all compounds that pass the chromatographic column can be ionized and therefore analyzed via the mass spectrometer. This is not the case for

high performance liquid chromatography (HPLC) and MS due to the incompatibilities of the two techniques (Ardrey, 2005). The main incompatibility is the mobile phase is a liquid being pumped through the column at a flow rate of typically 1 ml min^{-1}. Since the MS operates under vacuum conditions, the eluate can not be pumped directly in the source of the mass spectrometer and has to be removed beforehand. To achieve this efficiently, the electrospray and atmospheric pressure chemical ionization interfaces are most widely used currently.

The main advantage of LC-MS over GC-MS are that firstly, compounds do not need to be chemically altered prior to analysis. Secondly, compounds that are highly polar, thermo-unstable or have high molecular weights (e.g. oligosaccharides or lipids) can be quantified and separated (Nikolau, 2007). Additionally, a wide range of columns and elution procedures is available for the specific separation and detection of different compound classes. In conclusion, the combination of HPLC with mass spectrometry can help to identify and quantify additional primary and secondary metabolites and thereby complements the GC-MS approach.

2.3.3. Capillary electrophoresis - mass spectrometry (CE-MS)

Capillary electrophoresis (CE) is suited for the separation of polar and charged compounds, as compounds are separated on the basis of their charge-to-mass ratio. As this separation mechanism is different to other approaches such as reverse phase liquid chromatography (RPLC), CE can provide complementary information on the composition of a biological sample (Ramautar *et al.*, 2009). Further advantages of CE are fast and efficient separations without the need for extensive sample pretreatment, a low consumption of organic solvent and other reagents, and the use of simple fused-silica capillaries instead of expensive LC columns. On the other hand, the concentration sensitivity of CE is low due to the limited sample volume in the range of nanoliters that can be introduced into the capillary. This can be partly compensated with a coupling to a sensitive mass spectrometry (MS) based detector.

The coupling of CE-MS has been recognized as an attractive complementary technique for metabolomic studies. Initially, mainly targeted analyses have been performed but in recent years the potential of CE-MS for comprehensive metabolomics approaches has been recognized. The results of more than a decade in CE-MS based analysis have been reviewed by Monton and Soga (2007).

The range of applications of the main ionization and separation techniques is presented in Figure 2.7. Currently, no single method exists that is capable of covering the complete metabolome of an organism. Thus, integrated approaches combining the different techniques need to be realized for comprehensive analyses.

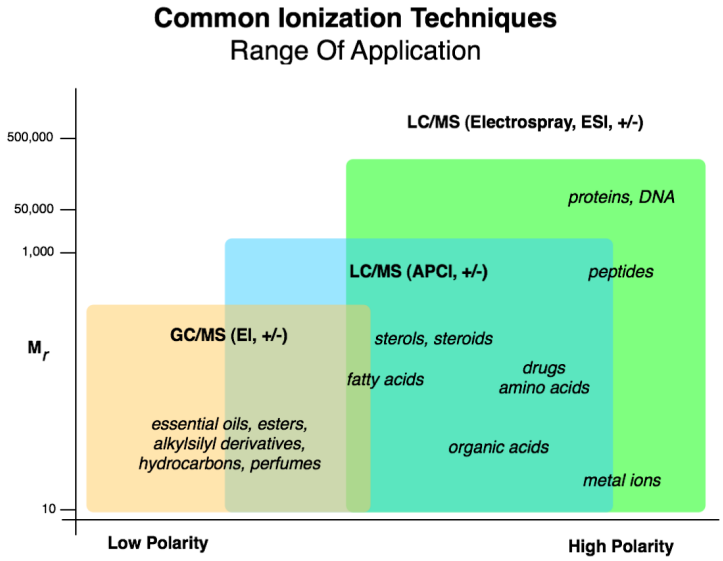

Figure 2.7.: The range of the different ionization and separation techniques with respect to mass and polarity of the analytes is presented. No single analytical method is capable of covering the complete metabolome quantitatively. For comprehensive metabolomics analyses, integrated approaches combining the different analytical platforms need to be employed. Figure adapted from Halket *et al.* (2005).

2.4. Structure of chromatographic data

The data that is produced by all hyphenated MS methods is typically mass spectral data together with a time dimension from the preceding compound separation technique (Villas Bôas, 2007). Thus, the structure of GC-MS, LC-MS, and CE-MS data is very similar. The mass spectra are obtained in regular intervals which means that a spectrum has been recorded for each time point in the chromatogram. A chromatogram represents specific values from the spectra; i.e. either single or extracted ions current (EIC) or the total intensity current of the mass spectra (TIC) and is plotted as a function of time. A complete data file can be seen as a matrix where spectral information spans the y-dimension and time is represented via the x-dimension. This is visualized in Figure 2.8: the grey-scale image in the upper right part illustrates the intensity values measured at each point.

A mass spectrum is presented in the lower part and can be stored in two ways, either as a continuum spectrum or in the centroid data format. The continuum format represents the most raw data format as all data points recorded by the spectrometer are stored. In centroided format, the spectrum is reduced to discrete mass-intensity pairs of the ions recorded. This format is commonly used, as it generates smaller data files. One drawback is that the masses of a centroid spectrum are recorded at a continuous scale. They can therefore not be aligned directly between consecutive scans but have to be binned to obtain a regular data matrix.

The bin width has to be chosen according to the accuracy of the mass spectral instrument.

The aim of the preprocessing of the raw data matrices that represent the chromatographic data is to find and identify the chromatographic peaks representing compounds present in the biological sample under study. A variety of signal processing and feature detection methods have been developed and will be presented in the following section.

2.5. Signal processing and feature detection

The general preprocessing strategy that transforms raw datasets from hyphenated mass spectrometry to quantitative metabolite information usually encompasses filtering, noise and baseline reduction of the raw data followed by feature extraction which corresponds to chromatographic peak detection in this context.

GC-, LC- or CE-MS datasets exhibit various sources of noise often covering or distorting true analyte signals. The aim of signal processing in general is the separation of the analyte signal from the different types of noise. The signals that are sought in the datasets under study typically exhibit a representative peak shape which often resembles a Gaussian distribution.

Peak detection is central and a variety of peak detection methods has been proposed for both chromatographic and mass spectral datasets over the years (Stein, 1999; Danielsson *et al.*, 2002; Andreev *et al.*, 2003; Du *et al.*, 2006). In general, the goal is to remove the baseline, de-noise the signal using e.g. smoothing filters and identify peaks above a given signal-to-noise (S/N) ratio. The intensities of extracted or total ion chromatograms are used and often information about the expected peak shape is integrated to improve the recovery rate. Furthermore, so called deconvolution (Stein, 1999) methods are necessary to separate co-eluting metabolites.

Various approaches for filtering noise can be applied to chromatographic signals. The methods range from simple median filtering in a specified window size to more elaborate signal processing techniques that increase the S/N ratio but do not affect the area under curve. One of these approaches is called *matched filtration* (Danielsson *et al.*, 2002). The general concept of matched filtration is the application of a filter function whose coefficients are equal to the expected shape of the signal. As described earlier, the Gaussian function is a simplified description of the peak shape obtained in chromatographic experiments. When such a filter is applied, a reduction of peaks whose widths are significantly less than the model peak shape can be observed. The application of matched filtration on LC-MS data was first reported by Danielsson *et al.* (2002). Additionally, matched filtration was recently extended by using the noise characteristics in areas without signal to improve the filter in an algorithm known as MEND (Andreev *et al.*, 2003). An implementation of the approach in the R system (Ihaka and Gentleman, 1996) has been made avail-

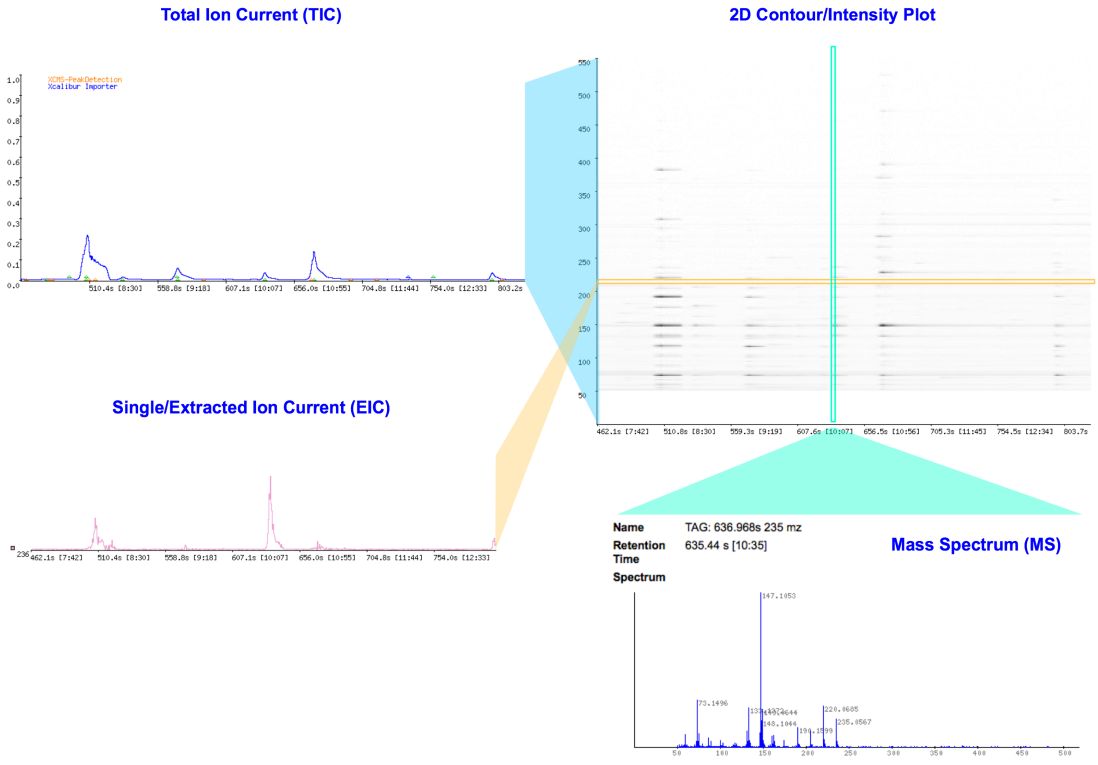

Figure 2.8.: Overview of the structure of chromatographic data. The data that is produced by all hyphenated MS methods are mass spectral data together with a time dimension from the preceding compound separation technique. Mass spectra are obtained in regular intervals, one mass spectrum recorded at a specific timepoint of the chromatographic separation process is presented in the lower part of the figure. The 2-dimensional plot at the top right part represents all measured data, intensities of individual ions are visualized using a grayscale mapping. The construction of a total (TIC) or extracted ion current or chromatogram (EIC) based on the mass spectral data is presented in the left part of the figure. An EIC represents the intensities of a single ion measured over time, the TIC represents the sum of all ions measured in the individual mass spectra.

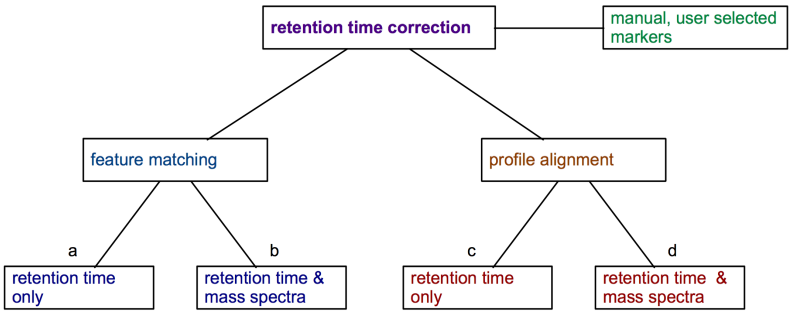

Figure 2.9.: Overview of chromatogram alignment approaches: Retention time correction can be achieved by selecting markers manually. This approach is time and labor intense and has various drawbacks. Hence, alignment algorithms use either the raw data for chromatogram alignment performing complete time series alignment (profile alignment) or make use of predicted chromatographic features (peaks) which have been previously detected by preprocessing steps (peak alignment, a.k.a. feature based alignment). In both variants, retention time information and additionally mass spectral data can be employed (Figure adapted from Robinson *et al.* (2007)).

able by Smith *et al.* (2006). Their XCMS system works both on LC- and GC-MS data.

Apart from the noise influence on single chromatograms, there are retention time drifts between consecutive measurements that can aggravate the comparison and analysis of multiple measurements. How these can be addressed is presented in the following section.

2.5.1. Chromatogram alignment

Due to instrument imperfections that are e.g. caused by small variations in the temperature program or mobile phase flow rates, a chromatographic run can not be reproduced even if the very same sample and instrument are used again. This is especially true for LC-MS measurements. Additionally, column aging and pollution by disproportional overrepresented analytes can lead to shifts in retention time and aberrations in peak compositions of recorded mass spectra. It is important to understand that such perturbations may affect the retention time of the studied analytes, however, the sequence of analyte occurrences is typically not altered. Nevertheless, sample matrix effects, that emerge from variations in sample composition, and the previously discussed difficulties in sample extraction and preparation influence the quality of a chromatographic separation (Robinson *et al.*, 2007; van Nederkassel *et al.*, 2006).

To compensate the aforementioned retention time drifts, chromatographic alignment can be performed. Easily identifiable retention standards can be added to the

samples yielding so called retention indices. Alternatively, alignments can be computed based on detected features (Robinson *et al.*, 2007) or by finding the maximal covariance between the chromatograms (Jonsson *et al.*, 2005). In any case, successful alignment simplifies the assignment of corresponding unknown peaks or mass signals across multiple measurements in the downstream analysis. Over the years, various approaches have been developed which work either on the complete measurement data, or employ previously detected features i.e. chromatographic peaks. A categorization of the methods is presented in Figure 2.9.

2.5.2. Continuous time series alignment

There exist several alignment techniques which correct retention time shifts on raw chromatogram data, among others dynamic time warping (DTW), parametric time warping (PTW), semi-parametric time warping (STW), fuzzy warping (FW), and correlation optimized warping (COW) (van Nederkassel *et al.*, 2006). However, COW has already been developed in 1997 by Nielsen *et al.* (1998) and has become a frequently used algorithm, which served as model for various newer algorithms, too. Although the original COW approach only supports pairwise alignment, it will be used to explain the principles of time series alignment techniques in the following.

Correlation optimized warping

The basic concept of Correlation Optimized Warping (COW) is to divide the chromatograms, of which one serves as target chromatogram, into a user-defined number of N sections. Each section is linearly stretched or compressed to match the target chromatogram as good as possible, which is accomplished by a warping function. Thereby the function shifts the end points of each section (except the outermost ones, which are fixed) within a user-specific range of size $[-s; +s]$, where s (called the slack parameter) is defined by the user. For each possible shift position, the section is then linearly interpolated and a correlation coefficient is calculated and stored. Starting from the first section, the warping function is iteratively called upon each section and with respect to previous shifts the correlation coefficients for further shifts are calculated. After all possible warping solutions are calculated and scored, the highest scoring warping solution is selected. The score of a warping solution is defined by the cumulative sum of its correlation coefficients of the contained sections (Nielsen *et al.*, 1998; van Nederkassel *et al.*, 2006). The computation time of the algorithm depends strongly on the two user-defined parameters N and s, where N usually depends on the number of peaks. Hence, chromatograms with many peaks or with great length differences where a high slack parameter must be used, lead to an unacceptably high computation time (van Nederkassel *et al.*, 2006). However, the major disadvantage of the algorithm emerges from the fact that the section for each chromatogram needs to be manually defined. The quality

of an alignment is therefore dependent on the experience of the user, who has to supply the proper position and amount of sections.

2.5.3. Feature based chromatogram alignment

The large distribution of feature based chromatogram alignment algorithms is certainly induced by their usually low computational time, as only an extract of data is used to determine an alignment. Peak detection belongs to the standard procedures of chromatogram analysis, which is performed in every metabolomic experiment, independent of the employed alignment tool. It is required for the identification of recorded analytes in both targeted and untargeted metabolite profiling as well as in metabolomics. The use of profitable data which is already available is a substantial argument to favor peak alignment. Peak alignments define alignments of peak lists instead of raw chromatogram data, as the name implies. A peak list L is defined as an ordered sequence $L = [p_1, p_2, \ldots, p_n], p_i < p_{i+1}$. However, a peak alignment can be transformed into a chromatogram alignment. Thereby aligned peaks serve as *anchor* points, between which the retention time of the respective chromatograms is interpolated using linear, polynomial, or spline interpolation. Thus, anchor points are comparable to the section end points in the COW alignment algorithm.

Dynamic programming approaches for chromatogram alignments

Various authors (Prakash *et al.*, 2006; Hoffmann and Stoye, 2009; Robinson *et al.*, 2007) developed dynamic programming based algorithms for chromatogram alignments. The general concept of these approaches is similar, Prakash and Hoffmann use profile data whereas Robinson employs feature lists. To present the general concept of dynamic programming based chromatogram alignments, the method by Robinson *et al.* is detailed in the following. In this approach, the alignment of two peak lists $L_X = [x_1, x_2, \ldots, x_n]$ and $L_Y = [y_1, y_2, \ldots, y_m]$ is defined as a list with one-to-one relations or gaps between peaks x and y, e.g.

$$A_{L_X L_Y} = [(x_1, y_1), (x_2, y_2), (x_3, -), (x_4, y_3), \ldots] \tag{2.1}$$

This example presents two cases analogous to pairwise sequence alignments: two peaks either *match* or *gaps* occur meaning that a peak from list L_X has no matching counterpart in L_Y or vice versa. The maximal length of $A_{L_X L_Y}$ is limited by $n+m$ and the alignment is obtained by backtracing a generated score matrix using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970).

It is important to note that in the case of Robinson's algorithm, the employed score function is composed of two terms. The first computes a mass spectral similarity and the second penalizes retention time deviations between the peak tuple.

$$score(x_i, y_j) = S(x_i, y_i) \times exp\left(\frac{(RT(x_i) - RT(y_j))^2}{2D^2}\right) \tag{2.2}$$

The function $S(x, y)$ returns the dot-product of the vector representations of the mass spectra associated to the two peaks x and y. The function $RT(x)$ returns the retention time of peak x in seconds and D represents the tolerated retention time deviation that can occur due to measurement errors. By varying D, the importance of the retention time deviations to the overall peak similarity score can be controlled. Robinson's algorithm represents an instance of class b presented in Figure 2.9. The normalized dot product (cosine) between the mass spectra vectors $MS(x)$ of length l of the compared peaks x_i and y_j is used as similarity measure and calculated as follows:

$$S(x_i, y_j) = \frac{\sum_{k=1}^{l} MS(x_i)_k \times MS(y_j)_k}{\sqrt{\sum_{k=1}^{l} MS(x_i)_k} \times \sqrt{\sum_{k=1}^{l} MS(x_i)_k}} \tag{2.3}$$

When the score matrix for L_X and L_Y has been computed, the traceback resulting from the application of dynamic programming gives the best alignment of the peaks. For this approach, a gap penalty G has to be defined additionally. For G, a fixed value between 0 and 1 is employed. Unlike in sequence alignment where affine gap-cost functions are used, the extension of existing gaps is not favorable to the creation of new gaps in the alignment of peak lists.

The construction of multiple alignments is detailed in the original publication (Robinson *et al.*, 2007) and will be sketched briefly. It follows the ideas of progressive multiple sequence alignments (Thompson *et al.*, 1994). In order to construct multiple alignments of peak lists $L_A, L_B, L_C, \ldots$, the authors first calculate all possible pairwise alignments between the peak lists. Then, from all pairwise alignments the alignment score for each pair of peak lists is calculated. Based on these scores, a guide tree is built which provides the similarity relationship between the peak lists. The peak lists are then progressively aligned following the branching order given by the guide tree resulting in a multiple alignment of all peak lists.

As the presented algorithm depends on previously detected peaks, the quality of the generated alignments is influenced by the peak detection method employed in the preprocessing. Automated peak detection remains a challenge, and any errors introduced at this stage will propagate through to the final alignment tables. Especially, if several peaks with similar mass spectra occur in a small retention time window, the method can generate misalignments.

In contrast, the reduction of the large chromatographic datasets to peak lists makes the approach very efficient. Furthermore, with the availability of multiple alignments, the processing of high-throughput metabolomics data is simplified. In the optimal case, the errors induced by chromatographic deviations and time shifts are compensated and all downstream analysis steps can benefit from the registered and aligned measurements.

2.6. Compound identification

After the detection and deconvolution of chromatographic peaks, the associated metabolites need to be identified. The mass spectra obtained through fragmentation by electron impact can be compared to annotated reference spectra from mass spectral databases (NIST, GMD). If chemical or other *soft* ionization methods were performed, exact molecular masses of the mother ions can be achieved. In this case, the Metlin database (Smith *et al.*, 2005) can be used to find candidates for the chemical identity of the measured compound. Database lookup can provide initial evidence for the chemical identity of the metabolite but additional information such as the retention time or retention index of the peak need to be taken into account. Analysts dealing with unknown compounds from complex environmental samples need to prove that a tentative identification is in fact the correct compound (Schymanski *et al.*, 2009).

2.6.1. Mass spectral reference databases

NIST

The National Institute of Standards and Technology (NIST) provides a standard reference mass spectral database. The so-called NIST/EPA/NIH library is the product of a multi-year, comprehensive evaluation and the main EI MS library contains 191,436 spectra. Together with the Retention Index Library which covers 293,247 Kovats RI values, information for 44,008 compounds is available. The spectra have been inspected by experienced mass spectrometrists and the chemical structures are checked for correctness and consistency, using both human and computer methods. Along with the mass spectra, the NIST database also provides references to IUPAC names and CAS registry numbers (Wiley, 2005).

The Human Metabolome Database (HMDB)

The Human Metabolome Database (HMDB) is a web-based bioinformatics/chemoinformatic resource with detailed information about human metabolites and metabolic enzymes. The HMDB contains MS spectra for some of the metabolite entries together with compound description, names and synonyms, chemical structure information, and physico-chemical data (Wishart *et al.*, 2007).

GMD Golm Metabolite Database

Kopka and co-workers (Kopka *et al.*, 2005; Schauer *et al.*, 2005) have compiled and maintain the Golm Metabolome Database (GMD) as an open access metabolome database. The goal of this database is to provide means for the unambiguous identification of metabolites in highly complex metabolite preparations from biological samples. The authors have established a collection of mass spectra, which comprise

frequently observed metabolites of either known or unknown exact chemical structure. The motivation is to pool the identification efforts currently performed in many laboratories around the world. The GMD provides public access to custom mass spectral libraries, metabolite profiling experiments as well as additional information and tools, e.g. with regard to methods, spectral information or compounds. The main goal will be the representation of an exchange platform for experimental research activities and bioinformatics to develop and improve metabolomics by multidisciplinary cooperation.

METLIN

Whereas the GMD focuses on GC-MS spectra, the Metlin database was established as a repository for exact masses of molecular or adduct ions obtained from LC-MS measurements (Smith *et al.*, 2005). METLIN is a freely accessible web-based data repository, which has been developed to assist in a broad range of metabolite research projects and to facilitate metabolite identification through mass analysis. METLIN includes an annotated list of known metabolite structural information that is easily cross-correlated with its catalog of high-resolution Fourier transform mass spectrometry (FTMS) spectra, tandem mass spectrometry (MS/MS) spectra, and LC-MS data.

2.6.2. De novo identification of compounds

With database lookup, compounds can only be identified if reference mass spectra exist in the previously described libraries. Hence, the de-novo identification through interpretation of metabolite mass spectra is highly sought (Böcker *et al.*, 2008). Identification can be achieved by experts using mass spectral classifiers to identify substructures and then building the matching molecule(s), either by hand or using structural generators such as MOLGEN (Benecke *et al.*, 1997). In any case, this process is time consuming, error prone and requires expert knowledge. A valuable piece of information in the process to identify yet unknown compounds is the molecular mass of the compound which can be obtained by highly accurate mass spectrometry instruments. One or more possible sum formulas of the unknown compound can be reconstructed based on the measured molecular mass as described in the following section.

2.6.3. Mass decomposition

The elucidation of the structure of unknown small molecules by mass spectrometry remains a challenge despite increased accuracy of the mass spectrometric devices.

Modern high-resolution mass spectrometry can determine the mass of sample molecules with an accuracy of 1-5 parts-per-million (ppm). As described in Section 2.2.4, the output of a mass spectrometer consists of peaks that correspond to the masses of the sample molecules together with their abundance. The natural isotopic

distributions of the elements result in several peaks in the output that correspond to the same type of sample molecule, the so called *isotope pattern*. The composition of the isotope patterns contains valuable information that can be used to identify the elemental composition of unknown compounds.

The first step towards the identification of the structure is the determination of the elemental composition of the unknown compound based on the measured ions. Efficient algorithms are available to enumerate all potential molecular formulas that sum up to the observed ion masses. Nonetheless, with growing molecular weight, the number of potential candidate formulas explodes (Böcker *et al.*, 2008, 2009). Apart from the exact enumeration algorithms, a set of seven heuristic rules has been developed to select the most likely and chemically correct molecular formulas. These heuristics include restrictions for the number of elements, LEWIS and SENIOR chemical rules, isotopic patterns, hydrogen/carbon ratios, element ratio of nitrogen, oxygen, phosphor, and sulfur versus carbon, element ratio probabilities and the presence of trimethylsilylated compounds (Kind *et al.*, 2007).

The authors evaluated their filtering rules on 6000 pharmaceutical, toxic and natural compounds and obtained a retrieval of the selected compound as top hit at 80% to 99% probability when assuming data acquisition with complete resolution of unique compounds and 5% absolute isotope ratio deviation and 3 ppm mass accuracy. Together with the efficient mass decomposition, this marks one possible approach for the elucidation of compounds not yet listed in mass spectral databases.

2.7. Data analysis strategies for metabolomics datasets

After the successful identification compounds based on mass spectral information and the quantification of the respective peaks, the experimental data can be represented as a data matrix with rows representing individual measurements and columns representing the detected compounds. The entries in the data matrix correspond to quantitative measurements of the compounds, e.g. the peak areas or intensities.

The reduction of experimental data to abundance matrices allows to employ various pre-treatment methods that have previously been established in other *Omics* techniques such as microarray analyses or quantitative proteomics. Nonetheless, metabolomics data have important properties and differ from e.g. transcriptomics data in several points.

2.7.1. Properties of metabolomics data

Metabolomics experiments obtain a snapshot of the metabolome that reflects the cellular state or phenotype for the experimental conditions under study. In an evaluation study, van den Berg *et al.* (2006) analyzed the properties of the datasets that are typically generated by metabolomics experiments. The authors identified dif-

ferences in the order of magnitudes between measured concentrations of abundant metabolites such as e.g. ATP and low abundant metabolites. This finding suggests that e.g. for bio-marker discovery, relative fold changes are more important than large absolute changes of metabolite concentrations and that the heteroscedasticity of the experimental data has to be accounted for in the analysis of metabolomics experiments.

If the experimental conditions included biologically induced variation, the aim of the experiment is often to detect differences in the fold changes in metabolite concentrations that are an effect of the induced variation. It is highly important to be able to distinguish the sought after changes from the uninduced biological and technical variation. Different strategies exist to limit the influence of technical variation. In order to compensate varying signal intensities due to changes in detector sensitivity or sample preparation, artificial internal standards can be used[2]. The conduction of replicate measurements allows to identify the biological variation and to employ tests for statistical significance of the observed changes.

