



Universiteit
Leiden
The Netherlands

Towards automated identification of metabolites using mass spectral trees

Rojas-Chertó, M.

Citation

Rojas-Chertó, M. (2014, June 19). *Towards automated identification of metabolites using mass spectral trees*. Retrieved from <https://hdl.handle.net/1887/26892>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/26892>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



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

Towards automated identification of
metabolites using mass spectral trees

Miquel Rojas-Chertó

2014

Towards automated identification of metabolites using mass spectral trees

Miquel Rojas-Chertó

Towards automated identification of metabolites using mass spectral trees

Miquel Rojas-Chertó

June 2014

PhD Thesis with summary in Dutch

ISBN: 978-90-74538-84-8

Chapter 1 and 6 ©Miquel Rojas-Chertó, 2011.

Chapter 2 © John Wiley & Sons, Ltd., 2011.

Chapter 3 and 5 © Oxford University Press, 2011-2012.

Chapter 4 © American Chemistry Society, 2012.

Typeset by $\text{\LaTeX}$

Printed by Off Page, Amsterdam, The Netherlands

Towards automated identification of metabolites using mass spectral trees

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 19 juni 2014
klokke 13:45 uur

door

Miquel Rojas-Chertó

Geboren te Tortosa, Spain
in 1978

Promotiecommissie

Promotor:

Prof. dr. Thomas Hankemeier

Co-promotor:

Dr. Theo Reijmers

Overige leden:

Prof. dr. Meindert Danhof (University Leiden)

Prof. dr. Ad Ijzerman (University Leiden)

Prof. dr. Jan van der Greef (University Leiden)

Prof. dr. Rainer Bischoff (University of Groningen)

Dr. Christoph Steinbeck (EMBL-EBI,UK)

The research described in this thesis was performed at the Division of Analytical Biosciences of the Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands.

This research has been financially supported by the Netherlands Metabolomics Centre.

A la meva família

Table of Contents

	Page
1. Introduction	9
2. Fragmentation trees for the structural characterisation of metabolites.....	41
3. Elemental Composition determination based on MS^n	69
4. Metabolomics identification using MS^n data: a new similarity approach for comparing mass spectral trees	99
5. Metitree: A web-application to organize and process multi-stage mass spec- trometry data	139
6. Summary.....	147
Samenvatting	153
Curriculum vitae	159
List of publications.....	163
Acknowledgments	167

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-throughput '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].

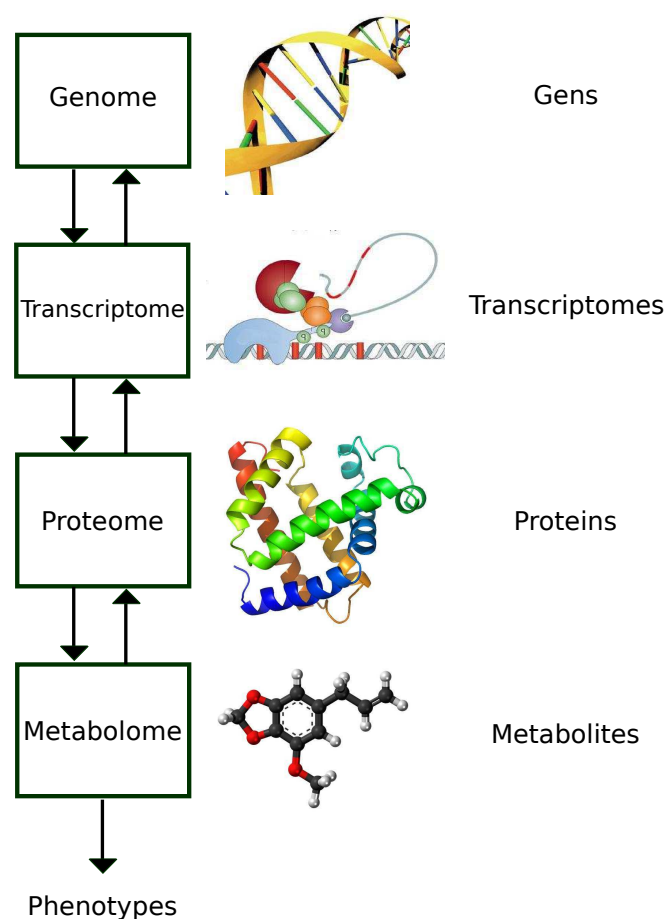


Figure 1.1: The role of metabolomics in the 'omics cascade'

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 & Chrysai-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 ^1H 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

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

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-

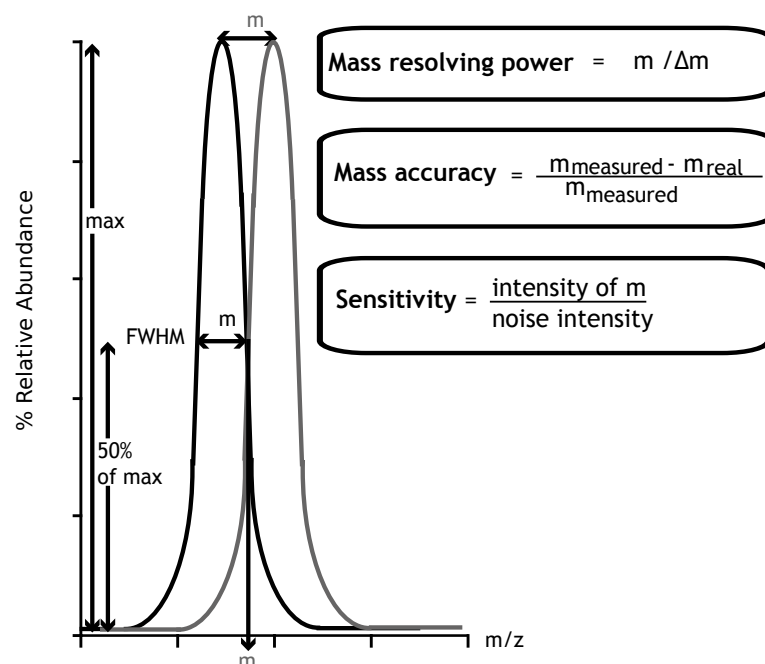


Figure 1.2: Parameters used to describe the performance of mass spectrometers

plete and 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, MSⁿ. 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 MSⁿ 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-MSⁿ 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 *et al.*, 2008], searching in MSⁿ spectral libraries [Oberacher *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.

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. It 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 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](http://www.hmdb.ca)) [Wishart *et al.*, 2008], and ChempSpider [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](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) [Spasić *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 properly 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-MSⁿ) 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 (MSⁿ). 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).

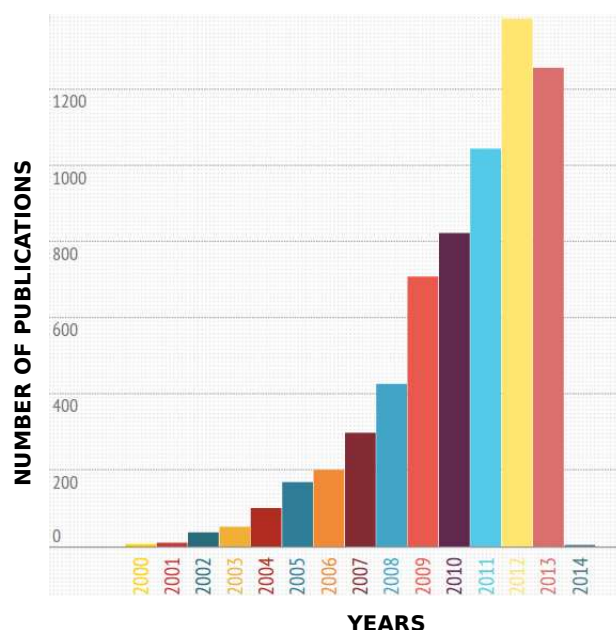


Figure 1.3: Number of publications listed by the PubMed for searches made on 'topic' for: metabolomics (search performed on 23th October 2013).

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 mzcoud database, or Mass Bank, fragmentation prediction sometimes leads to assigning a putative identity 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 identities 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 deprotonated 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 positive 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 (^1H), nitrogen (^{14}N), oxygen (^{16}O) and to lower degree phosphorus (^{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 , $(\text{NOPS})/\text{S}$) 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 composition [Dayringer & McLafferty, 1977]. The LEWIS and SENIOR rule [Senior, 1951] filters 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]. However, it is also a tool with which either the experimental isotopic abundances can be fitted to the theoretical isotopic pattern of a candidate elemental composition, or the number 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 identification [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^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 *et al.*, 2009]. The usage of isotope labelled (e.g. ^{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 *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^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 *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 *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 *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

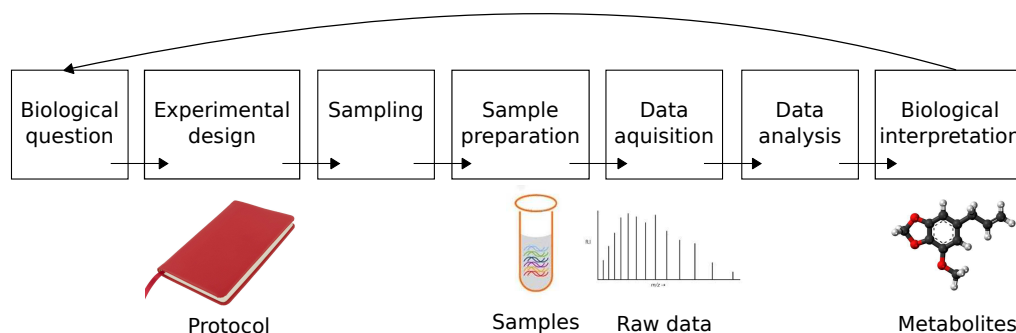


Figure 1.4: Biological interpretation

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 principle 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^n 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^n 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^n data is developed and described in **Chapter 4**. If no entry in the database with 'identical' MS^n 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^n data. A novel method was developed to compare MS^n 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^n fragmentation trees were used to demonstrate the performance of the tool comparing MS^n 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 MetiTree (<http://MetiTree.nl>) was built (**Chapter 5**). MetiTree offers the functionalities to organize, process, share, visualize, and compare MS^n 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.

Bibliography

- [Aebersold & Mann, 2003] Aebersold, R. & Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422** (6928), 198–207.
- [Afendi *et al.*, 2012] Afendi, F. M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L. K., Saito, K. & Kanaya, S. (2012) KNApSACk family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant & cell physiology*, **53** (2), e1.
- [Aliferis & Chrysai-Tokousbalides, 2010] Aliferis, K. A. & Chrysai-Tokousbalides, M. (2010) Metabolomics in pesticide research and development: review and future perspectives. *Metabolomics*, **7** (1), 35–53.
- [Alon & Amirav, 2009] Alon, T. & Amirav, A. (2009) Isotope abundance analysis for improved sample identification with tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM*, **23** (23), 3668–72.
- [Alpert, 2011] Alpert, A. J. (2011) HILIC at 21: Reflections and perspective. Forward. *Journal of chromatography. A*, **1218** (35), 5879.
- [Annesley, 2003] Annesley, T. M. (2003) Ion suppression in mass spectrometry. *Clinical chemistry*, **49** (7), 1041–4.
- [Ausloos *et al.*, 1999] Ausloos, P., Clifton, C. L., Lias, S. G., Mikaya, A. I., Stein, S. E., Tchekhovskoi, D. V., Sparkman, O. D., Zaikin, V. & Zhu, D. (1999) The critical evaluation of a comprehensive mass spectral library. *Journal of the American Society for Mass Spectrometry*, **10** (4), 287–99.

- [Boccard *et al.*, 2010] Boccard, J., Veuthey, J.-L. & Rudaz, S. (2010) Knowledge discovery in metabolomics: an overview of MS data handling. *Journal of separation science*, **33** (3), 290–304.
- [Bristow *et al.*, 2004] Bristow, A. W. T., Webb, K. S., Lubben, A. T. & Halket, J. (2004) Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries. *Rapid communications in mass spectrometry : RCM*, **18** (13), 1447–54.
- [Brown *et al.*, 2005] Brown, S. C., Kruppa, G. & Dasseux, J.-L. (2005) Metabolomics applications of FT-ICR mass spectrometry. *Mass spectrometry reviews*, **24** (2), 223–31.
- [Chen *et al.*, 2001] Chen, T., Kao, M.-Y., Tepel, M., Rush, J. & Church, G. M. (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology a journal of computational molecular cell biology*, **8** (3), 325–337.
- [Cui *et al.*, 2008] Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., Westler, W. M., Eghbalian, H. R., Sussman, M. R. & Markley, J. L. (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nature biotechnology*, **26** (2), 162–4.
- [Dayringer & McLafferty, 1977] Dayringer, H. E. & McLafferty, F. W. (1977) Computer-aided interpretation of mass spectra. STIRS prediction of rings-plus-double-bonds values. *Organic Mass Spectrometry*, **12** (1), 53–54.
- [de Hoffmann & Stroobant, 2007] de Hoffmann, E. & Stroobant, V. (2007) *Mass Spectrometry: Principles and Applications*. Wiley-Interscience.
- [De Vos *et al.*, 2007] De Vos, R. C. H., Moco, S., Lommen, A., Keurentjes, J. J. B., Bino, R. J. & Hall, R. D. (2007) Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, **2** (4), 778–791.
- [Dettmer *et al.*, 2007] Dettmer, K., Aronov, P. A. & Hammock, B. D. (2007) Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, **26** (1), 51–78.
- [Deutsch, 2008] Deutsch, E. (2008) mzML: A single, unifying data format for mass spectrometer output. *Proteomics*, **8** (14), 2776–2777.
- [Dudley *et al.*, 2010] Dudley, E., Yousef, M., Wang, Y. & Griffiths, W. J. (2010) Targeted metabolomics and mass spectrometry. *Advances in protein chemistry and structural biology*, **80**, 45–83.
-

- [Dunn, 2008] Dunn, W. B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical biology*, **5** (1), 011001.
- [Dunn *et al.*, 2005] Dunn, W. B., Bailey, N. J. C. & Johnson, H. E. (2005) Measuring the metabolome: current analytical technologies. *The Analyst*, **130** (5), 606–25.
- [Erve *et al.*, 2009] Erve, J. C. L., Gu, M., Wang, Y., DeMaio, W. & Talaat, R. E. (2009) Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination. *Journal of the American Society for Mass Spectrometry*, **20** (11), 2058–69.
- [Fancy & Rumpel, 2008] Fancy, S.-a. & Rumpel, K. (2008) GC-MS-Based Metabolomics. In *Biomarker Methods in Drug Discovery and Development Methods in Pharmacology and Toxicology* number 1. pp. 317–340.
- [Fiehn, 2001] Fiehn, O. (2001) Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks. *Comparative and Functional Genomics*, **2** (3), 155–168.
- [Fiehn, 2002] Fiehn, O. (2002) Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology*, **48**, 155–171.
- [Gates & Sweeley, 1978] Gates, S. C. & Sweeley, C. C. (1978) Quantitative metabolic profiling based on gas chromatography. *Clinical chemistry*, **24** (10), 1663–73.
- [Gay *et al.*, 2002] Gay, S., Binz, P.-A., Hochstrasser, D. F. & Appel, R. D. (2002) Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics*, **2** (10), 1374–91.
- [Giavalisco *et al.*, 2008] Giavalisco, P., Hummel, J., Lisec, J., Inostroza, A. C., Catchpole, G. & Willmitzer, L. (2008) High-resolution direct infusion-based mass spectrometry in combination with whole ¹³C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Analytical chemistry*, **80** (24), 9417–25.
- [Giavalisco *et al.*, 2009] Giavalisco, P., Köhl, K., Hummel, J., Seiwert, B. & Willmitzer, L. (2009) ¹³C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Analytical chemistry*, **81** (15), 6546–51.
- [Gibney *et al.*, 2005] Gibney, M. J., Walsh, M., Brennan, L., Roche, H. M., German, B. & Van Ommen, B. (2005) Metabolomics in human nutrition: opportunities and challenges. *The American Journal of Clinical Nutrition*, **82** (3), 497–503.
-

- [Goodacre *et al.*, 2004] Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G. & Kell, D. B. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in biotechnology*, **22** (5), 245–52.
- [Halket *et al.*, 2005] Halket, J. M., Waterman, D., Przyborowska, A. M., Patel, R. K. P., Fraser, P. D. & Bramley, P. M. (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *Journal of experimental botany*, **56** (410), 219–43.
- [Hamilton, 1963] Hamilton, P. B. (1963) Ion Exchange Chromatography of Amino Acids. A Single Column, High Resolving, Fully Automatic Procedure. *Analytical Chemistry*, **35** (13), 2055–2064.
- [He *et al.*, 2007] He, T., Quinn, D., Fu, E. & Wang, Y. K. (2007) Metabolome analysis by capillary electrophoresis-mass spectrometry. *Journal of chromatography B Biomedical sciences and applications*, **1168** (1-2), 43–52.
- [Horai *et al.*, 2010] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K. & Nishioka, T. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS*, **45** (7), 703–14.
- [Horning & Horning, 1971] Horning, E. C. & Horning, M. G. (1971) Metabolic profiles: gas-phase methods for analysis of metabolites. *Clinical chemistry*, **17** (8), 802–9.
- [Hu *et al.*, 2005] Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. & Graham Cooks, R. (2005) The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry JMS*, **40** (4), 430–443.
- [Jewison *et al.*, 2012] Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A. C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelnikov, I., Wilson, M. & Wishart, D. S. (2012) YMDB: the Yeast Metabolome Database. *Nucleic acids research*, **40** (Database issue), D815–20.
- [Josephs & Sanders, 2004] Josephs, J. L. & Sanders, M. (2004) Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers. *Rapid communications in mass spectrometry RCM*, **18** (7), 743–759.
-

- [Joshi-Tope *et al.*, 2005] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E. & Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, **33** (Database issue), D428–32.
- [Kaddurah-Daouk *et al.*, 2008] Kaddurah-Daouk, R., Kristal, B. S. & Weinshilboum, R. M. (2008) Metabolomics: a global biochemical approach to drug response and disease. *Annual review of pharmacology and toxicology*, **48**, 653–83.
- [Kameyama *et al.*, 2006] Kameyama, A., Nakaya, S., Ito, H., Kikuchi, N., Angata, T., Nakamura, M., Ishida, H.-K. & Narimatsu, H. (2006) Strategy for simulation of CID spectra of N-linked oligosaccharides toward glycomics. *Journal of proteome research*, **5** (4), 808–14.
- [Kanehisa, 2002] Kanehisa, M. (2002) The KEGG database. *Novartis Foundation Symposium*, **247**, 91–101; discussion 101–103, 119–128, 244–252.
- [Karp *et al.*, 2005] Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V. & López-Bigas, N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, **33** (19), 6083–9.
- [Katajamaa, 2007] Katajamaa, M. (2007) Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, **1158** (1-2), 318–328.
- [Keun *et al.*, 2002] Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Schlotterbeck, G., Senn, H., Niederhauser, U., Holmes, E., Lindon, J. C. & Nicholson, J. K. (2002) Analytical reproducibility in (1)H NMR-based metabolomic urinalysis. *Chemical Research in Toxicology*, **15** (11), 1380–1386.
- [Khoo & Al-Rubeai, 2007] Khoo, S. H. G. & Al-Rubeai, M. (2007) Metabolomics as a complementary tool in cell culture. *Biotechnology and applied biochemistry*, **47** (Pt 2), 71–84.
- [Kind & Fiehn, 2007] Kind, T. & Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, **8**, 105.
- [Kind *et al.*, 2009] Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S. & Fiehn, O. (2009) FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry*, **81** (24), 10038–10048.
-

