The handle http://hdl.handle.net/1887/26892 holds various files of this Leiden University dissertation

**Author:** Rojas-Chertó, Miquel  
**Title:** Towards automated identification of metabolites using mass spectral trees  
**Issue Date:** 2014-06-19
CHAPTER 1

Introduction
Biology has been experiencing a revolution with regards to an exponential increase of acquisition of highly relevant and rich biological data. Handling this huge amount of biological data for the so-called high-throughout ‘omics’ techniques (e.g. genomics, proteomics, transcriptomics, and metabolomics) resulted in the development of many new and improved analytical and bioinformatics/cheminformatics platforms [Romero et al., 2006]. Metabolomics has become one of the recent emerging ‘omics’ sciences and it deals with the composition and quantification of all (or at least many) endogenous and exogenous compounds with low molecular weight (metabolites) which are involved in all sorts of biochemical processes occurring in biological systems (i.e., cells, tissues, biofluids, or even the whole organism) [Fiehn, 2002, Gibney et al., 2005].

This introduction is divided into 5 sections describing general topics of the metabolomics field and relevant themes presented within this thesis. Section 1.1 provides a general overview of metabolomics, its related research fields and applications. The study of the metabolome (i.e. the complete collection of metabolites) requires execution of a variety of analytical platforms, of which many are based on mass spectrometry, which are described in Section 1.2. Section 1.3 describes the standard processing steps needed to analyze raw spectral data originating from these analytical platforms and the different processing strategies to handle them. Storage of these processed mass spectra data is essential to study biological systems to increase biological knowledge. In Section 1.4 an overview of the different metabolomics mass spectral databases are presented. Biological interpretation (e.g. referring and putting the results into context to literature describing previously executed metabolomics research) of the observed metabolome patterns is only possible when the identities of the measured metabolites are known. For targeted analytical platforms the identities are known but for untargeted platforms metabolite identification becomes an essential part of the whole metabolomics workflow. Section 1.5 describes the methods and tools used in metabolite identification. Section 1.6 ends with a short explanation of the relevance in biological interpretation. Finally, this introduction chapter concludes with the scope and aim of the research presented in this thesis.

### 1.1. Systems Biology and Metabolomics

In systems biology the perception that the biological system describes the conduct of its different components receives increasing attention [Kitano, 2002]. A biological system can be perturbed by many external causes: altering gene sequence, the transcription of genes, the expression and post-translational modification of proteins, and the composition and abundance of metabolites. The study of the entire biological system has been reinforced by the emergence of the different ‘omics’ tools (e.g. genomics, transcriptomics, proteomics and
metabolomics) and the associated generated high-throughput data [Romero et al., 2006]. Building models describing biology is an ambitious challenge because the biological processes are dynamic and depend heavily on the cell types, the organ, the organ-organ interactions, and the interaction between a system and the environmental conditions.

Among the ‘-omics’ technologies, metabolomics is the scientific study of chemical processes involving metabolites. The metabolome represents the collection of all metabolites, which are the end products of the cellular processes, in a biological cell, tissue, organ, or organism. Metabolites are key in linking the phenotype-genotype gap, since they reflect more directly the cellular physiological states as being the most downstream in the ‘omics’ family, as shown in Figure 1.1. Metabolomics enables the measurement of the state of a biological system at a specific moment in time within a particular genetic or environmental context reflected by the phenotypic change [Brown et al., 2005].

In human-based metabolomics, metabolites are commonly classified as endogenous or exogenous; where metabolites produced by the host organism are defined as endogenous and exogenous as metabolites that are coming from outside of the organism, such as food nutrients [Dunn, 2008]. In contrast, in plant-based metabolomics, it is more common...
to describe metabolites as being either ‘primary metabolites’, which are directly involved in growth, development and reproduction, and ‘secondary metabolites’, which are only indirectly involved in those processes, but play an important role in for example plant defense. The total size of the metabolome remains imprecise, however, several estimations have been proposed. Wishart [Wishart et al., 2009] quantifies several thousands of metabolites in humans, while Fiehn [Fiehn, 2002] estimates the number of metabolites in plants to be several hundred thousand.

The first studies of metabolites present in biological systems can be dated back to ancient China (1500-2000 BC), where doctors diagnosed diabetes by using ants as a detector of high glucose levels in human urine [Van Der Greef & Smilde, 2005]. However, it was Roger Williams in the late 1940s who introduced the concept of ‘metabolic pattern’ suggesting that humans might have different abundances of certain combinations of metabolites that can be detected in their biological fluids. He demonstrated it was possible, using simple paper chromatography, to identify characteristic metabolic patterns in urine and saliva, and related these with schizophrenia diseases [Gates & Sweeley, 1978]. Horning and Horning introduced the concept of ‘metabolic profile’ to describe the quantitative measurement of metabolite concentrations in urine [Horning & Horning, 1971]. Oliver proposed ‘metabolome’ as the complete set of small-molecule (< 1 kDa) endogenous metabolites in an organism [Oliver et al., 1998] and Nicholson defined ‘metabonomics’ as the ‘quantitative measurement of the dynamic multiparametric response of living systems to pathophysiological stimuli or genetic modification’ [Nicholson et al., 1999]. Fiehn extended the metabolome terminology to metabolomics as the comprehensive and quantitative analysis of all metabolites of an organism [Fiehn, 2001]. After the continuous evolution of these terms, finally the ‘metabolomics’ field concept has achieved more consensus and maturity, as observed by the formation of the Metabolomics Society in 2004 and its official journal Metabolomics in 2005.