2.7.2. Data pre-treatment methods

The necessary replicate measurements do increase the size of the experimental datasets which results in further requirements. Especially for low abundant compounds, the stable detection across all measurements of an experiment can be problematic and may lead to missing values if the abundances drop below the level of detection in the sample. Missing values may also occur due to noisy measurements or the overlap of a trace compound by compounds with higher abundance.

An important preprocessing step is therefore the treatment of missing values. In the most strict approach, columns or rows of the experimental data matrix containing missing values can be discarded. As this can reduce the number of available data drastically, missing value substitution is often preferable. Simple methods for missing value estimation in metabolomics experiments are e.g. the replacement by the mean or median of the metabolite concentrations over all samples or the imputation from nearest neighbors (Steinfath *et al.*, 2008).

After missing values have either been estimated or discarded, multivariate statistics and explorative data visualizations such as principal component and hierarchical clustering analyses can be employed. Furthermore, pairwise correlations between the measured metabolite levels can be computed (Steuer *et al.*, 2003b) in order to gain insight into changes of the metabolomic network under varying experimental conditions. As mentioned earlier, large differences in concentration of the metabolites can be observed. Yet, these differences are not proportional to the biological relevance of the individual metabolites but this distinction can not be made by typical data analysis methods.

With these limitations and requirements in mind, a variety of data pre-treatment methods have been evaluated (van den Berg *et al.*, 2006). The aim of the study

[2]Ribitol is often employed as internal standard in GC-MS analysis.

was to identify methods that can correct for aspects that hamper the biological interpretation of metabolomics datasets.

An initial step is the *centering* of the measured metabolite abundances. Centering converts all the abundances to fluctuations around zero instead of around the mean of the metabolite abundance. Thereby, differences in the offset between high and low abundant metabolites are transformed to fluctuations around zero. In a second step, *scaling* methods (pareto scaling, autoscaling, range scaling, vast scaling and level scaling) were applied on the metabolite abundances (Keun *et al.*, 2003). The aim of scaling is to compensate the fold differences between the metabolites by converting the data into differences in abundance relative to the scaling factor. Finally, the effect of *transformations* of the metabolite abundances using the *power* or *log* functions was examined.

The authors of the study state, that different aspects of the data are emphasized by different pre-treatment methods. In general, the choice for a pre-treatment method strongly depends on the biological question, the properties of the dataset and the data analysis method selected. As a general suggestion, the authors recommend the use of auto-scaling and range scaling for the explorative analysis of the dataset under study as both methods were able to remove the dependence of the rank of the metabolites on the average concentration and the magnitude of the fold changes. To evaluate the normalized and scaled results, an explorative data analysis was performed. The PCA (principal component analysis) of the range and auto scaled data showed biologically sensible results in their study.

2.7.3. Explorative data analysis

Many of the explorative and pattern-recognition strategies currently pursued in the analyses of *Omics* data are based on unsupervised techniques (Hastie *et al.*, 2009). In metabolomics, hierarchical cluster analysis (HCA) can be used to assess, in a multivariate manner, how similar sets of samples are to one another on the basis of their metabolite profiles (Goodacre *et al.*, 2004). For high dimensional metabolome data a typical optimization procedure is *simplification* or dimensionality reduction. Methods such as principal component analysis (PCA) or independent component analysis (ICA) (Hyvärinen and Oja, 2000) try to summarize the large body of metabolite data by a few parameters with minimal loss of information. The resulting low dimensional plots (ICA, PCA) and generated dendrograms (HCA) can be interpreted by the researcher. If metabolite profiles of known provenance are included in the clustering approaches, one can classify unknown samples by their closeness to the known datasets, a process referred to as *guilt by association* (Altshuler *et al.*, 2000).

2.7.4. Machine learning and classification

If the classification of the measurements is known *a priori*, supervised machine learning methods can be applied to train predictive models based on the observed

metabolite abundances. The recently released *caret* package (Kuhn, 2008), available from the BioConductor repository (Gentleman *et al.*, 2004), provides interfaces to classification and regression training and contains a variety of tools for developing predictive models. It furthermore gives access to the rich set of models available in R and simplifies as well as standardizes model training and tuning across a wide variety of modeling techniques.

As mentioned before, an important prerequisite for reliable metabolomics data analysis is a robust preprocessing of the raw datasets obtained from GC- or LC-MS measurements. Missing values, high correlation between the ratios or levels of single metabolites and near zero variance of metabolite abundances complicate the construction of predictive models for metabolomics datasets (Steuer *et al.*, 2003b). When these issues are addressed, state-of-the art classification methods and predictive models such as Support Vector Machines (SVM) and Random Forests can be applied for the analysis of metabolomics datasets.

Support vector machines

Support vector machines represent an intensely studied and theoretically well founded classification approach that has been successfully used for various bioinformatics problems in the past years (Boser *et al.*, 1992). SVMs have been applied for the functional classification of expression data from microarray experiments (Brown *et al.*, 2000), for the detection of homologous proteins (Hou *et al.*, 2003; Jaakkola *et al.*, 2000), and for the analysis of quantitative proteomics experiments. SVMs are ideally suited for two-class classification problems but can be generalized to multiple classes. Viewing input data as two sets of vectors in an n-dimensional space, a SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets.

Random forests

In supervised machine learning, a random forest is a classifier that consists of many decision trees. The output of the random forest is the class that is the mode of the classes output by individual trees. During the construction of a single classification tree, logical if-then conditions are determined for predicting or classifying cases. The original algorithm for inducing a random forest was developed by Breiman (2001). The method constructs a collection of decision trees with controlled variations. An advantage of the random forest approach is the fast learning process even for large number of input variables. The generation of highly accurate classification for many datasets could be shown. Especially beneficial for the detection of marker substances in the context of metabolomics research is the possibility to estimate the importance of variables (metabolite abundances) in determining classification.

Variable importance estimation

The importance of a variable for classification accuracy may be due to its (possibly complex) interaction with other variables. Hence, the definition of variable importance is generally problematic. Nonetheless, the importance of a variable is estimated in the random forest algorithm by computing the increase in prediction error when the data for that variable is permuted while all others are left unchanged. The percent increase in misclassification rate for a permuted variable does reflect the importance of that variable for the overall classification. To obtain importance estimates for the complete dataset, all input variables are consecutively and randomly permuted in the test set. Finally, when the aim of the analysis of metabolomics experiments is the detection of potential marker substances for two class experiments, the importance of a variable can directly be linked to its predictive power for the classification.

CHAPTER 3

State of the art in the analysis of metabolomics data

As described in the previous chapter, the generation of metabolomics data is largely performed via hyphenated mass spectrometry analyses. Mass spectra are consecutively acquired with a frequency determined by the scan rate of the MS instrument. Depending on the speed and accuracy of the MS instrument the scan rate can vary from single scans per second to up to 500 scans per second in modern Time-Of-Flight instruments such as the LECO Pegasus IV system. The size of the measured data is directly proportional to the scan rate and the duration of the chromatographic separation and ranges from few megabytes for simple Ion-Trap and LC-MS measurements to up to several gigabytes for GC/GC-MS measurements.

Different manufacturers of hyphenated MS systems such as LECO, Waters, Thermo Finnigan or Bruker Daltonics provide machine specific software solutions for the acquisition and storage of the measured chromatographic and mass spectral data. These datasets are typically stored in proprietary file formats that are only readable by vendor specific software systems.

Researches are limited to the closed source functionality of the system specific software solution. The historical focus of hyphenated-MS on targeted analysis methods becomes obvious since multivariate statistical features are missing in most of these applications. Nonetheless, for most of the software systems conversion tools exist that can export the proprietary datasets in open and standardized exchange formats.

Initially, the ANDI netCDF format was used to represent the chromatographic data but two XML based formats have been developed in the last years. A severe drawback is that the mzXML (Pedrioli *et al.*, 2004) and mzData (Orchard *et al.*,

2007) formats have been developed in parallel which are competing and redundant representations of the same raw data. Only recently, the mzML format (Deutsch, 2008) has been defined which aims at merging both efforts and providing a common standard for mass spectrometer output. Due to the novelty of mzML, few exporters are currently available from the vendors of mass spectrometry hardware.

3.1. Proprietary and vendor specific systems

3.1.1. Thermo Xcalibur

The Xcalibur® software provided by Thermo Scientific is a Microsoft® Windows based data system that provides instrument control and data analysis for Thermo Scientific mass spectrometers. The software provides automation functionality for quantitative processing and allows to review and rework the preprocessed GC- and LC-MS data using the so called Quan Browser® application. User-defined *Processing Setups* need to be generated for the identification and quantitation of compounds, therefore retention time intervals and a quantitation ion have to be defined for each compound. Apart from the integrated functionality, Xcalibur® can convert the proprietary RAW data format to netCDF files which allows to access the measured data in a standardized form.

3.1.2. LECO ChromaTOF

The current ChromaTOF® software by LECO Corp. (St. Joseph, MI) features signal deconvolution, automated peak detection functionality and a variety of calibration approaches. It is possible to perform a semi-quantitative analysis for the reporting of non-calibrated compounds with the software and to automatically export data to a variety of formats including CSV, netCDF and Raw file formats. In combination with the Pegasus® 4D GCxGC-TOFMS instrument, ChromaTOF® supports the analysis of two-dimensional gas chromatographic separation experiments.

3.1.3. Agilent Mass Hunter

The Mass Hunter® application is an integral part of the Agilent TOF software® system. The software works on electrophoretic and chromatographic mass spectral datasets with the focus on the extraction of information from single measurements. Various preprocessing functions are available to reduce data complexity, eliminate potential interferences, and generate a list of molecular features. In the software context, a *feature* is a discrete molecular entity defined by the combination of retention time and mass. Apart from typical visualizations of TIC-, EIC-, and Contour-Plots, Mass Hunter allows to compute potential molecular compositions of the detected features.

3.1.4. MassLynx

The MassLynx™ software acquires, analyses and manages information from mass spectrometry. MassLynx™ controls every mass spectrometric instrument from Waters (Milford, MA, USA). This includes the management of the sample and applied solvents and furthermore the control over the MS instrument and additional detectors. The software allows to detect nominal and exact masses and can also be applied for MS/MS analysis. In MS/MS, selected ions of a precursor mass spectrum (e.g. the molecular ion) are subjected to an additional fragmentation which is recorded in a subsequent mass spectrometer.

MassLynx organizes the experimental data in so called *Sample Lists*. The user employs the analysis methods of the software on these Sample Lists which simplifies the handling of results from GC-MS or LC-MS systems according to the manufacturer. Specific modules of the software systems allow to automatically quantify the samples or to qualitatively screen and identify compounds in the samples. According to the user manual, no support for multivariate statistics or integration with other *Omics* datasets is currently included in the system.

3.1.5. AMDIS

The AMDIS mass spectrometry analysis software (Stein, 1999) has been developed for the automatic extraction of pure component mass spectra from complex GC-MS data files. These spectra are used for identifying compounds based on a reference library. AMDIS incorporates methods for spectrum deconvolution and library searching with the addition of a variety of factors to account for noise and other characteristics of GC-MS data. One drawback of the freely available system is that it only works under Microsoft® Windows operating systems and does not provide methods apart from simple text files to store or manage the obtained results of GC-MS analyses.

3.2. Software from the scientific community

Apart from the need for open formats for storing raw measurement data, there is also a need for the consistent experimental description of metabolomics measurements. It has been repeatedly pointed out (Fiehn *et al.*, 2006, 2008) that the detailed description of experimental conditions and all relevant factors of the processing of experimental data is important for the interpretation, exchange, and reproducibility of metabolomics experiments.

Motivated by this need, a minimal set of reporting standards and best-practice recommendations (Sansone *et al.*, 2007) has emerged. Together with the definition of ontologies a standardized annotation of the experimental design and conditions has become feasible. This is of importance since experimental details are vital for the interpretation of the data and its re-use in related studies.

Apart from the proprietary systems described above, a growing number of tools and applications is being developed by the scientific community for the analysis of hyphenated MS datasets. The ArMet (Jenkins *et al.*, 2005) data model was a first proposal for the standardization of metabolomics software tools. It has been initially designed for plant metabolomics, and later has been extended to support e.g. microbial metabolomics experiments.

The specific features and functionality of tools are described in the following. The tools are ordered by the level of functionality they provide, i.e. tools dealing with raw data and data preprocessing are detailed at first and the collection ends with high-level analysis packages that mainly provide multivariate statistical analysis. The number of applications and tools in the scientific community for computational metabolomics has been growing rapidly over the past years. Therefore, the presented collection can hardly be complete. Nonetheless, the tools portrayed in this collection represent typical solutions for the main tasks in the analysis of metabolomics datasets.

3.2.1. SetupX and BinBase

Based on the ArMet data model, Scholz and Fiehn (2007) have presented the SetupX system, which allows for accurate description of the biological study design and accompanying meta-data in metabolomic databases. SetupX is an example for a web-based metabolomics LIMS system and is oriented towards metabolomics data from GC-MS measurements. For the analysis of the experimental data SetupX relies on the seamlessly integrated mass spectrometry database BinBase. The system has been implemented in Java and stores experimental datasets in a relational database system.

3.2.2. MZmine

The MZmine software contains methods for processing and visualizing mass spectrometry based molecular profile data. Support for the mzXML data format is provided. MZmine allows to perform batch processing for a large number of files and new methods for calculating peak areas using a post-alignment peak picking algorithm have been implemented. The implementation of Sammon's mapping and curvilinear distance analysis for data visualization and exploratory analysis complete the functionality of the Java application (Katajamaa *et al.*, 2006).

3.2.3. MET-IDEA

MET-IDEA is a data extraction tool solely available for the Microsoft® Windows operating systems. It only supports the netCDF input format. The implemented analysis method performs selected ion quantification and is capable of extracting semiquantitative data from raw data files, which allows for more rapid biological insight according to the authors (Broeckling *et al.*, 2006). The tool needs a user

specified list of ion and retention time pairs for the extraction of quantitative features from the raw chromatographic datasets.

3.2.4. MetAlign

The MetAlign software handles a broad range of accurate mass and nominal mass GC-MS and LC-MS data. It is capable of automatic format conversions, accurate mass calculations, baseline corrections, peak-picking, saturation and mass-peak artifact filtering, as well as alignment of up to 1000 data sets. A 100 to 1000-fold data reduction can be achieved by the preprocessing methods of the system. The MetAlign software output is compatible with most multivariate statistics programs but does not directly provide any statistical analysis features (Lommen, 2009).

3.2.5. XCMS and centWave

The LC-MS-based data analysis approach XCMS features metabolite profiling in bio-marker discovery, enzyme substrate assignment, drug activity/specificity determination, and basic metabolic research. It provides new data preprocessing approaches to correlate specific metabolites to their biological origin and incorporates novel nonlinear retention time alignment, matched filtration, peak detection, and peak matching. The method dynamically identifies hundreds of endogenous metabolites for use as standards and calculates a nonlinear retention time correction profile for each sample. Following retention time correction, the relative metabolite ion intensities are directly compared to identify changes in specific endogenous metabolites, such as potential bio-markers (Smith *et al.*, 2006). XCMS has been released as a BioConductor package, the extension of the system is therefore possible and recently a sensitive peak detection extension for LC-MS data termed *centWave* was presented (Tautenhahn *et al.*, 2008). The use of XCMS and centWave requires familiarity with the R programming language, no support for the annotation or management of experimental datasets is included.

3.2.6. TagFinder

The TagFinder (Lüdemann *et al.*, 2008) software does address modern metabolomics and fluxomics studies and especially focuses on non-biased metabolomic fingerprinting, footprinting and profiling experiments. The system does support NetCDF input files. In addition, preprocessed results from the LECO ChromaTOF® software system may be imported for downstream analysis. The generation of statistically accessible data matrices from large-scale GC-MS based metabolite profiling experiments relies on co-analysis of retention index (RI) marker substances within each chromatogram. So called *mass tags* are identified across all chromatograms of an experiment by TagFinder based on RI gaps in the sorted peak lists. Afterwards, clusters of intensity-correlated mass fragments are computed and the non-normalized intensities of the grouped masses are stored as data matrix.

The authors state that parameter settings for the scanning distance between mass tags and the minimum width of mass tags should be carefully adapted and revised for each new GC-TOF-MS profiling experiment.

3.2.7. MetaboAnalyst

MetaboAnalyst (Xia *et al.*, 2009) is a web-based metabolomic analysis tool. It supports data processing, data normalization, multivariate statistical analysis, and graphing. Rudimentary metabolite identification and pathway mapping features are provided. The main feature of MetaboAnalyst is the upload functionality of data matrices that are processed in a streamlined manner with functions from R and the BioConductor repository. Apart from the integration of visualization methods the classification performances of several machine learning algorithms (i.e. random forest, SVM) can be assessed. The annotation of identified metabolites is simplified through the connection with the Human Metabolite Database (HMDB).

3.3. Metabolic pathway repositories and visualization tools

Apart from the preprocessing and statistical analysis of metabolomics data, visualization of the generated data in their biochemical context is of central importance to support the interpretation by a researcher. Metabolic networks are typically employed to detail the connection between metabolites, proteins and genes. Initially, there were databases such as KEGG (Ogata *et al.*, 1999) or the different realizations of MetaCyc (Caspi *et al.*, 2008) that store information about the structure of such metabolic networks. These databases represent static knowledge of metabolic pathways of organisms from all three domains of life. The contained data have been collected and curated over the years of genomic research and can be presented using images of metabolic pathways linking metabolites and enzymes.

Several tools have been developed to visualize and analyze biological networks together with data obtained from functional genomics measurements. Most interesting in this context are tools that visualize experimental data in the form of biochemical networks. The authors of the VANTED system for advanced data analysis and visualization in the context of biological networks (Junker *et al.*, 2006) presented a comprehensive review of existing pathway visualization and mapping tools such as Cytoscape (Shannon *et al.*, 2003), MapMan (Thimm *et al.*, 2004), KaPPA-View (Tokimatsu *et al.*, 2005), PathwayExplorer (Mlecnik *et al.*, 2005), and the Viewer included in MetaCyc-related databases (Karp *et al.*, 2002) such as AraCyc (Mueller *et al.*, 2003). They pointed out that often only two conditions can be compared. In experiments designed to provide the basis for simulation in systems biology this is of limited use since often changes in metabolite concentration or transcript levels can only be understood if time series experiments are conducted and analyzed. It is also stressed by the authors that most tools are lim-

ited to transcriptomics datasets and only *Omics Viewer* (Paley and Karp, 2006), Cytoscape, and MapMan are designed to also display metabolite or other data. A severe limitation of some of the existing tools is their dependency on static maps, i.e. the data is mapped onto predefined pictures. This might be appropriate if the tools are being developed for a single organism or metabolic pathway but in general it clearly limits the re-usability of the approach.

KEGG Pathways and KEGG Markup Language

A major component of KEGG, the Kyoto Encyclopedia of Genes and Genomes, is the PATHWAY database which represents most of the known metabolic pathways (Ogata *et al.*, 1999). The database is continuously updated and consists of a collection of graphical diagrams, the so called pathway maps. In these maps, a box represents an enzyme and a circle a metabolic compound. The manually drawn and annotated pathway maps represent knowledge about the metabolism, genetic information processing, and cellular processes.

The KEGG Markup Language (KGML) is an XML-based exchange format and contains computerized information about graphical objects and their relations in the KEGG pathways. In KGML a *pathway* element is the root element that specifies one graph object. The nodes of the graph object are represented by the *entry* elements, whereas the *relation* and *reaction* elements specify the edges. An *entry* element contains information about a node of the pathway, like *id*, *name* and *type*. The *relation* element specifies a relationship between two proteins or protein and compound, which is indicated by an arrow. The *reaction* element describes the chemical reaction between substrates and products.

KaPPA-View

KaPPA-View is a web-based tool and was developed to represent quantitative data for individual transcripts as well as metabolites on plant metabolic pathway maps. The aim of the system is to support the generation of hypotheses of gene function in the metabolic pathways through an intuitive visualization of the transcripts and metabolites. The system uses SVG vector graphic images for the representation of the biochemical pathways and the experimental datasets that are mapped on the pathway representations (Tokimatsu *et al.*, 2005).

MapMan

MapMan is a user-driven tool that displays large datasets (e.g. gene expression data from *Arabidopsis thaliana* Affymetrix arrays) onto diagrams of metabolic pathways or other processes. It has been developed specifically for data generated in *Arabidopsis thaliana* experiments measuring transcript or metabolite levels. The visualization focus is on the display of experimental data in hierarchical and pre-defined pathway maps.

The functionality of KaPPA-View and MapMan can be accessed via web-applications. In general, a tendency to provide sophisticated analysis methods for functional genomics experiments and datasets through web-based frameworks can be observed. The advantage of web-based analysis tools compared to stand-alone applications is the ease of updates and the possibility to rapidly release new features. Apart from recent web-browsers no additional software needs to be installed by the user.

To summarize, tools such as KaPPA-View or MapMan focus on a limited set of organisms and user defined pathway maps for other organisms or related strains are not supported. Additionally, means to visualize metabolic pathway information is usually limited by the underlying pathway model as can be seen in the CellDesigner and KEGG pathways. Informative legends or additional user definable graphical elements that explain details are in general not supported.

Whereas most of the aforementioned tools allow to directly upload files with numerical results from *Omics* experiments in simple text-based files (CSV, TSV) or spreadsheets, the support to directly access *Omics* databases containing experimental results via e.g. Web Services is currently not well established.

3.4. Evaluation of existing systems

Whereas all of these publicly available systems are able to perform one or more preprocessing function such as peak detection, chromatogram alignments, compound identification or quantitation, the direct connection to statistical analysis tools and means to describe experimental designs and conditions is not addressed. Besides, the connection to results from transcriptomics and proteomics experiments and genome annotation data is absent in all systems. Most often, the systems are single user applications and therefore limit the exchange of the results and can not be included in automated workflows. Single user applications running solely on proprietary operating systems do typically represent bottlenecks in high-throughput analysis pipelines. For the open source systems XCMS and centWave no user interfaces are provide which means that researchers have to familiarize themselves with the R programming language and environment. Table 3.1 gives an overview of features and limitations of the existing systems.

Software	Analytical Platforms	OpenSource	GUI	API	Peak detection	Identification	Mult. Statistics
Thermo Xcalibur®	GC-MS	-	+	-	+	+	-
LECO ChromaTOF®	(GC-)GC-MS	-	+	-	+	+	-
Mass Hunter®	LC-, GC-MS	-	+	-	+	+	-
Mass Lynx®	LC-, GC-MS	-	+	-	+	+	-
AMDIS	GC-MS	-	+	-	+	+	-
MzMine	LC-MS	+	+	-	+	-	+
metIDEA	LC-, GC-MS	-	+	-	+	-	-
metAlign	LC-, GC-MS	-	+	-	+	-	-
XCMS	LC-MS, (GC-MS)	+	-	+	+	(+)	(+)
MetaboAnalyst	LC-, GC-MS, NMR	+	+	-	-	-	+

Table 3.1.: Tabular overview of analysis features of the presented open source and proprietary or vendor specific software systems. Most systems feature graphical user interfaces and provide methods for the detection and identification of chromatographic peaks. The support for multivariate statistics is currently missing in most systems. XCMS does only support the identification of compounds found in LC-MS experiments.

None of the described metabolomics analysis software systems and tools allows to visualize the generated results in a biochemical context. Yet, various tools to integrate and visualize multi-*Omics* datasets are available as presented before. Their main features and characteristics are subsumed in Table 3.2.

To conclude, a streamlined combination of metabolomics preprocessing, statistical analysis and visualization approaches is hardly possible and can only be achieved through the combination of various (proprietary) software tools and applications. Most of them are single user applications and require a local installation and access to the raw experimental data. The exchange of results can most often only be realized using spreadsheet representations of the generated datasets. This approach is limiting the comparability of findings across multiple experiments and laboratories.

As metabolomics is highly interdisciplinary, specialists from different fields may need to contribute to achieve a comprehensive analysis of a metabolomics experiment. This becomes especially important for the identification of yet unknown compounds that have e.g. been detected to vary significantly between healthy and diseased tissues and present potential biomarkers. Experts in both biochemistry and mass spectrometry are needed to contribute to the analysis of the same dataset and therefore an easy and structured exchange of the data and results is necessary (Sansone *et al.*, 2007; Fiehn *et al.*, 2006). Typically, single-user applications as the ones described above do not provide functionality to collaborate on experimental datasets.

Software	Pathway maps	OpenSource	User defined	Time Series	Multi *Omics*	Web Services
KEGG	+	-	+	-	+	+
CytoScape	+	+	+	+	+	-
KappaView	+	-	+	+	+	-
MapMan	+	-	+	+	+	-
MetaCyc	+	+	-	-	-	-
OME	+	-	+	+	+	-
VANTED	+	+	+	+	+	-

Table 3.2.: Tabular overview of the features of the presented pathway repositories and visualization tools. The presentation of *Omics* data on metabolic pathway maps is intuitive and therefore supported by all systems. Conversely, the combination of these datasets from multiple functional genomics datasources and the visualization of time series experiments is not possible with all systems. Support for Web Services is currently only provided by the KEGG database system.

CHAPTER 4

Requirements and System Design

The evaluation of the existing systems highlights that no free software system is available allowing to cover the complete process from raw data analysis and preprocessing of metabolomics data sets to multivariate statistics and the integration of the results in a multi-*Omics* context. Nonetheless, valuable tools exist for individual preprocessing tasks, e.g. peak detection, metabolite identification or chromatogram alignments. The availability of both open data standards for the raw MS datasets and a recommended database model for the description of metabolomics experimental data provide the basis for an integrated approach.

Based on the observation that metabolomics experiments can be conducted in a high throughput manner, streamlined and robust data processing strategies need to be established to cope with increasing data volumes and experimental datasets. Furthermore, various analytical platforms from hyphenated mass spectrometry are employed in metabolomics which are able to generate quantitative metabolite measurements. The number of individual instruments applicable for metabolomics research being available from the different vendors (Bruker Daltonics™, Thermo™, Agilent™, Waters™) is ever increasing. A comprehensive analysis system should therefore be able to integrate and support the future developments and instrument types.

These observations lead to a list of requirements for the design of the metabolomics analysis platform termed MeltDB.

- Structured storage and annotation of the data and meta-data generated in metabolomics experiments.

- Support for multiple projects and research collaborations.

- Only secured and authenticated user access.
- Scalable preprocessing functionality and flexible pipelines for the streamlined analysis of raw input data.
- Support for measurement data from various vendors and instruments.
- Integration of statistical analysis features and visualizations.
- Multi-Omics data integration and visualization

In the following, the general system design of MeltDB is developed in order to address the listed requirements.

4.1. System design

The first step is the definition of an object relational database model with the aim to store experimental data of metabolomics experiments together with descriptive meta-data and the applied preprocessing parameters and tools.

The design of the MeltDB object model was influenced by ArMet and the recommendations of the Metabolomics Standards Initiative (MSI) workgroup. Various classes of ArMet have been adopted and beyond that the MeltDB data model also supports user access control, a more flexible, ontology based metabolomic experiment annotation, and the possibility to integrate and parameterize preprocessing algorithms and methods that can be submitted to a compute cluster. The system is realized using a three tier architecture consisting of a database layer, the business logic layer and the presentation layer (Figure 4.1a). The O2DBI software (unpublished) was used to design the data model and generate an XML document that formally describes all classes and hierarchies. Based on this formal definition, a documented application programming interface (API) was created in both Java and Perl to provide the core functionality of the MeltDB software framework. The API supplies *create, retrieve, update* and *delete* (CRUD) functionality for all modeled classes of the data model. The core functionality can easily be extended and new methods can be added to the objects on demand. Furthermore, the auto-generated API is the basis of the business logic layer and provides an object relational mapping for all modeled classes of the MeltDB data model.