- [Kitano, 2002] Kitano, H. (2002) Systems biology: a brief overview. *Science (New York, N.Y.)*, **295** (5560), 1662–4.
- [Kopka *et al.*, 2005] 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. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics (Oxford, England)*, **21** (8), 1635–8.
- [Koulman *et al.*, 2007] Koulman, A., Tapper, B. A., Fraser, K., Cao, M., Lane, G. A. & Rasmussen, S. (2007) High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. *Rapid communications in mass spectrometry RCM*, **21** (3), 421–428.
- [Krishnan *et al.*, 2005] Krishnan, P., Kruger, N. J. & Ratcliffe, R. G. (2005) Metabolite fingerprinting and profiling in plants using NMR. *Journal of experimental botany*, **56** (410), 255–65.
- [Lampen *et al.*, 1994] Lampen, P., Hillig, H., Davies, A. N. & Linscheid, M. (1994) JCAMP-DX for Mass Spectrometry. *Applied Spectroscopy*, **48** (12), 1545–1552.
- [Last *et al.*, 2007] Last, R. L., Jones, A. D. & Shachar-Hill, Y. (2007) Towards the plant metabolome and beyond. *Nature reviews. Molecular cell biology*, **8** (2), 167–74.
- [Lemire & Busch, 1996] Lemire, S. W. & Busch, K. L. (1996) Calibration Point for Liquid Secondary Ion Mass Spectrometry Tandem Mass Spectra Measured with an EBqQ Hybrid Mass Spectrometer. *Journal of Mass Spectrometry*, **31** (3), 280–288.
- [Lenz & Wilson, 2007] Lenz, E. M. & Wilson, I. D. (2007) Analytical strategies in metabolomics. *Journal of Proteome Research*, **6** (2), 443–458.
- [Lewis, 1916] Lewis, G. N. (1916) The Atom and the Molecule. *J. Am. Chem. Soc.*, **38** (4), 762–785.
- [Marshall *et al.*, 1998] Marshall, A. G., Hendrickson, C. L. & Jackson, G. S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrometry Reviews*, **17** (1), 1–35.
- [Matsuda *et al.*, 2009] Matsuda, F., Yonekura-Sakakibara, K., Niida, R., Kuromori, T., Shinozaki, K. & Saito, K. (2009) MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. *The Plant Journal*, **57** (3), 555–577.
- [McLafferty *et al.*, 1974] McLafferty, F. W., Hertel, R. H. & Villwock, R. D. (1974) Computer identification of mass spectra. VI. Probability based matching of mass spectra. *Rapid*
-

- identification of specific compounds in mixtures. *Organic Mass Spectrometry*, **9** (7), 690–702.
- [McLafferty & Turecek, 1993] McLafferty, F. W. & Turecek (1993) *Interpretation of Mass Spectra*. University Science Books.
- [McLuckey & Wells, 2001] McLuckey, S. A. & Wells, J. M. (2001) Mass analysis at the advent of the 21st century. *Chemical Reviews*, **101** (2), 571–606.
- [Murray-Rust *et al.*, 2001] Murray-Rust, P., Rzepa, H. S. & Wright, M. (2001) Development of chemical markup language (CML) as a system for handling complex chemical content. *New Journal of Chemistry*, **25** (4), 618–634.
- [Navarro *et al.*, 2003] Navarro, J. D., Niranjana, V., Peri, S., Jonnalagadda, C. K. & Pandey, A. (2003) From biological databases to platforms for biomedical discovery. *Trends in biotechnology*, **21** (6), 263–8.
- [Nicholson *et al.*, 1999] Nicholson, J. K., Lindon, J. C. & Holmes, E. (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica; the fate of foreign compounds in biological systems*, **29** (11), 1181–9.
- [Oberacher *et al.*, 2009] Oberacher, H., Pavlic, M., Libiseller, K., Schubert, B., Sulyok, M., Schuhmacher, R., Csaszar, E. & Köfeler, H. C. (2009) On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *Journal of mass spectrometry JMS*, **44** (4), 485–493.
- [Oliver *et al.*, 1998] Oliver, S. G., Winson, M. K., Kell, D. B. & Baganz, F. (1998) Systematic functional analysis of the yeast genome. *Trends in biotechnology*, **16** (9), 373–8.
- [Pedrioli *et al.*, 2004] 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. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, **22** (11), 1459–66.
- [Quist *et al.*, 2011] Quist, J., Janssen, K. G. H., Vulto, P., Hankemeier, T. & Van Der Linden, H. J. (2011) Single-electrolyte isotachopheresis using a nanochannel-induced depletion zone. *Analytical Chemistry*, **83** (20), 7910–5.
-

- [Ramautar *et al.*, 2011] Ramautar, R., Mayboroda, O. A., Somsen, G. W. & De Jong, G. J. (2011) CE-MS for metabolomics: Developments and applications in the period 2008-2010. *Electrophoresis*, **32** (1), 52–65.
- [Rew & Davis, 1990] Rew, R. & Davis, G. (1990). NetCDF: an interface for scientific data access.
- [Rogers *et al.*, 2009] Rogers, S., Scheltema, R. A., Girolami, M. & Breitling, R. (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, **25** (4), 512–8.
- [Romero *et al.*, 2006] Romero, R., Espinoza, J., Gotsch, F., Kusanovic, J. P., Friel, L. A., Erez, O., Mazaki-Tovi, S., Than, N. G., Hassan, S. & Tromp, G. (2006) The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. *BJOG an international journal of obstetrics and gynaecology*, **113 Suppl** (5), 118–135.
- [Sana *et al.*, 2008] Sana, T. R., Waddell, K. & Fischer, S. M. (2008) A sample extraction and chromatographic strategy for increasing LC/MS detection coverage of the erythrocyte metabolome. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, **871** (2), 314–21.
- [Scholz & Fiehn, 2007] Scholz, M. & Fiehn, O. (2007) SetupX—a public study design database for metabolomic projects. In *Pacific Symposium On Biocomputing* number 1793-5091 (Print) LA - eng PT - Journal Article PT - Research Support, N.I.H., Extramural SB - IM pp. 169–180 University of California, Davis. Genome Center 451 E. Health Sci. Drive Davis, California 95616, USA.
- [Senior, 1951] Senior, J. K. (1951) Partitions and Their Representative Graphs. *American Journal of Mathematics*, **73** (3), 663.
- [Sheldon *et al.*, 2009] Sheldon, M. T., Mistrik, R. & Croley, T. R. (2009) Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *Journal of the American Society for Mass Spectrometry*, **20** (3), 370–6.
- [Smith *et al.*, 2005] Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R. & Siuzdak, G. (2005) METLIN: a metabolite mass spectral database. *Therapeutic drug monitoring*, **27** (6), 747–51.
- [Sparkman, 1996] Sparkman, O. D. (1996) Evaluating electron ionization mass spectral library search results. *Journal of the American Society for Mass Spectrometry*, **7** (4), 313–318.
-

- [Spasić *et al.*, 2006] Spasić, I., Dunn, W. B., Velarde, G., Tseng, A., Jenkins, H., Hardy, N., Oliver, S. G. & Kell, D. B. (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics*, **7** (1), 281.
- [Stein & Scott, 1994] Stein, S. E. & Scott, D. R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, **5** (9), 859–866.
- [Sud *et al.*, 2007] Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., Merrill, A. H., Murphy, R. C., Raetz, C. R. H., Russell, D. W. & Subramaniam, S. (2007) LMSD: LIPID MAPS structure database. *Nucleic acids research*, **35** (Database issue), D527–32.
- [Sumner *et al.*, 2007] Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reilly, M. D., Thaden, J. J. & Viant, M. R. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3** (3), 211–221.
- [Tolstikov & Fiehn, 2002] Tolstikov, V. V. & Fiehn, O. (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical biochemistry*, **301** (2), 298–307.
- [Tuniz *et al.*, 2004] Tuniz, C., Zoppi, U. & Hotchkis, M. (2004) Sherlock Holmes counts the atoms. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, **213** (null), 469–475.
- [Van Der Greef & Smilde, 2005] Van Der Greef, J. & Smilde, A. K. (2005) Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics*, **19** (5-7), 376–386.
- [van der Greef *et al.*, 2004] van der Greef, J., Stroobant, P. & van der Heijden, R. (2004) The role of analytical sciences in medical systems biology. *Current opinion in chemical biology*, **8** (5), 559–65.
- [van der Hooft *et al.*, 2011] van der Hooft, J. J. J., Mihaleva, V., de Vos, R. C. H., Bino, R. J. & Vervoort, J. (2011) A strategy for fast structural elucidation of metabolites in small volume plant extracts using automated MS-guided LC-MS-SPE-NMR. *Magnetic resonance in chemistry : MRC*, **49 Suppl 1**, S55–60.
-

- [Wang *et al.*, 2009] Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J. & Bryant, S. H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, **37** (Web Server issue), W623–W633.
- [Watkins & German, 2002] Watkins, S. M. & German, J. B. (2002) Toward the implementation of metabolomic assessments of human health and nutrition. *Current opinion in biotechnology*, **13** (5), 512–6.
- [Werner *et al.*, 2008] Werner, E., Heilier, J.-F., Ducruix, C., Ezan, E., Junot, C. & Tabet, J.-C. (2008) Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, **871** (2), 143–63.
- [Williams & Tkachenko, 2010] Williams, A. J. & Tkachenko, V. (2010). ChemSpider - Building an Online Database of Open Spectra.
- [Williams & Fleming, 2007] Williams, D. & Fleming, I. (2007) *Spectroscopy Methods in Organic Chemistry*. McGraw-Hill Education.
- [Wishart, 2008] Wishart, D. S. (2008) Applications of metabolomics in drug discovery and development. *Drugs in R&D*, **9** (5), 307–22.
- [Wishart *et al.*, 2009] Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazzyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J. & Forsythe, I. (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic acids research*, **37** (Database issue), D603–10.
- [Wishart *et al.*, 2006] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. & Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, **34** (Database issue), D668–72.
- [Wishart *et al.*, 2008] Wishart, D. S., Lewis, M. J., Morrissey, J. A., Flegel, M. D., Jeroncic, K., Xiong, Y., Cheng, D., Eisner, R., Gautam, B., Tzur, D., Sawhney, S., Bamforth, F., Greiner, R. & Li, L. (2008) The human cerebrospinal fluid metabolome. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, **871** (2), 164–73.
-

- [Wolf *et al.*, 2010] Wolf, S., Schmidt, S., Muller-Hannemann, M. & Neumann, S. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, **11** (1), 148.
- [Yao *et al.*, 2000] Yao, Z.-P., Wan, T. S. M., Kwong, K.-P. & Che, C.-T. (2000) Chiral Analysis by Electrospray Ionization Mass Spectrometry/Mass Spectrometry. 2. Determination of Enantiomeric Excess of Amino Acids. *Analytical Chemistry*, **72** (21), 5394–5401.
- [Young *et al.*, 2006] Young, N., Jewell, K., Block, D., Knox, C., Tang, P., Greiner, R. & Wishart, D. S. (2006) MetaboLIMS: A General Laboratory Information Management System for Metabolomics. In *In: Metabolomics Society Meeting*.
- [Zhang *et al.*, 2005] Zhang, H., Singh, S. & Reinhold, V. N. (2005) Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. *Analytical chemistry*, **77** (19), 6263–70.
- [Zolnai *et al.*, 2003] Zolnai, Z., Lee, P. T., Li, J., Chapman, M. R., Newman, C. S., Phillips, G. N., Rayment, I., Ulrich, E. L., Volkman, B. F. & Markley, J. L. (2003) Project management system for structural and functional proteomics: Sesame. *Journal of structural and functional genomics*, **4** (1), 11–23.
-

CHAPTER 2

Fragmentation trees for the structural characterisation of metabolites

Piotr T. Kasper^{1,2,*}, Miguel Rojas-Chertó^{1,2,*}, Theo Reijmers^{1,2}, Thomas Hanke-meier^{1,2}, Rob Vreeken^{1,2}

Rapid Communications in Mass Spectrometry 2011:26(19):2275-86

¹Netherlands Metabolomics Centre, Leiden, The Netherlands

²Division of Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands

*Contributed equally to this work.

2.1. Abstract

Metabolite identification plays a crucial role in the interpretation of metabolomics research results. Due to its sensitivity and widespread implementation, a favourite analytical method used in metabolomics is electrospray mass spectrometry. In this paper, we demonstrate our results in attempting to incorporate the potentials of multistage mass spectrometry into the metabolite identification routine. New software tools were developed and implemented which facilitate the analysis of multistage mass spectra and allow for efficient removal of spectral artefacts. The pre-processed fragmentation patterns are saved as fragmentation trees. Fragmentation trees are characteristic of molecular structure. We demonstrate the reproducibility and robustness of the acquisition of such trees on a model compound. The specificity of fragmentation trees allows for distinguishing structural isomers, as shown on the pair of isomeric prostaglandins. This approach to the analysis of the multistage mass spectral characterisation of compounds is an important step towards formulating a generic metabolite identification method.

2.2. Introduction

One of the central tasks of metabolomics is to identify metabolites in complex biological mixtures and to decode their structure. This is a challenging but essential task, because unless the identity of the studied metabolite is known, its quantitative data cannot be related to its biochemical role. This requires further developing and optimising the available analytical techniques in order to yield a robust metabolite identification platform.

Nuclear magnetic resonance (NMR)[Kwan & Huang, 2008, Robertson, 2005] and mass spectrometry (MS)[Scalbert *et al.*, 2009] are the methods most commonly used for the structural characterisation of chemical compounds. NMR offers a rapid and detailed analysis of the structure of the (un)known compound but the technique is severely limited due to its relatively low sensitivity. MS, on the other hand, offers high sensitivity and specificity [Villas-bo *et al.*, 2005] resulting in elemental formulas [Kind & Fiehn, 2007]. are the methods most commonly used for the structural characterisation of chemical compounds. NMR offers a rapid and detailed analysis of the structure of the (un)known compound but the technique is severely limited due to its relatively low sensitivity. MS, on the other hand, offers high sensitivity and specificity.

Obviously, an elemental formula is not specific enough to identify a metabolite. Its structure can be further characterised by gas-phase fragmentation reactions, e.g. Collision Induced Dissociation (CID). The resulting fragmentation spectrum reflects the structure of the precursor ion: the masses of the obtained product ions and their relative abundances char-

acterise the structure of the precursor ion and the experimental fragmentation conditions. In this way, a fragmentation spectrum offers a fingerprint of the molecular structure of the precursor, and as long as it can be reproducibly acquired it can be used to identify ionized molecules and fragment ions [McLafferty & Turecek, 1993].

The separation of metabolites prior to detection is often achieved using liquid chromatography (LC) or capillary electrophoresis (CE). Ionisation is mostly achieved through soft-ionisation techniques like, e.g. ElectroSpray Ionisation (ESI). The ions generated in the ESI source can be fragmented using CID. Regrettably, although the CID spectra are rich in information, it remains difficult to acquire data in a reproducible manner [Milman, 2005, Oberacher *et al.*, 2009]. This is mainly due to the fact that in beam-type instruments, the precursor ion's internal energy is difficult to control. More reproducible fragmentation spectra can be produced using ion traps [van der Hooft *et al.*, 2011], which require collisional cooling of the precursor ion for efficient trapping and selective (resonance) excitation. Furthermore, by using multistage MS experiments (MS^n), ion trap instruments can provide detailed information on the fragmentation, thereby helping to characterise the metabolites' structure.

Despite the growing popularity of versatile ion trap instruments, in-depth analysis of MS^n spectra remains difficult due to the lack of generic software tools. The challenge stems from the multidimensionality of MS^n data. The majority of the MS analysis software is well-suited for analysing spectra, but not for analysing one of the most important features of MS^n data: the precursor-product relations between the ions observed in separate MS^n spectra. The only software available at the moment which can be used to analyse and/or compare MS^n spectra is Mass Frontier (HighChem, Bratislava, Slovakia) [Sheldon *et al.*, 2009]. This proprietary software package, being not open-source, cannot be easily integrated into our specific workflow because it is designed to work only with the proprietary data format of one vendor. Furthermore, we wanted to remove spectral artefacts using the hierarchy of observed fragments. This would require software tools that can exchange data using common mass spectrometric data exchange formats such as mzXML [Pedrioli *et al.*, 2004] mzData [] or mzML [Martens *et al.*, 2011]. As a result, we decided to develop the necessary software ourselves. Our software package, called Multistage Elemental Formula generator (MEF) [Rojas-Chertó *et al.*, 2011], used the precursor-product ion relations in order to effectively and specifically extract the relevant data from multistage mass spectra.

Approaches to metabolite identification that use multistage MS fragmentation often require manual intervention by mass spectrometry experts [van der Hooft *et al.*, 2011, Konishi *et al.*, 2007, Cui *et al.*, 2000, Montoya *et al.*, 2009]. Recently, more automated approaches are reported that greatly facilitate this tedious analysis [Rojas-Chertó *et al.*, 2011, Jarussophon *et al.*, 2009, Scheubert *et al.*, 2011, Wolf *et al.*, 2010, Heinonen *et al.*, 2008].

Some of the methods focus on predicting fragmentation pattern in silico [Scheubert *et al.*, 2011, Wolf *et al.*, 2010, Heinonen *et al.*, 2008, Rasche *et al.*, 2011]. In contrast to these approaches we do not predict the hierarchy of the fragmentation trees. Similarly to the approach of Mass Frontier [Sheldon *et al.*, 2009] the hierarchy is derived from hierarchy of MS^n spectra, but the nodes of the fragmentation tree are fragment ions and not the fragmentation spectra as in Mass Frontier. In contrast to the MetFrag approach [Rasche *et al.*, 2011] the hierarchy of the ions in our approach is not calculated but observed in hierarchy of MS^n spectra. Scheubert [Scheubert *et al.*, 2011] applied Rasche's [Rasche *et al.*, 2011] method to predict MS^n spectra and demonstrated that the hierarchy of the fragment ions derived from hierarchy of MS^n spectra adds substantially to the model. As more tools become available for MS^n analysis and for fragmentation prediction, it is easier to link the MS fragmentation patterns of metabolites to their molecular structure. This, in turn, will greatly facilitate the generic use of MS^n spectra in the field of metabolite identification. Using our MEF tool, elemental formulas were unambiguously assigned to fragment ions [Rojas-Chertó *et al.*, 2011]. This tool uses hierarchy of the fragment ions in analogous way as previously reported approaches [Scheubert *et al.*, 2011, Böcker & Rasche, 2008, Böcker & Rasche, 2008]. Furthermore, the constraints derived from the ions hierarchy allowed us to discard irrelevant artefacts and to efficiently identify the peaks that were relevant for the precursor ion structure. In this way, we were able to store a hierarchical representation of the elemental composition of fragment ions as observed in MS^n spectra, together with the data characterising their MS signals, all in the form of a fragmentation tree. Our final aim is to develop a database based metabolite identification pipeline which will be reported separately in a later stage. This pipeline is though supported by a database filled with the above described MS^n fragmentation patterns. The use of this MS^n data, organised in fragmentation trees, will enable facile comparison of obtained results. We aim at using this approach in an on-line or at-line fashion. In such a set-up, where a compound with unknown structure elutes from and LC-system, one can either generate a fragmentation tree directly on-line or at-line after fraction collection and subsequent infusion through the nano-ESI interface. This fragmentation tree is subsequently evaluated against a database of fragmentation trees of known compounds/structures. Further bioinformatics based tools, which will be reported separately, will aid in (partial) recognition of fragmentation trees and using an 'in-house' structure generator, subsequent structure postulation for the unknown. This will complement a metabolite identification pipeline. However, before being able to assemble such a set-up, a robust acquisition and evaluation of MS^n data needs to be developed. Here we report on the development of this part of the envisioned pipeline. We studied the parameters involved in acquiring the fragmentation trees in order to evaluate their

robustness and reproducibility. Besides, we evaluate how these factors affect the topology of the resulting fragmentation tree. Furthermore, we assess the possibility of using this approach as well to discern between structurally related isomeric structures. For the latter we studied two isomeric prostaglandins and several eicosanoids.

2.3. Experimental

Materials and samples. Glutathione was purchased from Sigma (Sigma-Aldrich, Steinheim, Germany). Prostaglandin D2 and E2 and all other eicosanoids were obtained from Cayman Chemicals (Cayman Chemicals, Ann Arbor, MI, USA). All the samples were dissolved in 50% methanol 0.1% formic acid prior to acquisition. Samples were spun down (5 min, 15000 g) before being transferred into the 96-well sample plate of the NanoMate to prevent clogging of the nano-spray emitter by small particulate matter. Methanol (absolute, ULC/MS grade), water (ULC/MS grade) and formic acid (99%, LC/MS grade) were obtained from Biosolve BV, (Valkenswaard, the Netherlands).