At present, metabolomics studies have been applied in many different areas including drug development [Wishart et al., 2008], human health [Watkins & German, 2002], disease diagnosis [Kaddurah-Daouk et al., 2008], environmental science, environment toxicology [Aliferis & Chryssayi-Tokousbalides, 2010], nutrition and food science [Wishart, 2008], biological stress studies [van der Greef et al., 2004], functional genomics [Khoo & Al-Rubeai, 2007], and integrative systems biology [Goodacre et al., 2004].

One general aim of a metabolomics approach is to characterize biological indicators or profiles which can be used to interpret molecular mechanisms. The understanding relies on the identity of metabolites. Unfortunately, from the estimated hundreds of thousands of metabolites that exist in nature, the identity of a vast majority of them remains still unknown. Contrary to proteins, we can not deduce the structure of these metabolites from the genome
sequence [Gay et al., 2002]. The chemical structures of the metabolites are much more chemically variant. Therefore there is a substantial need to enlarge the list of quantified and identified metabolites and this is a major challenge in many metabolomics studies.

1.2. Analytical Instruments in Metabolomics

A perfect analytical platform to obtain quantitative data for identified metabolites might be characterized by: (i) performing direct sample analysis (no sample preparation is needed), (ii) covering all possible metabolite classes, (iii) being highly and equally sensitive to all compounds in the sample independently of their concentrations, (iv) generating reliable and reproducible results with a wide range of compounds, (v) automation of the complete process, and (vi) extracting always high-throughput data [Lenz & Wilson, 2007]. In spite of these clearly listed guidelines, none of the available analytical platforms can fulfill these all together resulting in a compromise between technological possibilities and functionality requirements.

At present, the two most common and efficient analytical techniques used to measure metabolites are Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS) based analytical methods, having both their own pros and cons [Dunn et al., 2005]. While NMR is classified as being a very robust, reproducible, and quantitative technique [Keun et al., 2002], MS is, on the other hand, known as being an extremely sensitive analytical technique [Dettmer et al., 2007]. NMR allows characterization of the chemical structure of compounds by registering the absorption of electromagnetic energy by the different atomic nuclei (such as $^1$H and $^{13}$C) by placing it in a strong magnetic field [Williams & Fleming, 2007]. Many examples have been demonstrating the efficiency using NMR in metabolomics [Krishnan et al., 2005]. MS instruments consist of three separate modules: ion source, mass analyzer, and detector. In MS ideally most of the molecules that enter the source get ionized, i.e. positively or negatively charged ions are created. Then the ions are accelerated into the mass analyzer, where they are separated according to their mass-to-charge ratio (m/z), and finally in the detector the arriving ions are registered and their number determined. From this data a mass spectrum (intensity vs m/z) can be generated, or an ion chromatogram of a certain m/z vs time can be reconstructed for quantification. In MS only compounds that are ionized will reach the detector and therefore will be detected. Most MS applications in metabolomics make use of a separation method prior to the ionization step. MS is a qualitative (i.e. for the identification of metabolites) and quantitative (i.e. for determining the amount of metabolites) technique. MS and NMR are complementary techniques that can be combined for e.g. efficient metabolite identification [van der Hooft et al., 2011]. However, MS has become the
Chapter 1: Introduction

technique of choice in many metabolomics studies because due to its major advantage for comprehensive metabolite profiling analysis of large number of low-abundance metabolites. Currently the most common separation techniques prior MS are gas chromatography (GC) [Dunn, 2008], liquid chromatography (LC) [Dunn, 2008], and capillary electrophoresis (CE) [He et al., 2007]. Separation techniques separate the molecules in the sample by passing through a separation column at different velocity. The compounds arrive at the end of the medium at different moments in time because due to their different interactions with the stationary phase (in the case of LC and GC) due to variation in chemical and physical properties to different electrophoretic mobility (CE). Such separation reduces the complexity of the data enormously and introduces an extra dimension (i.e. retention or migration time) that can be used for identification. Another important benefit of using a separation technique prior to MS is the reduction of ionization suppression [Annesley, 2003]. GC requires that the compounds are volatile and thermal stable, non-volatile metabolites can only be analyzed when prior to separation metabolites are derivatized during sample preparation [Fancy & Rumpel, 2008]. CE is designed to separate compounds based on their charge and size when they migrate through the interior of a small capillary filled with an electrolyte [Ramautar et al., 2011]. Despite the fact that CE-MS is a less frequent used for metabolomics compared to GC-MS and LC-MS, recent studies have demonstrated its potential [Ramautar et al., 2011]. In addition, new electrodriven separation approaches using nanochannels may emerge that allow ultimately better separation of metabolites [Quist et al., 2011]. The most important advantages of electrodriven separations are their very high resolving power, very small sample requirements, and their ability to separate cations, anions and uncharged molecules in a single analytical run. LC is the most versatile separation method; especially reversed-phase columns allow the separation of compounds covering a wide range of metabolite classes, however polar analytes are hardly retained. The sample is dissolved in an injection solution, which is introduced into constant flowing liquid, the 'mobile phase', and forced by a high pressure to pass through the column containing porous particles, which at their surface contain the 'stationary phase'. It should be mentioned various alternatives exist such as monolithic columns, etc, but they are not further discussed here. The specific time at which a compound elutes, called retention time, is determined by its interaction with the stationary phase. In case of complex samples, LC-MS can detect many peaks of many low concentrated compounds. LC-MS is considered the most versatile of the separation methods including normal phase (silica), reverse phase (C18,C8,C4, phenyl) [Tolstikov & Fiehn, 2002], hydrophilic interaction chromatography (HILIC) [Alpert, 2011], and ion exchange chromatography [Hamilton, 1963].