4.2. Data model

An overview of the top level classes and their interaction is given in Figure 4.1b. All modeled classes and the inheritance relations are represented in the Appendix. The classes *Chromatogram*, *ChromatogramGroup* and *Experiment* represent data from metabolomics experiments. Subclasses of *Chromatogram* allow to distinguish the instrument specific properties of chromatograms originating from LC-, GC-,

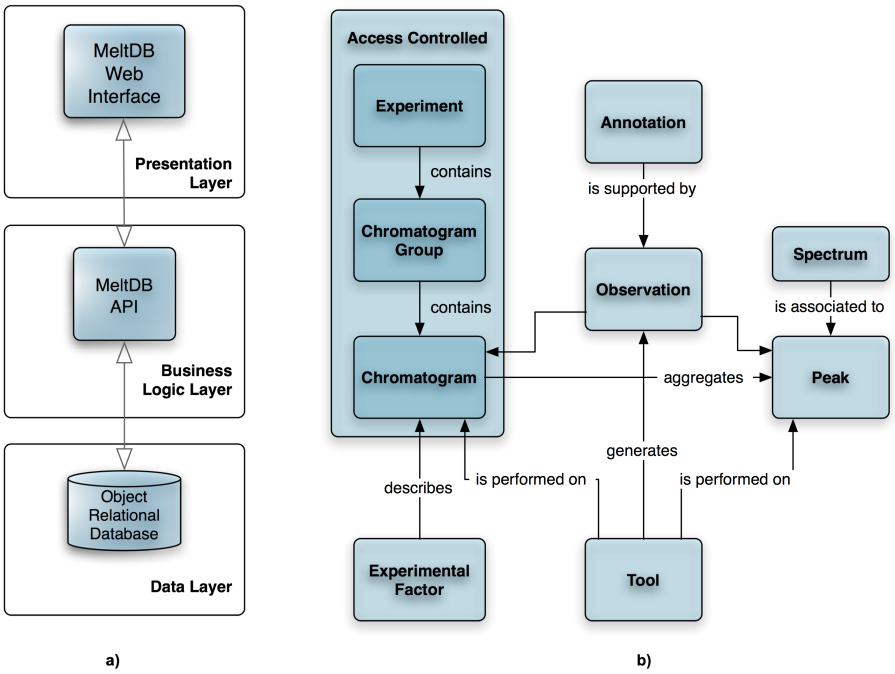

Figure 4.1.: The three tier architecture of the MeltDB framework (a) and a symbolic representation of the main data objects and their interaction (b). A representation of the complete data model of MeltDB can be found in the Appendix. The comprehensive HTML based API documentation is available via the MeltDB project web-page `http://meltdb.cebitec.uni-bielefeld.de`.

or CE-MS measurements. *ChromatogramGroup* objects aggregate chromatograms from biological or technical replicate sets.

In order to be able to support the different objectives of metabolomic experiments, four specialized subclasses of *Experiment* have been defined in accordance with the suggestions given by the MSI.

- Targeted Analysis: Detection and precise quantification of a single or a small set of target compounds within a metabolome sample.
- Metabolite Profiling: Detection and approximate quantification of a large set of target metabolites within a metabolite sample.
- Metabolomics: Detection, approximate quantification and tentative identification of as many of the compounds within a metabolome sample as possible.
- Metabolic Fingerprinting: Generation of a signature for a metabolome sample without regard for the individual compounds that it contains.

4.3. Experiment description

The MeltDB data model describes and annotates the experimental design using *ExperimentalFactors*. They represent e.g. growth conditions, quenching, extraction or sample preparation methods. Each chromatogram in MeltDB can be attributed with a list of these experimental factors, thereby the experimental conditions are annotated. The structure of the best practice recommendations of the MSI working group is integrated and can be extended dynamically. Once experimental factors have been defined in the MeltDB database, they can be reused for annotation of multiple chromatograms. The main classes for the annotation of metabolomics experiments are shown in Figure 4.2.

4.4. Software as a service

During the last years a tendency to replace the functionality of applications formerly running on single workstations through web-based applications can be observed. This concept has been termed 'Software as a Service' (SaaS) (Bennett *et al.*, 2000) and with the availability of web techniques such as AJAX and modern web toolkits, the look and feel and the responsiveness of the web applications is becoming comparable to applications running on local workstations. The SaaS approach provides various advantages over traditional applications. Updates of the web based application are centralized and there is no need to install software on the local machine apart from a recent web browser. Furthermore, the centralized data storage and backup is simplified and users can employ the functionality from everywhere and are not limited to a single analysis workstation in their laboratory. An important factor in an interdisciplinary field such as metabolomics is that researchers can

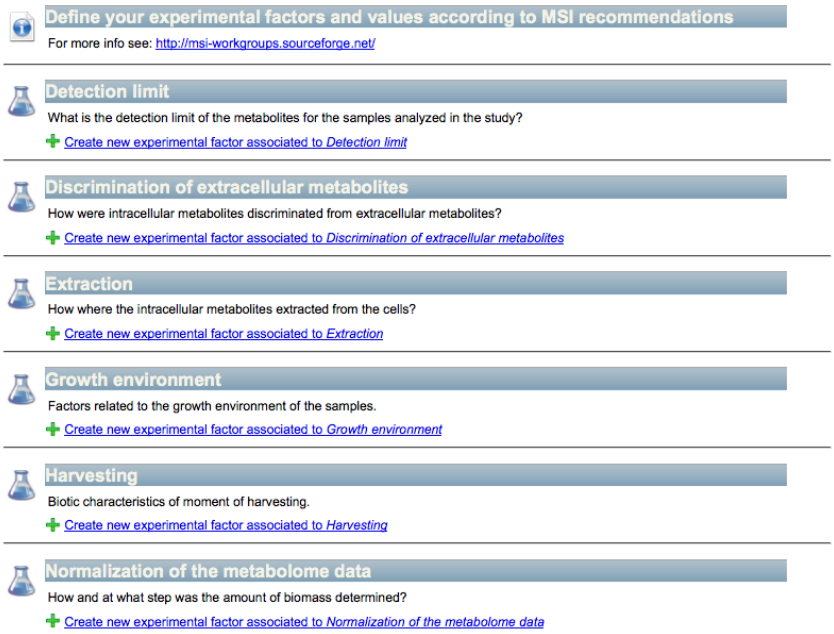

Figure 4.2.: Overview of the main classes representing the minimal information necessary for the annotation of metabolomics experiments as recommended by the MSI initiative. The MSI proposed a classification of information that has to be detailed for metabolomics experiments. Especially the sample preparation steps have a large influence on the composition and quantity of the metabolome and therefore harvesting, quenching, and extraction methods need to be specified to ensure the reproducibility of an experiment.

share experimental data and analyze their experiments in a collaborative manner. The evaluation of these advantages led to the decision to implement the MeltDB system using the SaaS concept.

4.5. Project management

The design of MeltDB as a web-based application has several implications. Predominantly, the access to the application and the stored datasets needs to be controlled via both a security and an access layer.

Providing a secure level of access control is essential for data privacy and thereby acceptance of the software system. The access control component of MeltDB should fulfill two authorization tasks. First, access to the individual project needs to be controlled, and second, access to the individual objects of interest is governed on a per-user basis. Furthermore, it should be possible to restrict the access to the available functionality of the MeltDB system for each user. Computationally intensive methods or tasks that can compromise data consistency are only allowed to developer and administrative users.

A role-based approach was chosen to accomplish this tasks. This concept has been previously employed in other functional genomics platforms implemented at the CeBiTec (Meyer *et al.*, 2003; Wilke *et al.*, 2003; Dondrup *et al.*, 2009) and proved to be well suited for the user management in distributed projects. While providing a high level of granularity to control access, it also keeps the system manageable by providing sets of predefined access rights for objects. A role provides a set of rights within the domain and by assigning a role to a user, the possible actions he can perform within a system are defined.

After the user authentication occurs at the login page of the MeltDB system, a personalized session is tracked. Access to experimental data is limited according to the predefined rights and roles i.e. a user with the role *Guest* has the right to review public datasets but does not have the right to *add experimental data*. Data security is ensured through MySQL permissions imposed on individual tables in the project databases. Higher level privileges e.g. the right to run and modify preprocessing pipelines on raw datasets are furthermore enforced by the business logic layer of MeltDB. The careful definition of user rights and roles will allow the use of the MeltDB system in a productive environment.

4.6. Access control model

The means of a relational database system are insufficient to control the access to individual database entries. The rights and privileges to read, update, create or delete can only be granted on complete tables or columns. To control individual entries in database tables, access control lists (ACL) can be employed to realize this additional layer of fine grained access control. Their application for distributed

functional genomic systems has already been described and implemented for the Emma2 transcriptomics system developed by Dondrup *et al.* (2009).

An ACL is generally a list of permissions attached to an object. The list defines who is allowed to access an individual object and which operations are accessible. In comparison to rights and privileges of the relational database system, ACLs offer a more fine grained control.

Even if the organization of metabolomics experiments and user access are comparable to the requirements for transcriptomics, an extension and adaptation of the described ACL functionality is necessary for the MeltDB system. To realize the concept of ACLs in MeltDB, the AC (Access Control) class needs to be introduced, which acts as interface and provides the required access control functionality for all experimental data stored in the MeltDB data model.

All classes inheriting the class AC will provide fine grained user and group rights similar to those of unix file systems. The owner of an *AC* object can grant the rights to read, write, or refer to his object to other MeltDB users and groups. The classes *AC::Chromatogram*, *AC::Experiment*, and *AC::ChromatogramGroup* and all their sub classes inherit the AC functionality. An overview of the relation modeled classes is given in Figure 4.3.

The functionality of the business logic layer of MeltDB implemented in Perl realizes the described functionality. As a consequence, public and trial projects for the MeltDB web server can easily be established and maintained.

The collaborative analysis of experimental data from metabolomics requires interdisciplinary exchange between researchers from different backgrounds. This process is made possible and simplified since the experimental data can be shared by granting the respective permissions to collaborating users and user groups. It is necessary to define the set of access permissions controlled by the ACLs which is presented in Table 4.1.

Each permission can be set to one out of two values, which can be either Yes or No. A combination of all permissions with an assigned value, a user or group, and an object constitutes a complete ACL. The resulting permission is derived from a combination of all ACLs assigned to the user for the given object; ACLs can be either directly assigned to a user or assigned by group memberships. As a result, many ACLs may apply for a given combination of user and object. In case of a conflict, the permissions are computed from all ACLs using a logical AND operation. The default access policy is No for each action in the action set. This results in an often desired behavior where a user cannot perform any action until it is allowed explicitly.

4.7. Tool concept

As presented in Section 3, various open software packages for data processing and analysis are available and new methods are developed on a regular basis in the computational metabolomics community. In order to integrate these tools and

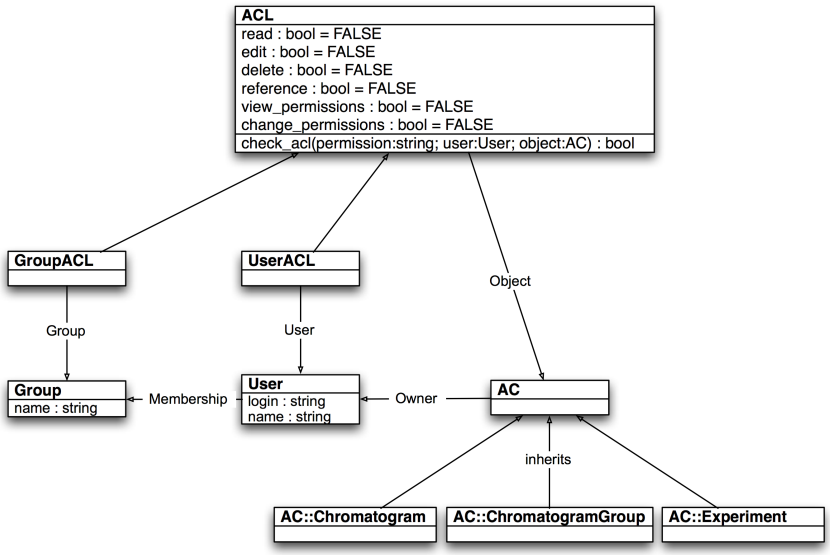

Figure 4.3.: UML-model of the database classes realizing the access control system of MeltDB. The user and group ACL classes represent the access control lists for users and groups of users and are both derived from the ACL superclass. The ACL class controls access to objects of the AC class hierarchy that represents all experimental data stored in the MeltDB system.

Permission	Description
Read	view the object and the information provided, like name, descriptions, referenced objects, etc.
Edit	change values, names and descriptions
Reference	re-use this object (mostly important for chromatograms and chromatogram groups) and link to it, a chromatogram can e.g. be referenced in another experiment
Delete	remove the object permanently
View permissions	view the ACLs assigned to this object
Change permissions	alter the permission settings by adding and removing ACLs

Table 4.1.: The set of available access permissions controlled by the ACLs in MeltDB.

methods, a general abstraction of a computational analysis method applicable to metabolomics data is specified in the MeltDB data model. Therefore, the *Tool* class is defined to represent an analysis package that can be integrated in the MeltDB system.

Each tool instance defines the type of data it can be performed on and which results in terms of objects in the MeltDB data model it generates. The available preprocessing methods do typically require a number of user defined parameters that are e.g. set via the command-line. For this, tool instances can aggregate *ToolParameter* objects that specify the type, name and range of the parameters. The support for the ToolParameter objects in combination with the Tool class makes it possible to store multiple parameterizations or instances of the same preprocessing method. Each one can be adapted to datasets from e.g. different instruments. As the input and output of any tool is strictly typed, the combination of tools to pipelines becomes possible.

The existing preprocessing functionality necessary for hyphenated mass spectrometry can be classified in peak detection, peak and compound identification, chromatogram alignments as well as importers for proprietary analysis methods. Thus, subclasses of the *Tool* class are defined to represent the respective properties of these distinct analysis tasks. As an example, one *Tool::PeakDetection* object in the MeltDB database is a parameterized instance of an integrated peak detection algorithm. The associated tool parameters define e.g. the noise window, the peak width and the baseline correction method.

Depending on the preprocessing method and the implemented algorithm, large computational requirements with respect to runtime and memory consumption can be expected. As researchers can select one or more suitable tool instances for the preprocessing of chromatograms or experiments, the submission of the actual computation to a compute cluster makes the analysis scalable and efficient. MeltDB utilizes the Distributed Resource Management Application API (DRMAA) [1] as a high-level API specification for the submission and control of jobs to Distributed Resource Management Systems (DRMS).

After submission, the combination of a tool and the data represented in the MeltDB data-model is represented as *Job* object. A job tracks the computational progress and stores potential errors and warning messages. An overview of the described classes is given in Figure 4.4.

Each realized tool class implements the abstract interface method *run*. This typically comprises a system call to the binary of the tool together with the specified tool parameters followed by a parsing step of the generated results. In order to connect these results with the chromatograms and experiments stored in the MeltDB database, *Peak*, *Observation*, or *Annotation* objects are created.

An observation associated to a chromatographic peak represents e.g. the computed retention index or a match to a mass spectrum from the GMD database. Whereas observations may contain hypothetical and contradicting information, an-

[1] http://www.ggf.org/documents/GWD-R/GFD-R.022.pdf

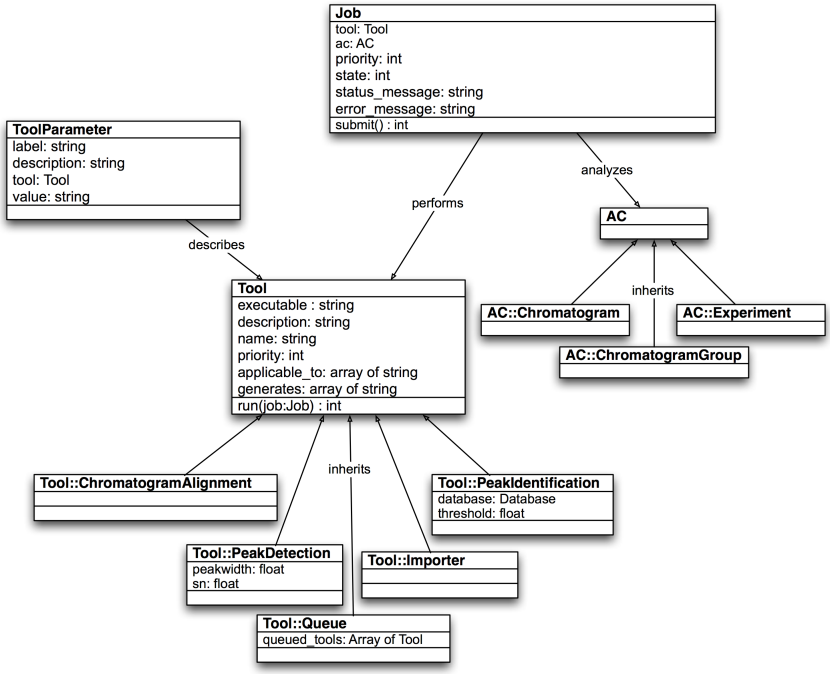

Figure 4.4.: UML-model of the database classes realizing the tool and job system of MeltDB. Tools can be performed on experimental data represented in the AC subclasses. The parameterization of a *Tool* instance can be defined via associated *ToolParameter* objects. The class *Job* implements the functionality to submit the combination of a *Tool* and *AC* object to a compute cluster using DRMAA and to track the progress of the computational analysis.

notations are used to aggregate approved information. Observations and annotations are linked to the creating tool. Thus, generated results are reproducible and transparent. The general concept of tools, observations and annotations is presented in Figure 4.1b.

An additional advantage of the presented observation/annotation concept is the possibility to store and compare the effect of different parameterizations of an algorithm or software package. Typically, the comparison is most intuitive for the researcher when the generated results, i.e *Observations* and *Annotations* are projected on TIC or EIC or visualizations of the raw chromatographic data as presented in Figure 2.8.

4.8. Statistical analysis

A wealth of multivariate statistical analysis methods is available via the free R project for statistical computing. Additionally, the freely available BioConductor repository already provides various analysis packages tailored to functional genomics. This led to the decision to integrate R functionality in MeltDB. Thereby, the support for multivariate statistical analysis of metabolomics experiments can be realized. The direct connection of the MeltDB API implemented in Perl with the R framework is possible via the free RSPerl interface[2]. Experimental data matrices stored in MeltDB can be converted into R data objects which can be analyzed directly within the R software. All generated R visualizations based on the experimental data matrices are rendered to static images in the PNG format which can be easily presented via the MeltDB web interface.

4.9. Data integration

An important prerequisite to systems biology and multi-*Omics* analyses is the integration of diversified experimental and annotation data. These datasets are typically stored in distributed life-science databases. The necessary integration of these data is often complicated by the lack of semantic knowledge about the content of specific database tables and inconsistencies in the description and labeling of identical datasets. Nonetheless, if a system has been realized that integrates various databases, problems may arise when changes or updates of the individual database structures are performed. If an embedded database is extended by new data and the data structure of the back-end needs to be changed, this usually affects the import procedures and the front-end applicability as well as the stability and integrity of the whole system (Philippi and Köhler, 2006).

To overcome these limitations, several strategies exist for the integration of heterogeneous datasources found in functional genomics research. One example is the BRIGEP approach described by Goesmann *et al.* (2005), which focuses on the inte-

[2] `http://www.omegahat.org/RSPerl`

gration of three functional genomics systems, namely GenDB (Meyer *et al.*, 2003), ProDB (Wilke *et al.*, 2003), and Emma Dondrup *et al.* (2003). All three systems share a similar software infrastructure based on a object relational database systems. To this end, a tight integration of the three systems was achieved on the level of the individual APIs. The previously implemented BRIDGE layer (Goesmann *et al.*, 2003) directly connects the APIs and allows to share objects that are stored in the object relational databases. As the API functionality encapsulates the data access completely, no direct connection to the database tables is necessary and changes and updates can be implemented in a transparent manner.

An alternative strategy was developed for the CoryneCenter platform Neuweger *et al.* (2007) for the integrated analysis of corynebacterial genome and transcriptome data. Here, a Web Services based approach to integrate heterogeneous software frameworks for functional genomics was employed. Web Services are software interfaces that interact via a network connection using XML-based messages. The XML messages either contain queries (i.e. function calls) or the corresponding results. The transfer of these messages is generally performed using the HTTP protocol. The message structure is described using SOAP (Simple Object Access Protocol). The higher level description of an entire Web Service is defined using the Web Service Description Language (WSDL). By following theses standards, any software implementing a SOAP interface can retrieve data directly from the provided service. Further information on the technical details of Web Services and SOAP has e.g. been detailed by Curbera *et al.* (2002). For CoryneCenter, the GenDB genome annotation system, the Emma transcriptomics suite and the CoryneRegNet data warehouse (Baumbach *et al.*, 2006) were connected to facilitate integrated analyses of gene regulatory networks based on microarray experiments.

From the experiences obtained during the development and implementation of these frameworks, the data integration strategy for MeltDB is derived. As the MeltDB API is designed using the O2DBI software, the requirements for a BRIDGE connection are automatically fulfilled.

Additionally, Web Services will also be employed to provide programmatic access to the experimental datasets stored in the MeltDB database. To make sure that only authenticated access is possible, the Web Services interface implements the previously described GPMS functionality. Thereby, a standardized and secure access to metabolomics data sets is possible. Furthermore, the seamless integration into other applications will provide added value for comprehensive data analyses on multi-*Omics* experiments.

CHAPTER 5

Implementation and Methods

Based on the formal design presented in the previous chapter, the implementation details of MeltDB and novel methods that could be realized with the system are presented. The structure of this chapter follows a possible workflow from raw data to the statistical analysis of numerical data matrices representing metabolite concentrations. The integration and visualization of the generated data via metabolic pathway maps concludes this chapter.

5.1. Supported input formats

MeltDB supports the established open data formats netCDF, mzData, and mzXML (Unidata, 2008; Orchard *et al.*, 2007; Pedrioli *et al.*, 2004) that are widely used in metabolomics and proteomics. A variety of conversion tools for these standards are available for vendor specific file formats[1]. Most of the raw datasets generated by hyphenated MS instruments can therefore be imported and analyzed using the system.

In order to limit the data being represented in the object relational database, unprocessed raw data in the supported data formats is stored as files and referenced together with meta information (format, sample volume, acquisition time, etc.). The *factory methods pattern* (Gamma *et al.*, 1995) is applied to create mass spectra and total or extracted ion chromatograms from the raw data of chromatograms. Factory methods define an interface for creating an object, but let the subclasses decide which class to instantiate. For each of the supported data formats (netCDF, mzDATA and mzXML), subclasses are implemented that generate the TIC, EIC

[1] `sashimi.sourceforge.net`, `tools.proteomics.org`

and Mass Spectra objects. Mass spectra can be retrieved from the raw data files by scan number or retention time, extracted ion chromatograms can be extracted for nominal m/z values or for floating m/z values together with a bin width parameter.

5.2. Raw data visualization

The availability of standardized interfaces to the different raw file formats provided by the MeltDB API allows to realize generic visualizations such as TIC views or chromatogram alignments for whole experiments as shown in Figure 5.1.

Visualization of raw data already allows researchers to assess the quality of single measurements or whole experiments. More beneficial is of course the combination of the visualizations with the result of the preprocessing and import functionality of MeltDB that will be presented in the following sections. Thus, the implementation of all raw data visualizations is designed to map additional information, e.g. the chemical identity of a detected feature.

5.2.1. Experimental overview

To obtain statistically significant results from metabolomics experiments, technical and biological replicate measurements are necessary. This implies that experiments typically contain large numbers of chromatograms. To obtain an overview of the raw data, the visualization of the chromatographic properties of dozens of measurements needs to be both concise and informative. Since the visualization of TIC plots contains the necessary information to identify qualitative differences but becomes too bulky for more than a few measurements, it is not suited to represent an overview of complex experiments with multiple replicates and conditions.

A brief but comprehensive chromatographic visualization realized using the MeltDB API employs the following transformation: The range of all TIC intensities found in a complete experiment is computed and mapped to gray scale intensities. White (RGB 1.0,1.0,1.0) represents the highest intensity found and black (RGB 0.0, 0.0, 0.0) represents the minimum intensity. The advantage to compute the intensity range across all measurements is that both relative differences can be visualized and the effect of varying sample amounts becomes immediately visible.

The peak intensities found in GC- and LC-MS measurements span several orders of magnitude and the largest peaks found in the analysis are rarely the most informative. Therefore, a scaling of all intensities is applied to improve the information content of the visualization. The power and log scaling are typical data transformations that emphasize small values (van den Berg *et al.*, 2006). After evaluation of the effect of both scalings for the visualization of various experiments, the log transformation was chosen as the default method for the comparative visualization of multiple chromatograms. As will be presented in Section 5.5.5, the overview visualization can also be enriched with additional information, e.g. the peaks de-

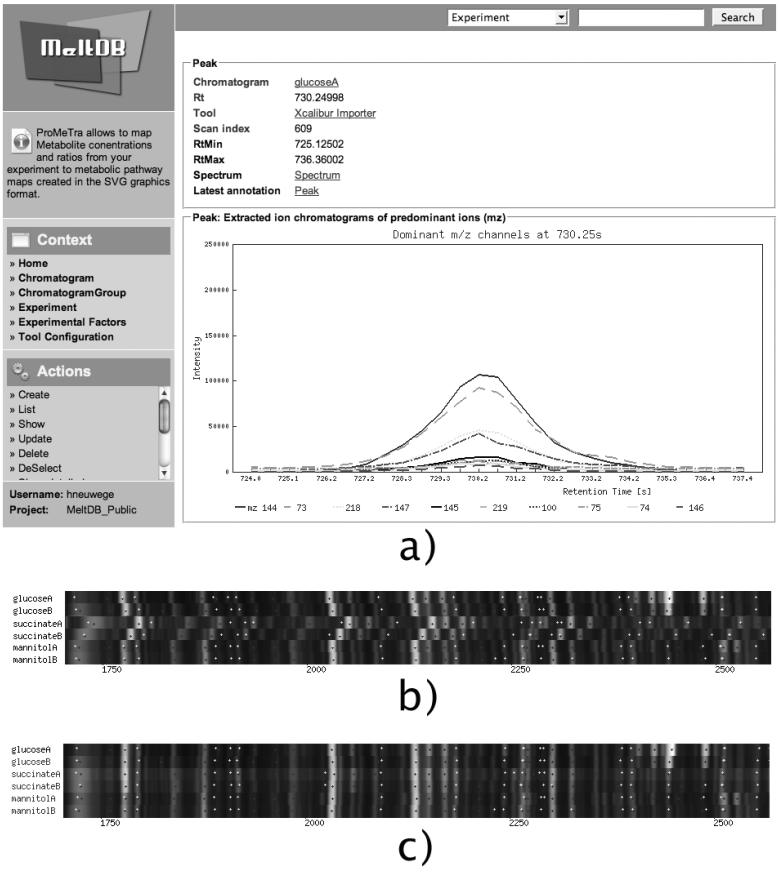

Figure 5.1.: The standardized access to the raw data stored in MeltDB allows to easily implement generic visualizations for whole Experiments, Chromatograms, or Peaks and Mass Spectra. The dominant mass channels of a chromatographic peak are shown exemplarily in the upper screenshot (a). Intensity coded TIC views of unaligned (b) and aligned chromatograms (c) can also be generated and allow to easily spot differences between measured peak intensities.

tected in the raw chromatograms. Additionally, links to connected entries in the MeltDB database are integrated automatically in the web based user interface.