Mass Spectrometry. MS and tandem MS experiments were performed on an LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Waltham, MA) controlled by Xcalibur software (version 2.0.7). The instrument was equipped with a TriVersa NanoMate (Advion, Ithaca, NY) nano-electrospray ion source. For the positive ionisation mode, the nitrogen pressure was set at 0.45 psi and the ESI voltage was 1.35 kV; for the negative ionisation mode the settings were 0.7 psi and 1.05 kV, respectively. The distance between the ESI chip and the capillary was approximately 5mm. The MS method was programmed in Xcalibur and consisted of 107 scan events: one full scan and 106 data-dependent tandem MS scans up to MS⁵. The method allowed for the fragmentation of the five highest peaks of the MS² and MS³ spectra and the three highest peaks of the MS⁴ (this method, with some minor modifications, was successfully demonstrated and reported on earlier [van der Hooft *et al.*, 2011]). All the spectra acquired were spectra combined from 3 μ scans. Full scan spectra (from 50 m/z units below to 50 m/z units above the molecular weight of the compound in question) were acquired with a resolving power of 30000 (FWHM at m/z 400). In the case of the MS²-5 spectra, this resolving power was reduced to 15000 (FWHM specified at 400 m/z) to speed up acquisition. Automatic Gain Control (AGC) was active at default settings. The fragmentation spectra were acquired for singly charged ions with a pre-cursor intensity threshold of 4500 ion counts. The isolation width for isolating the precursor ion varied from 1 to 3 m/z units and we used a normalised collision energy of 25% to 45%. Each measurement was performed in duplicate. The glutathione reference fragmentation spectrum (Figure 2.1) was acquired using the HCD cell on the LTQ-Orbitrap XL

Data processing. Orbitrap MS^n spectra were converted from Thermo's Xcalibur's own acquisition (*.RAW) files to mzXML format[Pedrioli *et al.*, 2004] using the ReAdW (version 4.3.1) conversion tool provided by the Institute for Systems Biology, Seattle, WA (available at <http://sourceforge.net/projects/sashimi/files/>). mzXML was chosen because it is open and it preserves the precursor mass attributes that point to the hierarchy of fragmentation spectra. Subsequently, mzXML files were analysed using XCMS software[Smith *et al.*, 2006](available at <http://masspec.scripps.edu/xcms/download.php>) in order to select mass peaks without losing the hierarchical relations between them. We used default XCMS settings for peak picking: the $mzGap$ was 0.2 m/z and the signal-to-noise ratio was 10. Each mass peak in the resulting table was assigned a precursor ion. We then used the Multi-stage Elemental Formula (MEF) generator [Rojas-Chertó *et al.*, 2011] (available at <http://abs.lacdr.gorlaeus.net/people/rojas-cherto>) to unambiguously assign elemental formulas to fragment ions and to neutral losses (with 6 ppm mass tolerance), as well as to remove spectral artefacts. The results were stored in a Chemical Markup Language[Murray-Rust & Rzepa, 2001] (CML) format which combines mass spectrometric and chemical information in a single exchange file. Each step of the analysis, as well as a detailed explanation of the MEF algorithm, can be found in [Rojas-Chertó *et al.*, 2011]. In order to distinguish the isomeric prostaglandins, we performed a hierarchical clustering analysis using the R software environment[R Development Core Team, 2008]. The fragmentation trees were represented as vectors of occurrences of elemental formula paths, and Euclidean distance was used as a similarity measure in mean linkage clustering. For clustering, we used the complete-linked Button-up algorithm[Hansen & Delattre, 1978].

The similarity measure of the pairs of isomeric molecules was calculated by applying the Tanimoto coefficient[Fligner *et al.*, 2002] using the CDK 2D-fingerprint library[Steinbeck *et al.*, 2003]. The Tanimoto coefficient was calculated by dividing the number of common EFPs observed for both metabolites by the total number of unique EFPs present for each metabolite minus the number EFPs present in both molecules.

2.4. Results and discussion

2.4.1. Acquisition of the fragmentation trees

The MS^n experiments were performed using an LTQ-Orbitrap mass spectrometer. Since this instrument has a high dynamic range in terms of mass accuracy, it allows assignment

id	name	lipid maps ID
8S-HETE	8S-hydroxy-5Z,9E,11Z,14Z-eicosatetraenoic acid	LMFA03060006
5S-HETE	5S-hydroxy-6E,8Z,11Z,14Z-eicosatetraenoic acid	LMFA03060002
8,9-EET	8,9-epoxy-5Z,11Z,14Z-eicosatrienoic acid	LMFA03080003
9-HETE	9-hydroxy-5Z,7E,11Z,14Z-eicosatetraenoic acid	LMFA03060089
20-HETE	20-hydroxy-5Z,8Z,11Z,14Z-eicosatetraenoic acid	LMFA03060009
15S-HETE	15S-hydroxy-5Z,8Z,11Z,13E-eicosatetraenoic acid	LMFA03060001
5,6-EET	5,6-epoxy-8Z,11Z,14Z-eicosatrienoic acid	LMFA03080002
12-HETE	12-hydroxy-5Z,8Z,10E,14Z-eicosatetraenoic acid	LMFA03060088
14,15-EET	14,15-epoxy-5Z,8Z,11Z-eicosatrienoic acid	LMFA03080005
11R-HETE	11R-hydroxy-5Z,8Z,12E,14Z-eicosatetraenoic acid	LMFA03060028
11,12-EET	11,12-epoxy-5Z,8Z,14Z-eicosatrienoic acid	LMFA03080004

Table 2.1: List of elemental formula paths constituting the fragmentation tree of glutathione.

of elemental formulas both to precursor ions and to their fragment ions. Since fragmentation is performed in a linear ion trap, the high yield for fragment ions (MS/MS efficiency) and the fast duty cycle facilitates extensive MS^n experiments in a relatively short period of time [McLuckey *et al.*, 1994]. Moreover, due to the selective resonance excitation of the precursor ion in the ion trap [March, 1997], the fragment ions obtained do not fragment and can be used as precursor ions for the next stage in the MS^n experiment. In this manner, the hierarchy of the MS^n spectra determines the hierarchy of the fragment ions (as shown in Figure 2.1).

To efficiently extract the hierarchy of these fragment ions, we used our own software, i.e. the Multi-stage Elemental Formula (MEF) generator [Rojas-Chertó *et al.*, 2011]). The precursor-product ion relations between all the mass peaks in the MS^n spectra (Figure 2.2a and 2.2b) were used as constraints in assigning elemental formulas to the individual fragment ions (Figure 2.2c). Specifically, the elemental formula of a fragment ion cannot contain more atoms of a certain element than its precursor ion, and a precursor ion cannot contain fewer atoms of an element than its fragment. Finally, the assigned elemental formula of a neutral loss and the elemental formula of the fragment have to add up to the elemental formula of the precursor [Rojas-Chertó *et al.*, 2011]. In order to unambiguously identify the hierarchical relation between precursor and product ions, we generated an Elemental Formula Path (EFP) for each ion in the fragmentation tree. An EFP is a list of elemental formulas assigned to consecutive precursor and product ions leading to a particular fragment ion (Figure 2.2d). In this way, a fragmentation tree can be represented as a collection of EFPs. Consequently, comparing the MS^n results of various compounds

ID	Elemental Formula Path	MSn	Elemental Formula	mass	Relative intensity	Mass error	SD
1	C ₁₀ H ₁₈ N ₃ O ₆ S ₁	1	C ₁₀ H ₁₈ N ₃ O ₆ S	308.091	100%	-0.5	0
2	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₁₁ N ₂ O ₃ S ₁	2	C ₅ H ₁₁ N ₂ O ₃ S	179.049	100%	1.3	0
3	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₁₁ N ₂ O ₃ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁	3	C ₅ H ₈ N ₁ O ₃ S	162.022	100%	0.2	0
4	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₁₁ N ₂ O ₃ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₄ H ₆ N ₁ O ₁ S ₁	4	C ₄ H ₆ N ₁ O ₁ S	116.016	27%	-0.3	1.4
5	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₁₁ N ₂ O ₃ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₅ H ₆ N ₁ O ₂ S ₁	4	C ₅ H ₆ N ₁ O ₂ S	144.011	100%	0.1	0
6	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₁₁ N ₂ O ₃ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₅ H ₆ N ₁ O ₂ S ₁ ... @C ₄ H ₆ N ₁ O ₁ S ₁	5	C ₄ H ₆ N ₁ O ₁ S	116.016	100%	-0.4	0
7	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁	2	C ₅ H ₈ N ₁ O ₃ S	162.022	33%	1.3	1.8
8	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₄ H ₆ N ₁ O ₁ S ₁	3	C ₄ H ₆ N ₁ O ₁ S	116.016	27%	-0.4	1.4
9	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₅ H ₆ N ₁ O ₂ S ₁	3	C ₅ H ₆ N ₁ O ₂ S	144.011	100%	0.1	0
10	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₅ H ₆ N ₁ O ₂ S ₁ @C ₄ H ₆ N ₁ O ₁ S ₁	4	C ₄ H ₆ N ₁ O ₁ S	116.016	100%	-0.5	0
11	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁	2	C ₈ H ₁₃ N ₂ O ₄ S	233.059	25%	1.7	1.4
12	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₇ H ₁₁ N ₂ O ₂ S ₁	3	C ₇ H ₁₁ N ₂ O ₂ S	187.054	100%	0.8	0
13	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₇ H ₁₁ N ₂ O ₂ S ₁ @C ₇ H ₈ N ₁ O ₂ S ₁	4	C ₇ H ₈ N ₁ O ₂ S	170.027	100%	0.5	0
14	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₇ H ₁₁ N ₂ O ₂ S ₁ @C ₇ H ₈ N ₁ O ₂ S ₁ @C ₆ H ₆ N ₁ S ₁	5	C ₆ H ₆ N ₁ S	124.022	100%	-0.4	0
15	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₈ H ₁₁ N ₂ O ₃ S ₁	3	C ₈ H ₁₁ N ₂ O ₃ S	215.049	32%	0.9	3.5
16	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₈ H ₁₁ N ₂ O ₃ S ₁ @C ₇ H ₁₁ N ₂ O ₂ S ₁	4	C ₇ H ₁₁ N ₂ O ₂ S	187.054	100%	0.8	0
17	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₈ H ₁₁ N ₂ O ₃ S ₁ @C ₇ H ₁₁ N ₂ O ₂ S ₁ @C ₆ H ₉ N ₂ S ₁	5	C ₆ H ₉ N ₂ S	141.048	100%	-0.1	0
18	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₈ H ₁₃ N ₂ O ₄ S ₁ @C ₈ H ₁₁ N ₂ O ₃ S ₁ @C ₇ H ₁₁ N ₂ O ₂ S ₁ @C ₇ H ₈ N ₁ O ₂ S ₁	5	C ₇ H ₈ N ₁ O ₂ S	170.027	82%	0.3	8.7
19	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₁₁ N ₂ O ₃ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₄ H ₆ N ₁ O ₁ S ₁ @C ₃ H ₆ N ₁ S ₁	5	C ₃ H ₆ N ₁ S	88.021	100%	-1.0	0
20	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₄ H ₆ N ₁ O ₁ S ₁ @C ₃ H ₆ N ₁ S ₁	4	C ₃ H ₆ N ₁ S	88.021	100%	-1.0	0
21	C ₁₀ H ₁₈ N ₃ O ₆ S ₁ @C ₅ H ₈ N ₁ O ₃ S ₁ @C ₅ H ₆ N ₁ O ₂ S ₁ @C ₄ H ₆ N ₁ O ₁ S ₁ @C ₃ H ₆ N ₁ S ₁	5	C ₃ H ₆ N ₁ S	88.021	100%	-1.0	0

Table 2.2: List of 11 isomeric eicosanoids.

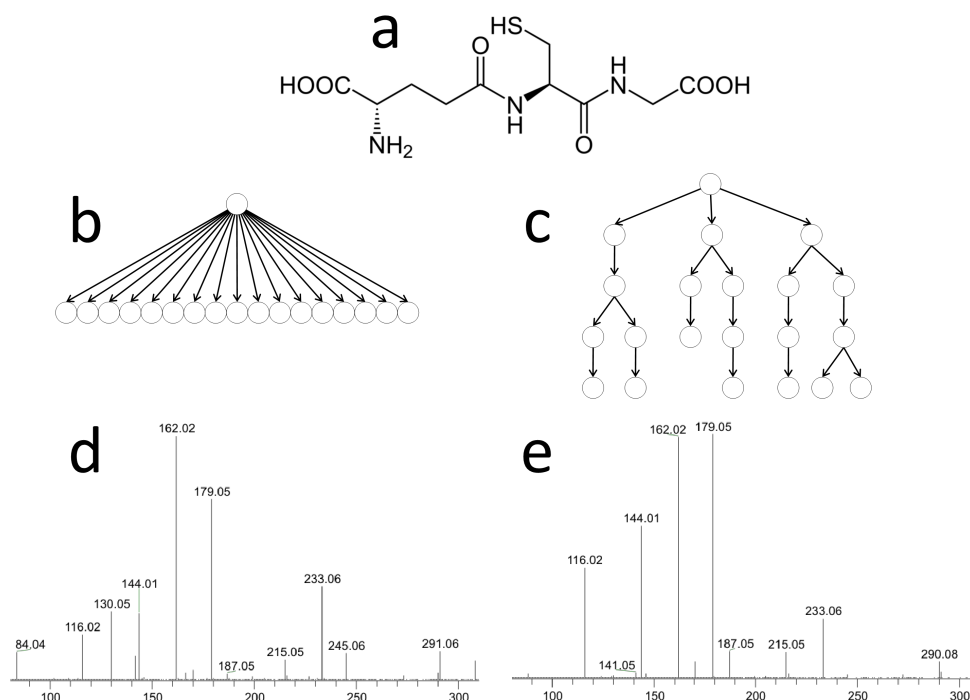


Figure 2.1: Comparison of MS/MS and MSⁿ fragmentation spectra of glutathione. Schematic hierarchical representation of precursor-product relations between ions derived from hierarchy of spectra in (b) MS/MS and (c) MSⁿ of glutathione(CHEBI:16856) (a). The corresponding spectra are (d) MS/MS spectrum and (e) total composite spectrum of MSⁿ experiment. The major differences between the two spectra lie in the different signal intensity ratio; most of the fragments (except 84 m/z) were observed in both spectra.

no longer requires a direct comparison between single fragmentation spectra of specific precursor ions, but one can compare individual fragmentation trees to see whether they include a particular EFP or a set of EFPs (Figure 2.1d).

In the process of assigning these elemental formulas, we discard the mass peaks that fail to satisfy the constraints derived from the hierarchical precursor-product relations. As a result, artefact peaks not satisfying the above-mentioned constraints, and originating from radio-frequency interference, electronic noise, or the side bands often observed in FT-MS systems [Mathur & O'Connor, 2009] are rejected. We optimised the MSⁿ acquisition protocol so as to yield fragmentation trees that consist of as many fragments as possible, and that can (easily) be reproduced. Since fragments convey structurally relevant information, it is better to have exhaustive 'wide and deep' fragmentation trees. In our approach we optimize spraying conditions, e.g. electrospray emitter voltage, solvent (see Experimental), temperature, etc. to obtain predominantly (de-) protonated molecules, depending on the mode of

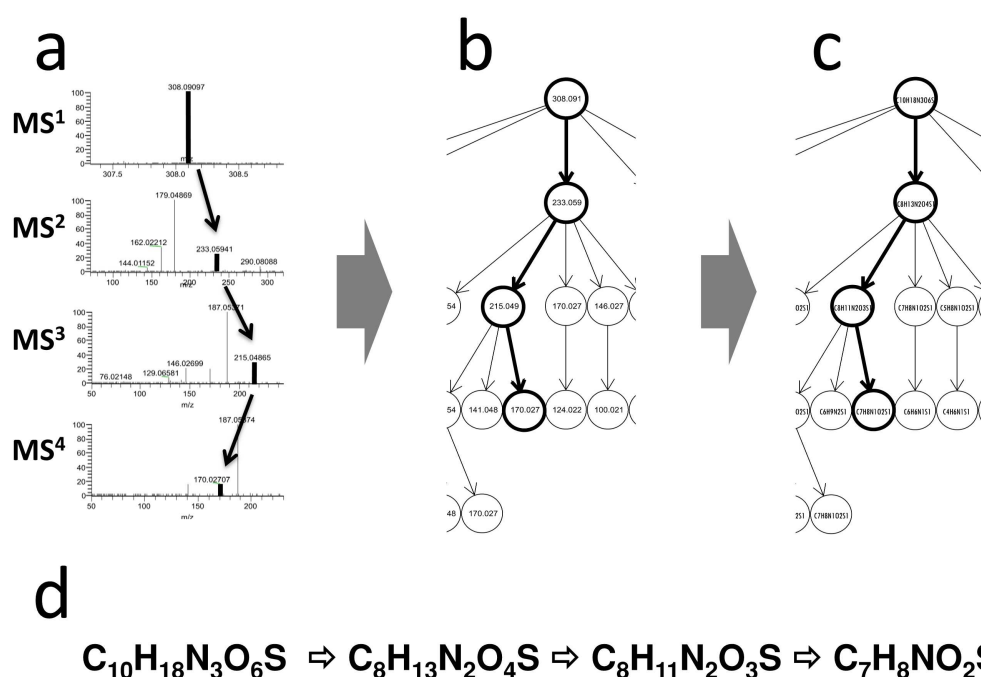


Figure 2.2: Schematic workflow of fragmentation tree creation shown on example of single elemental formula path. Peaks and their precursor-product relations depicted in (b) were extracted from MSⁿ spectra (a). Multistage Elemental Formula generator software was used to create a fragmentation tree (c). At this stage, we discarded peaks that were assigned an elemental formula that did not fit in the fragmentation tree. To facilitate unambiguous identification of the fragment ions within the tree they were described by an/their elemental formula path (EFP). The elemental formula path (EFP) (d) unambiguously identifies the fragment ion within the obtained fragmentation tree. The corresponding peaks, masses and elemental formulas are depicted in bold.

operation (positive ionisation (PI) or negative ionisation (NI)). These ions are subsequently used as precursor ions for fragmentation tree generation. In practice however, one will observe a variety of adducts or source fragments in (nano-)ESI, e.g. [M+Na]⁺, [M+NH₄]⁺, [M+H-H₂O]⁺ in the case of PI-mode or an [M+formate]⁻, [M+acetate]⁻ ion the case of NI mode of operation [Kassler *et al.*, 2011, Cech & Enke, 2002, Schug & McNair, 2003]. In some cases even, only ions are observed in either the PI- or the NI-mode. In these cases, only the obtained tree will be stored. The nature of the ESI-spectrum depends on e.g. the structure of the compound, and its proton affinity or gas-phase acidity, as well as the sol-

vent composition and pH. Especially, when compounds are eluted from an LC-system into an nano-esi source adducts tend to be formed. However, in most cases i) many of these adducts do produce after the 1st fragmentation step a (de-)protonated molecule, ii) the (de-)protonated molecule is mostly present next to the observed adducts and iii) the presence of this (de-)protonated is less influenced by actual spraying conditions in contrast to the e.g. $[M+Na]^+$ ion. Therefore, our initial approach focuses around the fragmentation of the $[M+H]^+$ or the $[M-H]^-$ ion. In a later stage we will evaluate how to incorporate fragmentation trees from initial precursor ions not being a $[M+H]^+$ or a $[M-H]^-$ ion. The fragmentation of ions strongly depends on mass spectrometric conditions, and these can be controlled using a number of parameters. Some parameters, such as resolution and isolation width, do not directly control the fragmentation process, but instead influence the detection and selection of ions, while other parameters do control the fragmentation process, such as collision energy, activation Q and activation time. Keeping other parameters at default settings, we compared fragmentation trees with different values for isolation width and collision energy, in order to assess the influence of these two parameters. In order to facilitate the analysis of MS^n spectra, the isolation width was adjusted so that we could, on the one hand, optimise sensitivity, and on the other hand isolate the mono-isotopic peak of the precursor ion without also isolating its ^{13}C isotopic peak. The absence of isotopic peaks in the fragmentation spectra prevented us from co-isolating fragment ions in subsequent MS^n stages, because their monoisotopic peaks were always at least 1 m/z unit apart (in singly charged ions). How these acquisition parameters (width of precursor ion isolation window and normalised collision energy) affected the detection of fragment ions is illustrated in Figure 3. A number of conclusions can be drawn from this figure. The first observation is that, as expected, the total ion count of the fragments observed in various MS^n stages decreases as the number of MS stages increases (Figure 2.3a and 2.3c). This is caused by a number of factors such as i) the loss of ions during the trapping, isolation and activation of both precursor and fragments ions in the different stages of the MS^n experiments, ii) the fact that ions below 1/3 of the pre-cursor ion m/z ratio are not trapped in the ion trap, and iii) the fact that ions below m/z 50 cannot be detected using the Orbitrap detector. Since we were not investigating the relative importance of these factors in this study, we can merely observe that the total ion intensity diminished as a result of multiple MS^n levels. The second observation is that on average, an isolation width of 2 m/z ($M \pm 1$ m/z) units led to higher total ion counts than an isolation width of 1 m/z ($M \pm 0.5$ m/z) unit (Figure 2.3a). This held for all the fragments observed. As a consequence of the higher peak intensities, the total number of EFPs detected in the fragmentation trees was, on average, 15 % higher when using a wider isolation width (Figure 3b). Thirdly, collision energy only had a marginal effect on the overall intensity (Figure 2.3c) and on the number of observed fragment ions and EFPs

(Figure 2.3d). A more detailed comparison of the effects of these acquisition parameters on the resulting fragmentation trees is shown in Figure 2.4. This figure plots the effect of the tested acquisition parameters on the relative intensities for particular EFPs. Clearly, the isolation width parameter did not significantly impact the relative intensity of the fragment ion peaks. Collision energy, on the other hand, did, as expected, influence these intensities, and consequently the ratio between fragment ions, although its influence turned out to be minor. In addition, we observed a small standard deviation of the relative intensity of the fragment ions (typically less than 2%, and max 9%). Basically, all tested collision energy settings yielded highly similar spectra.