GC and LC are coupled to different types of ion sources as each technique ionizes differently the compounds. GC is generally coupled in a gas-phase environment to an electron
1.2. Analytical Instruments in Metabolomics

ionization (EI) (a so-called hard ionization techniques) or chemical ionization (CI) module to ionize and (in the case of EI) fragment the compounds. GC-EI-MS results in a robust detection of a characteristic mass spectrum per compound wherein next to the parent ion also its fragments are recorded [Dunn, 2008].

Hard-ionization techniques shoot electron beams into the analyte to generate the (fragmented) ions. Alternatively, LC coupled to MS uses generally soft-ionization techniques to transform neutral (or possibly charged) compounds into charged molecular ions. This process can be achieved through different techniques such as electrospray ionization (ESI), atmospheric-pressure chemical ionization (APCI), atmospheric-pressure photoionization (APPI), fast atom bombardment (FAB), etc... Among them, ESI is the most common choice in LC-MS-based metabolomics studies. It is known for offering the analysis of a broader coverage of the complete metabolome and generally it keeps molecular ions intact, which is very helpful for the assignment of an initial identity by matching the m/z value measured with the masses of metabolites observed earlier [Sana et al., 2008]. One of the disadvantages associated with ESI is known as ionization suppression (a process in which a mixture of compounds compete for ionization causing that certain compounds in the mixture do not ionize) [Annesley, 2003]. As we commented earlier, once the compounds are ionized the mass analyzer separates ions according to their m/z values by applying magnetic and or electric fields. Commonly used mass analyzer in the metabolomics field include quadrupole mass filters/ion traps [Koulman et al., 2007], time-of-flight (TOF) [Kind et al., 2009], Orbitrap [Hu et al., 2005], and Fourier transform ion cyclotron (FT-ICR) [Marshall et al., 1998] equipment. Each analyzer has its own advantages (and disadvantages) and the performance can be described by listing several intrinsic parameters such as (i) the mass resolving power (or resolution) defined as the averaged mass-to-charge ratio associated with two adjacent mass signals of equal size and shape, (ii) the mass accuracy defined as the difference between the theoretical exact mass of an ion and its measured mass, (iii) speed of the analysis, (iv) the linear dynamic range defined as the concentration range showing linear dependence with the ion signal measured, and (v) the sensitivity defined as the ratio between the intensity level of the mass signal and the intensity level of the noise [McLuckey & Wells, 2001].

MS is a spectrometric method that allows the detection of the mass-to-charge ratios; depending on the ionization technique used this allows to derive the molecular mass (MM) of the detected metabolite from e.g. its protonated or deprotonated ion, or certain adducts, and by that, of the elemental composition. It should be noted that often the isotopic pattern is additionally used to derive the elemental composition. Usage of tables listing molecular masses of all known observed metabolites bellows assigning possible identity to the molecular mass of the measured metabolite. However, a big issue is here that the list is not com-
Chapter 1: Introduction

**Figure 1.2: Parameters used to describe the performance of mass spectrometers**

The molecular mass is detected with a certain uncertainty, i.e. within a possible mass window, so that multiple metabolites can fit the measured mass. For a certain mass accuracy and precision, without any additional information a restriction of possible candidate elemental compositions, and even more so, possible structures, is not possible. The tandem mass spectrometry technique provides characterization of additional structural information of the detected molecules. It is denoted as MS/MS or MS2 as the end result of performing two stages of MS analysis, i.e. separation of ions according to m/z, fragmentation of ions and subsequent separation of fragment ions. This can be achieved through multiple mass analyzers connected in-space (e.g. QqQ or QqTOF) or one single mass analyzer that performs several MS experiments in-time (e.g. ion trap). This last analytical technique facilitates that ions are selected/separated based on their mass-to-charge ratio, subsequently fragmented applying high-energy, and finally the resulting generated fragments recorded as a tandem mass spectrum [Aebersold & Mann, 2003], or, if more fragmentation experiments are conducted subsequently, MS$^n$. The fragmentation spectrum containing the masses of the parent ion and its fragments depends heavily on the structure of the ion fragmented, the energy applied and other experimental parameters [McLafferty & Turecek, 1993]. There are two different modes of ion activation for posterior fragmentation: either by collision-induced-dissociations (CID) and infra-red multiphoton dissociation (IRMPD) or chemical activation modes (electron capture dissociation) involving electron transfers. In CID the ions are collided against gas molecules making the ions break and fragment.
While single-stage MS/MS generates one set of fragment ion to characterize the molecular structure, multistage MS\textsuperscript{n} generates fragment ions from fragment ions (called spectral data) providing details about the fragmentation pathways. In principle these data allow structural annotation and better identification of the metabolites since the spectral data are likely to be (at least partly) unique for each metabolite. When using LC-MS\textsuperscript{n} one still encounters some challenges such as MS/MS usually collision-induced dissociation (CID) is less reproducible than fragmentation by electron ionization (i.e. GC-MS), analysis of the data is more complicated and takes more time [Werner \textit{et al.}, 2008], searching in MS\textsuperscript{n} spectral libraries [Oberacher \textit{et al.}, 2009] is less straightforward and standardized than other libraries.