5.3. User interface

The functionality of MeltDB can be accessed through a platform independent web interface. Perl CGI scripts running on an Apache web server dynamically create the HTML content and manage the authentication and the ongoing user sessions. The interactivity of the web application is increased through the use of mod_perl, JavaScript and AJAX which also results in fast access to all objects stored in the MeltDB database. In order to obtain a flexible and extendable web interface, the *Model-View-Controller* design pattern (Gamma *et al.*, 1995) is employed for the generation of the actual HTML content. Generic views applicable to all objects in the MeltDB database provide a tabular representation of e.g. the annotations associated to a peak. Specialized views realizing the same interface have been added to the system. Thus, the TICs and mass spectra of MeltDB objects such as chromatograms, peaks, or complete experiments can be visualized. Examples are presented in Figure 5.1.

Apart from the visualization of raw data and results, investigators can import, organize and annotate their experimental datasets according to the recommendations of the MSI. The user interface gives access to the preprocessing and import functionality included in MeltDB which will be presented in more detail in the following sections.

5.4. Preprocessing

After the import, organization and annotation of raw chromatographic datasets in the MeltDB database, data preprocessing is necessary to transform raw data into numerical matrices representing the quantified and identified metabolites in all measurements. A possible workflow from raw data to the statistical analysis of numerical data matrices representing metabolite concentrations is depicted in Figure 5.2. This figure does presents the functionality provided by MeltDB. In order to be able to support a wide variety of analytical platforms, no strict preprocessing pipeline is enforced by MeltDB. The system does in fact allow to complement results from vendor specific software systems such as ChromaTOF® or Xcalibur® with academic tools that e.g. perform peak detection (metaB, XCMS, centWave) or chromatogram alignments (XCMS, ChromA). The most simple approach to use MeltDB would be to import peaks that have already been quantified and identified by e.g. ChromaTOF® and use them as basis for further statistical analysis and data integration. Nonetheless, integrated methods can be employed by the user to e.g. detect peaks missed by vendor specific software systems or to compute chro-

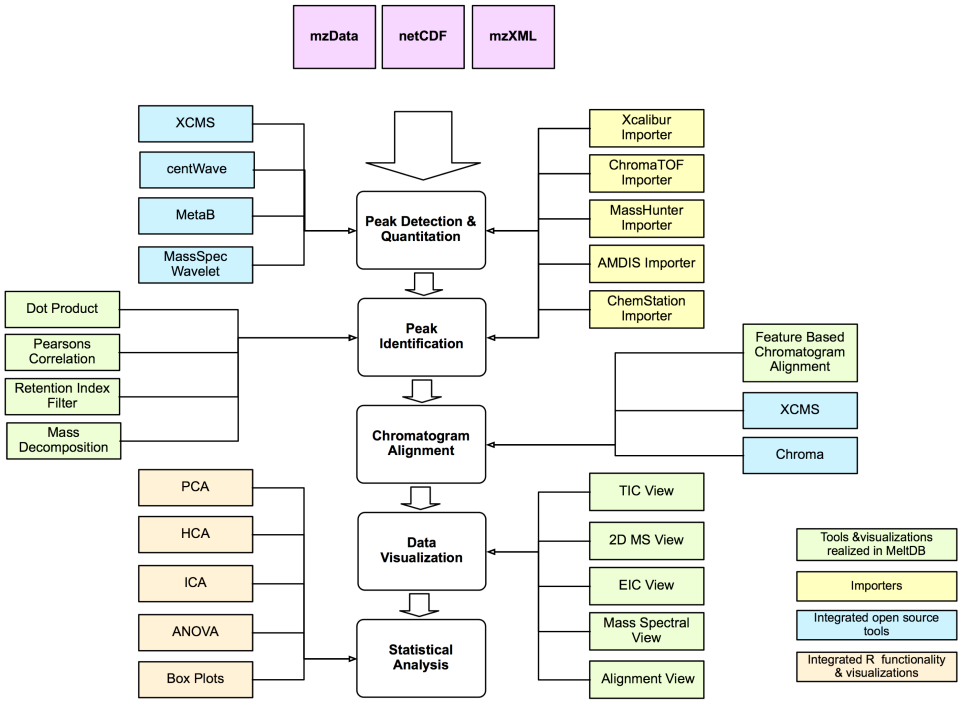

Figure 5.2.: preprocessing steps for metabolomics experiments together with contributing tools and importers integrated into MeltDB. Visualizations of raw data and statistical analyses are detailed as well.

matographic alignments as a basis for profiling analyses. MeltDB therefore allows to combine, compare and evaluate the results of different preprocessing methods.

5.5. Implemented Tools

With the MeltDB API providing access to Mass Spectra, EICs and TICs, the implementation of preprocessing methods and own tools is simplified and will be demonstrated by several examples in the following.

5.5.1. Peak detection

Smith *et al.* (2006) implemented in their XCMS application a peak detection method applying matched filtration in combination with a signal-to-noise threshold to extract de-noised and background corrected peaks. XCMS has been primarily developed for the profiling analysis of LC-MS datasets, but as the evaluation of the peak detection method on GC-MS measurements showed promising results, extensions were implemented to make the approach applicable for GC-MS chromatograms as well. Since, in contrast to LC-MS, multiple ions (50 - 250) can be measured for the same eluent in GC-MS with EI fragmentation (Lüdemann *et al.*,

2008), the original XCMS implementation fails to group these and treats them individually in the downstream analysis. To compensate this missing feature, detected peaks of extracted ion channels with overlapping retention time windows are analyzed for correlation of their abundance vectors in the MeltDB wrapper class *Tool::PeakDetection::XCMS*. This wrapper class also manages the execution of the XCMS functionality on a compute cluster and imports the generated results into the MeltDB database model. The background corrected intensity and area of the detected peaks is represented in MeltDB via *Observation* objects. To summarize, with the extended XCMS integration, peak detection for hyphenated mass spectrometry could be realized in MeltDB.

5.5.2. Mass spectral database search

Various methods for metabolite identification based on mass spectral comparison have been evaluated by Stein and Scott (1994). The recommended similarity measure S is the dot product between the vector representation of database and query spectra being calculated as follows:

$$S = \sum_j \hat{I}_{library,j} \hat{I}_{query,j} \tag{5.1}$$

$\hat{I}_{library,j}$ is the normalized intensity of the jth m/z bin of the library spectrum, and $\hat{I}_{query,j}$ that of the matching bin of the query spectrum. The length of both vectors is normalized to 1, and S is therefore always between 0 and 1. The latter value indicates identical spectra (Lam *et al.*, 2007).

Mass spectral libraries such as the NIST Wiley (2005) or GMD (Kopka *et al.*, 2005) can be parsed by MeltDB and are used to generate observations for the chemical identity of chromatographic peaks. The mass spectrum located at a peak apex is therefore extracted from raw data and compared against all database spectra. The similarity S is used to select database spectra above a given threshold $0 < t < 1$ for which the associated information such as compound name, Chemical Abstracts Service (CAS) Number or KEGG Compound ID are again connected to the peak via observations. The complete functionality is encapsulated in the *Tool::PeakIdentification::Cosine* class of the MeltDB API. As mass spectral information alone is insufficient for the unambiguous identification of structurally very similar compounds such as L-leucine and L-isoleucine, additional information such as the retention time of the eluents needs to be exploited.

5.5.3. Retention index computation

The GMD contains mass spectra for known compounds associated with information on Kovat's Retention Indices (RI) (Ettre, 1994). By combining the peak identification tool with a specialized retention index library, MeltDB is able to identify peaks representing retention indices (e.g. dodecane or pentadecane) that

have been added to a biological sample in GC-MS analysis. Relative to the retention time of these identified standards, retention indices can be computed by the *Tool::PeakIdentification::RI* tool in MeltDB for all other peaks detected in a chromatogram (RI_{peak}). By combining the deviation of retention indices from the database RI_{db} with the mass spectral similarity S, the number of false positive database matches for a given peak can be greatly reduced (Kopka *et al.*, 2005).

5.5.4. Compound identification

In order to combine the information from mass spectral similarity and the deviation of the measured Retention Index and the RI information found in the GMD database, a filtering function S' is employed.

$$S' = \frac{S}{e^{(\frac{RI_{peak}-RI_{db}}{w})^2}} \tag{5.2}$$

The function S' is implemented in the *Tool::PeakIdentification::AutoAnnotator* class to exclude mass spectral matches if a large difference of RI_{peak} and RI_{db} can be observed.

Through an adjustment of parameter w, the tolerated deviations of the retention indices according to the attributes of the chromatographic instrument can be controlled. The maximal deviation of the RI values for known compounds can be determined experimentally. For GC-MS measurements on a Thermo Finnigan Ion Trap instrument, RIs were found to deviate less than 30 units (Pühse, unpublished). The filtering of the potential matches to a query peak using S' with the retention index deviation parameter w efficiently removes false positive mass spectral matches found in the GMD.

The combination of mass spectral database search, the computation of Retention Indices and the integration of both results with the *Tool::PeakIdentification::AutoAnnotator* tool can be realized in MeltDB using the *Tool::Queue* functionality. The three described preprocessing methods are consecutively executed for single chromatograms or complete experiments by a *Tool::Queue* instance. Thereby, the analysis of metabolomics experiments can be standardized and simplified for the researcher.

5.5.5. Support for non-targeted profiling analysis

The established method to compute similarities between mass spectra has been implemented modularly and can be reused in various approaches. The first one presented here is the support for non-targeted profiling analysis in MeltDB. A general aim of profiling analyses is the identification and quantification of the metabolites present in an organism. For compounds with previously recorded mass spectra, the identification can be based on mass spectral similarity and retention time or retention index information. Chromatographic peaks that on the one hand can recurrently be detected in the experimental measurements but on the other hand

have no matching mass spectra can not be identified unambiguously. These peaks are therefore typically excluded in the generation of quantitative data matrices. To make the quantitative information of these peaks amenable to downstream statistical analysis, they need to be registered across the chromatograms of an experiment. Therefore, an association based on mass spectral similarity and common retention time or retention index characteristics needs to be constructed. Furthermore, a quantification of the peak area and the peak intensity has to be computed.

A corresponding strategy has been described in the TagFinder application by Lüdemann *et al.* (2008). As this software is limited to the netCDF input format and ChromaTOF® result files, it can not be included in the generic preprocessing pipelines of MeltDB.

Therefore, a novel tool to generate data for non-targeted metabolite profiling analysis was established that may either be performed on previously aligned chromatograms or on raw data alone. The only prerequisite is the initial detection of EIC peaks and the association of co-eluting peaks to common parent peak objects in the MeltDB data model. As this can be readily achieved by the previously described XCMS integration, the implementation of the non targeted profiling analysis was easily realized.

Method

For a given experiment consisting of a set $C = \{c_1, \ldots, c_n\}$ of n chromatograms, the initial step is the selection of a reference peak p from an arbitrary chromatogram c_i, $1 \leq i \leq n$. The retention time interval $I(p) = [rt(p) - \triangle t, rt(p_r) + \triangle t]$ is used to select proximal peaks in all other chromatograms c_j, $1 \leq j \leq n$, $j \neq i$, where the function $rt(p)$ returns the retention time of peak p. The value of $\triangle t$ is user defined and defaults to 10 seconds. If a multiple chromatogram alignment has been computed beforehand, the retention time window can be narrowed accordingly.

The result of the search in the retention time interval $I(p)$ is a list of peaks l_j found in each chromatogram c_j. Overall, one obtains the set L of $n-1$ peak lists, $L(p) = \{l_1, \ldots, l_{i-1}, l_{i+1}, \ldots, l_n\}$. In order to find the best matching peak from each list l_j, the mass spectrum at the apex of p is matched against the mass spectra at the peak apices of l_j using the previously described cosine score. If the mass spectral similarity of the best matching peak is above a conservative threshold t, this peak is labeled m_j and kept for further analysis. This results in a list $M(p)$ of up to $n-1$ associated peaks.

Quantitation of the reference peak p and the associated peaks in $M(p)$ is the next necessary step. The use of the TIC intensity or TIC area is problematic because of the influence of co-eluting peaks and potential background and noise that accumulates across several mass channels. In general, the use of a single representative EIC peak is recommended for quantification. To select a common EIC channel, the EIC peaks present in all peaks from $M(p)$ and p are counted and the group of the most abundant EICs is selected. In case that more than one such EIC group exists, the peak group with the highest median intensity is chosen. This

approach reduces the influence of chromatographic noise and aims at selecting the most robust feature to represent the peak group quantitatively.

As all subsequent statistical analysis methods are highly sensitive to missing data and outliers, the annotation of associated peaks will only occur if a sufficient subset of all chromatograms ($> 80\%$) in the experiment is covered.

Finally, a unique label is created for each registered peak group which contains the median retention time and the nominal mass of the EIC channel (e.g. 'TAG RT 1221.4s MZ 314'). Each peak is annotated using this label in MeltDB and the intensity and area of the selected EICs are represented as associated observations. Additionally, p_r and the annotated peaks are flagged such that they are not used in further iterations of the profiling approach. The next non-flagged reference peak is selected and the approach is repeated until all peaks are flagged.

Evaluation

The evaluation of the profiling approach was performed in comparison to a targeted analysis of human blood plasma samples using the Xcalibur® software with an established processing method defined by an expert researcher.

The manually curated processing method of the Xcalibur® analysis detects 43 metabolites in all of the 20 chromatograms of the experiment. Unfortunately, seven of these metabolite peaks are missed in more than five of the chromatograms analyzed, which further reduces the number of features that can be used in statistical analysis methods. After XCMS peak detection was performed on the dataset, more than 200 peaks could be identified for each individual chromatogram with a signal-to-noise threshold larger than five. This indicates that the Xcalibur® analysis is missing a large proportion of the metabolites in the sample. The untargeted profiling approach was able to unambiguously identify and quantify 122 of the more than 200 features detected by XCMS across all the 20 measurements. None of the annotated features had more than four missing peaks. A visualization of the results of both approaches mapped to the raw chromatographic data is presented in Figure 5.3. Whereas the Xcalibur® analysis does provide a metabolite identification, the profiling analysis is more comprehensive and gives more insight in the metabolic changes and differences of an experiment. The coverage and overlap of the detected peaks from both approaches is presented in Table 5.1.

Figure 5.3.: The TIC visualization the first part of 20 unaligned chromatograms of human blood plasma samples is enriched with information on compounds detected by an Xcalibur® analysis (red peak markers) and the automatically identified and quantified results of the non-targeted profiling analysis of MetDB (green peak markers). The non-targeted approach detects and quantifies various peaks that are missed by the Xcalibur® analysis. Blue peaks could not be annotated using the profiling approach because either the mass spectral similarity is to low or no common EIC for quantitation is present.

The manual inspection of the approximately twenty metabolites that could only be detected by Xcalibur® showed that the associated peaks featured predominantly small intensities below a signal-to-noise level of five and could therefore not be detected by the specified XCMS approach.

Nonetheless, the automated profiling approach covered all of the abundant metabolites of the Xcalibur® analysis and both methods were able to generate data matrices that showed clear metabolic differences (according to PCA and ICA analysis) between the two analyzed sample groups (data not shown). Interestingly, the non-targeted quantitation automatically selected the same EIC channels for several metabolites as specified in the Xcalibur® processing method defined by a human expert. The list of metabolites with matching TAGs is given in Table 5.2

5.5.6. Feature based chromatogram alignment (FBCA)

The computation of alignments is an optional processing step for profiling analyses and can be used to compensate drifts in the retention time of compounds occurring in the chromatograms of a large scale metabolomics measurement. Various algorithms have been proposed as described in Section 2.5.1, the feature based instances rely on peak detection methods (XCMS, Robinson's Method). As the quality of peak detection and quantitation is strongly dependent on the dataset under study, and the values of noise, peakwidth and smoothing parameters have a great influence, artifacts in the alignments can be produced even for theoretically optimal algorithms. Therefore, a robust feature based alignment approach that is independent of the previously employed peak detection method has been implemented using the MeltDB API.

Central to the approach is the search for chromatographic peaks having a characteristic and unique mass spectrum. In the following, two chromatograms X and Y are represented by their lists of detected peaks $L_X = [x_1, x_2, \ldots, x_n]$ and $L_Y = [y_1, y_2, \ldots, y_m]$. The peaks in the lists are ordered by their retention time. If an all-against-all comparison of the mass spectra of L_X and L_Y is computed using Equation 5.1, a similarity matrix M of size (n, m) is generated. A heatmap representation of such a matrix M is given in Figure 5.4. Now, characteristic peaks in L_X will only have one matching peak in L_Y exhibiting a high mass spectral similarity. The characteristic peaks represent unique substances occuring only once in both chromatograms and can therefore act as *anchor points* for chromatographic alignments. The first criterion for an anchor point is that the maximal similarity value in a row of the matrix is also the maximal similarity value in the respective column. If this criterion is met, the two matching peaks $x_i, 1 \leq i \leq n$, and $y_j, 1 \leq j \leq m$, are termed a *bidirectional best hit*. Especially in chromatographic areas with high background or for highly abundant compounds that saturate the detector, multiple peaks with very similar mass spectra can be detected. In Figure 5.4 this can be observed in the lower right part where multiple neighboring peaks have a high pairwise mass spectral similarity. To improve the robustness of the approach, the mass spectral similarity of potential anchor points has to fulfill an

Chromatogram	Only Xcalibur®	Both	Only Profiling
P_IP_14	18	31	91
P_IP_15	19	24	93
P_IP_16	17	28	88
P_IP_17	19	26	86
P_IP_18	16	27	90
P_IP_19	23	25	86
P_IP_20	22	27	86
P_IP_21	18	29	91
P_IP_23	21	28	87
P_IP_24	20	26	88
P_S_102	20	29	76
P_S_103	17	28	92
P_S_104	18	28	91
P_S_105	19	27	93
P_S_106	19	26	93
P_S_107	18	27	91
P_S_108	17	24	86
P_S_109	21	26	92
P_S_111	18	26	93
P_S_113	17	30	88
Total	377	542	1781

Table 5.1.: Comparison of the Xcalibur® and Profiling analysis with respect to covered features. The first column shows the name of the chromatogram, the second contains the number of features that could only be detected and identified in the Xcalibur® analysis. The third column shows the features that were identified by both tools whereas the last column gives the number of features identified by the MeltDB profiling analysis but not by Xcalibur®.

Compound	TAG (RT)	EIC
Alanine	536.3	116
Valine	706.2	144
Leucine	794.4	158
Iso-leucine	827.9	158
beta-Alanine	1035.2	248
Pyroglutamat	1181.7	156
Glutamat	1316.9	146
Glycerol-3-phosphate	1502.5	357
Tryptophan	2028.6	202
Sucrose (8TMS)	2424.0	361
Trehalose	2516.5	361

Table 5.2.: Features found by the MeltDB profiling analysis that exactly match the manually curated Xcalibur® results. The first column of the table lists the compound name specified by Xcalibur®, the median retention time of the matching tag is found in column two. In column three, the nominal mass of the ion used for quantitation by both Xcalibur® and the automated profiling is presented.

additional outlier criterion with respect to the values in column j and row i of the similarity matrix M.

Usually, an outlier can be found at the extreme values of a distribution, and the rejection of suspect observations needs to be based on an objective criterion. This can be achieved by using non-parametric tests for the detection of outliers such as Dixon's Q-test. It states if one (and only one) observation from a small set of replicate observations can be identified as outlier. The Q-test is based on the statistical distribution of *subrange ratios* of ordered data samples, drawn from the same normal population. Hence, a normal (Gaussian) distribution of data is assumed whenever this test is applied.

To test if the bidirectional best match also represents an outlier with respect to the mass spectral similarity, the computed similarities of peak x_i with all peaks in L_Y are arranged in ascending order:

$$s_1 < s_2 < \cdots < s_m$$

Note that s_m is the value found in the similarity matrix at position M_{ij} and the other values $s_1, \ldots, s_{m-1}$ originate from the remaining entries in row i of matrix M.

Then the **experimental Q-value** (Q_{exp}) statistic is calculated. This is a ratio defined as the difference of the suspect value from its nearest one divided by the range of the values (Q: rejection quotient). Thus, for testing s_m as possible outlier the following Q_{exp} value is used:

$$Q_{exp} = \frac{s_m - s_{m-1}}{s_m - s_1}$$

The obtained Q_{exp} value is compared to a critical Q-value (Q_{crit}) found in precomputed tables. This critical value corresponds to the confidence level (CL) the user requests (defaults to CL=95%). If $Q_{exp} > Q_{crit}$, then the match of x_i to y_j can be characterized as an outlier.

If the comparison of the mass spectral similarity of y_j to all peaks in L_X also satisfies the outlier criterion, the tupel x_i and y_j will be used as a robust anchor point for the generated alignment. An advantage over Robinson's algorithm is that the presented approach can be applied even if large deviations of retention times between chromatograms occur due to e.g. changes in the instrument setup.

Interpolation between anchor points

The retention times of the peak pairs of neighboring anchors are used to either compute a linear or cubic spline interpolation function for all intermediate scans. The cubic splines generate a continuous function from a set of nodes (the anchors) and in comparison to linear regression they ensure that the resulting function traverses through the nodes. To obtain a rough estimate for the parts of the chromatograms before the first anchor and after the last anchor, the extrapolation of the first and last segment of the interpolation function is used.

It is not trivial to specify an objective measure for the quality of an alignment, but an intuitive way to assess the results is a concise visualization of the aligned chromatograms. The implementation of the MeltDB Experiment TIC visualization was therefore extended to support the visualization of both pairwise and multiple alignments. In this approach, the anchor points are aligned and the generated interpolation function is used to project all intermediate TIC intensities at corresponding positions in the visualization. The results of the FBCA alignment tool employed on 30 GC-MS measurements of wheat samples are presented in Figure 5.5. In the interactive MeltDB web interface, the alignment visualization is enriched with peak markers that allow to access associated observations and annotations.

5.6. Importer

For preprocessing methods that can not be employed in an open source environment, integration has been achieved through individual importing functionality using the previously described Tool concept. For the following file formats importers have been implemented.

- AMDIS reports (Text format)
- Thermo Xcalibur® reports (XLS format)
- LECO ChromaTOF® reports (Text format)

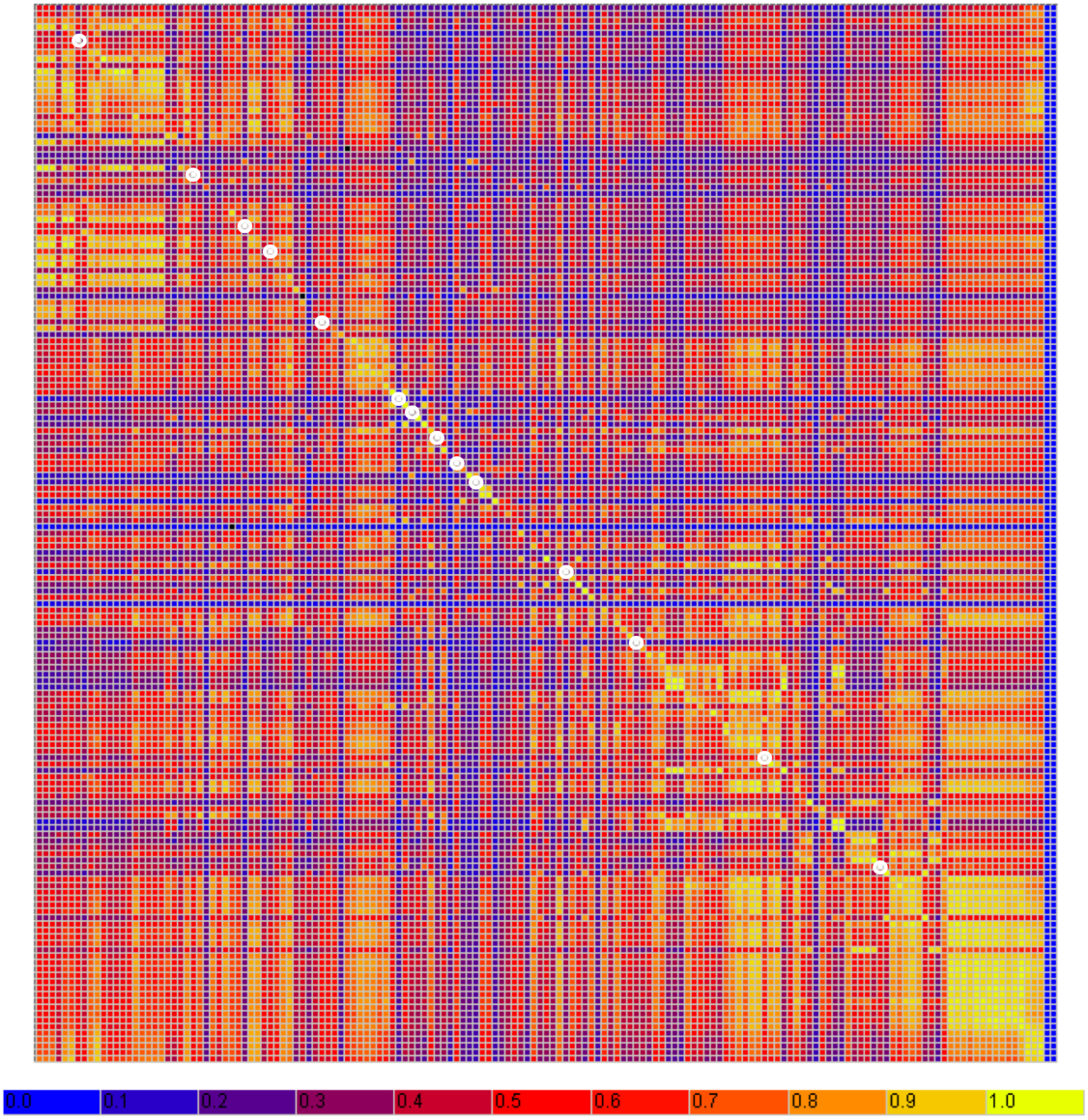

Figure 5.4.: The image shows the similarity matrix of mass spectra at predicted peak apices of two chromatograms. Detected anchor peaks fulfilling the bidirectional best hit and outlier criteria are are highlighted by white circles. A block of high peak similarities represented in the lower right part of the matrix represents neighboring peaks with very similar mass spectra. The bidirectional best hit criterion alone easily leads to mismatched anchors in such areas, using the additional *outlier* criterion will avoid anchors in these ambiguous regions.