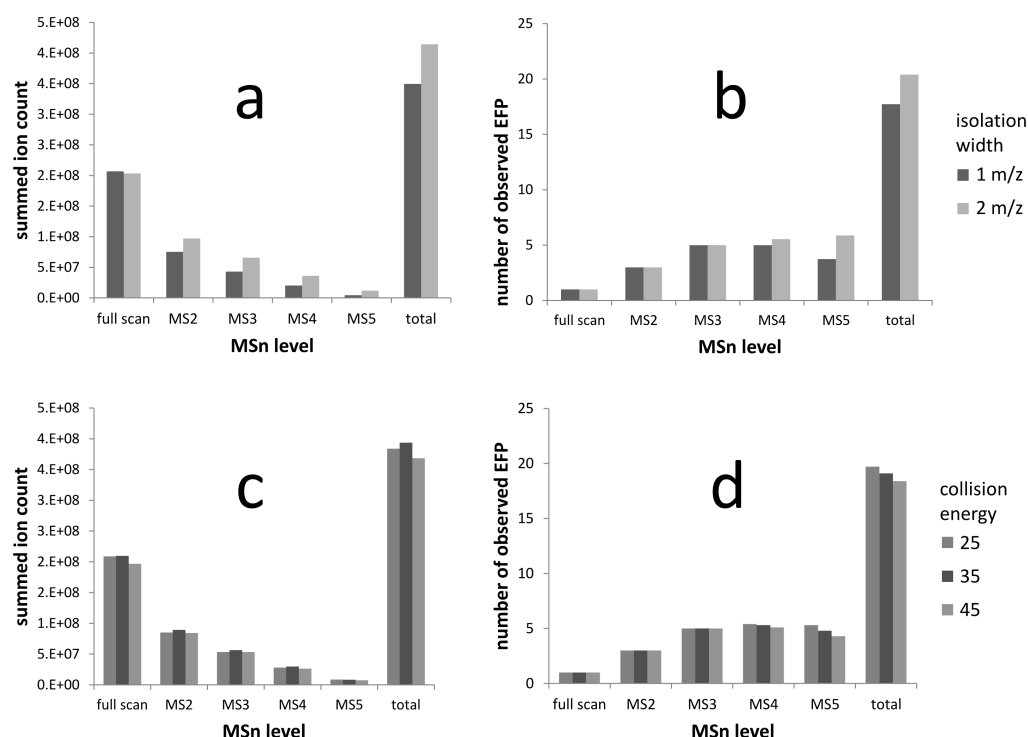


Figure 2.3: Influence of isolation width on summed ion count (a) and number of detected EFP (b), and influence of collision energy on summed ion count (c) and number of detected EFP (d) in MSⁿ spectra obtained for glutathione.

2.4.2. Reproducibility and robustness - the effects of the concentration of the analyte on the size and shape of the fragmentation tree

In order to study the influence of the analyte concentration on the reproducibility and robustness of the obtained fragmentation tree, trees were generated from MSⁿ spectra acquired with various concentrations of glutathione (ranging from 1 μ M to 1 mM). The absolute

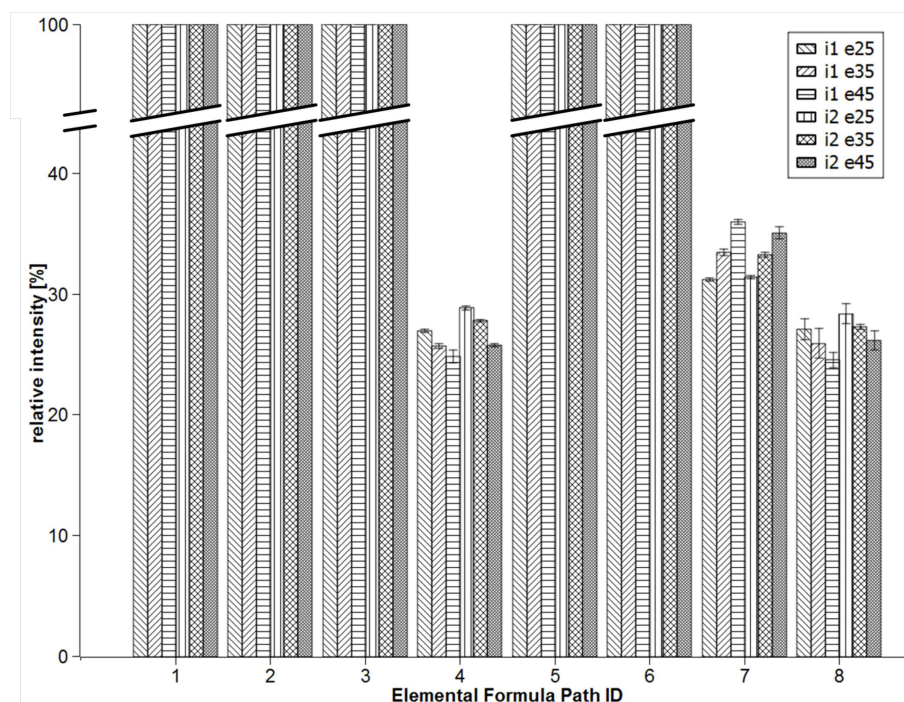


Figure 2.4: Comparison of fragmentation trees acquired using various mass isolation widths (i1 and i2, 1 m/z and 2 m/z respectively) and normalised collision energy (e25, e35, e45; normalised collision energy of 25%, 35% and 45% respectively) (shown on representative elemental formula paths from each level of MS^n spectra for clarity (see Table 1 for identity of elemental formula paths)). The relative intensity within a spectrum of according mass peaks is plotted with standard deviation.

intensities of fragment ions are directly related to the absolute intensity of their precursor ion, which in turn depends on the concentration of the analyte. The effect of the concentration of glutathione on the abundance of the observed fragment ions is illustrated in Figure 5a. As can be seen, the total summed intensity of the fragment ions in all MS^n levels was influenced by the absolute intensity of the precursor ion (MS^1), with one exception: fragment intensities for the two highest concentrations (1mM and 0.3 mM) seem to be the same even though the intensity of precursor ions differs. This is probably due to the fact that the number of charges which can be retained in the ion trap is limited. The analyte concentration also influenced both the total size of the obtained fragmentation tree and the total number of fragments detected in a tree as can be seen in Figure 5b. From 21 elemental formula paths observed for 1mM glutathione, 20 were observed for all glutathione concentrations in the range from 10 μM to 1 mM. Only the two lowest concentrations tested (3 μM and 1 μM) yielded smaller fragmentation trees, consisting respectively of 18 and 13 EFPs. The number of detected ions was most reduced in MS^4 and MS^5 spectra (reduction by almost

50%), but it was also reduced in MS³ and MS² spectra. Clearly, for all higher concentrations the instrument was able to compensate for the lower ion abundances by longer ion accumulation times, yielding fragment-rich MSⁿ spectra. Moreover, a comparison of the relative peak intensity obtained for each detected EFP demonstrated that the ratios between peak abundances can easily be reproduced across the whole range of glutathione concentrations (data not shown). As expected, the largest deviation was observed for low intensity peaks, which were characterised by the lowest signal-to-noise ratio. The results show that fragmentation tree topology can be acquired in a robust way and the analyte concentration has a minor effect on the overall arrangement of the fragmentation tree.

2.4.3. Specificity of analysis - fragmentation tree structure characteristic for isomeric prostaglandins

Isomerism is commonly observed in metabolites and specific isomers often have a unique biological function in a living organism. Therefore, a successful metabolite identification method, selective enough to discern between isomeric structures, is essential for understanding the biochemical roles of each individual isomer. Although tandem mass spectrometry inherently cannot distinguish enantiomers [Sawada, 1997] - for that we might consider derivatisation with a chiral label or separation with chiral chromatography prior to MS analysis - it can, in principle, differentiate between individual constitutional isomers and/or diastereoisomers [Gaucher & Leary, 1998]. In order to assess the feasibility of discerning between isomers on the basis of an analysis of their fragmentation trees, the above-mentioned protocol was used on two isomeric prostaglandins: D2 and E2 (Figure 6a and 6b, respectively). Prostaglandins are important mediators of (patho)-physiological effects [Shimizu, 2009]. Although chemically very similar, prostaglandin D2 (PGD₂) and prostaglandin E2 (PGE₂) have different biological functions. It is challenging but crucial for studies on biological systems to be able to distinguish such closely related structures reliably in a wide range of concentrations. We checked the repeatability and robustness of the acquisition of a fragmentation tree for both prostaglandins over a range of concentrations, viz. from 10 nM to 100 μ M, in the NI-mode. This polarity was used because the protonated molecule is not observed in positive ion mode due to predominant water loss [Nithipatikom *et al.*, 2003]. Since, as demonstrated above for glutathione, the analyte concentration influences the size of the resulting fragmentation tree, it was important in the case of prostaglandins to establish whether their concentration and the size of the resulting fragmentation trees interfered with the analysis aimed at distinguishing the two isomers. The analysis of the obtained fragmentation trees shows that both prostaglandins yield a similar number of fragment masses in their fragmentation trees (14 for PGD₂ and 18 for

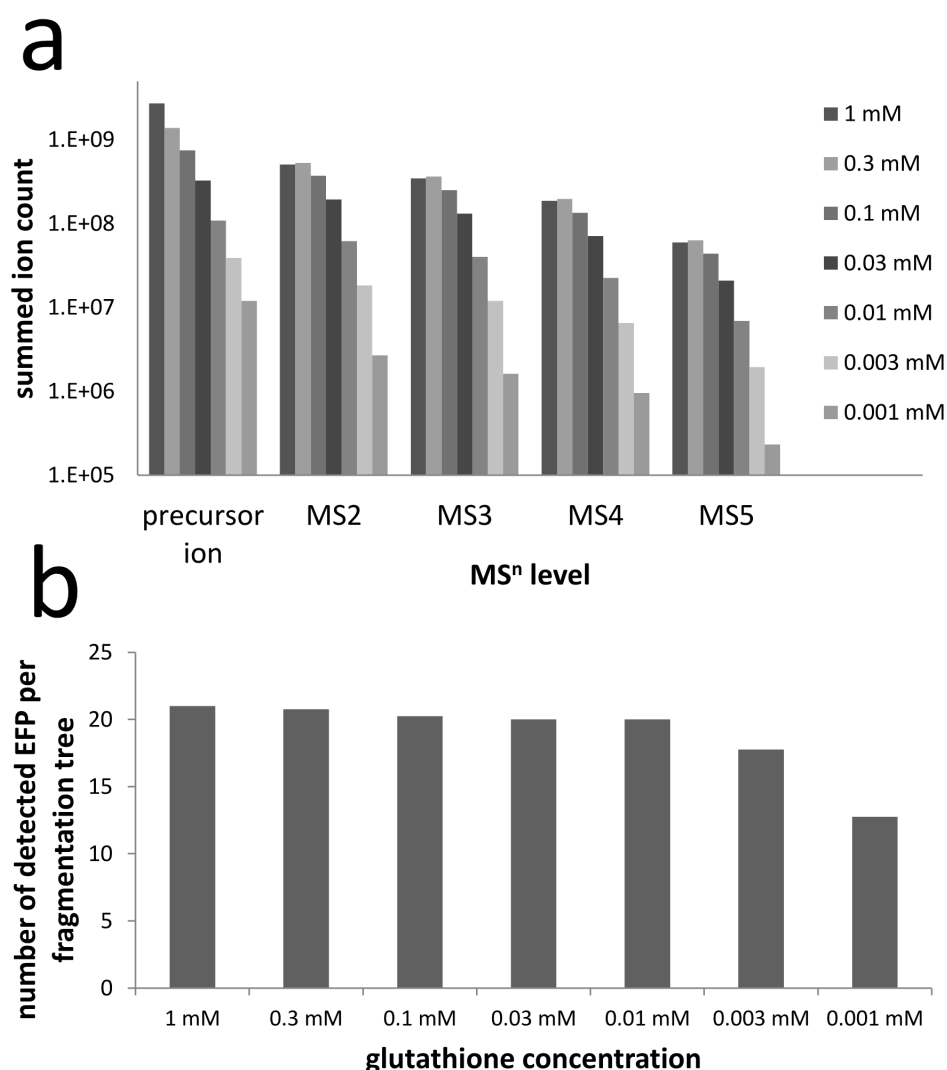


Figure 2.5: Influence of glutathione concentration on (a) the sum ion count of the mass peaks detected in its fragmentation tree (the ion count is plotted on logarithmic scale) and (b) the number of elemental formula paths detected in glutathione's fragmentation tree.

PGE2). These fragments constitute 25 elemental formula paths in the fragmentation tree of PGD2 and 31 elemental formula paths in the tree of PGE2 (Figure 6c). Despite their structural similarity, only 13 elemental formula paths are detected for both prostaglandins; 12 of the observed elemental formula paths are characteristic for PGD2 and 18 for PGE2. In total, 43 unique elemental formula paths are detected, consisting of 32 unique elemental formulas of fragment ions. Because the number of EFPs that we detect is higher than the number of observed fragment masses (as some fragment masses were observed in more than one spectrum), we postulate that the number of characteristic features for each

prostaglandin is higher than the number of characteristic features (peaks) that were reported in tandem MS spectra[Margalit *et al.*, 1996]. This means that our method is more specific than a method relying on tandem MS spectra. Comparing the fragmentation trees obtained from various concentrations of prostaglandins reveals that lowering the concentration influences the tree topology while preserving the characteristic elemental formula paths which distinguish between the isomers. Obviously, the fragmentation trees obtained from higher concentrations consist of more elemental formula paths than the fragmentation trees from lower concentrations (Figure 6c). In addition, as can be seen, the number of observed characteristic elemental formula paths for each prostaglandin decreases with the concentration. The differences between the fragmentation trees of PGD2 and PGE2 acquired from various concentrations are visualised in a clustered heatmap (Figure 6c). As a result of the above mentioned detection of characteristic EFPs, the fragmentation trees of the two prostaglandins form separate clusters. In the case of the two lowest concentrations (0.03 μ M and 0.01 μ M (H and I, respectively)) we only observe EFPs that are common to the two prostaglandins. These results suggest that with concentrations of 0.1 μ M and higher, isomeric prostaglandins can be distinguished unambiguously. Furthermore, it shows that minor fluctuations in the topology of a fragmentation tree between repetitions are negligibly smaller compared to the differences observed between isomeric metabolites, which proves the selectivity of our approach.

2.4.4. Specificity of analysis - distinguishing isomers

The approach, demonstrated above on the pair of isomeric prostaglandins, was evaluated on a set of 11 isomeric eicosanoids (Table 2.2), constituting 55 pairs of isomers. The similarity of fragmentation trees within each pair was given as a Tanimoto coefficient [Fligner *et al.*, 2002] (see Figure 2.7). The identical analysis of the fragmentation trees of isomeric prostaglandins (Figure 2.6) yielded Tanimoto coefficients of more than 70% for replicate measurements, more than 60% for similar fragmentation trees and less than 30% for dissimilar fragmentation trees (data not shown). These findings are a good indication for the interpretation of the data on the above mentioned eicosanoids. It can be seen in Figure 2.7 that for majority of fragmentation tree pairs the isomers can be distinguished: 48 pairs out of 55 yielded a Tanimoto coefficient lower than 30%. The remaining 7 pairs yielded Tanimoto coefficient lower than 70% implying dissimilarity. These results support the conclusion that using fragmentation trees as in our approach results in differentiation of positional isomers. Especially, in the case of isomeric structures one can envision that having actual annotations of elemental formulae of fragment ions, neutral losses and as well chemical structures of fragment ions in the database would be highly desirable. In order to accomplish this we store fragmentation trees consisting of mass spectrometric

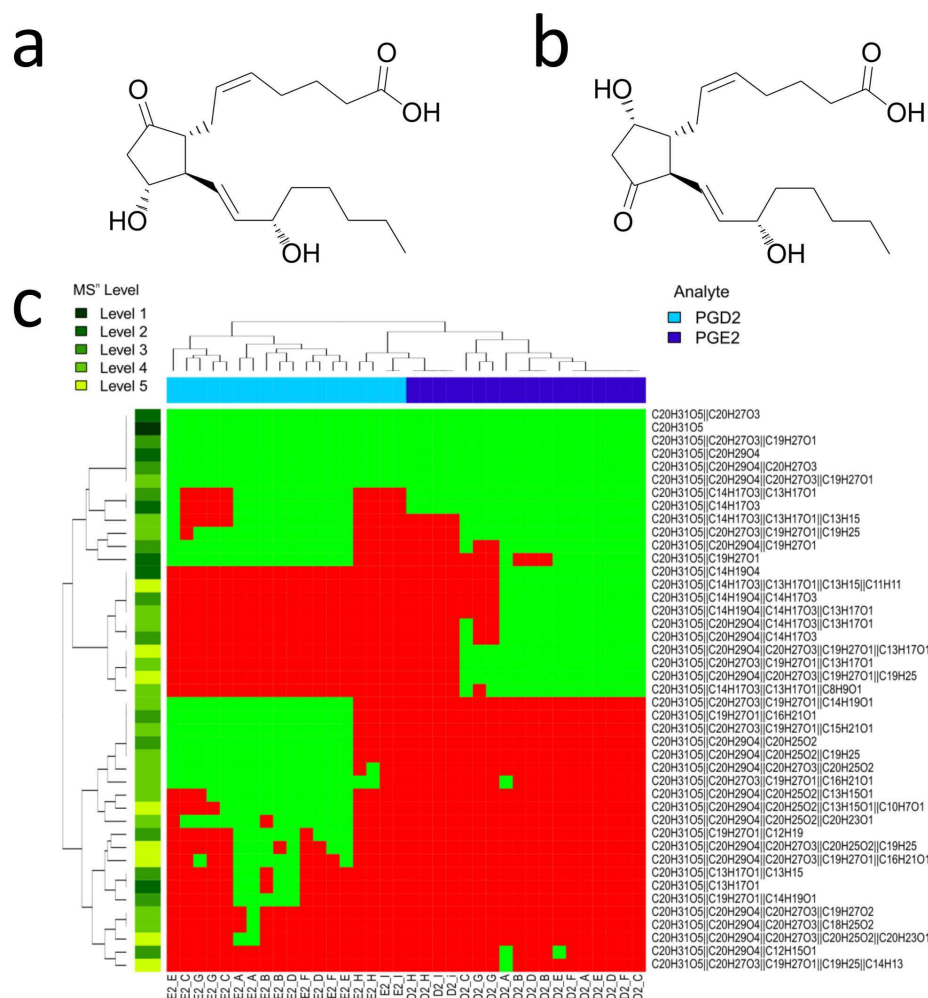


Figure 2.6: Chemical structure of prostaglandin E2 (CHEBI:15551) (a) and D2 (CHEBI:15555) (b) and clustered heatmap analysis of fragmentation trees acquired from various concentrations of prostaglandin E2 and prostaglandin D2 (green blocks denote detected, red blocks un detected EFPs). Fragmentation trees were acquired in duplicate from each concentration (denoted with letters A-I: A - 100 μ M, B - 30 μ M, C - 10 μ M, D - 3 μ M, E - 1 μ M, F - 0.3 μ M, G - 0.1 μ M, H - 0.03 μ M, I - 0.01 μ M). Both prostaglandins form separate clusters due to the observed characteristic EFPs, except in the case of the two lowest concentrations (H and I), where we observed no characteristic EFPs.

data annotated with elemental formula of fragment ions and neutral losses, using Chemical Markup Language (CML). CML supports various chemical concepts, such as reactions and molecules, and makes it possible to store the chemical structure of each ion in InChi format. To assist further structure elucidation structural annotation of these fragmentation trees will be added in a later stage. Such an endeavour to manually annotate a fragmentation tree collection could be developed by experts and the mass spectrometric society as an open

	8S-HETE	5S-HETE	8,9-EET	9-HETE	20-HETE	15S-HETE	5,6-EET	12-HETE	14,15-EET	11R-HETE
11,12-EET	17%	14%	26%	18%	14%	10%	7%	66%	11%	46%
11R-HETE	21%	12%	38%	16%	21%	16%	12%	18%	17%	
14,15-EET	17%	14%	17%	7%	18%	59%	13%	17%		
12-HETE	23%	12%	23%	9%	19%	20%	15%			
5,6-EET	12%	10%	23%	16%	10%	15%				
15S-HETE	21%	20%	16%	7%	17%					
20-HETE	15%	16%	15%	10%						
9-HETE	10%	7%	47%							
8,9-EET	61%	14%								
5S-HETE	35%									

Figure 2.7: Similarity of the fragmentation trees of isomeric eicosanoids given as Tanimoto coefficient. 11 eicosanoids (see Table 2) constitute 55 pairs of isomers. The similarity lower than 30% allows for unambiguous distinguishing of fragmentation trees.

project. Information on the actual structures, or the most likely structures, will allow for (sub-)structure recognition and immensely assist in the identification endeavour of unknowns. This will hugely impact the way in which we interpret fragmentation spectra.