Metabolites are so much structurally different resulting in a wide range of physicochemical properties that one single analytical method does not provide the coverage of the whole metabolome, and a set of comprehensive diverse analytical techniques is necessary to cover in principle a wide range of metabolites.

\section*{1.3. Mass Spectra Data Handling in Metabolomics}

Extracting the relevant information of the overwhelming amount of data generated from an analytical platform has become an important issue in the metabolomics research field. This challenge even increases because the complete set of metabolites is characterized by largely variable chemical properties like molecular weight, polarity or solubility and the wide dynamic range of concentrations at which they occur in the biological system being studied. Data handling can be further separated into data processing and data analysis. The data processing stage consists of processing of raw data with methods which process the signals of the acquired spectra and posterior combination of the data of several measurements. The aim is to transform the raw data into an easy-to-use rather clean and less complex data format such as peak or compound lists per sample. The subsequent data analysis step focusses on the statistical analysis and interpretation of the processed data [Katajamaa, 2007].

However, before any data processing or analysis can be done, it is important to take all possible sources of know experimental variation into account. All sorts of experimental variation caused by errors during sample preparation, calibration, inclusion and detection of various kinds of contaminants, instrumental drift and detector saturation can lead less reliable raw data. Therefore, analytical and biological replicates should be measured to significantly aid to identify the sources of these errors and to reduce the variation due to these errors.

Raw data consist often of multiple-processed files, generated by a mass spectrometer
and stored in usually a vendor specific format. Generally, the company provides software packages to process the raw data and convert them to more general formats with, however, the risk of losing certain information. Unfortunately, this incompatibility of the raw data to be processed directly by other software packages, limits the control over the data and to extract all information which should be available in principle, resulting in poorer analysis accuracy despite the efforts put in proper acquiring and preprocessing the data. Nowadays several open-formats such as ASCII text, binary netCDF [Rew & Davis, 1990], JCAMP-DX [Lampen et al., 1994], mzML [Pedrioli et al., 2004], mzXML [Pedrioli et al., 2004], mzML [Deutsch, 2008], and CML [Murray-Rust et al., 2001] exist. Currently mzML has joined the latest XML format and it is intended to replace both the mzData and the mzXML format and to remain as the standard format to be used in mass spectrometry since it retains the best technical attributes of the previous formats.

In summary, preprocessing aims to reduce noise and remove artefacts, to reduce the complexity of the spectra, and/or to make spectra more comparable to allow ultimately the quantitative or qualitative analysis of the data. To be able to interpret a mass spectrum of a certain peak (and therefore compound) and compare spectra of such individual peaks across runs, the raw data must be first converted into a mass peak list for each spectrum of interest. Different steps that need to be performed are:

1. **Baseline Correction** which removes the baseline slope and offset from a spectrum.
2. **Filtering** which removes or reduces contaminants from the data.
3. **Outlier Screening** eliminate peaks which display too much deviation from the majority of their replicates (analytical or biological).
4. **Time Alignment** correcting for drifts occurring in retention time dimension to enable data comparison across samples.
5. **Data Binning** allowing data dimensionality reduction by grouping the measured data into a limited number of bins.
6. **Deconvolution** regrouping ions coming from the same metabolite.
7. **Centroiding or Peak Detection** combining multiple m/z values corresponding to a given ion into a single peak feature.
8. **Normalization peaks intensities** reduces the systematic variation of LC-MS data.

After pre-processing, the LC-MS raw data are represented by a peak list, or a compound list when identities could be assigned to the peaks. The aim of the subsequent statistical analysis step is the detection of relevant peaks which are biological significant, i.e. which
intensities/concentrations are modified between different (biological) groups of samples. LC-MS based platforms can yield a large amount of information on biological extracts, generally detecting thousands of features corresponding to parent ions, in-source fragments, and adducts of metabolites. The statistical analysis of this wealth of information can be achieved with both univariate and multivariate analysis methods. The choice of the most appropriate data analysis strategy for a given data set constitutes an important issue. The univariate approach assumes that the biological effect of interest is influenced only by one (or more) individual metabolites (or parameters). In such an approach, for each measured feature/peak the significance is calculated and thus the most relevant peaks/variables are identified to explain the difference between pre-defined (biological) groups. If should be noted that measures have to be taken to prevent possible false positives due to the large number of peaks, features and/or metabolites detected. Commonly used univariate techniques are for example a t-test, fold-change analysis, Wilcoxon rank-sum test, and analysis of variance (ANOVA). The multivariate analysis assumes that the biological effect of interest is associated with a combination of multiple peaks/features. Often, multivariate analyses are applied for visualizing the complete peak list in one single plot as a first analysis. Multivariate analysis can be further categorized into supervised and unsupervised techniques [Boccard et al., 2010]. Unsupervised methods provide a visual representation of the data by reducing the dimensionality of the data without using any prior knowledge, and are therefore useful as a first explorative data analysis. An example of unsupervised methods is principal component analysis (PCA). Supervised methods include prior knowledge about the data during statistical analysis. Examples to differentiate two classes with such an statistical methods in metabolomics are support vector machines (SVM), artificial neural networks (ANN), decision trees, and partial least squares-discriminant analysis (PLS-DA).