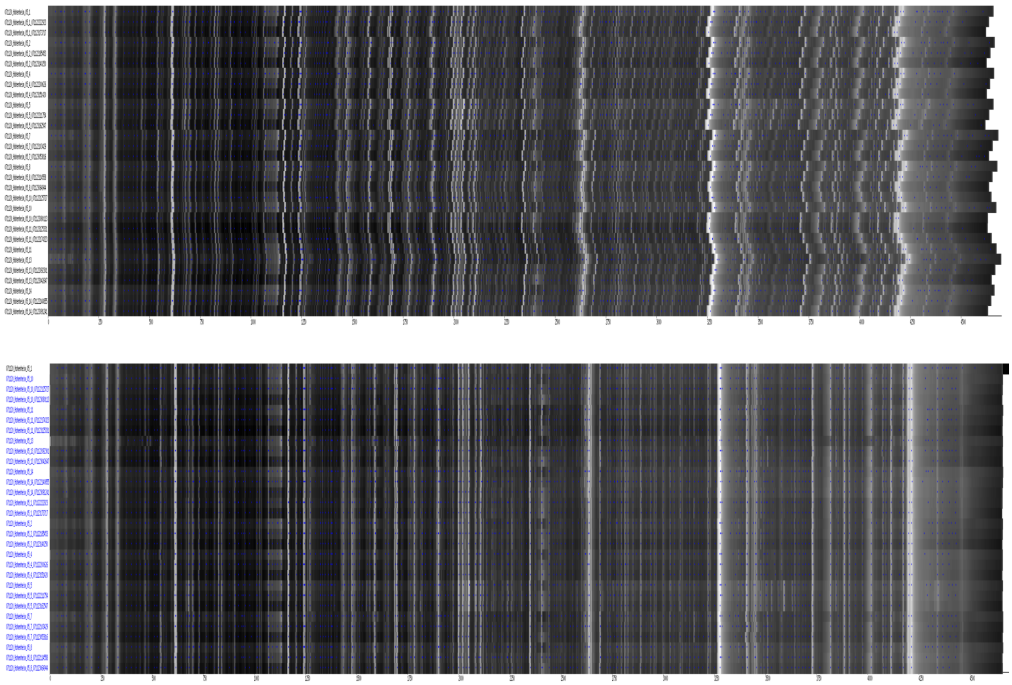

Figure 5.5.: Visualization of unaligned and aligned GC-MS measurements using the MeltDB TIC visualization. The upper part shows 30 unaligned GC-MS measurements, the lower part represents the same measurements being aligned and interpolated using the MeltDB FBCA tool.

- MassHunter reports (Text format)
- ChemStation peak lists (Text format)

The importers transform information on the chromatographic peaks predicted by external software tools into the MeltDB data model. Information associated to peaks such as chemical identity, peak area and intensity are represented using observations and annotations. The extensible importer concept realized in the MeltDB data model makes it easy to import additional text based vendor specific formats. Together with the support for netCDF, mzXML and mzData metabolomics experiments conducted on various GC-MS and LC-MS instruments can be analyzed using MeltDB.

5.7. Implemented statistical analysis features

As described in the previous chapter, R has been directly integrated with MeltDB using the RSPerl interface. This avoids the cumbersome and error prone conversion of data tables from proprietary software packages into a format that can be interpreted by a statistic software framework. Furthermore, experimental factors assigned to a MeltDB experiment are projected to the R representations of the data and allow interpretation of the visualizations in the context of the experimental design. The data matrix featuring normalized peak areas and intensities is e.g. used for the box plot visualization presented in Figure 5.6a. All statistical methods can be executed using the web interface and additional criteria such as the treatment of missing values, the scaling of the values and the exclusion of certain metabolites or whole chromatograms can be controlled by the user. The recommended scaling methods evaluated by van den Berg *et al.* (2006) have directly been implemented in the MeltDB API. The visualization of independent and principal component analysis (ICA and PCA) (Hyvärinen and Oja, 2000), the results of a hierarchical cluster analysis (HCA) 5.6b or the results of a pairwise correlation analysis of the pool sizes of the measured metabolites can easily be generated and exported in either PDF or PNG format. The following list gives an overview of the statistical and explorative analysis methods available via the MeltDB web interface:

- Student's t-test
- Analysis of Variances (ANOVA)
- Hierarchical Cluster Analysis (HCA)
- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Boxplots

- Volcano Plots
- Support Vector Machines
- Random Forest

One of the main goals of the R integration in MeltDB is the simplification of the integration of additional analysis and visualization functionality provided through novel BioConductor packages. The generation of data matrices amenable to statistical analysis is therefore standardized via the MeltDB API. In addition, a generic user interface that allows to customize the submitted data for each experiment stored in MeltDB has been implemented. The combination of these methods allows to easily embed new statistical analysis functionality in the MeltDB web application. Since the data matrices generated in current metabolomics experiments are typically in the range of a few hundred metabolites and measurements, most multivariate methods can be performed on the fly.

5.7.1. Visualization of PCA and ICA results

The static images of exploratory data visualizations that the R framework provides are sufficient for datasets containing a limited set of measurements and metabolites or features. However, these two-dimensional visualizations become unreadable if hundred of measurements and metabolites are available. To improve the interactive analysis of PCA and ICA results, a Java based 3D Viewer application was implemented. The Java Web Start application is started via the MeltDB web interface and accesses the generated PCA and ICA results in order to display them in a 3-dimensional space. The data representation can be freely moved in the three dimensional space. Replicate groups can be hidden and information on individual features and measurements can be retrieved via the interactive user interface. The results of a PCA analysis of the replicate measurements for four time points of a fermentation experiment are illustrated in Figure 5.7.

5.7.2. Metabolite correlation analysis

As described in Section 2.6.2, the *de novo* identification of yet unknown compounds is a major task in metabolomics research. In order to identify potentially interesting target compounds for further analysis, a profiling approach is beneficial since the researcher can reduce the list of hundreds of peaks in complex samples to those exhibiting a certain behavior e.g. due to concentration changes or a high correlation of the pool sizes to known compounds.

The correlation analysis of peak data as described previously (Steuer *et al.*, 2003a,b) has been applied to observe fluctuations in metabolic networks by Morgenthal *et al.* (2006). The authors could show that direct transfer of functional connections as in transcriptomics, where concertedly regulated genes show a correlated expression profile, is not possible in metabolite analysis *ab initio*.

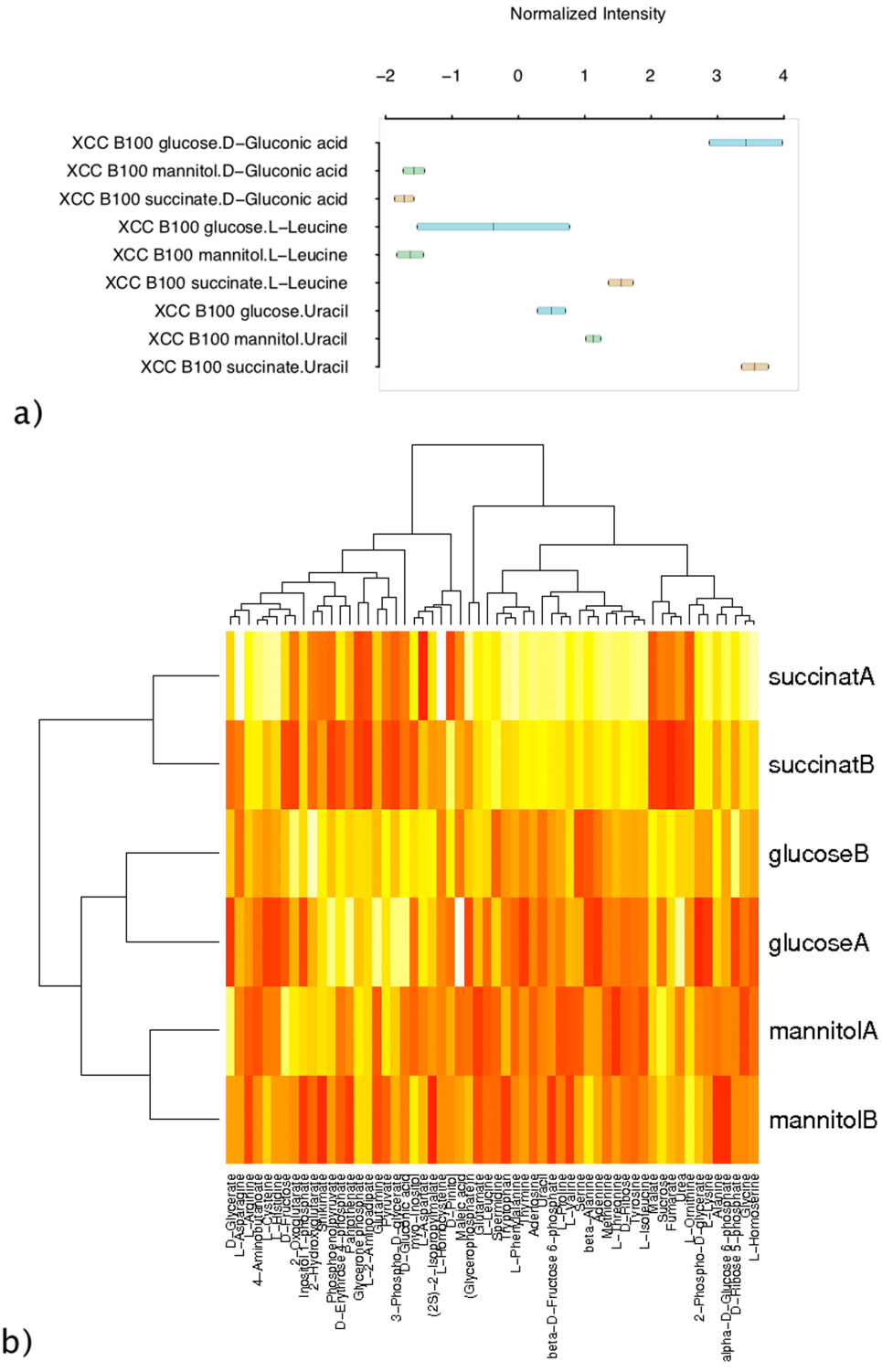

Figure 5.6.: MeltDB supports a collection of statistical and explorative data visualizations. The normalized and log scaled pool data for selected metabolites can e.g. be visualized using box-plots (a). The web interface also allows to visualize the results of hierarchical cluster analysis (HCA) on both metabolites and chromatograms (b). Here, the Euclidean distance function and the complete linkage method are applied.

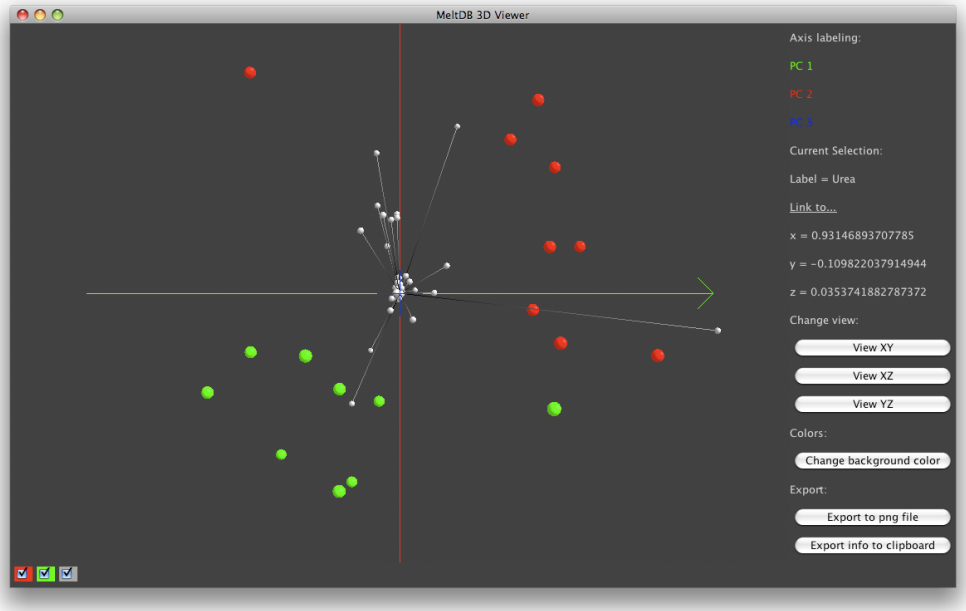

Figure 5.7.: To improve the interactive analysis of the results of a PCA and ICA, a Java based 3D Viewer application was implemented. The Java Web Start application can be started via the MeltDB web interface and accesses the generated PCA and ICA results and renders them using Java 3D functionality. The view of the data-set can be zoomed and rotated in the three dimensional space. The measurements of two replicate groups of human blood plasma samples are represented as green and red spheres in the screenshot. Each replicate group can be hidden and information on individual features and measurements can be retrieved via the interactive user interface. In addition to the measurements, the metabolites that contribute most to the principal components and may explain the visualized clusters are presented as gray arrows.

Nonetheless, for metabolites that are directly connected by enzymatic reaction in metabolic pathways such as the citrate cycle, high correlation between pool sizes can be observed. To simplify the detection of linear metabolite correlations for the researcher, the MeltDB interface offers a correlation analysis method implemented using R.

Method

Pearson's correlation coefficient is used as a measure of linear correlation of metabolite pool sizes and is computed for the vectors that represent either the area or intensities of identified compounds or mass spectral tags obtained by the MeltDB data analysis functionality. The resulting symmetric data matrix contains values between -1 and 1 and can be represented in form of a heatmap as shown in Figure 5.8. The Euclidean distance between the individual columns and rows in this matrix is used to compute a hierarchical clustering (complete linkage). The results of the clustering are represented as dendrograms that represents the similarity between the metabolites. When displaying hierarchical clusters as dendrograms, at each merge a decision is needed to specify which subtree should go on the left and which on the right. Since, for n observations there are $n-1$ merges, there are $2^{(n-1)}$ possible orderings for the leaves in a cluster tree, or dendrogram. In the algorithm used here, the subtrees are ordered such that the tighter cluster is on the left (the last, i.e., most recent, merge of the left subtree is at a lower value than the last merge of the right subtree). Single observations are the tightest clusters possible. The ordering of the leafs in the dendrograms presented in Figure 5.8 is then also used to reorder the columns and rows of the presented correlation matrix. Thereby, blocks of metabolites with high pairwise correlation values are observable.

Evaluation

The correlation analysis of the normalized metabolite measurements of a fermentation experiment of the Lysine production strain of *C. glutamicum* is able to highlight the correlation of the metabolites Fumarate, Malate, Citrate, and Pyruvate as shown in Figure 5.8. These metabolites are part of the citrate cycle and as such closely connected through orchestrated enzymatic reactions. A simplified overview of the citrate cycle with the detected metabolite correlation is presented in Figure 5.9.

The web interface of the correlation analysis allows to furthermore combine features found in the non-targeted profile analysis described above with already quantified and identified compounds. If high correlation between a yet unknown feature and known metabolites can be found, this evidence can indicate the chemical identity of the unknown compound as it might be closely connected through metabolic pathways and enzymes.

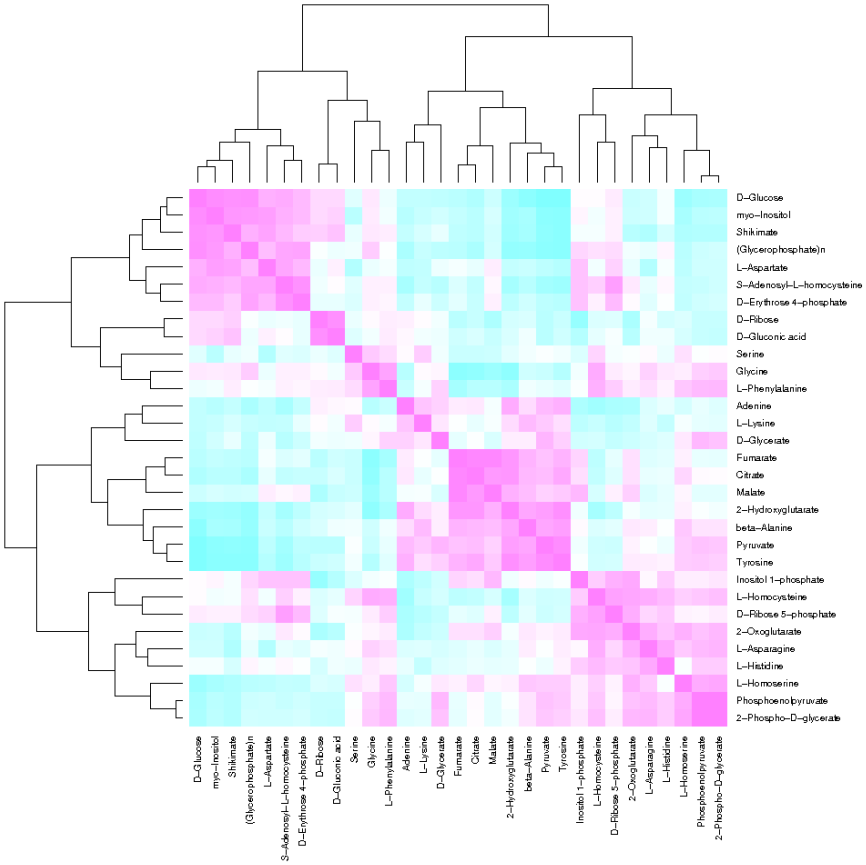

Figure 5.8.: A heatmap of the computed correlation values, high correlation values are represented in pink, anti-correlation is presented in blue. The correlation analysis is able to detect the strong correlation of the metabolites Fumarate, Malate, Citrate, and Pyruvate which are present in the citrate cycle of *C. glutamicum*. For the analysis only metabolites that could be identified in all measurements of the fermentation experiment were used, normalization to dry weight and the internal standard ribitol was applied.

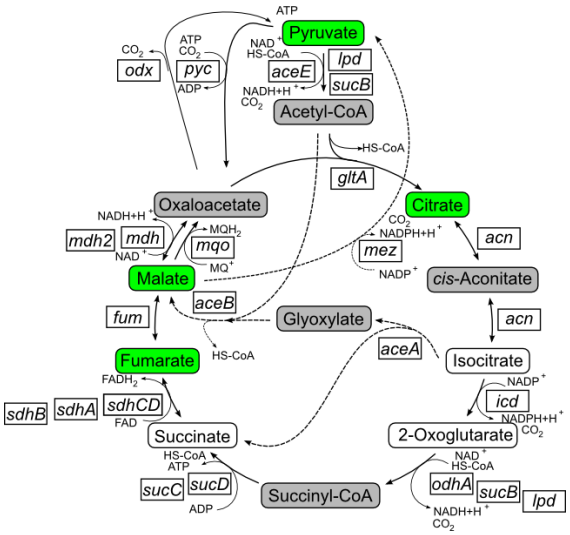

Figure 5.9.: A simplified version of the citrate cycle of *C. glutamicum* presenting the correlated metabolites Fumarate, Malate, Citrate, and Pyruvate.

5.8. Mass decomposition for the identification of metabolites

Apart from the analysis of correlation between known compounds and repeatedly detected but unidentified peaks, a second approach to the elucidation of the chemical identity of compounds based on their mass spectrum has been realized in the MeltDB system. The mass decomposition approach that aims at identifying the chemical identity of a measured compound by generating sum formulas that have the exact or very similar mass as the one measured by the mass spectrometric instrument.

Böcker and Lipták (2005) presented an efficient algorithm to enumerate all sum formulas or molecules consisting of a given alphabet Σ of atoms (e.g. CHNOPS) of size n that sum up to a query mass M in consideration of a measurement error ε. To transform the problem into the integer knapsack problem, all real-valued monoisotopic masses of the atoms in the selected alphabet are converted to integer values. By using a *blowup factor* $b \in \mathbb{R}$ and the function $\phi(a) := \lceil ba \rceil$, the integer masses $a_1 < a_2 < \cdots < a_n$ associated to the atoms are generated in ascending order.

Afterwards, all molecules with monoisotopic mass in the interval $[l, u] \subset \mathbb{N}$ with $l := \phi(M - \varepsilon)$ and $u := \phi(M + \varepsilon) + \delta u$ need to be generated. The value δu is added to the upper bound to compensate rounding errors.

$$a_1 c_1 + a_2 c_2 + \cdots + a_n c_n \in [l, u] \tag{5.3}$$

One solution vector $c = (c_1, \ldots, c_n)$ satisfying 5.3 is termed *compomer* and may only contain non-negative integers. The efficient generation of these compomers for each value in the interval $[l, u]$ can be performed using the FIND-ALL algorithm described by Böcker and Lipták (2005).

One drawback of the original approach is that a large proportion of biochemically infeasible or improbable sum formulas are generated which have to be either filtered or ranked in succeeding steps if the final goal is to identify the correct sum formula for a yet unknown compound. This is the case even if high mass accuracy ($1-5ppm$) of the instrument is given. Kind *et al.* (2007) have presented seven rules which can be used to limit the search space of possible sum formulas and reduce the number of formulas to be tested by e.g. matching isotope pattern distributions. The authors do not present an efficient implementation of the sum formula generation but use an existing proprietary software. Analysis of the mass decomposition algorithm (Böcker and Lipták, 2005) shows that it is easily possible to integrate an *upper bounds* criterion for the number of atoms in the sum formulas defined by Kind *et al.* (2007).

Method

In the following, the improved algorithm is presented that makes use of this upper bounds criterion to circumvent unnecessary calls to the recurrence especially at early stages of the sum formula generation.

For producing all witnesses of the integer query mass $M' = \phi(M)$, the recursive FIND-ALL function is used. The pseudocode is given in Algorithm 1. As in the simple backtracing using the dynamic programming tableau, the method maintains a current compomer c, an index i, and a current mass m. At step i, the entries $c_n, c_{n-1}, \ldots, c_{i+1}$ of compomer c have already been filled in, and the remaining mass $m = M' - \sum_{j=i+1}^{n} c_j a_j$ will be decomposed over $\{a_1, \ldots, a_i\}$. The invariant at the call of FIND-ALL(m, i, c) is that mass m is decomposable over $\{a_1, \ldots, a_i\}$ and $c_j = 0$ for $j = i, i-1, \ldots, 1$. The function $\mathrm{lcm}(x, y)$ that is used in the pseudo-code returns the least common multiple of x and y.

FIND-ALL furthermore uses a precomputed Extended Residue Table (ERT). The two-dimensional table of size na_1 contains for each $r = 0, \ldots, a_1 - 1$ and each $i = 1, \ldots, n$, the smallest number $n_{r,i}$ congruent r modulo a_1 such that $n_{r,i}$ is decomposable over $\{a_1, \ldots, a_i\}$. The ERT can be computed in $O(na_1)$ time and the construction algorithm is detailed in the Appendix. For the improved variant of the FIND-ALL algorithm, the original ERT implementation is used.

In the original version, the upper bound for the number of iterations of the for loop is based only on the value l which is the least common multiple of the current mass a_i and a_1 divided by a_1. During the generation of candidate molecules, this can result in compomers that e.g. solely consist of phosphorus atoms. Such compounds can not exist in nature and should be excluded beforehand if the aim is the identification of existing but yet unidentified biochemical compounds. To circumvent unnecessary iterations of the for loop and thereby reduce the number

Algorithm 1 Algorithm FIND-ALL mass M' , index i , compomer c

```
if i = 1 then
  c_1 ← M/a_1; output c; return;
else
  lcm ← lcm(a_1, a_i);
  l ← min(occurrences(M', i), lcm/a_i); {min of occurrence of a_1 in natural
  molecules and least common multiple }
  for j = 0 to l - 1 do
    c_i ← j;
    m ← M' − j a_i; {start with j pieces of a_i}
    r ← m  mod a_1;
    lbound ← ERT(r, i - 1);
    while m ≥ lbound do
      FIND-ALL(m, i - 1, c);
      m ← m - lcm;
      c_i ← c_i + l;
    end while
  end for
end if
```

of recurrent calls to the FIND-ALL function, the new bounded version has been implemented. In the improved version of the algorithm, l is now limited to maximal occurrence of atom i from Σ for a given query mass M' in natural occurring compounds. Therefore, the *occurrences* function is called, which uses the index i of the current atom and the query mass M' to retrieve the upper bound from a pre-computed lookup table. If no entry for the combination of i and M' can be detected in the lookup table, *occurrences* returns ∞ and the value of l is computed as in the original algorithm. This ensures that the algorithm can also be applied for larger molecules and unusual atom alphabets.

To obtain biochemically reasonable bounds for the *occurrences* lookup table, an analysis of the KEGG compounds database was conducted beforehand. The database contains a comprehensive set of sum formulas of biochemically relevant compounds together with their molecular masses. For all compounds with molecular mass up to 1000 Dalton, the maximal occurrences of the typical atoms in biochemical compounds were obtained from the annotated sum formulas. The database was split into mass intervals of size 100 in order to be able to benefit from atom number limits even for compounds of low molecular mass. The results are presented in Table 5.3.

Evaluation

In order to evaluate the runtime improvement of the newly implemented algorithm compared to the unbounded version, 200 compounds from the KEGG compound

Mass	H	C	N	O	F	Si	P	S	Cl	Br
< 100	14	7	4	4	1	1	1	2	2	1
< 200	28	15	6	7	6	1	2	5	4	2
< 300	44	22	8	10	7	3	3	5	6	3
< 400	54	28	8	14	7	3	3	5	7	3
< 500	68	34	9	18	9	3	4	5	10	3
< 600	72	42	10	21	9	3	5	5	12	4
< 700	88	48	13	24	9	3	6	5	12	4
< 800	94	55	13	27	9	3	7	5	12	4
< 900	98	63	13	27	9	3	7	5	12	4
< 1000	100	63	13	31	9	3	7	5	12	4

Table 5.3.: The KEGG compound database lists the sum formulas and masses of more than 15000 biochemically relevant molecules. The analysis of all compounds with masses below 1000 Dalton shows that the number of individual atoms in the sum formulas are mostly below 10. The table contains the maximal number of atoms that can be found in molecules in the listed mass intervals. Using this information as upper bound for the frequencies in the recurrences of the mass decomposition algorithm, the number of mass decompositions per second can be greatly improved without losing sensitivity. In the original article, the limited atom alphabet CHNOPS was used. With the bounded algorithm, the negative runtime effect of the extension of the atom alphabet is reduced, especially since the remaining atoms F, Si, Cl and Br occurring in biochemically relevant compounds are never exceeding 12 occurrences per valid sum formula.

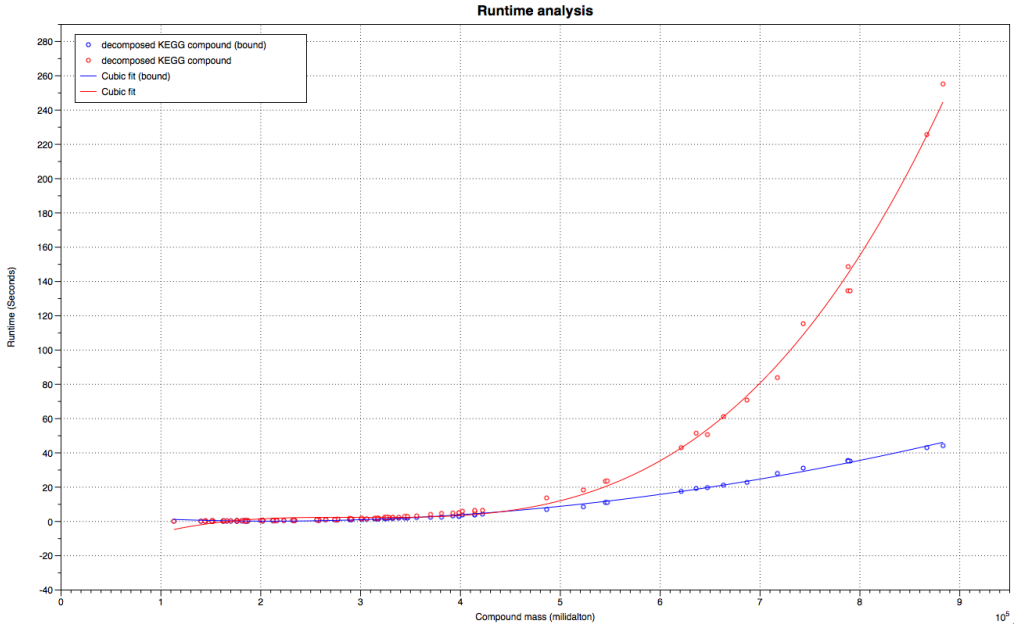

Figure 5.10.: The improved algorithm was compared to the original mass decomposition approach. For compounds of molecular mass close to 1000 Dalton, 7-fold runtime improvements can be observed. As presented in the graph, the relation of the runtimes of both methods is not linear. Especially for large compounds, the improved variant becomes most beneficial.

database with specified sum formulas and known molecular mass were chosen in the mass interval from 0 to 1000 Dalton. For each of these compounds mass decomposition was performed using the original and the improved algorithm with a maximal mass deviation of ± 0.05 Dalton. Runtime information was obtained using a 64 bit 2600 MHz AMD system running Solaris OS 5.10. Figure 5.10 shows the runtime in relation to the molecular mass of the compound. It can be observed that the bounded version is up to 7 times faster, the improvement is observable especially for large molecules and masses.