2.5. Conclusion

Multistage mass spectrometry has, due to the popularity of ion trap instruments, become a very powerful technique for structural characterisation in e.g. metabolomics. Although the technique is well known, until now the resulting MS^n data could not be straightforwardly analysed, and the results were only accessible as collections of related fragmentation spectra. We demonstrate that the use of constraints derived from the precursor-product ion relations of the ions observed in MS^n spectra not only allows to efficiently remove artefacts, it also allows us to assign unambiguously, elemental formula to each relevant fragment ion. Reproducibly representing MS^n spectra as fragmentation trees allows for facile comparison of the fragmentation data of individual metabolites. This is extremely beneficial in view of our envisioned database based metabolite identification pipeline. Although we only demonstrate this approach by means of data of several compounds, fragmentation trees (PI and NI in most cases) of approx. 500 individual compounds (including ca. 100 isomers) present in human bio-fluids have in the meantime been acquired and are being evaluated. These compounds belong to a wide set of compound classes and span a large part of the (human) metabolome.

The adoption of this approach will greatly depend on accessibility of public

databases which store and exchange annotated fragmentation tree data. The data format used must accommodate both mass spectrometric and chemical data. To our knowledge, the only data format fulfilling this requirement is Chemical Markup Language (CML)[Murray-Rust & Rzepa, 2001]. The reproducibility and robustness of the acquisition of fragmentation trees suggests that they can potentially be used in computer-aided generic metabolite identification methods. However, in order to fully evaluate the feasibility of this approach, between-lab reproducibility must be assessed.

2.6. Acknowledgement

The authors acknowledge Dr. Agnieszka Kraj and Dr. Rob van der Heijden for their efforts in the early stages of the project. The authors are grateful to Prof. Dr. Nico Nibbering and Dr. Ronnie van Doorn for their input. Justin van der Hooft and Ric de Vos are thanked for their input in the application of the approach reported here. This project was (co)financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research.

2.7. Supporting Information

2.7.1. Library to process and compare MSⁿ data

The Java library to process and compare MSⁿ data is available as an open source project from Sourceforge at <http://sourceforge.net/projects/samsn>. The README file contains a tutorial explaining the features and the command lines to use.

```
sn = 1 . Signal to noise threshold
mzgap = 0.5 . Minimal distance between adjacent peaks
rint = 0.0, 0.05, 0.1, 0.2 . Relative intensity threshold. We process with 4
different values to simulate a dilution series.
acc = 15 . Maximal accuracy range set in ppm
rules = RDBE . (Ring Double Bond Equivalents) Constraint rules applied to
the formula occur = 0.4 . Minimum occurrence to appear in all repetitions
within one file to be accepted as a fragment.
ec = C0..10,H1..20. Elements to be included, together with the upper-
/lower-limit of the number of atoms. They will depend on the compound
to be analyzed. E.g. ec=C0..10,H1..20 means that the range of the carbon
atom is set between 0 and 10 and the range of the hydrogen atom is set
between 1 and 20.
```

MS1 isotope pattern information was not used for assigning the parent ion and fragments. All MSⁿ data was processed with the above tool using the following command line:

```
> java -jar sams.jar -occur=0.4 -sn=10 -mzgap=0.2 -rint=0 -acc=6 -
ec=[MY_ELEMENTS] -rules=[RDBER] -imzXML filename.mzXML -ocml
filename.cml process
```

2.7.2. The noise and artefact removal

The noise detection/removing is not a feature of MEF software. After peak picking performed by XCMS (with adjustable signal-to-noise threshold) the remaining noise and arte-

	8S-HETE	5S-HETE	8,9-EET	9-HETE	20-HETE	15S-HETE	5,6-EET	12-HETE	14,15-EET	11R-HETE
11,12-EET	0.43	0.45	0.45	0.16	0.29	0.32	0.31	0.77	0.36	0.41
11R-HETE	0.18	0.19	0.17	0.05	0.15	0.15	0.11	0.11	0.13	
14,15-EET	0.51	0.56	0.54	0.18	0.32	0.86	0.30	0.30		
12-HETE	0.41	0.46	0.40	0.09	0.24	0.32	0.20			
5,6-EET	0.34	0.32	0.39	0.19	0.31	0.25				
15S-HETE	0.53	0.60	0.50	0.12	0.35					
20-HETE	0.39	0.42	0.38	0.11						
9-HETE	0.17	0.16	0.33							
8,9-EET	0.92	0.73								
5S-HETE	0.78									

Figure 2.8: Dot-product comparison of total composite spectra generated from MSⁿ spectra of 11 eicosanoids (see Table 2) constituting 55 pairs of isomers. See Figure 7 for fragmentation tree comparison.

facts are removed as a result of peak assignment. MEF assigns the set of possible (within the allowed mass tolerance) elemental compositions to each peak and then insures consistency with elemental compositions of the precursor and consequent fragments. The noise is not detected and it is removed by removal of non-relevant peaks (peaks not belonging to the metabolite).

2.7.3. The noise and artefact removal

To emphasize the capability of the experimental setup the composite spectra of isomers listed in Table 2.2 were compared using algorithm of dot-product comparison [Stein & Scott, 1994]. MSⁿ spectra were summed and converted into total composite spectra with Xcalibur version 2.0.7 (Thermo Fisher Scientific, Waltham, MA). These composite spectra were then binned and the dot-product was calculated for each pair of isomers. The results are demonstrated in Figure 2.8. Although the isomers can be discerned from each other using this approach, the comparison of fragmentation trees using Tanimoto coefficient gives sharper distinction between isomers (Figure 2.2).

2.7.4. The Supporting Data

- The spectra analyzed in this paper (mzXML format) can be downloaded from <http://analyticalbiosciences.leidenuniv.nl/people/kasper>
- The MEF software used in the analysis can be downloaded from <http://abs.lacdr.gorlaeus.net/people/rojas-cherto>

Bibliography

- [Böcker & Rasche, 2008] Böcker, S. & Rasche, F. (2008) Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics (Oxford, England)*, **24** (16), i49–i55.
- [Cech & Enke, 2002] Cech, N. B. & Enke, C. G. (2002) Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews*, **20** (6), 362–387.
- [Cui *et al.*, 2000] Cui, M., Song, F., Zhou, Y., Liu, Z. & Liu, S. (2000) Rapid identification of saponins in plant extracts by electrospray ionization multi-stage tandem mass spectrometry and liquid chromatography/tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, **14** (14), 1280–1286.
- [Fligner *et al.*, 2002] Fligner, M. A., Verducci, J. S. & Blower, P. E. (2002) A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics*, **44** (2), 10.
- [Gaucher & Leary, 1998] Gaucher, S. P. & Leary, J. A. (1998) Stereochemical differentiation of mannose, glucose, galactose, and talose using zinc(II) diethylenetriamine and ESI-ion trap mass spectrometry. *Analytical Chemistry*, **70** (15), 3009–3014.
- [Hansen & Delattre, 1978] Hansen, P. & Delattre, M. (1978) Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association*, **73** (362), 397–403.
- [Heinonen *et al.*, 2008] Heinonen, M., Rantanen, A., Mielikäinen, T., Kokkonen, J., Kiuru, J., Ketola, R. A. & Rousu, J. (2008) FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid communications in mass spectrometry RCM*, **22** (19), 3043–3052.

- [Jarussophon *et al.*, 2009] Jarussophon, S., Acoca, S., Gao, J.-M., Deprez, C., Kiyota, T., Draghici, C., Purisima, E. & Konishi, Y. (2009) Automated molecular formula determination by tandem mass spectrometry (MS/MS). *The Analyst*, **134** (4), 690–700.
- [Kassler *et al.*, 2011] Kassler, A., Pittenauer, E., Doerr, N. & Allmaier, G. (2011) CID of singly charged antioxidants applied in lubricants by means of a 3D ion trap and a linear ion trap-Orbitrap mass spectrometer. *Journal of mass spectrometry JMS*, **46** (6), 517–528.
- [Kind & Fiehn, 2007] Kind, T. & Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, **8**, 105.
- [Konishi *et al.*, 2007] Konishi, Y., Kiyota, T., Draghici, C., Gao, J.-M., Yeboah, F., Acoca, S., Jarussophon, S. & Purisima, E. (2007) Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products. *Analytical chemistry*, **79** (3), 1187–97.
- [Kwan & Huang, 2008] Kwan, E. E. & Huang, S. G. (2008) Structural Elucidation with NMR Spectroscopy: Practical Strategies for Organic Chemists. *European Journal of Organic Chemistry*, **2008** (16), 2671–2688.
- [March, 1997] March, R. E. (1997) An Introduction to Quadrupole Ion Trap Mass Spectrometry. *Journal of Mass Spectrometry*, **32** (February), 351–369.
- [Margalit *et al.*, 1996] Margalit, A., Duffin, K. L. & Isakson, P. C. (1996) Rapid quantitation of a large scope of eicosanoids in two models of inflammation: development of an electrospray and tandem mass spectrometry method and application to biological studies. *Analytical Biochemistry*, **235** (1), 73–81.
- [Martens *et al.*, 2011] Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpf, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A. & Deutsch, E. W. (2011) mzML—a community standard for mass spectrometry data. *Molecular cellular proteomics MCP*, **10** (1), R110.000133.
- [Mathur & O'Connor, 2009] Mathur, R. & O'Connor, P. B. (2009) Artifacts in Fourier transform mass spectrometry. *Rapid communications in mass spectrometry : RCM*, **23** (4), 523–9.
- [McLafferty & Turecek, 1993] McLafferty, F. W. & Turecek (1993) *Interpretation of Mass Spectra*. University Science Books.
-

- [McLuckey *et al.*, 1994] McLuckey, S. A., Van Berkel, G. J., Goeringer, D. E. & Glish, G. L. (1994) Ion trap mass spectrometry of externally generated ions. *Analytical Chemistry*, **66** (13), 689A–696A.
- [Milman, 2005] Milman, B. L. (2005) Towards a full reference library of MS(n) spectra. Testing of a library containing 3126 MS2 spectra of 1743 compounds. *Rapid communications in mass spectrometry : RCM*, **19** (19), 2833–9.
- [Montoya *et al.*, 2009] Montoya, G., Arango, G. J. & Ramírez-Pineda, J. R. (2009) Rapid differentiation of isobaric and positional isomers of structurally related glycosides from *Phytolacca bogotensis*. *Rapid communications in mass spectrometry RCM*, **23** (21), 3361–3371.
- [Murray-Rust & Rzepa, 2001] Murray-Rust, P. & Rzepa, H. (2001) Chemical markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *J. Chem. Inf. Comput. Sci*, **41** (5), 1113–1123.
- [Nithipatikom *et al.*, 2003] Nithipatikom, K., Laabs, N. D., Isbell, M. A. & Campbell, W. B. (2003) Liquid chromatographic-mass spectrometric determination of cyclooxygenase metabolites of arachidonic acid in cultured cells. *Journal Of Chromatography B Analytical Technologies In The Biomedical And Life Sciences*, **785** (1), 135–145.
- [Oberacher *et al.*, 2009] Oberacher, H., Pavlic, M., Libiseller, K., Schubert, B., Sulyok, M., Schuhmacher, R., Csaszar, E. & Köfeler, H. C. (2009) On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *Journal of mass spectrometry JMS*, **44** (4), 485–493.
- [Pedrioli *et al.*, 2004] 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. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, **22** (11), 1459–66.
- [R Development Core Team, 2008] R Development Core Team (2008). R: A Language and Environment for Statistical Computing.
- [Rasche *et al.*, 2011] Rasche, F., SvatosìĚ, A., Maddula, R. R. K., BoìŁttcher, C. & BoìŁlcker, S. (2011) Computing Fragmentation Trees from Tandem Mass Spectrometry Data. *Analytical Chemistryhemistry*, **83** (4), 1243–1251.
-

- [Robertson, 2005] Robertson, D. G. (2005) Metabonomics in Toxicology: A Review. *Toxicological Sciences*, **85** (2), 809–822.
- [Rojas-Chertó *et al.*, 2011] Rojas-Chertó, M., Kasper, P. T., Willighagen, E. L., Vreeken, R., Hankemeier, T. & Reijmers, T. (2011) Elemental Composition determination based on MSn. *Bioinformatics*, **27** (17), 2376–2383.
- [Sawada, 1997] Sawada, M. (1997) Chiral recognition detected by fast atom bombardment mass spectrometry. *Mass spectrometry reviews*, **16** (2), 73–90.
- [Scalbert *et al.*, 2009] Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B. S., Van Ommen, B., Pujos-Guillot, E., Verheij, E., Wishart, D. & Wopereis, S. (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics : Official journal of the Metabolomic Society*, **5** (4), 435–458.
- [Scheubert *et al.*, 2011] Scheubert, K., Hufsky, F., Rasche, F. & Böcker, S. (2011) Computing Fragmentation Trees from Metabolite Multiple Mass Spectrometry Data. *Journal of computational biology*, **18** (11), 377–391.
- [Schug & McNair, 2003] Schug, K. & McNair, H. M. (2003) Adduct formation in electrospray ionization mass spectrometry II. Benzoic acid derivatives. *Journal of chromatography. A*, **985** (1-2), 531–9.
- [Sheldon *et al.*, 2009] Sheldon, M. T., Mistrik, R. & Croley, T. R. (2009) Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *Journal of the American Society for Mass Spectrometry*, **20** (3), 370–6.
- [Shimizu, 2009] Shimizu, T. (2009) Lipid mediators in health and disease: enzymes and receptors as therapeutic targets for the regulation of immunity and inflammation. *Annual review of pharmacology and toxicology*, **49** (1), 123–50.
- [Smith *et al.*, 2006] Smith, C. A., O'Maille, G., Want, E. J., Abagyan, R. & Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using non-linear peak alignment, matching, and identification. *Analytical chemistry*, **78** (3), 779–87.
- [Stein & Scott, 1994] Stein, S. E. & Scott, D. R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, **5** (9), 859–866.
- [Steinbeck *et al.*, 2003] Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003) The Chemistry Development Kit (CDK): an open-source Java library
-

for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, **43** (2), 493–500.

[van der Hooft *et al.*, 2011] van der Hooft, J. J. J., Mihaleva, V., de Vos, R. C. H., Bino, R. J. & Vervoort, J. (2011) A strategy for fast structural elucidation of metabolites in small volume plant extracts using automated MS-guided LC-MS-SPE-NMR. *Magnetic resonance in chemistry : MRC*, **49 Suppl 1**, S55–60.

[Villas-bo *et al.*, 2005] Villas-bo, S. G., Smedsgaard, J. r. & Nielsen, J. (2005) Mass spectrometry in metabolome analysis. *Building*, **24** (5), 613–646.

[Wolf *et al.*, 2010] Wolf, S., Schmidt, S., Muller-Hannemann, M. & Neumann, S. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, **11** (1), 148.

CHAPTER 3

Elemental Composition determination based on MSⁿ

Miguel Rojas-Chertó^{1,2}, Piotr T. Kasper^{1,2}, Egon L. Willighagen^{1,3,4}, Rob Vreeken^{1,2},
Thomas Hankemeier^{1,2} and Theo Reijmers^{1,2}

Bioinformatics 2011:27(17):2376-83

¹Netherlands Metabolomics Centre, Leiden, The Netherlands

²Division of Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands

³Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

⁴Plant Research International, Wageningen UR, Wageningen, The Netherlands

3.1. Abstract

Identification of metabolites is essential for its use as biomarkers, for research in systems biology, and for drug discovery. The first step before a structure can be elucidated is to determine its elemental composition. High resolution mass spectrometry, which provides the exact mass, together with common constraint-rules, for rejecting false proposed elemental compositions, can not always provide one unique elemental composition solution. The Multi-stage Elemental Formula (MEF) tool is presented in this paper to enable the correct assignment of elemental composition to compounds, their fragment ions, and neutral losses that originate from the molecular ion by using multi-stage mass spectrometry (MS^n). The method provided by MEF reduces the list of predicted elemental compositions for each ion by analyzing the elemental compositions of its parent (precursor ion) and descendants (fragments). MS^n data of several metabolites were processed using the MEF tool to assign the correct elemental composition and validate the efficacy of the method. Especially the link between the mass accuracy needed to generate one unique elemental composition and the topology of the MS^n tree (the width and the depth of the tree) was addressed. This method makes an important step towards semi-automatic de novo identification of metabolites using MS^n data.

3.2. Introduction

Metabolomics is the extensive examination of the metabolic phenotype, the metabolome, under a specific set of conditions. It has emerged as a functional key to understand the behavior of complex biological systems, including cells, tissues, and biofluids. Currently, the essential challenges addressed in metabolomics are identification (full coverage of the metabolome with structural characterization), and quantification (accurately measure concentrations at low abundance levels). Metabolite identification to enable proper biological interpretation is the first step in the research of biomarkers, systems biology, and drug discovery [Butcher *et al.*, 2004, Dunn, 2008, Kind & Fiehn, 2010].

ElectroSpray Ionization (ESI) enabled Mass Spectrometry (MS) to become the most essential analytical tool used in both quantitative and qualitative metabolomic research [Cui *et al.*, 2000]. In combination with liquid or gas chromatography, MS is the widely accepted standard method [Dunn & Ellis, 2005] for the analysis of metabolomic samples. In this way each compound in the measured sample is characterized by a retention time and a mass value.

However, this characterisation is not sufficient for metabolite identification, and the first step towards elucidation of the chemical structure from an unidentified compound is the de-

termination of its elemental composition. An elemental composition expresses which and how many atoms constitute a particular chemical compound. Although each elemental composition has a unique molecular weight (mass), a molecular weight does not have a unique elemental composition. This situation is complicated by the instrumental precision, and the accuracy of the mass measurements limits this viability. Nowadays, mass spectrometry instrumentation such as time-of-flight, ion cyclotron resonance or magnetic sectors, are able to measure mass-to-charge ratios (m/z) with a mass accuracy up to 1 ppm (parts per million) meaning that a mass of 100 Da is measured accurately up to four decimal places. In addition, modern desktop computers together with available software tools make it possible to generate and verify instantaneously all theoretical possible chemical elemental compositions for a given mass. Regrettably, several studies have shown that even high mass accuracy (smaller than 1 ppm) does not necessarily result in one unique assigned elemental composition [Kind & Fiehn, 2006, Kim *et al.*, 2006]. The number of possible elemental compositions increases exponentially with the mass and the set of chemical elements taken into account. Generally, available methods that derive the elemental composition from a given mass use the following three steps, generation, filtering and matching. The generation step generates a candidate list with all possible elemental compositions enumerated systematically. The filtering step rejects all the elemental compositions that do not satisfy certain rules. The matching step compares the theoretical isotope patterns with the experimental one. The best match is reflected as the most probable elemental composition. Normally, the elemental composition annotation process starts when the user provides a constraint set consisting of the mass of the ion, the set of chemical elements to include and exclude, the limit range of number of atoms for each element, and the mass tolerance (mass error window). There are two different approaches for generating the elemental composition and these depend on the search model used. First there is the deterministic search that enumerates all possible elemental compositions. This approach is computationally intensive but checks the complete solution space [Dromey & Foyster, 1980]. Next to this approach is the local search approach where the investigated solution space is restricted. An example of this approach is developed by [Zhang *et al.*, 2005] who based the generation of possible elemental compositions on the optimization of the match between the theoretical and observed isotope patterns. Constraint-rules are applied to limit the number of possible elemental composition candidates and they are generally based on empirical knowledge. They are well documented and their limitations have been analyzed elsewhere [Kind & Fiehn, 2007]. Several examples of these chemical rules are the rings-plus-double-bonds equivalent (RDBE) [Dayringer & McLafferty, 1977] or double-bond equivalent (DBE), LEWIS and SENIOR rule [Senior, 1951] and the nitrogen rule. The application of heuristic rules [Kind & Fiehn, 2007] is a different approach to exclude non-valid

elemental compositions. These constraints are derived from the analysis of chemical structure databases. The extracted rules heavily depend on the quality and diversity of the data in the database from which the rules were determined. As a consequence, those who do not have access to rich chemical structure databases will have limited success in extracting reliable rules. Fortunately, recent initiatives like Blue Obelisk movement [Guha *et al.*, 2006] or Science Commons are encouraging open source, open data, and open standards which facilitate better access for scientists. Next to the mass, the observed isotope pattern is used to eliminate elemental compositions from the candidate list because different elemental compositions will have different mass spectral isotopomeric abundances. The isotope distribution provides information which is unique for a given elemental composition. Hence, the experimental isotope abundance pattern of a metabolite's mass spectrum can be compared with the theoretical one [Stoll *et al.*, 2006] to remove false candidates and produce a final list of results sorted according to the degrees of similarity, called the 'hit-list'. However, insufficient intensity, limited accuracy, overlapping isotope patterns and co-isolation, complicate this approach and make extraction of the isotope pattern from the experimental data a difficult task. Furthermore, the peak intensities and masses sometimes are not accurate because often MS data is acquired using a malfunctioning centroiding algorithm [Erve *et al.*, 2009, Gu *et al.*, 2006]. For these reasons it is a challenge to search for additional rules which improve the efficacy of determining the elemental composition.