1.4. Databases in Metabolomics

Optimal usage of the vast amount of information generated from metabolomic experiments requires the development of databases for the storage and distribution of different type of metabolomics data. Databases are resources that facilitate data analysis and allow to retrieve relevant information for data interpretation. Currently there are a number of databases used in metabolomics; however this number is still very limited compared to genomics or proteomics. The databases can be distinguished into two groups according to their type of data: (1) metabolite centric databases and (2) study centric database. While the last group stores study specific data such as which metabolites are detected in which sample, what the design of the study is, and what other parameters and data are available, the first group consists of (a) general compound databases, (b) reference spectral databases,
(c) species specific metabolite profile databases, and (d) metabolite pathway databases.

General metabolite databases contain all kinds of physico-chemical properties of metabolites and are usually consulted to obtain the exact mass and/or elemental composition of certain known metabolites, or specific information on metabolites. Three relevant examples are PubChem [Wang et al., 2009], Human Metabolome Database (HMDB, http://www.hmdb.ca) [Wishart et al., 2008], and Chemspider [Williams & Tkachenko, 2010] databases (all freely accessible via internet).

Reference spectral databases are intended to be used as a tool for proper identification of compounds by comparing the spectra generated from an unknown compound to a spectral library or a database of reference compounds. The metabolites in the database of which the spectra match best with the spectrum of the unknown metabolite are the most probable identities of the unknown metabolite. The success of the search depends on the comprehensiveness and quality of the spectral data in the database [Ausloos et al., 1999]. So far, there are several reference spectral libraries and databases available that provide metabolomics related information. The most representative spectral libraries or databases are the US National Institute of Science and Technology database (NIST, http://www.nist.gov/srd/nist1a.htm), the Golm Metabolite Data-base (GMD) [Kopka et al., 2005], MassBank [Horai et al., 2010], METLIN [Smith et al., 2005], HMDB [Wishart et al., 2008], and the Madison Metabolomics Consortium Database (MMCD) [Cui et al., 2008].

Species specific metabolite databases are specific to a certain chemical class of metabolites, species, biofluid, and/or tissue in a given state of a biological system. Their data may consist of physical and chemical properties of metabolites, but in any case biological properties of metabolites, pathway information as well as their associated disease information, and often quantitative data of the metabolites present in the corresponding biofluids, tissues, or organs. The databases are mainly used for biological interpretation purposes, since they can be used to determine the biological functions of specific metabolites. Examples of these databases are the HMDB [Wishart et al., 2008], HMDB cerebrospinal fluid (CSF) metabolome database [Wishart et al., 2008], LIPID MAPS Structure Database (LMSD) [Sud et al., 2007], DrugBank [Wishart et al., 2006], and Yeast Metabolome Database (YMDB) [Jewison et al., 2012].

Metabolic pathway databases are repositories of biochemical pathways and reactions relating genes, enzymes and metabolites. They provide a description of which metabolites are involved in which biological reactions and processes. Examples of these databases are KNApSAcK [Afendi et al., 2012], KEGG [Kanehisa, 2002], BioCyc [Karp et al., 2005], and Reactome [Joshi-Tope et al., 2005].

LIMS is a software designed to manage laboratory information that offers data tracking support such as sample receipts, users, experimental protocols, instruments, raw data,
data processing, experiment results, and data reporting (McDowall 1988). In metabolomics several platforms have been implemented. These include SetupX [Scholz & Fiehn, 2007], Sesame LIMS [Zolnai et al., 2003], MetaboLIMS [Young et al., 2006], and Metabolomic Modeling (MeMo) [Spasìc et al., 2006].

The constant extraction of new information means that metabolite centric as study centric databases are continuously growing. This requires a significant effort to keep each database up-to-date and reliable. To proper manage, handle, and retrieve the data in the study centric database it is key to capture all relevant information in a given biological sample for data analysis and interpretation [Navarro et al., 2003]. Ideally all database should be comprehensive, user-friendly, well-annotated, and publicly available.

In the eighties, GC-MS was historically the one of the few methods-of-choice when performing metabolomic studies [Kopka et al., 2005]. Mass spectral databases were initially constituted with data acquired on GC-MS systems due to the high reproducibility of the spectra obtained from different instruments, running in different labs, used by different operators at different moments in time. Although MS spectra acquired using MS/MS fragmentation after ESI or APCI ionization are not as reproducible as GC-MS spectra acquired using electron impact ionization, also new databases are emerging dedicated to MS and MS/MS spectra acquired with LC-MS/MS. Liquid chromatography coupled to a Triple quadrupole tandem mass spectrometer, Quadrupole time-of-flight mass spectrometer (LC-MS/MS), Ion trap, and Orbitrap of Fourier transform mass spectrometry (FTMS) (LC-MS\textsuperscript{n}) are the established analytical techniques for profiling, quantifying, and identifying the metabolome due to their high selectivity and sensitivity, minimal needs for sample preparation, capacity to separate complex mixtures, and their capability to characterize full scan MS and product ion scans (MS\textsuperscript{n}). However, it is shown that fragmentation spectra vary a lot between the different types of mass analyzers available and even for instruments of the same mass analyzer type but originating from different vendors the fragmentation data is often not reproducible [Bristow et al., 2004]. Furthermore, on the chromatography side, the use of retention parameters in LC is a challenging task by the variety of stationary phases of columns available and used and the different eluent gradients which can be used to provide a good separation of the analytes of interest. It should be mentioned that for GC analysis the type of column and temperature gradient used is much better standardized. Hence the collection of a universal library for LC-MS has been so far limited and there is an urgent need to create spectral libraries acquired with atmospheric ionization as used in LC-MS taking into account the acquisition information provided by different kinds of instruments. The first step towards the production of a universal spectral library is the development of a standard method capable of obtaining reproducibility product ion spectra. Several steps can be followed to minimize instrument-dependent variability, such as summation of several
spectra acquired at different collision energies [Josephs & Sanders, 2004]. An alternative is to establishing a calibration point [Lemire & Busch, 1996] where instrumental conditions are monitored until the relative abundances of two standard product ions spectra are equal.