The improvement of the algorithm is not limited to the runtime of the algorithm since the filtering of the potential candidate formulas has to be applied afterwards to remove infeasible formulas from the generated candidate list. Simple filters can generally be computed in constant time (Degree of Unsaturation, Lewis check, Senior check, Heteroatom rule) (Kind *et al.*, 2007) but especially the computation of theoretical isotope patterns needs at least $O(n * K^2)$ for each sum formula with n being the size of the atom alphabet and K representing the length of the computed distribution (Böcker *et al.*, 2006). The matching and scoring of the measured and simulated isotope patterns takes additionally $O(K)$ time. Every infeasible sum formula that is filtered out implicitly by the improved algorithm does not need

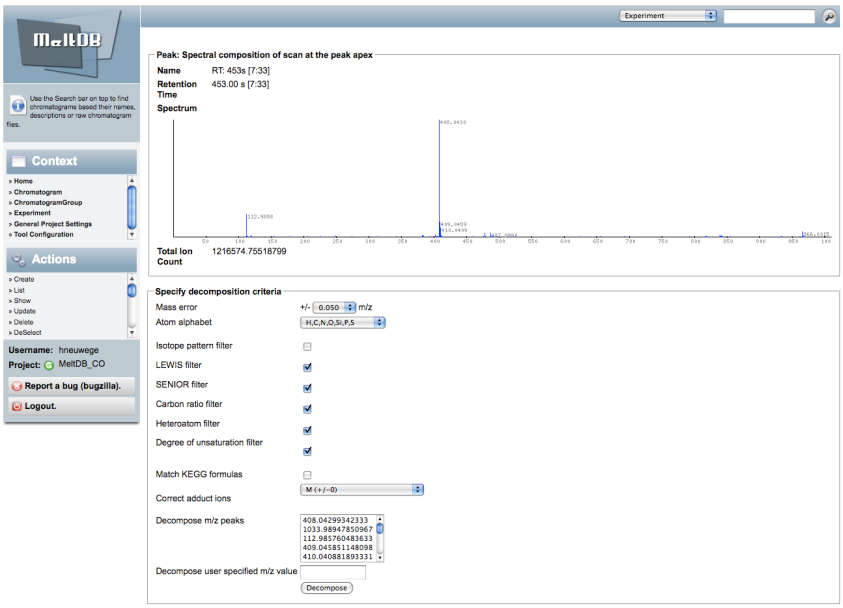

Figure 5.11.: The improved mass decomposition functionality is included in the MeltDB web interface and can be applied to every detected chromatographic peak. The implementation automatically extracts the dominant ions from the mass spectrum. The interface allows to specify the expected mass error of the instrument and to correct the effect of potential adduct ions which is especially important in LC-MS measurements. Isotope pattern filters and matches to the KEGG compound database can be activated additionally.

to be post processed in the filtering stage which furthermore improves the overall runtime of the method.

To make the improved functionality available to researchers, it is integrated directly into the MeltDB web interface and can be applied on every detected chromatographic peak. The user interface presented in Figure 5.11 allows to specify the expected mass error of the instrument and to correct the effect of potential adduct ions, which is especially important in LC-MS measurements. Several filters can be activated to reduce the number of computed sum formulas. As the implementation does automatically extract and sort the dominant ions found in the mass spectrum, the access to the mass decomposition is greatly simplified for the researcher.

After the efficient generation and filtering of the sum formulas, each one is compared against the KEGG compound database. Matching sum formulas are highlighted and both compound name and synonyms are presented to the user.

5.9. Manual annotation functionality and user defined References

If no matching mass spectra can be found in the GMD or NIST databases, the integrated methods for the computation of mass spectral similarity can also be applied to find matching peaks in other publicly available chromatograms organized in MeltDB. If an annotated peak with matching mass spectra, retention time or retention index could be detected, the information may be inherited in order to annotate the yet unknown peak. Thereby the number of 'unknowns' can be reduced step by step. Knowledge generated once can be reused for the identification of the same compound in further analyses.

5.10. Data integration

Data integration in the MeltDB system is achieved on various levels. The previously mentioned KEGG compound database (Kanehisa *et al.*, 2006) is regularly imported through direct access to the KEGG FTP server. Relevant terms and relations of the compounds are thereby directly represented in the MeltDB database model. The current version of the KEGG compound database contains $\approx$ 15000 entries which act as a controlled vocabulary in MeltDB for compounds relevant to biological systems. References to other metabolite databases (CheBI, CAS) and the connection to metabolic pathways are included in the MeltDB database representation.

Apart from the controlled vocabulary, the KEGG compound database does furthermore provide an association of metabolites to reactions, enzymes, pathways, and genes. It is therefore possible to link metabolic experiments with existing genome projects stored in e.g. the GenDB genome annotation system. To connect both systems, the use of SOAP based Web Service technology has been chosen.

Gene annotations containing EC numbers can be requested via the GenDB Web Service for an prokaryotic organism under study. Thereby, metabolic pathway representations from the KEGG database can be enriched together with qualitative information on detected metabolites (Figure 6.2). The visualization of qualitative data on KEGG pathways is feasible, but for time series experiments or quantitative datasets, the predefined pathway representations of KEGG are of limited use since the static images are easily overcrowded. For the integrated visualization of quantitative datasets, several alternatives are available as described in Section 3.3. Nonetheless, none of the existing tools allows the direct import of quantitative metabolomics data via Web Services.

To provide a multi-*Omics* visualization platform based on Web Service data exchange, the existing core functionality of the MeltDB web interface and visualization features was re-applied and extended towards a second application called ProMeTra that will be described in the following sub-sections. The generic implementation of the models, views and controllers being implemented for the MeltDB

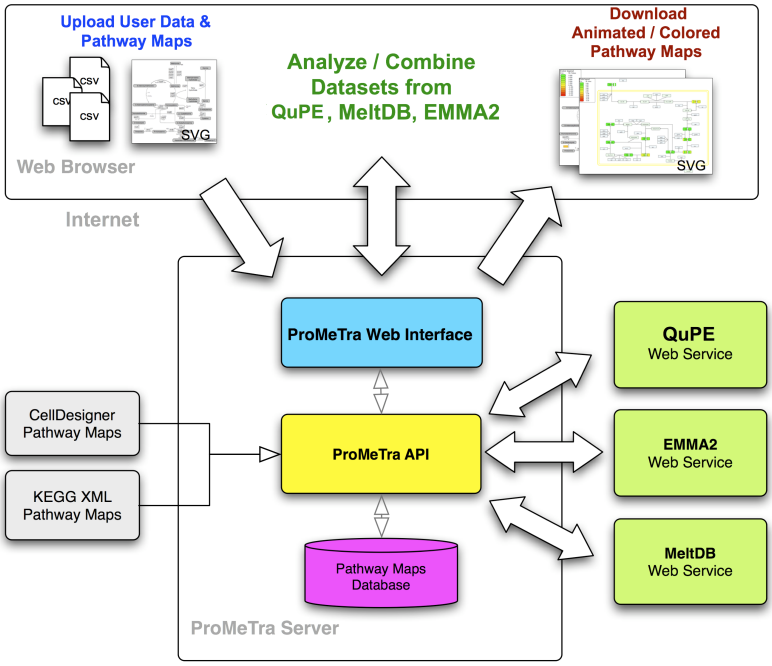

Figure 5.12.: ProMeTra is a web-based system for the integration and visualization of *Omics* datasets. The connection to existing functional genomics platforms such as MeltDB, QuPE, and Emma2 is realized via SOAP-based Web Services. Researchers can upload their own metabolic pathway maps in an annotated SVG (Scalable Vector Graphics) format and employ the ProMeTra functionality to render quantitative information originating from transcriptomics, proteomics or metabolomics experiments onto these images. The web interface allows the download of the enriched SVG images. Additionally, methods to upload quantitative results in spreadsheet formats are provided.

system could be re-used for the fast development of the novel application (Neuweger *et al.*, 2009).

5.10.1. Multi-omics data integration

Comparable to the MeltDB system, the functionality of ProMeTra can be accessed through a platform independent web application. ProMeTra does furthermore act as an interactive interface to the experimental datasets stored in the *Omics* platforms MeltDB for metabolomics, Emma2 for transcriptomics (Dondrup *et al.*, 2009), and QuPE for proteomics (Albaum *et al.*, 2009) that contain experimental results. ProMeTra features user access control and offers a public account to the scientific community, users can directly log into ProMeTra and upload own datasets and pathway maps.

A researcher can employ the preprocessing and visualization functionality of ProMeTra via the web interface. Apart from a recent web browser that supports SVG images (e.g. Firefox or Safari) no additional software needs to be installed. Users of Microsoft® Internet Explorer need to install a SVG viewer plugin from Adobe which is freely available.

Users can also upload their own datasets in a text based CSV format that can easily be generated by any spreadsheet application. Details of the supported data formats and the organization of Excel files can be found in the online documentation. The uploaded data files in Excel or CSV format are only stored during a ProMeTra session and are automatically deleted afterwards to ensure the privacy of experimental data. In contrast to the temporarily stored data files, user defined pathway images enriched with information on the presented genes, transcripts, proteins or metabolites can be stored on the ProMeTra server persistently. Every user can decide if his pathways are made public, delete or update his uploaded pathway images via the ProMeTra web interface. Information on the pathway maps are stored in an object relational database on the server. User defined SVG pathway maps can e.g. be generated using the freely available Inkscape software.

The core of ProMeTra is an object oriented API that provides access to the pathway maps and the experimental data sets. The main classes are *DataFactory*, *Element* and *Color*. Subclasses of the interface *DataFactory* are responsible for retrieving experimental data from data sources. Based on the numerical range of the experimental data a mapping of various color gradients (e.g. red-yellow-green) is computed by instances of the *Color* class. The functionality to enrich annotated SVG elements in pathway or genome maps is encapsulated in the *Element* class. It provides XML parser functionality to access and extend the DOM (Document Object Model) tree of any SVG image. The Element class inherits all methods of the XML::DOM::Element class and adds animation and coloring methods.

5.10.2. Web Services and external data integration

It could already be shown how Web Services can be used to connect heterogeneous software frameworks in functional genomics (Neuweger *et al.*, 2007). MeltDB and Emma2 provide SOAP based Web Services written in Perl which provide access to normalized quantitative data from metabolomics and transcriptomics experiments. QuPE offers Java based and WSDL specified methods to obtain the preprocessed experimental datasets originating from quantitative proteomics experiments. ProMeTra is the first web-based system to make use of this functionality and supports the main functional genomics techniques in one system. For researchers that do not have the possibility to analyze their data using the described web-based systems, a simple CSV and Excel based data import via the ProMeTra web interface is provided as well.

5.10.3. Visualization and Animation features

ProMeTra supports SVG images that are extended by annotations for genes, proteins or metabolites. The images in the open and user readable data format SVG can be uploaded to the web-server via the ProMeTra web interface. A set of high quality pathways for the industrial amino acid producer *Corynebacterium glutamicum* is available and used in the following application example. Metabolic pathways can either be designed and submitted by the user or can be converted via ProMeTra functionality from SBML files defined in CellDesigner. An SBML to SVG converter that already includes the mapping of the elements to the KEGG compound database and includes annotated gene locus tags has been developed. The mapping of numerical experimental data such as concentrations and ratios is done through a color encoding and rectangles in the SVG image representing genes, proteins or metabolites are subdivided or animated alternatively. Therefore the DOM tree of the SVG image is extended by ProMeTra such that child elements are added to the respective rectangles. Animations are realized such that the background color of the elements representing genes, proteins or metabolites changes over time. In both cases, the user defined layout is preserved. ProMeTra offers different color gradients to encode the values of the submitted experimental datasets. The M-Value based color gradient ranging from red (-5) to green (+5) being common for microarray and metabolite ratios is used as default. Further color gradients can easily be defined via the flexible ProMeTra API and if data with larger absolute values is submitted to the ProMeTra system, the mapping of the value range to the color gradient is computed dynamically.

It has been pointed out that the representation of *Omics* data on metabolic pathways is most intuitive to the researcher but other concepts of data visualization can be addressed in ProMeTra. Annotated bacterial genomes present at the NCBI genome repository can be transformed into so called GenomeMaps. Therefore, GenBank (Benson *et al.*, 2008) files of the available replicons are parsed using BioPerl and SVG images (the GenomeMaps) are generated automatically. A GenomeMap represents each annotated coding sequence of the replicon as rectangle in a grid. The order of the rectangles is determined by the chromosomal position of the stop codon of the respective coding sequence and the rectangles are labeled by the associated locus tag or the gene name if present. The grid is filled row by row starting at the top left position for the first gene after the origin of replication. GenomeMaps have been generated for more than 400 bacterial genomes and are available through the ProMeTra web application. A GenomeMap enriched with the data of transcriptional changes during a fermentation experiment of *C. glutamicum* is presented in Figure 5.13.

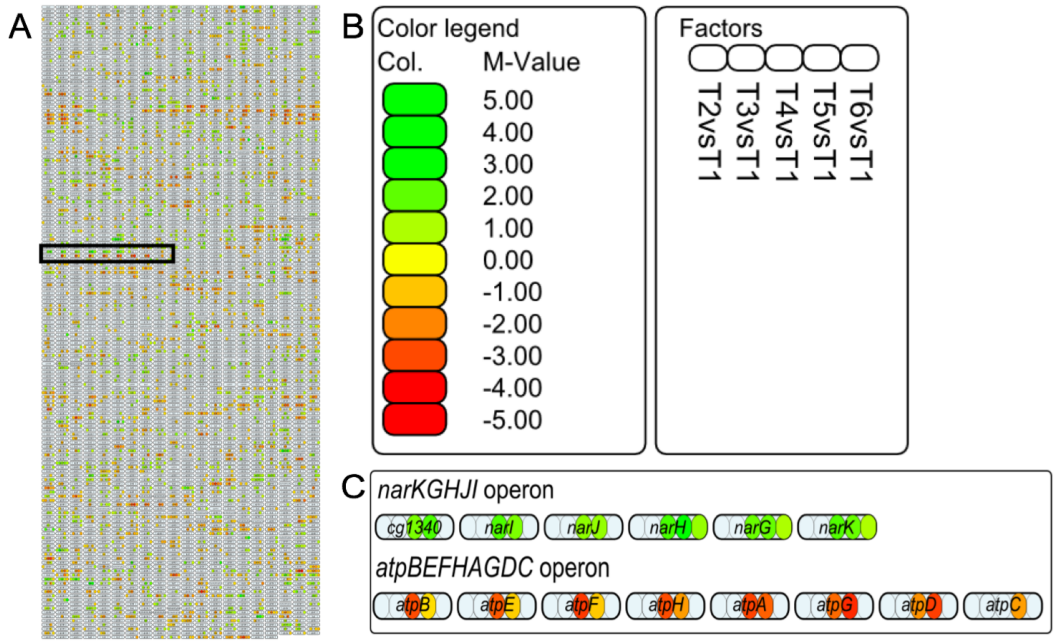

Figure 5.13.: Genome scale ProMeTra visualization of relative transcript abundances during fermentation of the L-lysine producer *C. glutamicum* DM1730. The zoomed out GenomeMap (A) contains all coding regions of *C. glutamicum* as rounded rectangles in chromosomal order. Colored boxes represent the transcript ratios of each gene of time points t2 to t6 in comparison to t1 using the color mapping in the legend (B). The SVG format allows to zoom into interesting areas of the GenomeMap as presented for the *nar* and *atp* operons. The genes of each of the two operons show a corresponding expression profile (C).

CHAPTER 6

Application examples

The previous chapters presented the implementation of the web-based system for the storage, administration, analysis and integration of metabolomics datasets. Now, the realized analysis system allows researchers to cover the computational aspects of metabolomics experiments in a structured and platform independent manner. The following application examples of metabolomics experiments will highlight several of the analysis features of MeltDB.

6.1. Xcc B100 grown on three different carbon sources

The first application example is a metabolomics experiment conducted on the bacterium *Xanthomonas campestris* pv. campestris B100 and will present visualization and analysis features provided by MeltDB.

The genus *Xanthomonas* mainly consists of phyto-pathogens of wide host range. On their ability to infect different host-plants the genus is subdivided into species and pathovars, a classification system that has also been verified by 16S rDNA sequence analysis. One of these species, *Xanthomonas campestris* pv. campestris (Xcc), is the causal agent of black rot disease in crucifers. Xcc secretes numerous lytic enzymes and an exopolysaccharide (EPS) to facilitate its pathogenic and saprophytic life style. But not only the pathogenic properties put this rod shaped, gram-negative bacterium in the focus of interest; the exopolysaccharide, so-called xanthan gum, produced by Xcc is also of economical importance. Xanthan gum finds a variety of industrial uses as stabilizer in foods, cosmetics, paints, and as biolubricant in oil drilling. As biotechnological product it therefore takes 1% of the

world market shares and is estimated to have an annual turnover of 400 Mill. US$ (2004). The strain used in this work is *Xanthomonas campestris* pv. campestris B100 (Xcc B100). The proteome of this strain has been widely researched and only recently the genome of Xcc B100 has been published (Vorhölter *et al.*, 2008). But the EPS production and regulation still needs to be thoroughly scrutinized on the metabolomic plain. Under certain conditions Xcc is able to channel up to 80% of the available carbon source into the xanthan gum production.

Xanthomonas campestris pv. campestris B100 was cultivated in Vincent-Minimal-Medium (VMM) in shaken flasks. The medium was supplemented with 1% (w/v) of glucose, mannitol, and succinate, respectively. Two biological replicates for each carbon source were cultivated. Sample preprocessing, derivatization and GC-MS measurements were conducted as described previously (Barsch *et al.*, 2004). The resulting chromatograms were converted to netCDF format using functionality of the Xcalibur® software and imported into the MeltDB system. Annotation of the chromatograms was done in accordance with the recommendations of the MSI and the chromatograms were organized in replicate groups according to their carbon source in MeltDB. Peak detection, identification and quantification was performed using a manually defined preprocessing method defined in the Xcalibur® software. The MeltDB importer tool was used to transfer these results into the data model and to link identified metabolites to the KEGG compound database.

An initial comparative inspection of the raw datasets was performed using the *Experiment TIC plot* provided by the MeltDB web interface (Figure 5.1b)). Although retention time deviations of some seconds were observed between the individual chromatograms, most of the predicted peaks of the Xcalibur® preprocessing method are present in all samples.

After the MeltDB chromatogram alignment was performed and visualized, deviations in peak intensities between the different replicate groups became evident. To analyze the differences in more detail, peak areas were normalized in each chromatogram relative to the area of the ribitol peak (set to 100). The data matrix computed within MeltDB was used for further statistical analysis such as e.g. simple boxplot visualizations shown in Figure 5.6a. The heatmap visualization of a hierarchical clustering on both metabolites and chromatograms shown in Figure 5.6b highlights that the replicates of the three groups (glucose, mannitol, and succinate) cluster together. The dendrogram associated to the chromatograms included in this feature also shows that the glucose and mannitol groups are more similar to each other than to the succinate group. Higher abundances of 3-phospho-glycerate, L-2-aminoadipate, and glycerone phosphate in the glucose and mannitol approach, and the higher amounts of L-leucine and uracil in the succinate approach seem to be the main reason why the mannitol and glucose groups differ from the succinate group. The main differences between the glucose and mannitol groups are found in the higher metabolic pools of gluconic acid, when glucose is the sole carbon source. Box-plots of the normalized intensities for these metabolites are depicted in Figure 5.6a.

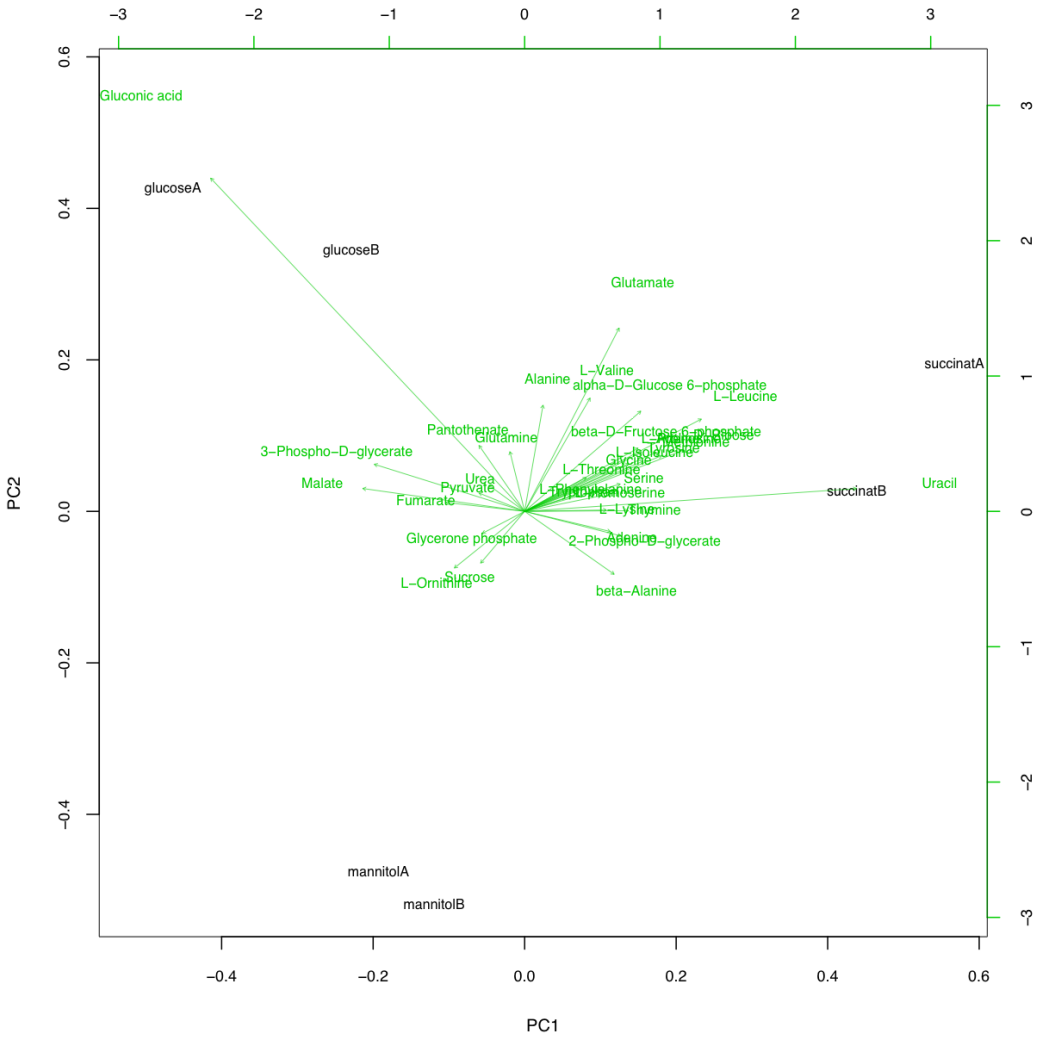

Figure 6.1.: This *biplot* visualization of the PCA analysis represents both the scores of the observations and the loadings of the variables of a matrix of multivariate data on the same plot. The scores of the observations are represented for the first and second principal component using the x- and y-axis. The green axis labels to the right and top are used to represent the loadings of the variables. In this case, the measurements (observations) are shown in black and the metabolites (variables) are represented using green arrows. It can be observed that the replicate measurements for succinate, mannitol and glucose cluster together. Furthermore, the biplot allows to identify the metabolites contributing to the clustering of the measurements. A high abundance of gluconic acid in the measurements of the glucose group can be observed whereas uracil can be predominantly found in the succinate measurements.

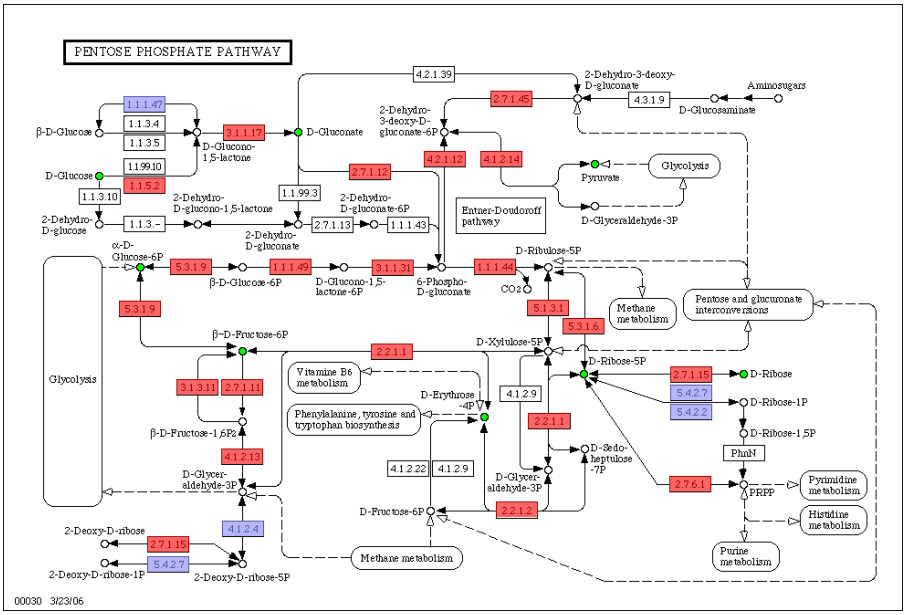

Figure 6.2.: Enhanced MeltDB visualization of the pentose phosphate pathway from KEGG. Identified metabolites from the chromatograms of the actual experiment together with the genome annotation data of Xcc B100 stored in GenDB are colored. Detected metabolites are marked in green, enzymes identified and annotated in Xcc B100 by a human annotator are shown in red, enzymes for which at least potential homologous sequences could be detected in Xcc B100 using BLAST searches against the KEGG database are marked in blue.