An new approach to introduce constraints in the search for the unique elemental composition is to use multi-stage mass spectrometry (MS^n) information. The technology used in multi-stage fragmentation mass spectrometry permits, by consecutive isolation and fragmentation of ions under low-energy collision-induced dissociation (CID), the creation of a set of hierarchical linked mass spectral data, as shown in Figure 1a. Whenever the number of ions is significant large, each new generated fragment ion can be isolated and applied to new collisions. Through this procedure, a new mass spectrum of fragment ions is obtained. The hence created mass spectral tree data, with its interlinked ion relations, gives a richer description of the measured compound than the data obtained from single-stage MS or Tandem MS. The obtained tree topology more accurately characterizes the analyzed compound. It should be noted that the MS^n approach has not been developed to obtain an elemental composition of the molecular ion alone. The MS^n approach will become important for the identification of unknown metabolites if more MS^n data in reference databases will become available. The MS^n approach also suffers from similar issues as mentioned previously when using isotopic patterns (insufficient intensity, overlapping mass spectral trees) but especially when isotopic information is not available, because of limited accuracy, the MS^n approach may be of use.

Many approaches have been designed for metabolite identification using MS^n , but they

often involve manual intervention by mass spectrometry experts and consequently are a very time consuming task [Cui *et al.*, 2000, Konishi *et al.*, 2007, Jarussophon *et al.*, 2009]. Due to the complexity of processing experimental high resolution MS^n data, it is to be preferred that for interpretation of this type of data a systematic computational process is used. E.g. solving the elemental compositions for fragment ions is rather complex because for the fragment ions in MS^n spectra not the same constraint rules can be applied as for molecular ions. Nowadays more techniques are becoming available enabling automated processing of MS^n or MS/MS data like SmartFormula3D (Bruker Daltonics) [Tyrkkö *et al.*, 2010] and the method proposed by [Rasche *et al.*, 2011].

This article presents a new method to determine the elemental composition of a certain compound using MS^n data. It is shown that applying this approach to experimental MS^n data results in an unique elemental composition of the parent ion, their fragments and neutral losses, for several metabolites analyzed in our lab. Additionally, the dependency between mass accuracy and tree topology is analyzed and used to device guidelines for creating spectral trees with sufficient information to uniquely determine the molecular formula.

3.3. Approach

3.3.1. Algorithm constraining elemental composition generator

The multi-stage elemental formula (MEF) tool allows the determination of the elemental composition of ions and neutral losses from experimental MS^n data. First, the hierarchical mass spectral data needs to be preprocessed and translated to a fragmentation tree representation. We define a fragmentation tree as a hierarchical organization of ions in a graphical form, such as shown in Figure 3.1. The tree in Figure 3.1 shows colored nodes that define the ions/fragments, while the edges reflect the fragmentation reactions occurring. Fragments originating from a precursor ion are called child nodes and these are situated below the precursor ion/parent node in the tree. Each child node has only one parent. The so-called root node, shown at the top of the tree, has no parents and generally is the protonated or deprotonated molecular ion. All fragment ions that originate from the same parent ion with the same acquisition time, belong to the same experimental scan of the mass spectrum. We define the neutral losses as the residue of the fragmentation product which is not detectable by the mass spectrometer. We can calculate the mass of the neutral losses as the difference between the mass of the fragment and its precursor, if both masses are known. Note that the neutral loss masses do not have to correspond to one unique chemical structure. It could also express the sum of the masses for different neutral

losses from consecutive fragmentations. The number of possible elemental compositions for a given mass depends heavily on the upper and lower limit of the number of atoms admitted to be present in the chemical formula. By narrowing the range of the number of atoms of each chemical element in the elemental composition, the list of theoretical possible elemental composition candidates decreases. The MEF tool, that we developed, is based on constraining for each chemical element the upper and lower limit (defined as the range) of number of atoms admitted to be present in the elemental composition. These ranges are derived from the elemental compositions of the precursor and fragments. Figure 3.3 shows an example of how an elemental composition range is generated from a list of elemental compositions. The range is extracted using the highest and lowest number of atoms in the complete elemental composition list. When a certain chemical element is not present in one theoretical possible elemental composition on the candidate list, the lower limit value is set to 0. When all elemental compositions in the list do not contain a certain element both the upper and lower limit values are set to 0. Repetition of this process for all possible precursor-fragment-child combinations produces new constraints that are used in the next cycle to further decrease the list of candidates. The procedure for solving the elemental composition of the precursor, the fragment ions, and the neutral losses using fragmentation trees is summarized briefly in Figure 3.2. So the process begins by defining the input data for the MEF: the masses of all ions and the relations between them (the fragmentation pattern). The pre-processing paragraph in the Methods section describes in more detail how this information is extracted from the raw MS^n data. The first step, before any analysis can begin, is the correction of the ion masses. Depending on the detection mode used during acquisition (positive or negative mode) the masses of the ions must be adjusted (extracting or adding the mass of an electron) to make a valid comparison between the masses calculated from the theoretical-generated elemental compositions and the experimental mass. Next, an elemental composition generator generates and lists all possible theoretical elemental compositions for each ion and the neutral losses. The set of input parameters needed to generate the elemental compositions consists of the mass, the mass accuracy (Δm), the set of chemical elements to be present in the elemental composition and the range of the number of atoms for each element. If for some ion no chemical formula is derived, this particular ion and its fragments and the fragments of the fragments are removed from further analysis. For all ions and neutral losses a candidate list of potential elemental compositions is generated, which is subsequently used to extract the chemical formula ranges. These ranges correspond with the upper and lower number of atoms admitted for each chemical element in the candidate list (for an example see Figure 3.3). Once an elemental composition range has been obtained for the fragment ions and the neutral losses, three constraint rules are applied: the precursor consistency rule, the

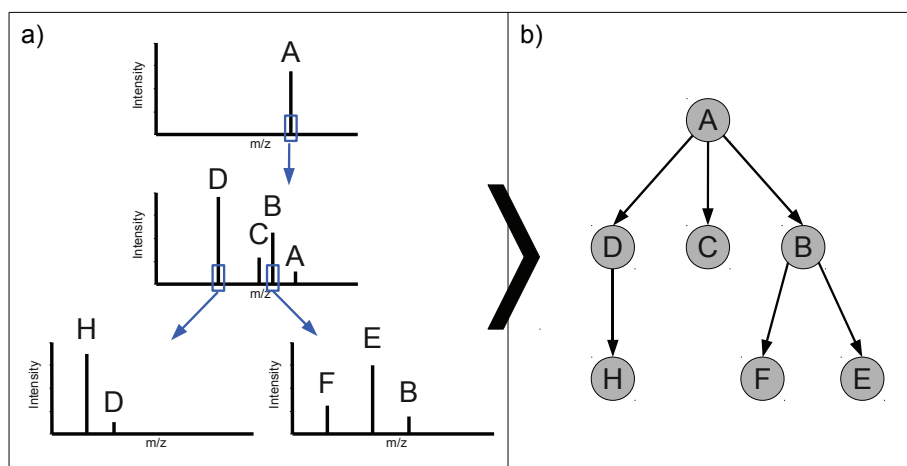


Figure 3.1: Correlation between the spectral tree graph representation (a) and the fragmentation tree graph representation (b). In a spectral tree the nodes are defined by spectra while in a fragmentation tree the nodes are characterized by ions and the edges by fragmentation paths.

fragment consistency rule, and the combinatorial consistency rule. These rules have the function to reduce the number of generated elemental compositions on each candidate by consulting the elemental compositions of the parent and the fragments.

The precursor consistency rule validates an elemental composition of a certain ion according to which elemental compositions are assigned to its precursor ion. The precursor consistency rule defines that the number of atoms per element in the elemental composition of an ion can not exceed the upper limit defined by the elemental composition range of the precursor ion. In other words, a fragment can not contain more atoms of a certain chemical element than the precursor does.

The fragment consistency rule validates the elemental composition from the parent point of view. An elemental composition is considered valid when for all chemical elements the number of atoms for a specific element is higher than the lower limit of the elemental composition range of the fragment(s). A precursor ion can not contain less atoms of a certain chemical element than any of its fragments ions.

The combinatorial consistency rule uses the concept of conservation of mass. In a chemical reaction the total mass of the reactants is equal to the total mass of the products. As fragmentation reactions are also chemical reactions, it is applied here as well but not on the mass level but on the elemental composition level. The combinatorial consistency rule validates the elemental composition by checking if certain combinations of elemental compositions are present in the different candidate lists. There are 3 ways to do this.

- The elemental composition of a parent ion is accepted if at least once it is found as the

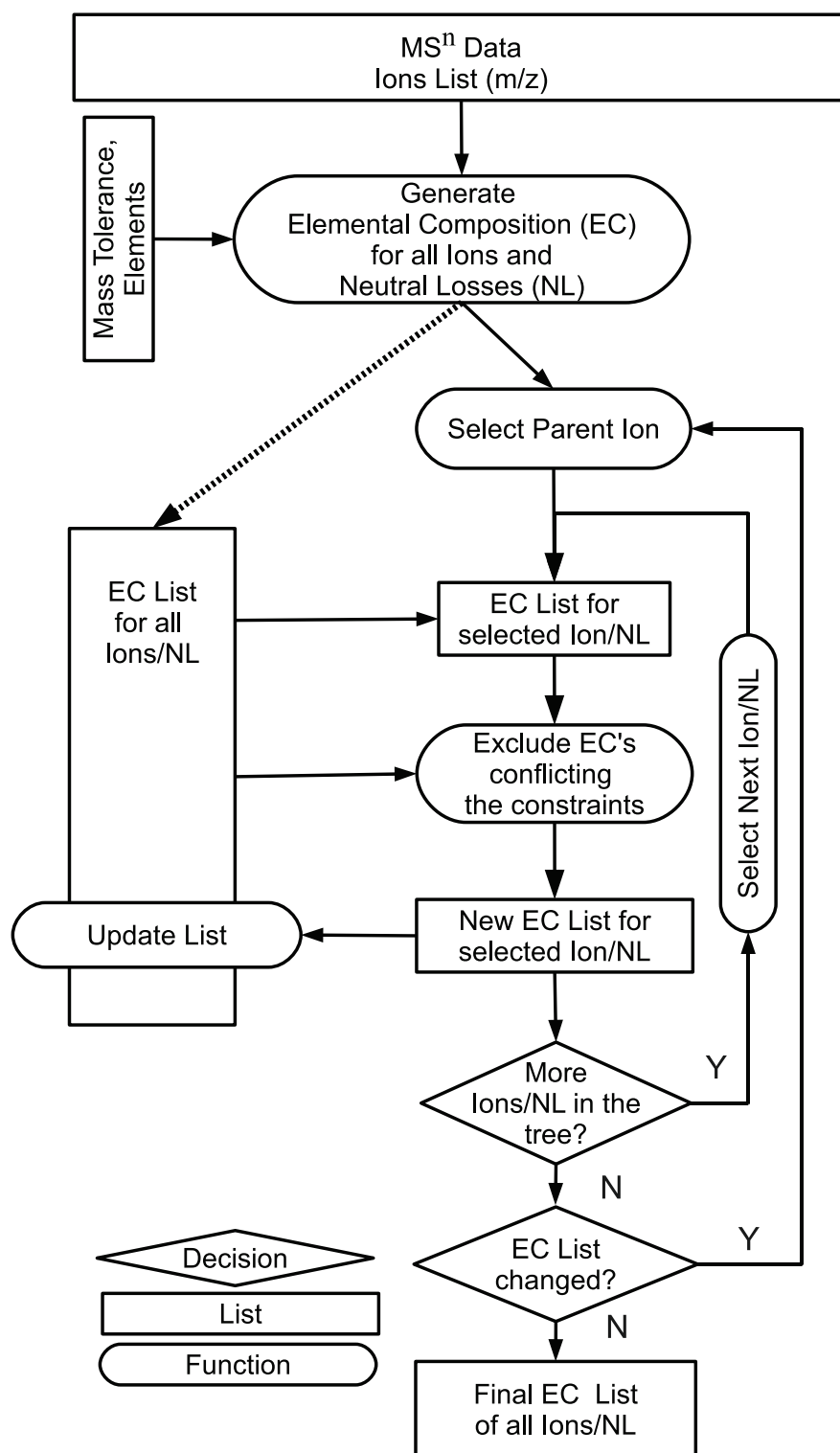


Figure 3.2: Flow diagram of the MEF method extracting the elemental composition of all ions and neutral losses given MS^n data.

389.185179 +/- 5ppm		C	H	N	O	S	P
$C_{22}H_{23}N_5O_2 \Rightarrow$		22	23	5	2	--	--
$C_{19}H_{26}N_4O_5 \Rightarrow$		19	26	4	5	--	--
$C_{22}H_{32}N_1O_1S_2 \Rightarrow$		22	32	1	1	2	--
$C_{19-22}H_{23-32}N_{1-5}O_{1-5}S_{0-2} \Leftarrow$		C_{19-22}	H_{23-32}	N_{1-5}	O_{1-5}	S_{0-2}	P_{0-0}

Figure 3.3: The example above shows how the elemental composition range is derived from an elemental composition list when a mass of 389.185179 Da. is provided with a mass tolerance of 5ppm.

sum of one elemental composition from a fragment and one elemental composition of the neutral loss. If no such combination is found the elemental combination of the parent ion is removed from the list.

- The elemental composition of a fragment is accepted if at least once it is found as the difference of one elemental composition from a parent and one elemental composition of the neutral loss. If no such combination is observed the elemental composition of the fragment is removed.
- The elemental composition of a neutral loss is accepted if at least once it is found as the difference of one elemental composition from a parent and one elemental composition of a fragment. If no such combination is found the elemental composition of the neutral loss is removed.

These constraints are applied to each ion. When all ions are analyzed the process is repeated for all ions starting from the parent node. This will be stopped when the list of candidate elemental compositions for all ions and neutral losses has not changed anymore. As soon as one or several elemental compositions are removed from the candidate list of a certain ion this may result in removal of elemental composition in neighboring ions. Table 3.1 displays an example how the total number of possible elemental compositions is decreasing in each loop. These numbers are obtained after applying the MEF tool to MSⁿ data acquired for the compound Threonine. For simplification only information of part of the nodes is shown. Any change to the list with potential elemental compositions will produce a new constraint in the next loop and will influence the lists of other ions. Finally, the iteration finishes when all ions conclude with the assignment of one unique elemental composition.

Fragment			Cycle				
Ion ID	Precursor ID	mass	1 st	2 nd	3 rd	4 th	5 th
1		120.06530	30	10	2	1	
2	1	102.0556	6	4	2	1	
3	2	56.049 3	3	3	2	1	
4	2	84.044 7	7	6	2	1	
5	1	74.060 19	13	4	2	1	

Table 3.1: Total number of elemental compositions obtained for each ion in each cycle during the application of the MEF tool for Threonine. The results were obtained using a mass tolerance of 10ppm and the following set of atoms; C, H, N and O.

3.4. Methods

3.4.1. Experimental section

A group of twelve compounds (see supplementary material about compounds) with known structure within a mass range of 150-450 Da was used to demonstrate the MEF method. The metabolites used in these experiments were purchased from Sigma (Sigma-Aldrich, Steinheim, Germany) and were of highest available purity. All compounds were dissolved at concentration of 0.1 mM. The solvent was 1:1 methanol:water (v/v) containing 0.1 % formic acid. Solvents were of UPLC/MS quality and were purchased from Biosolve (Valkenswaard, The Netherlands). Mass spectra for these twelve compounds were obtained using a Finnigan LTQ-Orbitrap (Thermo Electron Corp.). The MSⁿ experiments were recorded using a data-dependent scanning function with the criteria to select the five most intense ions detected for MS² and the three most intense ions for the rest of the MSⁿ levels. For signal averaging, the mass spectrometer was set with five microscans. The Orbitrap was operated at 30,000 resolution, normalized collision energy of 35 % and an isolation window of 1 Th.

3.4.2. Pre-processing

The information needed to start the MEF method consists of the exact masses, intensities for each peak, the specific acquisition times and the precursor scan. This information was extracted from the raw data using different existing external tools that were adopted to handle MSⁿ data. Ultimately a pipeline for automated processing of multi-level mass spectral tree data was created by connecting these different tools. The raw MS data files (binary

files) were converted to mzXML format [Pedrioli *et al.*, 2004] using ReadW software which is provided by the Institute for Systems Biology (the ISB) (see supplementary material for the mzXML files). We chose for the mzXML format because it is vendor-independent and lists the precursor mass attributes that are used to find the relations between the different MS spectra (the hierarchical links of the spectral tree). The information about the relation between the different fragments of the spectral tree turned out to be not present in NetCDF (ASTM E2078-00 'Standard Guide for Analytical Data Interchange Protocol for Mass Spectrometric Data') files (another often used format to transfer mass spectral information from the analytical platform to data analysis software). For extraction of the ion peaks and finding relations between fragments of MS data, we used the freely available XCMS software [Smith *et al.*, 2006]. XCMS reads mzXML files and identifies ion features (a specific m/z at a specific acquisition time and the precursor scan). With this information it is possible to link specific fragment ion formation to the parent ion creating a hierarchical fragmentation path (fragmentation tree). The MS^n data were peak-detected and noise-reduced to exclude signals related to noise which could interfere in the analysis. The settings used to process all the MS data were the default XCMS settings. The final result is a table containing the information about the ion fragment peaks and their precursor ions which is used to initiate the MEF method for the extraction of the elemental compositions.

3.4.3. Data storage

It is important that after any information retrieval the outcome is stored for posterior handling. Here the fragmentation pattern needed to be stored. There are several formats to store a chemical reaction representation, for example formats based on SMILES like mrv [Bode, 2004], the connection table based formats, such as Symyx molfiles and rxn files, and the markup-based format, Chemical Markup Language (CML) [Murray-Rust *et al.*, 2001]. At the moment we generate fragmentation trees from MS^n data which represents sequences of reactions. Each ion (reactant and product) is characterized by its elemental composition. Thus, we have chosen to use the CMLReact [Holliday *et al.*, 2006], an extension of CML, to connect the reaction components. CML has the ability to share all general XML features and unifies all available information for Internet publishing and computer processing. It supports various chemical concepts, such as molecules, reactions, spectra, and other chemical data and data sources. The reaction is represented by molecular species behaving as reactants and products using the appropriate tags. In our application we describe the ions only with their elemental composition, each product or reactant is defined with the tag `elementalformula`. When a fragment has more than one elemental candidate, these are put into a list.

Elemental Formula Generator code

The generator for the automated extraction and handling of chemical formula given a molecular mass has been developed separately. It is available as part of the Chemistry Development Kit [Steinbeck *et al.*, 2003] (CDK) library. CDK is an open source Java library for chemoinformatics and bioinformatics which provides code for calculating QSAR descriptors, applying 2D and 3D modeling techniques, defining reaction mechanisms, etc. To generate elemental compositions first the mass accuracy, the set of the chemical elements and the maximal and minimal limit of the number of atoms to be present in the formula must be specified. The exact working of the algorithm generating all mathematical possible chemical element combinations is described in the article of [Dromey & Foyster, 1980]. Furthermore, we integrated in the rCDK [Guha, 2007] package features to access certain functionalities needed for generating elemental compositions. The rCDK package provides an interface to CDK for R users, making the MEF method available in the R statistics software, for example, for direct integration with XCMS.

3.5. Results & Discussion

The MEF method facilitates the analysis of MS^n data, which is not sufficiently explored in the mass spectrometry field yet. In a first series of experiments, we explore what mass accuracy is needed to resolve the elemental composition using MS^n data varying the set of chemical elements in the elemental composition and the MS^n level taken into account. Of our particular interest was to see to what extent the incorporation of the additional information in MS^n data, widens the ppm accuracy needed to uniquely identify the molecular formula corresponding to the measured mass.

To determine the needed accuracy value for obtaining a single elemental composition, the MEF method was repeatedly run with different values set as mass tolerance for all ions. The loop started with a rather large mass tolerance value of 180 ppm and we stopped decreasing it until the unique and correct elemental composition was found for the fragmented compound.