1.5. Metabolomics Identification

One of the main bottlenecks in metabolomics is the identification of metabolites. Increasing interest in the profiling and identification of the complete metabolome has led to a lot of improvement in the development and robustness of the analytical techniques (Figure 1.3).

![Figure 1.3: Number of publications listed by the PubMed for searches made on 'topic' for: metabolomics (search performed on 23th October 2013).](image)

Identification of metabolites is however much more challenging than identification of peptides and proteins. Proteins consist of a linear sequence of a limited set of usually 20 different amino acids. Usage of MS/MS spectra allows the identification of these individual amino acids together with information how they are connected. When compared with databases containing amino acids sequences the peptide/protein can be identified. It should be mentioned that proteins can be modified, but this is not within the scope of this introduction. Contrary to proteins, a metabolite is only characterized by containing a certain combination of chemical elements (e.g. C, H, O, S, N, and P) and only in particular cases you can observe repeated patterns of functional groups. Because metabolites chemically and physically differ so much it is very difficult to predict their fragmentation patterns by ap-
plying all sort of general fragmentation rules. In combination with reference fragmentation databases, for example MassFrontier software (HighChem Ltd.), the recently announced mzcloud database, or Mass Bank, fragmentation prediction sometimes leads to assigning a putative identify to unknown metabolites.

Verification of a putative identity is both a cost and time expensive job, because verification of a putative identity is still mainly done manually. The identification requires at least two independent and orthogonal type of characteristics (i.e. retention time, accurate mass, tandem mass spectrum, etc.) of the unknown compound to be compared to known compounds obtained under identical experimental conditions. The Metabolomics Standards Initiative has established four different levels to accept the identification or validation of a metabolite [Sumner et al., 2007]. They range from level 1 identification, for a rigorous identification, to unidentified signals at level 4. They have been summarised as:

1. Level: identified compounds that the proposed compound has at least two identical, independent and orthogonal, experimental characteristics as a reference compound.

2. Level: putatively annotated compounds is obtained by comparing it with physiochemical properties and/or spectral libraries of reference standards.

3. Level: putatively characterized compound when the similarity is based upon an analogous chemical class of compounds.

4. Level: unknown compound is a metabolites that can only be differentiated and quantified based upon spectral data.

Metabolomics uses either targeted or non-targeted analytical methods to study the whole or specific parts of the metabolome. Targeted metabolomics aims to measure and profile a limited preselected set of compounds [Dudley et al., 2010]. Its limitation is that they require commercial availability of reference compounds [Last et al., 2007], and many metabolites of interest might not be accessible. However, when there is no a priori knowledge about which metabolites are the most relevant/significant for a certain study addressing a certain biological study, often a non-targeted strategy is followed [De Vos et al., 2007]. The generated data contain an extended number of features (signals of unknown identity) which afterwards can be analysed by statistical methods. The preliminary goal of the non-targeted strategy is to provide novel information (not being constrained by only considering the known, identified metabolites) on which metabolite features express significant difference between the samples. However, ultimately also those discriminative or significant compounds need to be identified.

Overall, identification in LC-MS is the process where a m/z signal is assigned with a metabolite (or analyte) identity. The ability to assign metabolite identifities depends on the
ability to combine different experimental parameters of LC-MS analysis such as retention
time, accurate mass, isotopic pattern, fragmentation pattern, etc.

The identification process starts with a peak formed either from a molecular ion, a de-
protonated ion, an adduct, a naturally occurring isotope molecular ion or a fragment of
a metabolite. The first step is the assignment of a single, correct elemental composition
to each m/z peak in a spectrum. It is a first step because it provides a simple, efficient
and automatable way to search chemical and metabolite databases. Although, there has
been considerable improvement in the analytical techniques to measure accurate mass
spectra, it has been shown that even with an accuracy of less than 1 ppm, the resolution
and accuracy is in many cases not sufficient for unambiguous assignment of a unique
elemental composition. And there are demonstrations of the relevant influence of spectral
accuracy of molecular ions on elemental compositions calculations [Erve et al., 2009].