A *biplot* (Gabriel, 1971) of the PCA analysis presented in Figure 6.1 indicates e.g. a high abundance of gluconic acid in the glucose group. Since gluconic acid is a direct intermediate of the pentose phosphate pathway (Figure 6.2), one may assume that glucose is not only catabolized via the glycolysis and pentose and glucuronate interconversions, but in contrast to the mannitol group also via this pathway. On the other hand there were no differences between the glucose and mannitol groups in the amount of 3-phospho-glycerate and 2-phospho-glycerate, which indicates that Xcc B100 metabolizes mannitol and glucose with seemingly the same mass flux via the glycolysis. The higher abundance of glycerone phosphate in the mannitol group again shows that mannitol degradation not only differs from glucose metabolism in the low utilization of the pentose phosphate pathway, but also in the higher usage of the glucuronate interconversions, since glycerone phosphate is an intermediate found in the periphery of this pathway. Since succinate is directly internalized into the citrate cycle, it is not catabolyzed via any of the above pathways. Rather glucose is anabolized via gluconeogenesis, which explains the higher number of differences of the succinate group to the other two.

These findings are facilitated by MeltDB through a mapping of the detected metabolites onto the metabolic pathway maps provided by the integrated KEGG database. As MeltDB is also connected with the functional genomics packages GenDB and Emma2 via web-services and the BRIDGE layer (Goesmann *et al.*, 2005), current gene annotation data from the XCC annotation project is dynamically added to the pentose phosphate pathway (Figure 6.2). Thus, MeltDB supports the analysis of metabolomic experiments in a knowledge based environment and simplifies the interpretation of experimental results in a biological context. These findings are part of the first publication of the MeltDB system (Neuweger *et al.*, 2008).

6.2. Analysis of human heart and blood plasma samples

MeltDB was designed to provide a seamless connection of the experimental datasets to the R software and integrates multivariate analysis and visualization features for metabolomics datasets via the web interface. The extension of the analysis functionality towards classification and variable importance analysis became evident during the first experiments conducted on human blood plasma samples and human heart tissue samples. Here, the metabolic differences between healthy and diseased samples were sought. Each measured sample was clearly labeled and therefore techniques of supervised machine learning could be applied. The following will detail both how the MeltDB system could be extended to address the novel requirements and how the analysis functionality could be employed in the search for changes in metabolite levels being characteristic for human heart insufficiency patients.

For each of the two classes (healthy, heart insufficiency patient) ten biological replicates were measured using a GC-MS (Ion Trap) instrument. The experimental data that was used for the evaluation of the profiling analysis described in Section 5.5.5 is reapplied in the following study. As described earlier, the Xcalibur® analysis identified and quantified 36 metabolites in all of the 20 measurements.

Manual effort for the definition of the compound library as well as the manual detection and quantitation of missed peaks in the Xcalibur® software provides a valuable reference dataset. To find metabolic differences between the two sample groups, several analysis features of MeltDB were employed. The first method was the analysis of variance performed on all normalized peak intensities of the 36 metabolites. Of these, the differences between the compounds Citrate and D-glucose were most significant according to the results of an ANOVA analysis presented in Table 6.1.

To examine the performance of machine learning and classification on this metabolomics dataset, state-of-the-art classifiers such as support vector machines, random forest and neural networks were employed. A standardized interface to classification and regression analysis in R is provided through the *caret* package

Compound	p-value
Citrate	0.00000
D-Glucose	0.00001
D-Ribose	0.00002
Salicylate	0.00013
(2S,3R)-3-Hydroxybutane-1,2,3-tricarboxylate	0.00064
Tryptophan	0.00080

Table 6.1.: An analysis of variance (ANOVA) of 36 metabolites detected in 20 human blood plasma samples from two groups was conducted using MeltDB. For the metabolites presented in the table, the normalized peak intensities exhibit significant differences ($p-value < 0.001$) between the two sample groups.

(Kuhn, 2008) from the BioConductor repository. *Caret* simplifies the training and evaluation of various classifiers based on experimental data. As MeltDB organizes single measurements in groups and experiments, the distinction between healthy and diseased samples can directly be extracted from the experimental description.

The resulting data matrix is labeled with the classes (heart insufficiency, healthy) from the experiment and the classification performance of *SVM*, *random forest* and *neural network* models are computed. The implemented *caret* analysis pipeline repeatedly splits the data matrix into a training and a test set to estimate the model performance. Centering and scaling is performed on both sets using the predictor means and standard deviations from the training set. Afterwards, highly correlated metabolites and those that exhibit near zero variance are excluded from both training and test sets. All models are trained under varying tuning parameters and for each classifier, the optimal model is reported. For the dataset under study, an optimal classification could be achieved by all three classifiers. This indicates that the measured metabolic profile of the blood plasma samples differs considerably between the healthy individuals and the heart insufficiency patients.

6.2.1. Variable importance estimation

The identification of the metabolites or variables with the most influence on the trained model is of great interest as they point to the main metabolic differences and potential marker substances for the two sample classes. The variable importance analysis provided by *caret* allows to extract variable importance estimations from the trained models. For the random forest model, the values of each predictor variable are permuted and the change in prediction accuracy before and after the permutation is used as measure for the variable's importance presented in Figure 6.3.

A clear separation of both sample groups is possible. This finding is also supported by the more comprehensive profiling analysis implemented in MeltDB. Af-

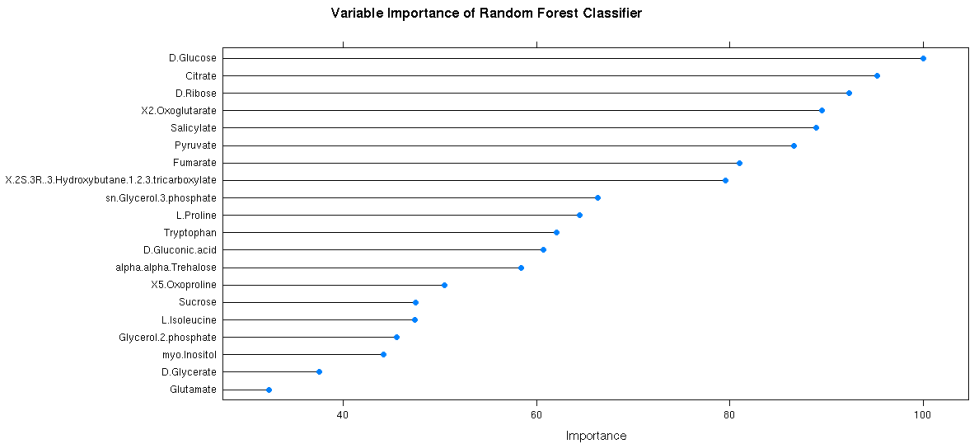

Figure 6.3.: Variable importance estimation based on the trained random forest model of the classified blood plasma dataset. Apparently, metabolites that already show significant differences in the ANOVA analysis are also important for the classification using random forest. The most important metabolite is set to value 100 and others are scaled accordingly.

ter applying the previously described profiling analysis of MeltDB that identifies common EIC channels for the unambiguous quantitation and annotation of these peaks, a second dataset with 122 features was available for more detailed classification analysis. Again, a clear separation of the two sample groups could be achieved (data not shown). The results of the MeltDB analysis of the blood plasma samples in combination with samples from human heart tissue samples have been presented at the twenty ninth annual meeting of the international society for heart and lung transplantation (Gezelbash *et al.*, 2009). Further experiments will be conducted in the future to increase the number of replicate measurements and to analyze the detected differences in more detail.

6.3. Analysis of a multi-omics fermentation experiment of Corynebacterium glutamicum

Furthermore, MeltDB was applied for the analysis of a fermentation experiment of the L-lysine producing strain *Corynebacterium glutamicum* DM1730. During fermentation, oxygen supply was switched off in order to perturb the system and observe its reaction. At six different time points, transcript abundances, intracellular metabolite pools, as well as extracellular glucose, lactate, and L-lysine levels were determined.

The interpretation and visualization of the results of this complex experiment was facilitated by the previously described ProMeTra software. Both transcrip-

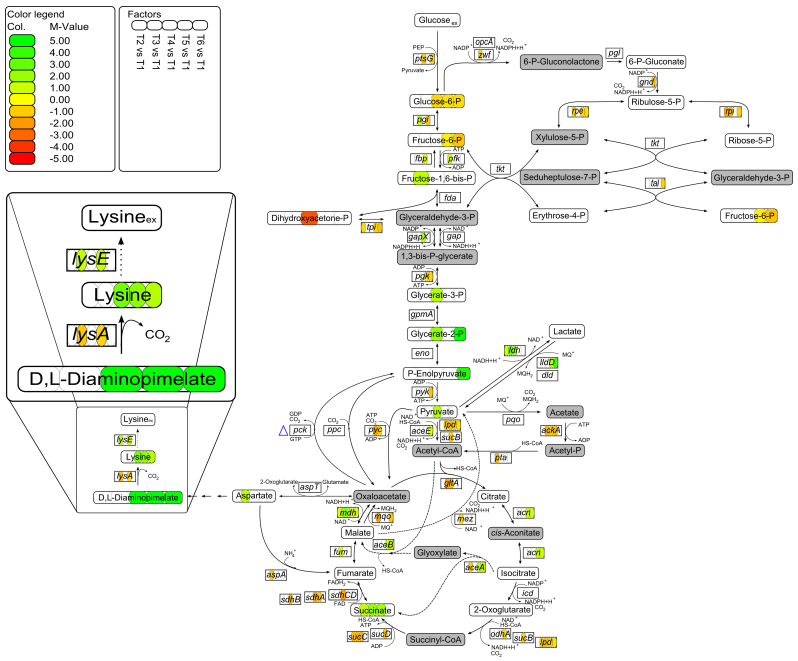

Figure 6.4.: ProMeTra visualization of relative metabolite pools and transcript abundances during fermentation of the L-lysine producer *C. glutamicum* DM1730. The metabolite pool ratios are shown in rounded rectangles and the transcript ratios are shown in boxes as mean values (M-values) of time points t2 to t6 in comparison to t1. The values in the range of -5 to 5 are shown in a color code from green to red. The pathway map shows the glycolysis, pentose phosphate pathway, tricarbolic acid cycle and the lysine pathway for *C. glutamicum* with a zoomed L-lysine pathway. The pools of the metabolites shown in grey were not determined.

tome and metabolome data were visualized on a metabolic pathway map. Visual inspection of the combined data confirmed existing knowledge but also delivered novel correlations that are of potential biotechnological importance.

The ProMeTra tool uses the measured values for the relative expression values of the transcripts and the relative pool sizes of the metabolites to map them onto a previously defined pathway map as presented in Figure 6.4. Metabolome analysis was performed by GC-MS using the protocol described by Plassmeier *et al.* (2007). The pathway map shows the main metabolic pathways of *C. glutamicum* from glucose uptake to L-lysine excretion. The L-lysine pathway is shown in a short version, as most of the metabolites between L-aspartate and L-lysine are not identified with GC-MS, because of missing reference substances and non-volatile metabolites. The values of measurements, metabolites, and transcripts are shown in color code from green (upregulated) to red (downregulated). Only those values are displayed that had an error probability less than 5% in a Student's t test. For each measurement

the value is given relative to that at the first timepoint (logarithmic growth). Even if neither transcript levels precisely predict enzyme activities nor metabolite pools do this for fluxes, a high number of correlations could be identified that correspond with actual knowledge on bacterial metabolism. The interpretation of the complex experiment was simplified through the visualization of the results on the presented metabolic pathway representation. These and additional findings have been presented in more detail in the ProMeTra publication (Neuweger *et al.*, 2009).

6.4. Statistical analysis of cell harvesting methods

MeltDB was employed for the evaluation of a new technique for harvesting of *C. glutamicum* cells for metabolome analysis on the basis of size exclusion chromatography (SEC). This classical technique can be used for separation of cells and extracellular compounds. Residual analysis demonstrates that this method effectively depletes extracellular compounds. In order to test the potential of this method, SEC was compared with common methods used for harvesting of *C. glutamicum* cells for metabolome analysis, by name cold methanol quenching, fast centrifugation and fast filtration. Fermentations were performed using the wild type and lysine production strain DM1730 in minimal medium. In addition, the strain DM1730 was grown in complex medium CASO bouillon. All samples were analyzed with a GC-MS instrument and subjected to MeltDB analysis.

To assess the global differences in the sizes of metabolite pools resulting from the three methods, an unsupervised clustering analysis available through the MeltDB analysis system was performed. The initial compound detection of the measured samples has been performed using the Xcalibur® software. Both the intensity and the area of chromatographic peaks of identified reference compounds were computed by Xcalibur®. The results were imported into the MeltDB software for further analysis. Normalization to dry weight and the internal standard (ribitol) has been performed using MeltDB functionality. To additionally compensate differences between the three fermentations, all measured intensities and areas have been transformed to represent the relative deviation from the median value of the wild type measurements. The resulting data matrix was analyzed via the integrated functionality of the statistical software system R. For the hierarchical cluster analysis (HCA) presented in Figure 6.5, the Ward clustering algorithm was employed and the Euclidean distance function was used on the normalized data matrix described above. Ward clustering generally aims at finding compact, spherical clusters. As expected, replicate groups were clustered first, but also the two groups of the CASO measurements with centrifugation and MeOH treatment were most distant to the remaining measurements.

Different clusters could be distinguished by the HCA, corresponding to the different strains and harvesting techniques. The compact clustering of individual harvesting methods revealed a high reproducibility. The ratios of the lysine production strain DM1730 in comparison to the wild type revealed that each harvesting

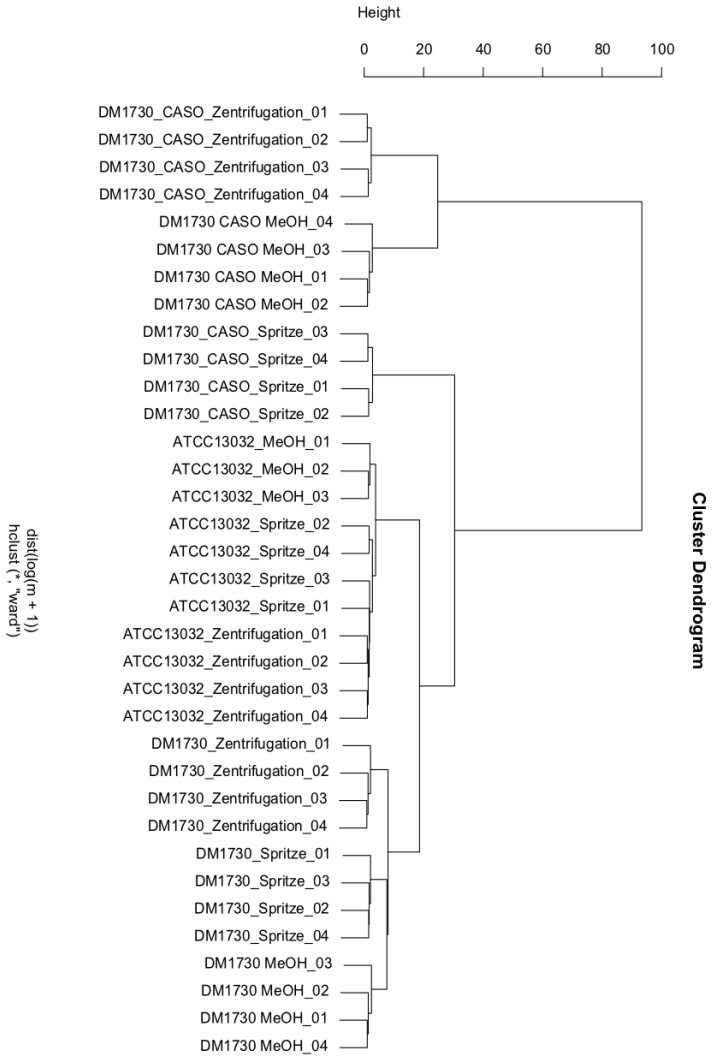

Figure 6.5.: For the hierarchical cluster analysis the Ward clustering algorithm was employed with an Euclidean distance function. The Cluster Dendrogram shown in the figure represents the differences between the measurements. It can be observed that all nine replicate groups form compact clusters and the two groups of the CASO measurements with Centrifugation and MeOH treatment are most distant to the remaining measurements.

method produced the same distance in the HCA corresponding to comparable pool size ratios for most of the metabolites. Unsurprisingly, cultivation in complex media resulted in large distance differences in the HCA when comparing the same strain harvested with centrifugation or cold methanol in comparison to the cells grown in minimal media. This is most likely due to the presence of high amounts of unremoved media components in the samples. In contrast, samples harvested with the SEC method are more similar to the samples derived from minimal media than to those from complex medium harvested with the other two methods. While these are still the outliers in the minimal media cluster (most likely due to shifts in internal pool sizes of amino acids), this indicates a much more stringent removal of media compounds in samples collected with the SEC method.

The statistical analysis supported by MeltDB revealed that the SEC method is the most suitable method for harvesting when intracellular pool sizes are to be measured. It completely removes extracellular compounds, in contrast to fast centrifugation and cold methanol quenching. The SEC method turned out to be usable for fermentations under production conditions, handling well even high cell densities. The results of this analysis together with the description of the newly developed cell harvesting methods is currently prepared for publication by Marcus Persicke *et al.* (in preparation).

CHAPTER 7

Discussion

MeltDB is the first web based system to support the annotation of metabolomic experiments using the recently published recommendations of the metabolomics standards initiative. The system is currently applied for the analysis of metabolomics experiments on various prokaryotes (*C. glutamicum*, *S. meliloti*, *Xanthomonas campestris* pv. campestris), higher organisms (*Chlamydomonas reinhardii*, *Triticum* spp.) and human blood and heart tissue samples as described in Chapter 6. The metabolomics datasets stored and analyzed in MeltDB originate from a variety of GC-MS and LC-MS instruments and several different work groups and institutions apply the system.

This indicates that the requirements for the structured storage and annotation of metabolomics experiments have successfully been addressed. The system can be employed for the analysis of metabolomics experiments. To this date, more than 50 users in multiple projects and research collaborations all over the world have stored and annotated raw data in MeltDB projects.

The organization of various projects and the large number of users was simplified by the integration and customization of the Generalized Project Management System (GPMS) being developed by the Bioinformatics Resource Facility at CeBiTec, Bielefeld University. The web based user interface and the secured access to the experimental datasets makes the MeltDB system applicable in distributed research projects and allows to analyze sensitive data since all communication from and to the web server is encrypted using state-of-the-art SSL technology. Especially the possibility to analyze experimental results in a collaborative approach is unique to MeltDB since most of the other existing software solutions for metabolomics are single user applications running on workstations. As these are typically overstrained by experiments consisting of hundreds of measurements, the scalable nature of the

MeltDB system with the connection to a compute cluster and the storage of generated results in an object relational database model allows to easily cope with growing experimental datasets and user numbers. Compared to the previously existing software tools for the analysis of metabolomics experiments, MeltDB is the only system that features a platform independent user interface together with the support for a scalable compute cluster architecture. It distinguishes itself from other web based solutions such as MetaboAnalyst by supporting various input formats, flexible preprocessing pipelines, and importers for all widely used vendor specific export formats. Interactive and user configurable data visualizations for both raw data and the results of multivariate statistical analyses are also unique to MeltDB in the presented comprehensiveness. Via MetaboAnalyst, researchers may only download an archive of PNG images generated by a variety of pre-configured R functions that can be computed for the uploaded data matrices. In comparison to the features of existing systems for metabolomics analysis and visualization presented previously in Tables 3.1 and 3.2, the combination of MeltDB and ProMeTra is more comprehensive and also offers improved data integration strategies.

The availability of the source code of the system led to a collaboration with the Center for Mathematical Modeling (CMM), at the University of Santiago, Chile, where a second installation was realized during a visit. The work group develops support for CE-MS based metabolomics based on the MeltDB data model and the already implemented analysis functionality of MeltDB.

Apart from this external collaboration which spurs the advancement of the MeltDB project, an evolutionary development of the system led to several new tools and visualizations. The feedback from the user community and experimental requirements have driven the realization of novel analysis methods and the adaptation and refinement of existing ones. Although various methods have been developed to support the emerging requests and requirements, the core data base model as well as the user access and tool concept could be efficiently applied for all analyzed experiments. No changes to the initial data model apart from the inclusion of CE-MS measurements had to be made.

The careful design of the MeltDB data model and the automatically generated functionality to store, retrieve, update, and delete objects from the database was just the starting point for the core API development. Raw data access was encapsulated in Factory Interfaces and especially the simplified and standardized access to mass spectral data allowed to rapidly implement novel analysis methods and evaluate existing algorithms. As this functionality does mainly concern the software developers, the user community will typically not be affected by such details of the software development project.

Yet, the use of customizable analysis-pipelines within the project showed substantial benefits for a standardized execution of the computational part of the experimental metabolomics analysis. The provided MeltDB functionality can readily be used. Furthermore, experienced users of the system can also parameterize the integrated and implemented preprocessing methods in MeltDB to obtain the optimal set of methods and to choose the optimal parameters. The results of different

parameterizations can be observed and compared through the visualization of the generated results on the raw data presented in Section 5.2. But even if MeltDB allows to compare different parameterizations of a tool, the need for knowledge, at least on a basic level, about the applicable methods cannot be compensated by the software alone.

Taking into account the evolution of the hyphenated mass spectrometry technology observed in recent years, it is questionable, whether a single analysis software can satisfy all the different emerging technological requirements. A software solution such as MeltDB may therefore only be complete with respect to the supported standards and data sources. Nonetheless, it could be shown that the integration of novel analysis features can easily be realized using the existing software infrastructure.

The preceding application examples detail that the system is capable of performing heterogeneous tasks encountered in metabolomics research thus far and opens up the possibility for rapid development of novel analysis and preprocessing features. Even though, it has to be assumed that there will be more extensions required than could be envisioned at the moment.

Hence, extensibility has been a key feature of MeltDB already by design. Several extensions could already be made, such as the efficient mass decomposition algorithms and the classification and machine learning extension, delivering another unique feature of this system compared to other analysis tools. In addition, methods such as the integrated metabolite correlation analysis can lead to the formulation of novel hypotheses. Even if the automatic detection of hidden metabolic associations by the correlation analysis on metabolomics data alone is open to question, the observations generated by MeltDB can give evidence for the chemical identity of yet unknown compounds. Still, it remains the task of the researcher to validate the generated hypothesis together with the integration of metabolic pathway information and e.g. genome annotation data. To support these efforts, MeltDB already provides the connection to metabolic pathway repositories and the prokaryotic genome annotation projects in GenDB.

The decision to seamlessly integrate both R and BioConductor functionality has paid off. Simple integration of multivariate statistics together with explorative data visualization allows to perform state-of-the-art analysis methods (ICA, SVM, Random Forest) on any experimental dataset stored in MeltDB.

With respect to the integration of data from different sources, MeltDB has contributed to the multi-*Omics* platform that is in the meanwhile established at the Center for Biotechnology. It is e.g. possible to automate integrated data analysis tasks by the use of Web Services. The merits of data integration could be demonstrated by the evaluation of MeltDB experiments in combination with expression-data stored in the Emma 2 system that were combined and analyzed on metabolic pathway maps using the ProMeTra application.

It should be emphasized, that the approach of data integration of different *Omics* data sources is also feasible without any specialized software. Genome annotations as well as expression data can be downloaded from public resources manually. The

added value of a software framework consisting of several linked applications is the maintenance of consistent references between biological sequence entries and other relevant objects within the framework. But, most importantly, it is the automation of an otherwise tedious task, which makes Web Service-based data integration of MeltDB and ProMeTra so attractive.

The web based approach of MeltDB makes it possible to use the system in distributed research projects and across institutes all over the world. In contrast to ordinary repositories or data warehouses that only provide data, MeltDB combines both, data and applicable methods and algorithms. Thereby, the system allows to share expertise on the stored data together with data analysis functions and optimized tools between the participating institutions. The system makes it possible to compare the results achieved by different laboratories in a standardized manner. Furthermore, methods for data analysis can be evaluated directly within the analysis environment based on a solid foundation of real-world data.

All in all, MeltDB is a comprehensive system for the integrated analysis of hyphenated mass spectrometry measurements and marks an important advance in the computational analysis of metabolomics experiments.

CHAPTER 8

Conclusion and Outlook

MeltDB was designed as a platform-independent software for the analysis and integration of metabolomics experiments. During the evolutionary development process, various aspects of the system have been extended. It was possible to integrate support for additional MS based analytical platforms. Part of the flexibility can be attributed to the generic data model that has been designed initially. The software infrastructure implemented for MeltDB made it possible to efficiently realize a second application. The ProMeTra system is based on the generic web based architecture and concepts realized for MeltDB initially. Both systems are already featured in a current review on the visualization of *Omics* data for systems biology (Gehlenborg *et al.*, 2009).

From the user perspective, MeltDB provides a large set of methods for metabolomics data analysis. Furthermore, it provides unique features for management, retrieval and exchange of the contained data. With its integrative capabilities of other functional genomics resources, a useful combination of methods has been established, that allows to push metabolomics research forward. Consequently, MeltDB is a versatile systems biology platform to support the efficient and large-scale analysis of metabolomics experiments.

Yet, several extensions to the system can be envisioned. The number of approaches for the elucidation of the chemical identity of yet unknown metabolites in existing MeltDB experiments has already been increased through the integration of novel algorithms and analysis concepts detailed in the Implementation Chapter. Nonetheless, an additional emphasis should be put on this aspect in continuing work. Especially, since the profiling analysis shows that there are various yet unidentified compounds showing interesting changes in metabolite pool level under the experimental conditions as described in Chapter 6.

Currently, most of the mass spectral databases for EI fragmentation provide only nominal mass resolution (GMD, NIST). Therefore, the support for mass spectra with higher mass resolution could be integrated in the MeltDB system to improve the identification and especially the differentiation of similar compounds. The MeltDB database already contains hundreds of measurements from various instruments, several of them containing multiple reference compounds. These datasets could be an appropriate pool for the generation of a highly resolved mass spectral reference database. To make use of these high resolution data, improved methods for the comparison of the mass spectra need to be established.

The support for metabolic flux analyses is another logical extension of the MeltDB platform. Isotope labeled carbon (^{13}C) or nitrogen (^{15}N) atoms are typically used in metabolic flux analysis studies. The integration of such atoms in compounds of interest through the cellular metabolism leads to shifts of the measured isotope patterns in the corresponding mass spectra. Currently, the identification and quantitation of these shifts is already supported by the MeltDB API. Together with the integration of GenDB and KEGG, a possible generation of draft models and stoichiometric matrices of metabolic pathways for prokaryotes is in reach. Thus, MeltDB could be an ideal platform for the generation of quantitative metabolite labeling data being the prerequisite for modeling and simulation approaches.

The application of MeltDB in various national and international cooperations and research projects leads to increasing data volumes that need to be stored and analyzed. To cope with the increasing number of users and experiments, extensions and adaptations to the compute infrastructure for storage and computation might become mandatory. Due to the three tier and cluster based architecture of MeltDB, additional compute nodes and data storage capacities can easily be integrated.

All in all, the presented MeltDB system is an important milestone for the analysis of metabolomics experiments and offers the necessary flexibility to address future challenges in computational metabolomics research. Metabolomics is already applied for the system-wide analysis of biological systems and recent advances have added to the comprehensiveness of the approach.Thus, the field represents an important building block of systems biology. With the integration of high throughput data from metabolomics experiments, the next step toward a holistic understanding of cellular behavior could be accomplished.