In the first experiment MS^n data of 5-hydroxy-lysine was analyzed. 5-hydroxy-lysine was chosen because it does fragment in multiple high mass fragments and spectra can be acquired till MS level 5. With XCMS software a peak list of 12 mass fragments was extracted. The MEF calculations were executed using four different sets of chemical elements: CHNO, CHNOS, CHNOSP and CHNOSPSi. We also included in the analyses Si, a heavier atom than C, N, P and S, to show the effect of the MEF tool of generating possible candidates when the atoms are more different in mass.

The results, see Figure 4, show that inclusion of additional MS^n levels in the extraction of the elemental composition has a high influence on the mass tolerance needed. For the fragments with low masses (found in the highest MS^n levels) a relative short list of candidate elemental compositions is generated with the consequence of putting stronger constraints on the precursor ions. As a consequence the needed mass tolerance value to end up with one unique elemental composition can be higher (less accurate mass data is needed). There is a direct relation between the number of nodes (fragment ions) in the MS tree and the number of edges (fragmentation reactions). The number of nodes in a MS tree depends on the depth of the tree (the MS^n level) and the width of the tree (fragment ions on a certain MS level). More edges in the MS tree lead to more dependencies between elemental formulas list, ultimately leading to stronger constraints and a less stronger need for high accurate MS data. Another important factor influencing the outcome using the MEF tool is the set of different chemical elements to be included in the elemental composition calculations. Figure 3.4 shows that a higher mass accuracy is needed to assign one unique chemical formula when more different chemical elements are taken into account.

In the second experiment mass spectral tree data was acquired and processed for twelve different metabolites, with masses between 150 and 450 Da, containing the following chemical elements: C, H, N, O, S and P. For all metabolites MS spectra were acquired up to MS level 5. Each metabolite fragments in a different way resulting in an unique fragmentation tree topology for each of them. Again the efficacy of the MEF tool was tested. For all metabolites the mass accuracy needed to determine the correct elemental composition of the parent ion was studied. Different metabolites with different molecular weights were compared while including different MS levels in the calculations. Figure 3.5 summarizes the results of this experiment. When information from more MS^n levels (depth of the tree) was included in the determination of the elemental composition all metabolites show that a less tight mass tolerance is needed. Again, taking advantage of the fact that more fragments are participating in the constraining process, the MS data has to be less accurate to determine the unique molecular formula. This approach shows that MS level is a relevant factor to help with the assignment of the elemental composition. From Figure 5 we can conclude that taking into account high MS levels allows us to set as a parameter higher mass tolerance error (or the instrument does need less mass accuracy) to determine the correct elemental composition using the MEF tool. To get one single elemental composition for the 12 different metabolites shown in Figure 5 on average MS^1 data of 1.83 ppm accuracy is needed while if MS^n data is available on average 35.58 ppm accuracy is needed. This effect is stronger for metabolites with a low molecular weight. Fragmentation of low weight metabolites result in fragments having relatively short elemental composition candidate lists ultimately leading to stronger constraints in the application of the MEF tool. The MEF tool accepts

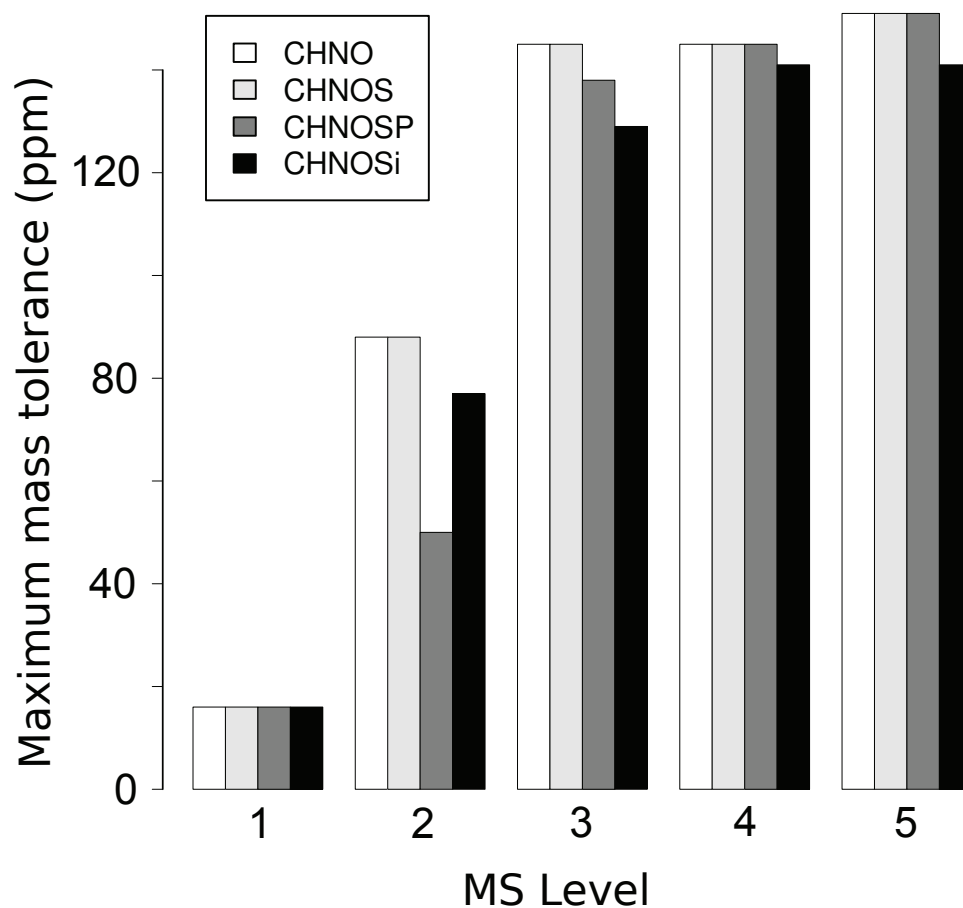


Figure 3.4: The mass tolerance needed to generate one unique elemental composition for 5-hydroxy-lysine. Different sets of chemical elements and MS levels were taken into account. The range limit of atoms for each element are between 0-50 for C, N, O, S, P, Si, and 0-100 for H.

as input mzXML files which makes the approach vendor and/or instrument independent. It can be applied both to high accurate instruments with limited fragmentation abilities (e.g. QTOF instruments limited to MS^2) and low accurate instruments using extensive MS levels (non-FTMS (ion trap) with MS^n).

The next exercise was executed to find out what the most informative edges in the fragmentation tree are for generating the molecular formula. Acquiring mass spectral tree data is time consuming and the aim of this exercise was to see which part of the tree gives the most relevant information for the MEF tool to ultimately guide data acquisition. As such, the results of this experiment would reduce the number of MS^n measurements to be taken. There are two approaches explored here to limit this number: First the acquisition of an MS^n tree can be tuned to generate deep or wide spectral trees depending on the experimental conditions set. The second acquisition approach is the selection of high or low ion masses for generation of the next fragment. To investigate both options, the tree topology template represented in Figure 3.6 was used. It consists of five nodes (representing the fragments), which can be extracted several times as subtree from all available fragmentation trees. Characteristic for this template is that the right side is formed by ions with high masses and the left side by ions with low masses. Using this template we wanted to see which scenario needs less mass tolerance error to generate the correct elemental composition. The three scenarios selected were: a) wide/parallel mode, b) linear/serial mode with high masses, and c) linear/serial mode with low masses. The δ values are the masses of the neutral losses calculated from the difference between the masses of the precursor ion and the fragment ion. The template on the left side of Figure 3.6 was found 133 times in all fragmentation trees from the twelve metabolites.

Figure 3.7 shows in a histogram how many times a certain scenario was able to find the correct elemental composition using less accurate MS^n data than the other scenarios. Scenario c) from Figure 3.6 clearly is the best data acquisition strategy to follow. In most cases it is best to follow the linear approach and to go as deep as possible in the fragmentation tree. Furthermore, when there are multiple fragments available to choose from the fragmentation tree that contains the lowest masses is to be preferred.

In another exercise scenario c was further investigated. Two situations were considered where the $\delta 1$ value was fixed and the $\delta 2$ value divided into a high and low mass situation. Like previously the best MEF results were obtained when the difference in mass between the precursor and fragment is as high as possible.

During the analysis of mass spectral tree data with the MEF tool we encountered typical situations that lead to no assigned chemical formula for certain ions. In many cases the selected ion turned out to be a false peak because of electrical or chemical noise or the mass tolerance applied to the ion was smaller than the experimental accuracy. When a valid

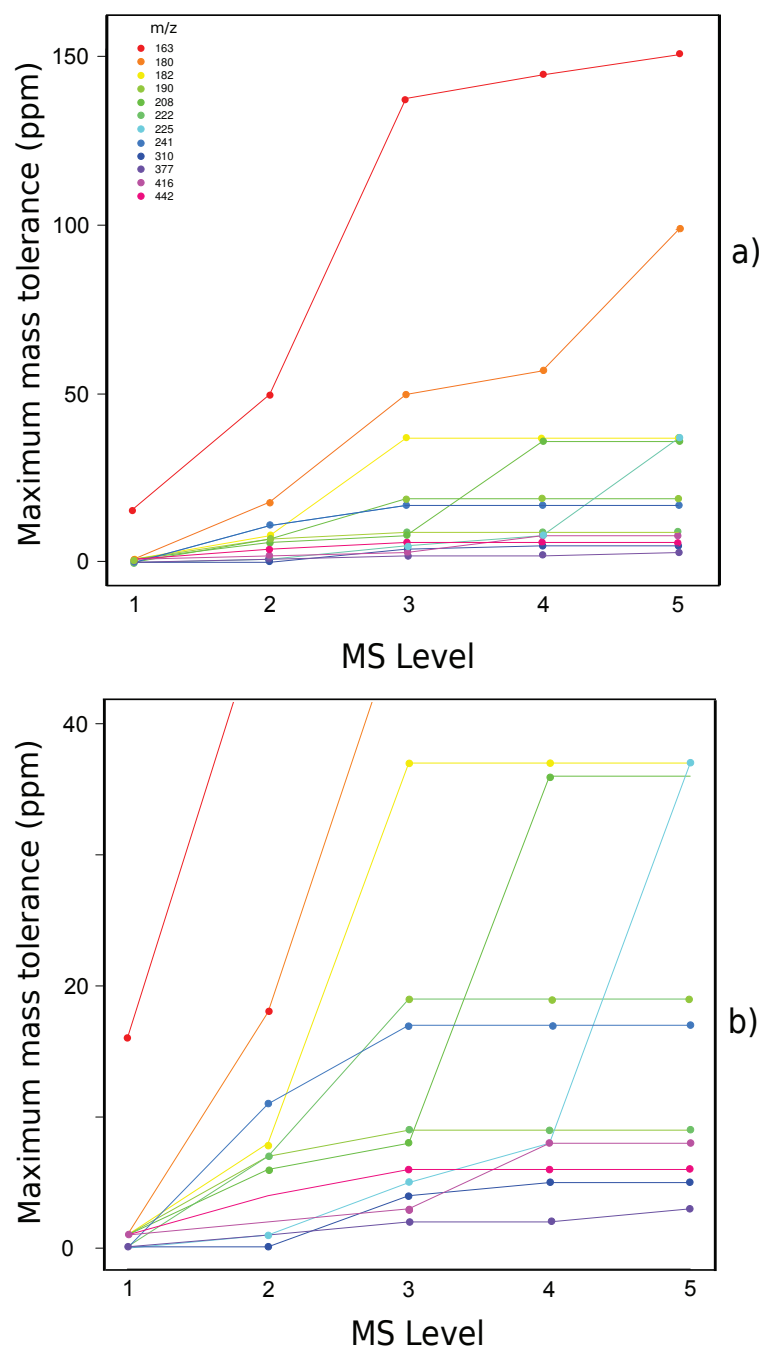


Figure 3.5: The mass tolerance (mass error window) needed to obtain one unique elemental composition for 12 different metabolites. Different MS levels are taken into account. Figure b) is an enlargement of figure a). All masses or m/z given are those of the $[M+H]^+$ ions.

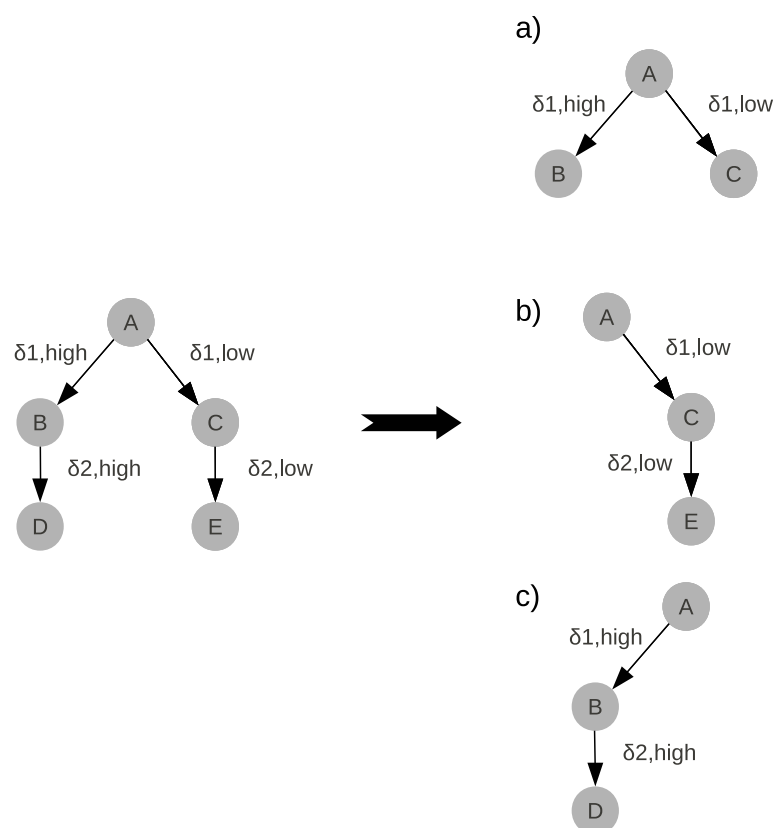


Figure 3.6: Tree topology template used to check what the most informative nodes and edges are in an acquired fragmentation tree. Three possible scenarios to study are extracted. They represent the acquisition of fragmentation trees with; a) wide/parallel mode, b) linear/serial model with high masses, and c) linear/serial mode with low masses. The ions are represented by a circle. The δ values are the masses of the neutral losses calculated from the difference between the mass of the precursor ion and the fragment ion. The circle C and E are the fragments with the highest mass while B and D are the fragments with the lowest mass.

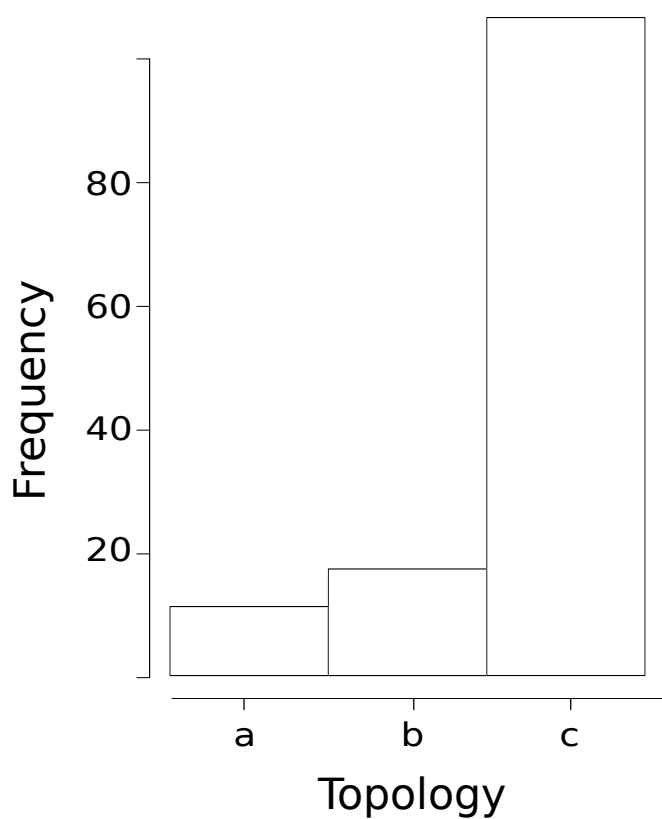


Figure 3.7: A histogram showing the distribution of the best topology between the scenario a), b), and c). The best topology is this that needs the lowest mass accuracy to generate one unique elemental composition.

elemental composition would be assigned to a false peak this ultimately could lead to a false elemental composition assignment of the top parent ion. Although we did not encounter this in the data we processed the next version of the MEF tool will contain features that allow the identification of these false peaks as well.

The runtime of the calculation does not only depend on the molecular mass of the compounds but also on the number of fragments and neutral losses that are available in the spectral tree. The aim of the development of the MEF tool is a proof of concept to demonstrate the power of MS^n to determine the elemental composition of the molecular ion, fragment ions and neutral losses rather than an efficiently performing algorithm. The runtime is between seconds and minutes.

3.6. Conclusion

Due to fast recent instrumentation developments, MS^n has become a powerful tool for the characterization of metabolites. A new method was developed that can be applied for the processing and analysis of multi-stage mass spectrometry (MS^n) data. This method resolves the chemical elemental composition of a compound using constraints extracted from the predicted elemental composition of its fragments and requires a lower mass accuracy than conventional methods. The viability of the Multi-stage Elemental Formula (MEF) method was tested with experiments on real MS^n data of several metabolites. The method does not only list the elemental composition of the parent ion but also of its fragments and the neutral losses. The results presented here show that the method assigns very efficiently the correct elemental composition while reducing the needed accuracy to middle mass tolerance depending on the chemical structure of the metabolite and the topology of the fragmentation tree analyzed. For 5-hydroxy-lisine the mass tolerance needed to solve the elemental composition jumps from 16 ppm to 150 ppm when additional fragmentation tree information is added as a constraint. This approach shows that the MS level is a relevant factor to help with the assignment of the elemental composition. If MS^n spectra are acquired to a higher level, the maximum mass tolerance to determine the correct elemental composition using the MEF tool is getting higher. To obtain the elemental composition of a protonated molecule (or adduct), this approach lowers the requirements with regards to mass accuracy of a mass spectrometer if MS^2 or higher MS level spectra can be obtained, and can be combined with the isotopic pattern of the protonated molecule (or adduct). This decreasing need for highly accurate data to solve the elemental composition by adding fragmentation tree information could help for those groups which can not effort an expensive instrument with powerful resolution power. Alternative, it will reduce the time needed to perform an identification. The output with the fragmentation pattern containing the el-

emental compositions is stored in a CML [Murray-Rust *et al.*, 2001] file format waiting for a proximate future for a standard exchange file format specific for metabolites. Currently, the MEF method provides a list of elemental composition candidates ordered according to the difference to the measured mass. In future work, it is planned to implement additional constraint rules into the method and also apply isotope abundance analysis to increase the identification accuracy. As we are able to characterize the fragments and neutral losses with the elemental composition, we are moving towards getting better understanding of the fragmentation patterns and use this information to identify the 'correct structure' of unknown compounds.

3.7. Acknowledgement

This study was financed by the research programme of the Netherlands Metabolomics Centre (NMC) which is a part of The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. We thank the editor and the anonymous reviewers for their helpful comments.