As a consequence restrictive criteria are required to remove the number of false posi-
tive elemental composition proposals. For instance, the pre-selection of expected chemical
elements and the maximum number of atoms are required. Usually metabolites consist of
carbon (\(^{12}\)C), hydrogen (\(^{1}\)H), nitrogen (\(^{14}\)N), oxygen (\(^{16}\)O) and to lower degree phosporus
\((^{31}\)P) and sulphur (\(^{32}\)S) atoms. Also Na and K adducts must be taken into account. Another
criterion to limit the number of false positives is the application of different heuristic and
chemical rules such as the ‘Golden Rules’ defined by Kind and Fiehn [Kind & Fiehn, 2007].
Some examples of these rules are the nitrogen rule, the octet rule, and the rings-plus-
double-bonds equivalent (RDBE), the LEWIS and SENIOR rule, expected ratios between
elements (H/C, (NOPS)/S) and chemical element probabilities. The nitrogen rule states
that an odd nominal molecular mass of a compound contains an odd number of nitrogen
atoms [de Hoffmann & Stroobant, 2007]. This rule becomes unreliable for masses above
500u [Werner et al., 2008]. The octet rule, formulated nearly one hundred years ago by
LEWIS [Lewis, 1916], defines the number of possible chemical bonds per atom type based
on the electronic distribution of the atoms involved. The double-bond rule specifies the
maximum number of rings and double bonds in the structure given an elemental com-
position [Dayringer & McLafferty, 1977]. The LEWIS and SENIOR rule [Senior, 1951] fil-
ters elemental compositions on the basis of atom valence considerations. Relative isotopic
abundance (RIA) measurements are being used in mass spectrometric measurements for
age determination, forensics, and food authenticity monitoring [Tuniz et al., 2004]. How-
ever, it is also a tool with which either the experimental isotopic abundances can be fit-
ted to the theoretical isotopic pattern of a candidate elemental composition, or the num-
ber of certain atoms in the elemental composition can be calculated. Several studies have
been shown that isotope patterns are relevant to increase confidence in metabolite iden-
tification [Giavalisco et al., 2008]. Another approach which can be used to filter elemental
compositions is the analysis of the fragments generated using the multistage mass spectrometry (MS\textsuperscript{n}) technique. Elemental composition can be excluded from the analysis of elemental compositions of lower mass fragments [Alon & Amirav, 2009]. Another alternative is to use predefined biochemical reactions or transformations together with a probabilistic statistic model to produce a list of possible elemental composition candidates given [Rogers \textit{et al.}, 2009]. The usage of isotope labelled (e.g. \textsuperscript{13}C) material as internal standard is also an efficient method of obtaining information about the identity of certain compounds. The comparison of the monoisotopic masses from unlabelled and labelled compounds gives access to the number of both C and N atoms, limiting the number of possible elemental compositions [Giavalisco \textit{et al.}, 2009].

The next step after assigning the elemental composition is the determination of the structure of the metabolite on the basis of its MS spectrum (especially electron ionization spectra) or fragmentation spectrum (MS/MS or MS\textsuperscript{n}). That is often achieved via search against a compound database or mass spectral library. The aim of a library search is either to obtain the correct structure present in the database as one unique hit or to retrieve partial structural fragments of the unknown metabolite which may allow to gain some information about the class or maybe some part of the structure of a molecule. The total number of compounds or reference spectra entries is an extremely important characteristic of a database. This database density will reduce the likelihood to generate false positives identities [Matsuda \textit{et al.}, 2009]. Compound databases are used to match the calculated/observed elemental composition against elemental composition of metabolites stored in the databases. In this way a putative identity is generated for the observed elemental composition. However this approach cannot provide highly confident identification results because a single elemental composition can still result in many different chemical structures, each representing a different metabolite, or a chemical compound in general. An additional comparison between retention indices stored in the database and the observed retention index can lead to the distinction of compounds having similar mass spectra. On the other hand mass spectral libraries can be used where an experimental mass spectrum is compared against a collection of recorded mass spectra that are stored [Halket \textit{et al.}, 2005]. The number of available MS/MS libraries obtained with atmospheric pressure ionization is small compared to the number of available electron ionization libraries, which is mainly due to the fact that MS/MS spectra are not as reproducible as electron ionization spectra. This limits the building of robust MS/MS spectral databases.

Search algorithms for electron ionization spectra were developed in the ninetieths [Sparkman, 1996], and these include the INCOS algorithm, probability-based matching (PBM) [McLafferty \textit{et al.}, 1974], and the dot-product [Stein & Scott, 1994] spectral similarity. Similar approaches are used for matching MS/MS spectra of small molecules.
[Halket et al., 2005]. They have in common that they all measure the correlation between a query spectrum and a spectrum in the mass spectral library. The library spectrum with the highest correlation is considered to give the most probable identification. Ion traps can be used to generate multi-stage mass spectra ($MS^n$) by consequently fragmenting precursor and all its product ions. It is shown that similar fragmentation patterns can be linked to similar substructures [Sheldon et al., 2009]. This can aid to elucidate pieces of the molecule structure although you would not find a complete match in the searched database. Determination of the complete stereochemical configuration is usually not obtained from analysing the MS information only but a separation technique using a chiral column is needed. It is possible to determine the chirality of molecules when ESI-MS/MS is combined with chiral selector agents [Yao et al., 2000].

There exists still a big gap between the chemical compounds currently covered in metabolomics and their respective measured mass spectra. This space could be filled with mass spectra generated by computers. The big challenge is that this in-silico algorithm should predict accurately its mass fragments and their abundances. Some successes are obtained for molecules with certain structural scaffolds showing consistent fragmentation patterns. These include lipids, oligosaccharides [Zhang et al., 2005], glycans [Kameyama et al., 2006], and peptides [Chen et al., 2001]. Nevertheless, an increase is seen of the algorithms generating in-silico fragmentation spectra for general metabolites [Wolf et al., 2010]. The most straightforward approach and final conclusion to obtaining confirmation of the identity of metabolites in a biological sample is to test commercially available standard compounds on the same analytical experiment using MS/MS spectra and retention time as is suggested by the Metabolomics Standards Initiative.