APPENDIX A

ERT construction & Data model

Algorithm 2 Construction algorithm of the Extended Residue Table (ERT) used in the FIND-ALL algorithm described by Böcker and Lipták (2005). The algorithm computes the ERT of a weighted alphabet $a_1, \ldots, a_k$ with the smallest mass a_1 in runtime $O(ka_1)$. The function $\gcd(x, y)$ returns the greatest common divisor of x and y.

for $i = 1$ to k; **do**
  $ert(0, i) = 0$;
  for $r = 1$ to $a_1 - 1$ **do**
    $ert(r, i) = \infty$;
  end for
end for
for $i = 2$ to k **do**
  $d \leftarrow \gcd(a_1, a_i)$;
  for $t = 0$ to $d - 1$ **do**
    find $n = \min\{\ \mathrm{ert}(q, i-1) | q \bmod d = t\}$; $\mathrm{ert}(n \bmod a_1, i) \leftarrow n$;
    if $n < \infty$ **then**
      for 1 to $a_1/d - 1$ **do**
        $n \leftarrow n + a_i; r \leftarrow n \bmod a_1$;
        $n \leftarrow \min\{n, \mathrm{ert}(r, i-1)\}$; $\mathrm{ert}(r, i) \leftarrow n$;
      end for
    end if
  end for
end for

Figure A.1.: Visualization of all classes and their inheritance relations defined in the object-relation database model of MeltDB. The design has been realized using the O2DBI software and a formal definition of the class hierarchies in XML format is available. Core APIs in Perl and Java together with programmer documentation are generated based on this XML model. In addition, the SQL database structure for the object-relational representation of all classes is available.

A.1. Abbreviations

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
CE	Capillary Electrophoresis
CE-MS	Capillary Electrophoresis coupled to Mass Spectrometry
EIC	Extracted Ion Current or Extracted Ion Chromatogram
GC	Gas Chromatography coupled to Mass Spectrometry
GC-MS	Gas Chromatography
GMD	Golm Metabolite Database
HCA	Hierarchical Cluster Analysis
HPLC	High Performance Liquid Chromatography
ICA	Independent Component Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC	Liquid Chromatography
LC-MS	Liquid Chromatography coupled to Mass Spectrometry
MSI	Metabolomics Standards Initiative
MS	Mass Spectrum or Mass Spectrometry
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
SEC	Size Exclusion Chromatography
SOAP	Simple Object Access Protocol
SVM	Support Vector Machine
SVG	Scalable Vector Graphics
TIC	Total Ion Current or Total Ion Chromatogram
WSDL	Web Service Definition Language
XCC	*Xanthomonas campestris* pv. campestris
XML	Extensible Markup Language

Bibliography

Albaum S., Neuweger H., Fränzel B., Lange S., Mertens D., Trötschel C., Wolters D., Kalinowski J., Nattkemper T., Goesmann A.: Qupe - a rich internet application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. *Bioinformatics*, in press.

Altshuler D., Daly M., Kruglyak L.: Guilt by association. *Nat Genet*, 26(2):135–7, (2000).

Andreev V. P., Rejtar T., Chen H.-S., Moskovets E. V., Ivanov A. R., Karger B. L.: A universal denoising and peak picking algorithm for lc-ms based on matched filtration in the chromatographic time domain. *Anal Chem*, 75(22):6314–6326, (2003).

Ardrey R. E.: *Liquid chromatography, mass spectrometry - An introduction*. Wiley, Chichester [u.a.] (2005).

Barsch A., Patschkowski T., Niehaus K.: Comprehensive metabolite profiling of sinorhizobium meliloti using gas chromatography-mass spectrometry. *Funct Integr Genomics*, 4(4):219–230, (2004).

Baumbach J., Brinkrolf K., Czaja L., Rahmann S., Tauch A.: Coryneregnet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. *BMC Genomics*, 7(1):24, (2006).

Bedair M., Sumner L. W.: Current and emerging mass-spectrometry technologies for metabolomics. *Trends in Analytical Chemistry*, 27(3):1–13, (2008).

Benecke C., Grüner T., Kerber A., Laue R., Wieland T.: Molecular structure generation with molgen, new features and future developments. *Fresenius' Journal of Analytical Chemistry*, 359:23–32, (1997).

Bennett K., Layzell P., Budgen D., Brereton P., Macaulay L., Munro M.: Service-based software: The future for flexible software. *Proceeding of the Seventh Asia-Pacific Software Engineering Conference*, APSEC 2000:214–221, (2000).

Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Wheeler D. L.: Genbank. *Nucleic Acids Res*, 36(Database issue):D25–30, (2008).

Bino R. J., Hall R. D., Fiehn O., Kopka J., Saito K., Draper J., Nikolau B. J., Mendes P., Roessner-Tunali U., Beale M. H., Trethewey R. N., Lange B. M., Wurtele E. S., Sumner L. W.: Potential of metabolomics as a functional genomics tool. *Trends Plant Sci*, 9(9):418–425, (2004).

Böcker S., Letzel M. C., Lipták Z., Pervukhin A.: Decomposing metabolomic isotope patterns. *Proc. of the 6th Workshop on Algorithms in Bioinformatics*, pages 1–12.

Böcker S., Letzel M. C., Lipták Z., Pervukhin A.: Sirius: decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–24, (2009).

Böcker S., Lipták Z.: Efficient mass decomposition. In: *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 151–157, ACM, New York, NY, USA (2005).

Böcker S., Lipták Z., Martin M., Pervukhin A., Sudek H.: Decomp–from interpreting mass spectrometry peaks to solving the money changing problem. *Bioinformatics*, 24(4):591–3, (2008).

Boser B., Guyon I., Vapnik V.: A training algorithm for optimal margin classifiers. *Proceedings Fifth ACM Workshop on Computational Learning Theory*, pages 144–152.

Breiman L.: Random forests. *Machine Learning*, 45(5-32):1–28, (2001).

Britz-Mckibbin P., Nishioka T., Terabe S.: Sensitive and high-throughput analyses of purine metabolites by dynamic ph junction multiplexed capillary electrophoresis: a new tool for metabolomic studies. *Analytical sciences : the international journal of the Japan Society for Analytical Chemistry*, 19(1):99–104, (2003).

Broeckling C. D., Huhman D. V., Farag M. A., Smith J. T., May G. D., Mendes P., Dixon R. A., Sumner L. W.: Metabolic profiling of medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J Exp Bot*, 56(410):323–36, (2005).

Broeckling C. D., Reddy I. R., Duran A. L., Zhao X., Sumner L. W.: Met-idea: data extraction tool for mass spectrometry-based metabolomics. *Anal Chem*, 78(13):4334–4341, (2006).

Brown M., Grundy W., Lin D., Cristianini N.: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, (2000).

Buchholz A., Hurlebaus J., Wandrey C., Takors R.: Metabolomics: quantification of intracellular metabolite dynamics. *Biomol Eng*, 19(1):5–15, (2002).

Caspi R., Foerster H., Fulcher C. A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S. Y., Shearer A. G., Tissier C., Walk T. C., Zhang P., Karp P. D.: The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 36(Database issue):D623–31, (2008).

Curbera F., Duftler M., Khalaf R., Nagy W., Mukhi N., Weerawarana S.: Unraveling the web services web: an introduction to soap, wsdl, and uddi. *IEEE Internet computing*, 6(2):86–93, (2002).

Danielsson R., Bylund D., Markides K. E.: Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectrain liquid chromatography-mass spectrometry. *Analytica Chimica Acta*, 454:167–184, (2002).

Deutsch E.: mzml: a single, unifying data format for mass spectrometer output. *Proteomics*, 8(14):2776–7, (2008).

Dondrup M., Albaum S. P., Griebel T., Henckel K., Jünemann S., Kahlke T., Kleindt C. K., Küster H., Linke B., Mertens D., Mittard-Runte V., Neuweger H., Runte K. J., Tauch A., Tille F., Pühler A., Goesmann A.: Emma 2–a mage-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics*, 10:50, (2009).

Dondrup M., Goesmann A., Bartels D., Kalinowski J., Krause L., Linke B., Rupp O., Sczyrba A., Pühler A., Meyer F.: Emma: a platform for consistent storage and efficient analysis of microarray data. *J Biotechnol*, 106(2-3):135–46, (2003).

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

Dunn W. B., Ellis D. I.: Metabolomics: Current analytical plattforms and methodologies. *Trends in Analytical Chemistry*, 4:285 –294, (2005).

Ettre L.: New, unified nomenclature for chromatography. *Chromatographia*, 32(7-8):521–526, (1994).

Fernie A. R., Trethewey R. N., Krotzky A. J., Willmitzer L.: Metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol Cell Biol*, 5(9):763–9, (2004).

Fiehn O.: Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics*, 2(3):155–68, (2001).

Fiehn O.: Metabolomics–the link between genotypes and phenotypes. *Plant Mol Biol*, 48(1-2):155–171, (2002).

Fiehn O., Kopka J., Dörmann P., Altmann T., Trethewey R. N., Willmitzer L.: Metabolite profiling for plant functional genomics. *Nat Biotechnol*, 18(11):1157–61, (2000).

Fiehn O., Kristal B., van Ommen B., Sumner L. W., Sansone S.-A., Taylor C., Hardy N., Kaddurah-Daouk R.: Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *OMICS*, 10(2):158–163, (2006).

Fiehn O., Wohlgemuth G., Scholz M., Kind T., Lee D. Y., Lu Y., Moon S., Nikolau B.: Quality control for plant metabolomics: reporting msi-compliant studies. *Plant J*, 53(4):691–704, (2008).

Fraser P. D., Pinto M. E., Holloway D. E., Bramley P. M.: Technical advance: application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. *Plant J*, 24(4):551–8, (2000).

Gabriel K. R.: The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, (1971).

Gamma E., Helm R., Johnson R., Vlissides J.: *Design Patterns. Elements of Reusable Object-Oriented Software.*. Addison-Wesley (1995).

Gehlenborg N., Baliga N. S., Goesmann A., Hibbs M. A., Kitano H., Kohlbacher O., Neuweger H., Schneider R., Tenenbaum D., Gavin A.-C.: Visualization of omics data for systems biology. *Nature Methods*, submitted.

Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A. J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J. Y. H., Zhang J.: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, (2004).

Gezelbash, Niehaus K., Neuweger H., Schwientek P., Morshuis M., Schulte-Eistrup S., Koerfer R., Milting H.: Metabolomic profiling of the terminal failing human myocardium pre and post mechanical unloading by ventricular assist devices. *Journal of Heart and Lung Transplantation*, 28(2):82, (2009).

Goesmann A., Linke B., Bartels D., Dondrup M., Krause L., Neuweger H., Oehm S., Paczian T., Wilke A., Meyer F.: Brigep–the bridge-based genome-transcriptome-proteome browser. *Nucleic Acids Res*, 33(Web Server issue):W710–W716, (2005).

Goesmann A., Linke B., Rupp O., Krause L., Bartels D., Dondrup M., McHardy A. C., Wilke A., Pühler A., Meyer F.: Building a bridge for the integration of heterogeneous data from functional genomics into a platform for systems biology. *J Biotechnol*, 106(2-3):157–67, (2003).

Goodacre R., Vaidyanathan S., Dunn W. B., Harrigan G. G., Kell D. B.: Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol*, 22(5):245–52, (2004).

Gross J. H.: *Mass spectrometry - A Textbook*. Springer, Berlin [u.a.] (2004).

Halket J. M., Waterman D., Przyborowska A. M., Patel R. K. P., Fraser P. D., Bramley P. M.: Chemical derivatization and mass spectral libraries in metabolic profiling by gc/ms and lc/ms/ms. *J Exp Bot*, 56(410):219–43, (2005).

Halket J. M., Zaikin V. G.: Derivatization in mass spectrometry–1. silylation. *European journal of mass spectrometry (Chichester, England)*, 9(1):1–21, (2003).

Hall R. D.: Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol*, 169(3):453–68, (2006).

Hastie T., Tibshirani R., Friedman J. H.: *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, NY (2009).

Hoffmann N., Stoye J.: Chroma: Signal based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics*.

Hou Y., Hsu W., Lee M., Bystroff C.: Efficient remote homology detection using local structure. *Bioinformatics*, 17(19):2294–2301, (2003).

Huhman D. V., Sumner L. W.: Metabolic profiling of saponins in medicago sativa and medicago truncatula using hplc coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry*, 59(3):347–60, (2002).

Hyvärinen A., Oja E.: Independent component analysis: algorithms and applications. *Neural Netw*, 13(4-5):411–430, (2000).

Ihaka R., Gentleman R.: R: a language for data analysis and graphics. *Journal of computational and graphical statistics*.

Jaakkola T., Diekhans M., Haussler D.: A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*.

Jenkins H., Johnson H., Kular B., Wang T., Hardy N.: Toward supportive data collection tools for plant metabolomics. *Plant Physiol*, 138(1):67–77, (2005).

Jonsson P., Johansson A. I., Gullberg J., Trygg J., A J., Grung B., Marklund S., Sjöström M., Antti H., Moritz T.: High-throughput data analysis for detecting and identifying differences between samples in gc/ms-based metabolomic analyses. *Anal Chem*, 77(17):5635–5642, (2005).

Junker B. H., Klukas C., Schreiber F.: Vanted: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109, (2006).

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K. F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M.: From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–D357, (2006).

Karp P. D., Riley M., Paley S. M., Pellegrini-Toole A.: The metacyc database. *Nucleic Acids Res*, 30(1):59–61, (2002).

Katajamaa M., Miettinen J., Oresic M.: Mzmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636, (2006).

Kell D. B., Brown M., Davey H. M., Dunn W. B., Spasic I., Oliver S. G.: Metabolic footprinting and systems biology: the medium is the message. *Nat Rev Microbiol*, 3(7):557–65, (2005).

Keun H., Ebbels T., Antti H., Bollard M.: Improved analysis of multivariate data by variable stability scaling: application to nmr-based .... *Analytica Chimica Acta*.

Kienitz H. H., Aulinger F.: *Massenspektrometrie*. Verl. Chemie, Weinheim/Bergstrae (1968).

Kind T., Tolstikov V., Fiehn O., Weiss R. H.: A comprehensive urinary metabolomic approach for identifying kidney cancerr. *Anal Biochem*, 363(2):185–95, (2007).

Kopka J.: Current challenges and developments in gc-ms based metabolite profiling technology. *J Biotechnol*, 124(1):312–22, (2006).

Kopka J., Fernie A., Weckwerth W., Gibon Y., Stitt M.: Metabolite profiling in plant biology: platforms and destinations. *Genome Biol*, 5(6):109, (2004).

Kopka J., Schauer N., Krueger S., Birkemeyer C., Usadel B., Bergmüller E., Dörmann P., Weckwerth W., Gibon Y., Stitt M., Willmitzer L., Fernie A. R., Steinhauser D.: Gmd@csb.db: the golm metabolome database. *Bioinformatics*, 21(8):1635–1638, (2005).

Kuhn M.: Building predictive models in r using the caret package. *JSS Journal of Statistical Software*, 28(5):1–26, (2008).

Lam H., Deutsch E. W., Eddes J. S., Eng J. K., King N., Stein S. E., Aebersold R.: Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7(5):655–667, (2007).

Lockhart D. J., Dong H., Byrne M. C., Follettie M. T., Gallo M. V., Chee M. S., Mittmann M., Wang C., Kobayashi M., Horton H., Brown E. L.: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80, (1996).

Lommen A.: Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem*, 81(8):3079–86, (2009).

Lüdemann A., Strassburg K., Erban A., Kopka J.: Tagfinder for the quantitative analysis of gas chromatography–mass spectrometry (gc-ms)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737, (2008).

McLafferty F. W.: *Interpretation of mass spectra*. University Science Books (1993).

Meyer F., Goesmann A., McHardy A. C., Bartels D., Bekel T., Clausen J., Kalinowski J., Linke B., Rupp O., Giegerich R., Pühler A.: Gendb–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, 31(8):2187–2195, (2003).

Mlecnik B., Scheideler M., Hackl H., Hartler J., Sanchez-Cabo F., Trajanoski Z.: Pathwayexplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res*, 33(Web Server issue):W633–7, (2005).

Monton M. R. N., Soga T.: Metabolome analysis by capillary electrophoresis-mass spectrometry. *Journal of Chromatography A*, 1168(1-2):237–46; discussion 236, (2007).

Morgenthal K., Weckwerth W., Steuer R.: Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *BioSystems*, 83(2-3):108–17, (2006).

Mueller L. A., Zhang P., Rhee S. Y.: Aracyc: a biochemical pathway database for arabidopsis. *Plant Physiol*, 132(2):453–60, (2003).

Needleman S. B., Wunsch C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53, (1970).

Nesbitt C. A., Zhang H., Yeung K. K.-C.: Recent applications of capillary electrophoresis-mass spectrometry (ce-ms): Ce performing functions beyond separation. *Analytica Chimica Acta*, 627(1):3–24, (2008).

Neuweger H., Albaum S. P., Dondrup M., Persicke M., Watt T., Niehaus K., Stoye J., Goesmann A.: Meltdb: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, 24(23):2726–32, (2008).

Neuweger H., Baumbach J., Albaum S., Bekel T., Dondrup M., Hüser A. T., Kalinowski J., Oehm S., Pühler A., Rahmann S., Weile J., Goesmann A.: Corynecenter - an online resource for the integrated analysis of corynebacterial genome and transcriptome data. *BMC systems biology*, 1:55, (2007).

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

Nielsen N., Carstensen J., Smedsgaard J.: Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805(1–2):17–35, (1998).

Nikolau B. J. H.: *Concepts in plant metabolomics*. Springer, Dordrecht (2007).

Nobeli I., Ponstingl H., Krissinel E. B., Thornton J. M.: A structure-based anatomy of the e.coli metabolome. *J Mol Biol*, 334(4):697–719, (2003).

Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M.: Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34, (1999).

Oliver S. G., Winson M. K., Kell D. B., Baganz F.: Systematic functional analysis of the yeast genome. *Trends Biotechnol*, 16(9):373–8, (1998).

Orchard S., Montechi-Palazzi L., Deutsch E. W., Binz P.-A., Jones A. R., Paton N., Pizarro A., Creasy D. M., Wojcik J., Hermjakob H.: Five years of progress in the standardization of proteomics data 4th annual spring workshop of the hupo-proteomics standards initiative april 23-25, 2007 ecole nationale suprieure (ens), lyon, france. *Proteomics*, 7(19):3436–3440, (2007).

Paley S. M., Karp P. D.: The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Res*, 34(13):3771–8, (2006).

Pedrioli P. G. A., Eng J. K., Hubley R., Vogelzang M., Deutsch E. W., Raught B., Pratt B., Nilsson E., Angeletti R. H., Apweiler R., Cheung K., Costello C. E., Hermjakob H., Huang S., Julian R. K., Kapp E., McComb M. E., Oliver S. G., Omenn G., Paton N. W., Simpson R., Smith R., Taylor C. F., Zhu W., Aebersold

R.: A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*, 22(11):1459–1466, (2004).

Philippi S., Köhler J.: Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet*, 7(6):482–8, (2006).

Plassmeier J., Barsch A., Persicke M., Niehaus K., Kalinowski J.: Investigation of central carbon metabolism and the 2-methylcitrate cycle in corynebacterium glutamicum by metabolic profiling using gas chromatography-mass spectrometry. *J Biotechnol*, 130(4):354–63, (2007).

Poole C. F.: *The essence of chromatography*. Elsevier Science Ltd (December 2002).

Prakash A., Mallick P., Whiteaker J., Zhang H., Paulovich A., Flory M., Lee H., Aebersold R., Schwikowski B.: Signal maps for mass spectrometry-based comparative proteomics. *Mol Cell Proteomics*, 5(3):423–32, (2006).

Ramautar R., Somsen G. W., de Jong G. J.: Ce-ms in metabolomics. *Electrophoresis*, 30(1):276–91, (2009).

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

Roessner U., Wagner C., Kopka J., Trethewey R. N., Willmitzer L.: Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J*, 23(1):131–42, (2000).

Sansone S.-A., Fan T., Goodacre R., Griffin J. L., Hardy N. W., Kaddurah-Daouk R., Kristal B. S., Lindon J., Mendes P., Morrison N., Nikolau B., Robertson D., Sumner L. W., Taylor C., van der Werf M., van Ommen B., Fiehn O.: The metabolomics standards initiative. *Nat Biotechnol*, 25(8):846–848, (2007).

Schauer N., Steinhauser D., Strelkov S., Schomburg D., Allison G., Moritz T., Lundgren K., Roessner-Tunali U., Forbes M. G., Willmitzer L., Fernie A. R., Kopka J.: Gc-ms libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett*, 579(6):1332–7, (2005).

Scholz M., Fiehn O.: Setupx–a public study design database for metabolomic projects. *Pac Symp Biocomput*, 12:169–180, (2007).

Schymanski E. L., Meringer M., Brack W.: Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Anal Chem*, 81(9):3608–17, (2009).

Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., Ideker T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, (2003).

Shevchenko A., Jensen O. N., Podtelejnikov A. V., Sagliocco F., Wilm M., Vorm O., Mortensen P., Shevchenko A., Boucherie H., Mann M.: Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci USA*, 93(25):14440–5, (1996).

Smith C. A., O'Maille G., Want E. J., Qin C., Trauger S. A., Brandon T. R., Custodio D. E., Abagyan R., Siuzdak G.: Metlin: a metabolite mass spectral database. *Ther Drug Monit*, 27(6):747–751, (2005).

Smith C. A., Want E. J., O'Maille G., Abagyan R., Siuzdak G.: Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, 78(3):779–787, (2006).

Soga T., Ohashi Y., Ueno Y., Naraoka H., Tomita M., Nishioka T.: Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J Proteome Res*, 2(5):488–94, (2003).

Soga T., Ueno Y., Naraoka H., Ohashi Y., Tomita M., Nishioka T.: Simultaneous determination of anionic intermediates for bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal Chem*, 74(10):2233–9, (2002).

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

Stein S., Scott D.: Optimization and testing of mass spectra library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom*, 5:859–866, (1994).

Steinfath M., Groth D., Lisec J., Selbig J.: Metabolite profile analysis: from raw data to regression and classification. *Physiologia Plantarum*, 132:150–161, (2008).

Steuer R., Kurths J., Fiehn O., Weckwerth W.: Interpreting correlations in metabolomic networks. *Biochem Soc Trans*, 31(Pt 6):1476–8, (2003a).

Steuer R., Kurths J., Fiehn O., Weckwerth W.: Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026, (2003b).

Sumner L. W., Mendes P., Dixon R. A.: Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6):817–36, (2003).

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

Thimm O., Bläsing O., Gibon Y., Nagel A., Meyer S., Krüger P., Selbig J., Müller L. A., Rhee S. Y., Stitt M.: Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, 37(6):914–939, (2004).

Thompson J. D., Higgins D. G., Gibson T. J.: Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, (1994).

Tokimatsu T., Sakurai N., Suzuki H., Ohta H., Nishitani K., Koyama T., Umezawa T., Misawa N., Saito K., Shibata D.: Kappa-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiol*, 138(3):1289–300, (2005).

Tolstikov V. V., Fiehn O.: Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal Biochem*, 301(2):298–307, (2002).

Toyo'oka T.: *Modern Derivatization Methods for Separation Science*. Wiley (Jan 2000).

Unidata: Unidata netcdf (2008).

van den Berg R. A., Hoefsloot H. C., Westerhuis J. A., Smilde A. K., van der Werf M. J.: Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(1):142, (2006).

van Nederkassel A. M., Daszykowski M., Eilers P. H. C., Heyden Y. V.: A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118(2):199–210, (2006).

Villas Bôas S. G.: *Metabolome analysis - An introduction*. Wiley, Hoboken, NJ (2007).

Vorhölter F.-J., Schneiker S., Goesmann A., Krause L., Bekel T., Kaiser O., Linke B., Patschkowski T., Rückert C., Schmid J., Sidhu V. K., Sieber V., Tauch A., Watt S. A., Weisshaar B., Becker A., Niehaus K., Pühler A.: The genome of xanthomonas campestris pv. campestris b100 and its use for the reconstruction of metabolic pathways involved in xanthan biosynthesis. *J Biotechnol*, 134(1-2):33–45, (2008).

Weckwerth W.: Metabolomics in systems biology. *Annual review of plant biology*, 54:669–89, (2003).

Weckwerth W., Loureiro M. E., Wenzel K., Fiehn O.: Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc Natl Acad Sci USA*, 101(20):7809–14, (2004).

Wiley: Nist/epa/nih mass spectral library 2005 (Jan 2005).

Wilke A., Rückert C., Bartels D., Dondrup M., Goesmann A., Hüser A. T., Kespohl S., Linke B., Mahne A., M. an d McHardy, Pühler A., Meyer F.: Bioinformatics support for high-throughput proteomics. *J Biotechnol*, 106(2-3):147–156, (2003).

Wishart D. S., Tzur D., Knox C., Eisner R., Guo A. C., Young N., Cheng D., Jewell K., Arndt D., Sawhney S., Fung C., Nikolai L., Lewis M., Coutouly M.-A., Forsythe I., Tang P., Shrivastava S., Jeroncic K., Stothard P., Amegbey G., Block D., Hau D. D., Wagner J., Miniaci J., Clements M., Gebremedhin M., Guo N., Zhang Y., Duggan G. E., Macinnis G. D., Weljie A. M., Dowlatabadi R., Bamforth F., Clive D., Greiner R., Li L., Marrie T., Sykes B. D., Vogel H. J., Querengesser L.: Hmdb: the human metabolome database. *Nucleic Acids Res*, 35(Database issue):D521–6, (2007).

Wolfender J.-L., Ndjoko K., Hostettmann K.: Liquid chromatography with ultraviolet absorbance-mass spectrometric detection and with nuclear magnetic resonance spectroscopy: a powerful combination for the on-line structural investigation of plant metabolites. *Journal of Chromatography A*, 1000(1-2):437–55, (2003).

Xia J., Psychogios N., Young N., Wishart D.: Metaboanalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*, 37(Web Server issue):652–660, (2009).

Acknowledgments

First of all, I wish to thank Prof. Dr. Jens Stoye and Prof. Dr. Karsten Niehaus for the opportunity to do my PhD thesis under their supervision. I am very grateful for their support, fruitful and critical discussions and inspiration. Also, I am very grateful to Dr. Alexander Goesmann for the opportunity to develop MeltDB within the Computational Genomics group and the BRF and for his overall support in pushing this project forward. I also thank all the people of the BRF and the Computational Genomics group, for proof-reading the manuscripts, technical support, and providing a positive work atmosphere.

I would like to acknowledge the International Graduate School in Bioinformatics and Genome Research for funding the PhD project, the executive directors Dr. Dirk Evers and Dr. Susanne Schneiker-Bekel, and all PhD students of the graduate school for their support.

Diploma, Bachelor and Master Theses have been supervised in the course of this work and I would like to thank the following students and PhD candidates for their help and contributions to MeltDB and ProMeTra: Nils Hoffmann, Jörn Winnebald, Daniel Dörr, and Leonhard Jonathan Stutz.

I wish to thank Dr. Michael Dondrup, Dr. Jan Baumbach, Marcus Persicke, Tony Watt, Sarah Schatschneider and Tobias Thüte for the pleasant collaboration, for providing measurement data, and for their excellent feedback and fruitful discussions regarding the features of the MeltDB system.

Furthermore, I would like to express my gratitude to all my friends who have helped me on the way, especially I wish to thank Dr. Robert Kasper for proof-reading and interdisciplinary friendship and Stefan Albaum for countless corrections and expert advice.

I am very grateful to my parents for their love and trust and the support throughout my studies.

Finally, I wish to thank my girlfriend Britta for her amazing and invaluable support.

Bielefeld, November 2009

Heiko Neuweger

ERKLÄRUNG

Ich, Heiko Neuweger, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel benutzt habe.

Bielefeld, den 9. November 2009

Heiko Neuweger