### 1.6. Biological interpretation

![Diagram of biological interpretation](Image)

**Figure 1.4: Biological interpretation**
Once the relevant metabolite identities are assigned and metabolites are quantified, biological conclusions need to be drawn. Actually, the metabolomics identification is a prerequisite for the biological interpretation of the data and/or data analysis. Metabolite identification is therefore a critical step in the metabolomics pipeline (Figure 1.4 shows the general overview). The pipeline starts with a biological question engaged in a biochemical context and it ends with the biological interpretation. However, the end of the pipeline is connected to the biological question at the start of the pipeline, since the biological interpretation of the identified metabolites generates new knowledge which leads to new questions. Dedicated tools for the biological interpretation of metabolomics data are limited. Some of them are free available like the KEGG pathway database (http://www.genome.jp/kegg/pathway) or the Nutritional Metabolomics Database (http://www.nugowiki.org). Additionally, others like in the HMDB database metabolites are described briefly in a ‘MetaboCard’ designed to contain chemical, biochemistry, and clinical data. Besides these digital approaches existing knowledge in the literature, or by an expert, is still maybe the most efficient tool used to put the identified metabolite into an appropriate biological context.
1.7. Scope

In this postgenomic era there is a specific need to have a better understanding of the human biochemistry and physiology, where the metabolism plays a central role (Metabolomics). Mass spectrometry is often used for profiling these metabolites, which however are often challenging to identify. Multi-stage mass spectrometry (MS^n) is a promising approach in the annotation and structural elucidation of these metabolites. However, following a manual approach will be too time consuming to assign or elucidate the identity of many metabolites. Therefore there is a urgent need for computational tools specifically designed for the processing and interpretation of high mass resolution MS^n data in a fast and efficient way.

The goal of this thesis is to develop a novel semi-automatic approach for the identification of relevant human metabolites in body fluids and tissues using MS^n data. The tools are to be integrated into a workflow and validated for assigning identities to unknown metabolites present in databases but also to unknown metabolites not presented in a library. The pipeline should be available to the metabolomics community. The research of this thesis focusses on the identification of human and plant metabolites, but are in principally applicable also for other scientific fields.

In Chapter 2 a new multi-stage mass spectrometry (ESI-MS^n) method is developed. The MS^n data acquisition protocol was developed to obtain reproducible and robust MS^n data. Furthermore, the influence of the acquisition parameters on the resulting data was studied to verify the robustness of the method. It was investigated whether the MS^n method can be a powerful tool to discern metabolites with similar elemental formula. Multi-stage MS data of a pair of isomeric prostaglandins were acquired and analysed to demonstrate the specificity of fragmentation trees in distinguishing structural isomers. The focus of the next chapter was the optimization of the assignment of the elemental composition to an unknown metabolite, the first step when analysing MS^n data to identify metabolites.

A tool was developed that enables the correct assignment of the elemental composition to molecular ions, their fragment ions, and neutral losses of MS^n data Chapter 3. The final goal was not only the assignment of elemental compositions but also the detection and elimination of artefacts, which were observed to be sometimes rather dominantly present when acquiring MS^n data with an LC-ion trap-Orbitrap MS system. The developed tool reduces efficiently the list of possible elemental composition candidates for each ion by analysing the elemental composition of its parent (precursor ion) and descendants (fragments). Furthermore, the correlation between mass accuracy and the topology of the fragmentation tree was analysed. After processing MS^n data the resulting data needed to be analysed.
A novel approach for this is described in Chapter 4. First a search algorithm was developed to compare experimental MS\textsuperscript{z} data against a given mass spectral library and assigns which trees are most similar to the experimental spectral tree. During the process of identification it is relevant to determine whether the MS\textsuperscript{z} data of the unknown compound is already present in a mass spectral library. This can be achieved by matching the unknown fragmentation tree against those stored in the library(ies). A new method to compare MS\textsuperscript{z} data is developed and described in Chapter 4. If no entry in the database with 'identifical' MS\textsuperscript{z} are present in the libraries, one would like to identify molecules present in databases with structures similar to the unknown compound on the basis of similarity of the MS\textsuperscript{z} data. A novel method was developed to compare MS\textsuperscript{z} data based on detecting the presence of certain combinations of fragments and neutral losses in the fragmentation tree (fingerprints). Two different libraries (plant and human metabolites) containing 867 reference MS\textsuperscript{z} fragmentation trees were used to demonstrate the performance of the tool comparing MS\textsuperscript{z} data.

To make all the in-house developed tools (Chapter 3 and Chapter 4) freely available and easily accessible from any computer, the web-application MetiT ree (http://MetiT ree.nl) was built (Chapter 5). MetiT ree offers the functionalities to organize, process, share, visualize, and compare MS\textsuperscript{z} data through a web browser.

In Chapter 6, a summary of the research conducted of this thesis is given, conclusions drawn and future perspective of metabolite identification discussed.


identification of specific compounds in mixtures. Organic Mass Spectrometry, **9** (7), 690–702.


Bibliography