3.8. Supplementary Data

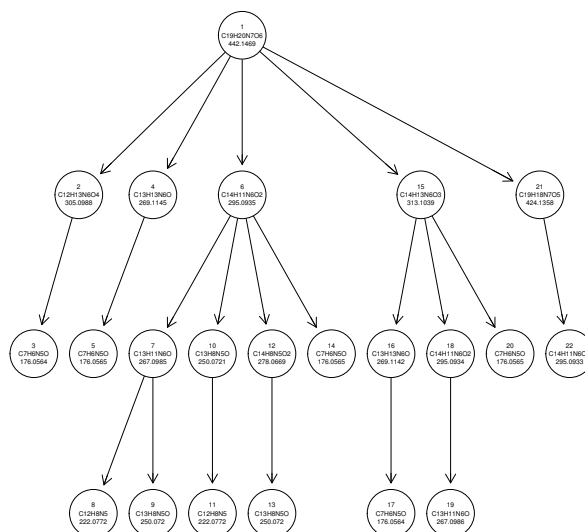


Figure 3.8: Compound InChI=1/C19H19N7O6/c20-19-25-15-14(17(30)26-19)23-11(8-22-15)7-21-10-3-1-9(2-4-10)16(29)24-12(18(31)32)5-6-13(27)28/h1-4,8,12,21H,5-7H2,(H,24,29)(H,27,28)(H,31,32)(H3,20,22,25,26,30)/t12-/m0/s1

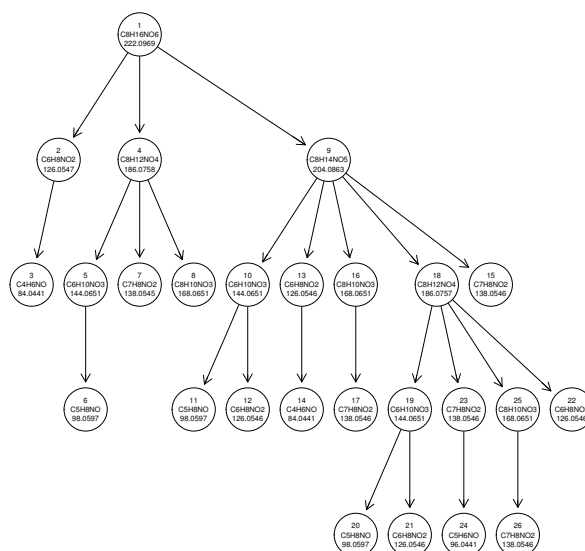


Figure 3.9: Compound InChI=1/C8H15NO6/c1-3(11)9-5-7(13)6(12)4(2-10)15-8(5)14/h4-8,10,12-14H,2H2,1H3,(H,9,11)/t4-,5-,6+,7-,8+/m1/s1

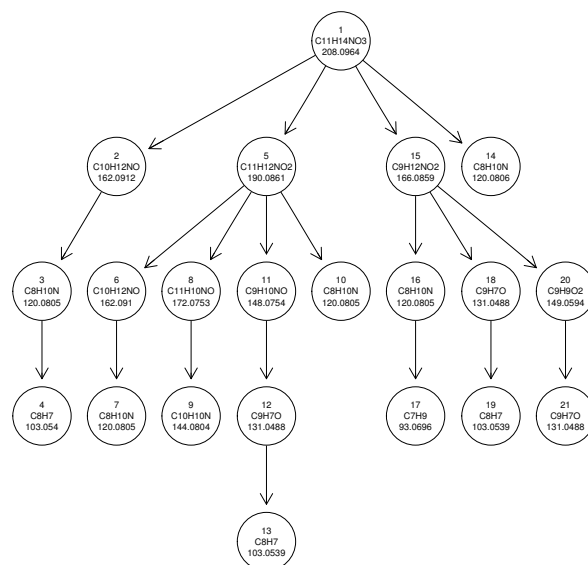


Figure 3.10: Compound InChI=1/C11H13NO3/c1-8(13)12-10(11(14)15)7-9-5-3-2-4-6-9/h2-6,10H,7H2,1H3,(H,12,13)(H,14,15)/t10-/m0/s1

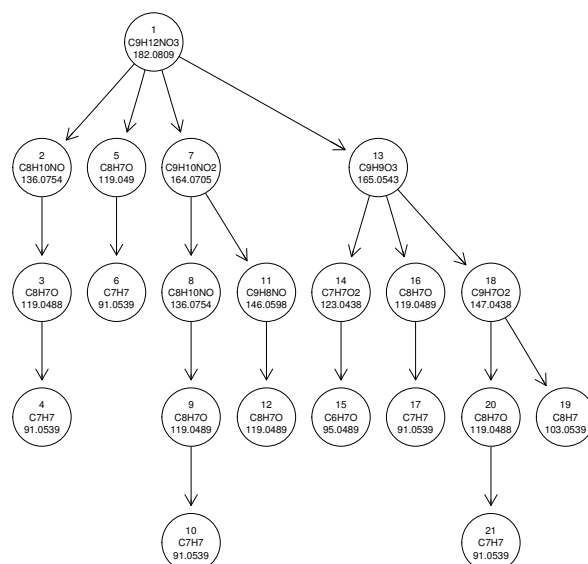


Figure 3.11: Compound InChI=1/C9H11NO3/c10-7(9(12)13)5-6-3-1-2-4-8(6)11/h1-4,7,11H,5,10H2,(H,12,13)

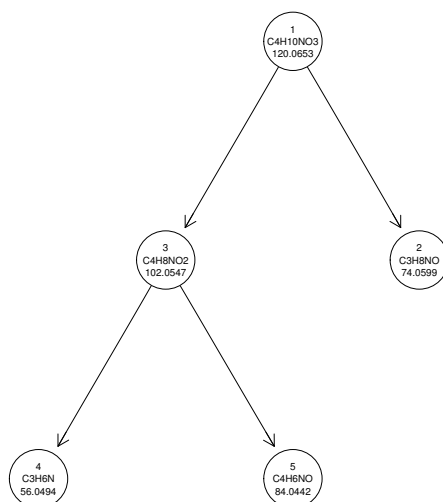


Figure 3.12: Compound InChI=1/C4H9NO3/c1-2(6)3(5)4(7)8/h2-3,6H,5H2,1H3,(H,7,8)/t2-,3+/m1/s1

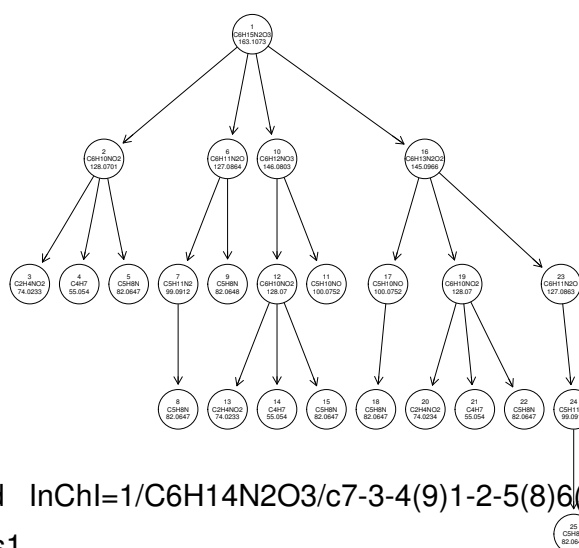


Figure 3.13: Compound InChI=1/C6H14N2O3/c7-3-4(9)1-2-5(8)6(10)11/h4-5,9H,1-3,7-8H2,(H,10,11)/t4-,5+/m1/s1

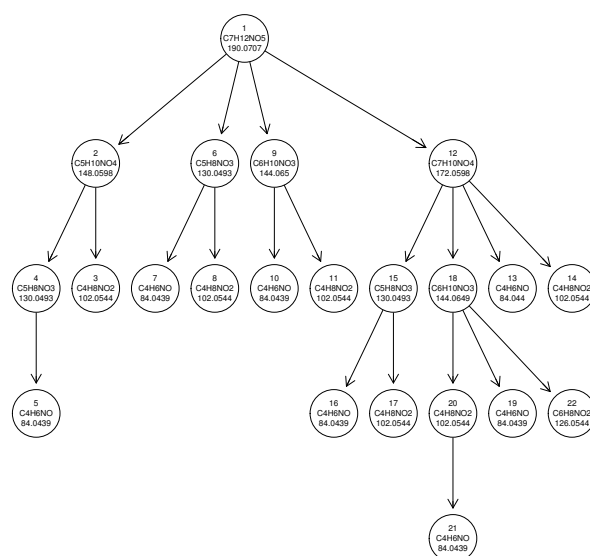


Figure 3.14: Compound InChI=1/C7H11NO5/c1-4(9)8-5(7(12)13)2-3-6(10)11/h5H,2-3H2,1H3,(H,8,9)(H,10,11)(H,12,13)

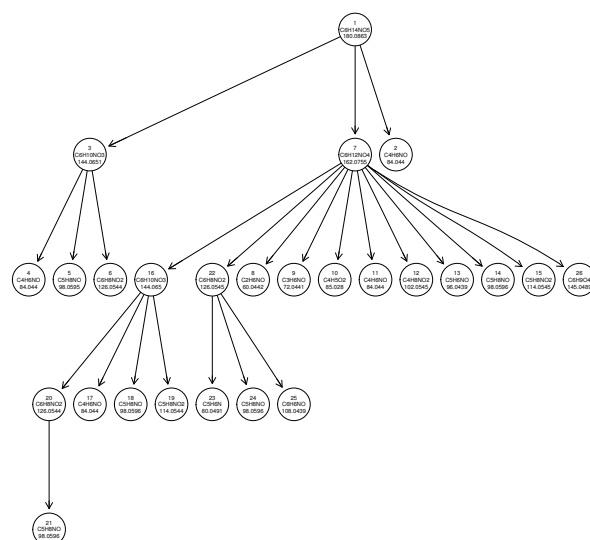


Figure 3.15: Compound InChI=1/C6H13NO5/c7-3-5(10)4(9)2(1-8)12-6(3)11/h2-6,8-11H,1,7H2/t2-,3-,4-,5-,6-/m1/s1

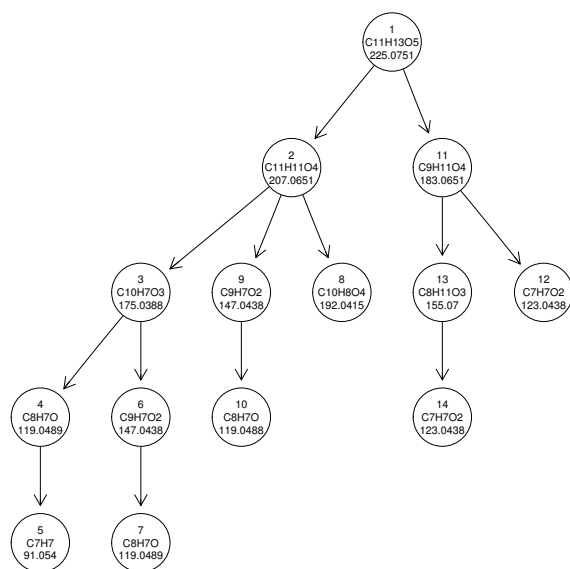


Figure 3.16: Compound InChI=1/C11H12O5/c1-15-8-5-7(3-4-10(12)13)6-9(16-2)11(8)14/h3-6,14H,1-2H3,(H,12,13)

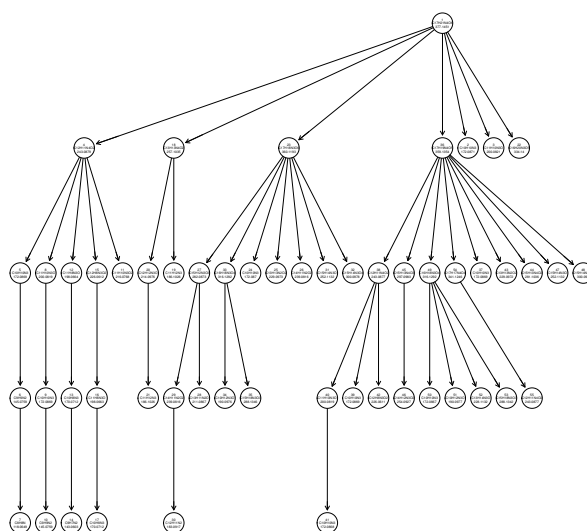


Figure 3.17: Compound InChI=1/C17H20N4O6/c1-7-3-9-10(4-8(7)2)21(5-11(23)14(25)12(24)6-22)15-13(18-9)16(26)20-17(27)19-15/h3-4,11-12,14,22-25H,5-6H2,1-2H3,(H,20,26,27)/t11-,12+,14-/m0/s1

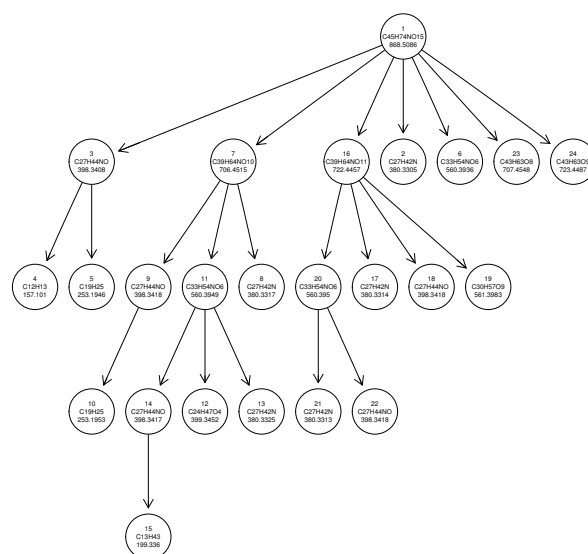


Figure 3.18: Compound InChI=1/C45H73NO15/c1-19-6-9-27-20(2)31-28(46(27)16-19)15-26-24-8-7-22-14-23(10-12-44(22,4)25(24)11-13-45(26,31)5)57-43-40(61-41-37(54)35(52)32(49)21(3)56-41)39(34(51)30(18-48)59-43)60-42-38(55)36(53)33(50)29(17-47)58-42/h7,19-21,23-43,47-55H,6,8-18H2,1-5H3/t19-,20+,21-,23-,24+,25-,26-,27-,28-,29+,30+,31-,32-,33+,34-,35+,36-,37+,38+,39-,40+,41-,42-,43+,44-,45-/m0/s1

Bibliography

- [Bode, 2004] Bode, J. W. (2004) Reactor ChemAxon Ltd. *Journal of the American Chemical Society*, **126** (46), 15317–15317.
- [Butcher *et al.*, 2004] Butcher, E. C., Berg, E. L. & Kunkel, E. J. (2004) Systems biology in drug discovery. *Nature biotechnology*, **22** (10), 1253–9.
- [Cui *et al.*, 2000] Cui, M., Song, F., Zhou, Y., Liu, Z. & Liu, S. (2000) Rapid identification of saponins in plant extracts by electrospray ionization multi-stage tandem mass spectrometry and liquid chromatography/tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, **14** (14), 1280–1286.
- [Dayringer & McLafferty, 1977] Dayringer, H. E. & McLafferty, F. W. (1977) Computer-aided interpretation of mass spectra. STIRS prediction of rings-plus-double-bonds values. *Organic Mass Spectrometry*, **12** (1), 53–54.
- [Dromey & Foyster, 1980] Dromey, R. G. & Foyster, G. T. (1980) Calculation of elemental compositions from high resolution mass spectral data. *Analytical Chemistry*, **52** (3), 394–398.
- [Dunn, 2008] Dunn, W. B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical biology*, **5** (1), 011001.
- [Dunn & Ellis, 2005] Dunn, W. B. & Ellis, D. I. (2005) Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, **24** (4), 285–294.
- [Erve *et al.*, 2009] Erve, J. C. L., Gu, M., Wang, Y., DeMaio, W. & Talaat, R. E. (2009) Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications

- for elemental composition determination. *Journal of the American Society for Mass Spectrometry*, **20** (11), 2058–69.
- [Gu *et al.*, 2006] Gu, M., Wang, Y., Zhao, X.-G. & Gu, Z.-M. (2006) Accurate mass filtering of ion chromatograms for metabolite identification using a unit mass resolution liquid chromatography/mass spectrometry system. *Rapid communications in mass spectrometry : RCM*, **20** (5), 764–70.
- [Guha, 2007] Guha, R. (2007) Chemical Informatics Functionality in R. *Journal of Statistical Software*, **18** (5), 1 – 16.
- [Guha *et al.*, 2006] Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. & Willighagen, E. L. (2006) The Blue Obelisk-interopability in chemical informatics. *Journal of chemical information and modeling*, **46** (3), 991–8.
- [Holliday *et al.*, 2006] Holliday, G. L., Murray-Rust, P. & Rzepa, H. S. (2006) Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. *Journal of chemical information and modeling*, **46** (1), 145–57.
- [Jarussophon *et al.*, 2009] Jarussophon, S., Acoca, S., Gao, J.-M., Deprez, C., Kiyota, T., Draghici, C., Purisima, E. & Konishi, Y. (2009) Automated molecular formula determination by tandem mass spectrometry (MS/MS). *The Analyst*, **134** (4), 690–700.
- [Kim *et al.*, 2006] Kim, S., Rodgers, R. P. & Marshall, A. G. (2006) Truly 'exact' mass: Elemental composition can be determined uniquely from molecular mass measurement at ~ 0.1 mDa accuracy for molecules up to ~ 500 Da. *Science*, **251**, 260–265.
- [Kind & Fiehn, 2006] Kind, T. & Fiehn, O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, **7**, 234.
- [Kind & Fiehn, 2007] Kind, T. & Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, **8**, 105.
- [Kind & Fiehn, 2010] Kind, T. & Fiehn, O. (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews*, **2** (1-4), 23–60.
- [Konishi *et al.*, 2007] Konishi, Y., Kiyota, T., Draghici, C., Gao, J.-M., Yeboah, F., Acoca, S., Jarussophon, S. & Purisima, E. (2007) Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products. *Analytical chemistry*, **79** (3), 1187–97.
-

- [Murray-Rust *et al.*, 2001] Murray-Rust, P., Rzepa, H. S. & Wright, M. (2001) Development of chemical markup language (CML) as a system for handling complex chemical content. *New Journal of Chemistry*, **25** (4), 618–634.
- [Pedrioli *et al.*, 2004] 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. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, **22** (11), 1459–66.
- [Rasche *et al.*, 2011] Rasche, F., SvatosìĚ, A., Maddula, R. R. K., BoìŁttcher, C. & BoìŁcker, S. (2011) Computing Fragmentation Trees from Tandem Mass Spectrometry Data. *Analytical Chemistry*, **83** (4), 1243–1251.
- [Senior, 1951] Senior, J. K. (1951) Partitions and Their Representative Graphs. *American Journal of Mathematics*, **73** (3), 663.
- [Smith *et al.*, 2006] Smith, C. A., O'Maille, G., Want, E. J., Abagyan, R. & Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using non-linear peak alignment, matching, and identification. *Analytical chemistry*, **78** (3), 779–87.
- [Steinbeck *et al.*, 2003] Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Wilhagen, E. (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, **43** (2), 493–500.
- [Stoll *et al.*, 2006] Stoll, N., Schmidt, E. & Thurow, K. (2006) Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization FTMS. *Journal of the American Society for Mass Spectrometry*, **17** (12), 1692–9.
- [Tyrkkö *et al.*, 2010] Tyrkkö, E., Pelander, A. & Ojanperä, I. (2010) Differentiation of structural isomers in a target drug database by LC/Q-TOFMS using fragmentation prediction. *Drug testing and analysis*, **2** (6), 259–270.
- [Zhang *et al.*, 2005] Zhang, J., Gao, W., Cai, J., He, S., Zeng, R. & Chen, R. (2005) Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **2** (3), 217–30.
-

CHAPTER 4

Metabolomics identification using MS^n data: a new similarity approach for comparing mass spectral trees

Miguel Rojas-Chertó^{1,2}, Julio E. Peironcely^{1,2,3}, Piotr T. Kasper^{1,2}, J.J.J. van der Hooft^{1,4,5}, R.C.H. de Vos^{1,4,6}, Rob Vreeken^{1,2}, Thomas Hankemeier^{1,2} and Theo Reijmers^{1,2}

Analytical Chemistry 2012;84(13):5524-34

¹Netherlands Metabolomics Centre, Leiden, The Netherlands

²Division of Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands

³TNO, Zeist, The Netherlands

⁴Laboratory of Biochemistry, Wageningen University

⁵Plant Research International, Wageningen University and Research Centre

⁶Centre for Biosystems Genomics

4.1. Abstract

Multi-stage mass spectrometry (MS^n) generating so-called spectral trees is a powerful tool in the annotation and structural elucidation of metabolites and is increasingly used in the area of accurate mass LC/MS-based metabolomics to identify unknown, but biologically relevant compounds. As a consequence, there is a growing need for computational tools specifically designed for the processing and interpretation of MS^n data. Here, we present a novel approach to represent and calculate the similarity between high mass resolution mass spectral fragmentation trees. This approach can be used to query multiple stage mass spectra in MS spectral libraries. Additionally the method can be used to calculate structure-spectra correlations and potentially deduce substructures from spectra of unknown compounds. The approach was tested using two different spectral libraries composed, of either human or plant metabolites, which currently contain 872 MS^n spectra acquired from 549 metabolites using Orbitrap FTMSⁿ. For validation purposes for 282 of these 549 metabolites additional replicate 765 MS^n spectra acquired with the same instrument were used. Both the dereplication and de-novo identification functionalities of the comparison approach are discussed. This novel MS^n spectral processing and comparison approach increases the probability to assign the correct identity to an experimentally obtained fragmentation tree. Ultimately, this tool may pave the way for constructing and populating large MS^n spectral libraries that can be used for searching and matching experimental MS^n spectra for annotation and structural elucidation of unknown metabolites detected in untargeted metabolomics studies.

4.2. Introduction

Metabolomics emerges from the need to study and understand the function of the genes through their end products, so-called metabolites. Over the last years mass spectrometry (MS) has proven itself as a powerful technology for the detection and annotation of compounds and became important for analyzing the metabolome of any organism [Kind & Fiehn, 2010]. Depending on the nature of the sample and the information scientists want to extract from it, mostly two different MS ionization techniques are used, i.e. hard and soft ionization. Both ionization techniques can be used separately [Hernández *et al.*, 2011, Grange *et al.*, 2002] as well as combined [Portolés *et al.*, 2011]. Hard ionization methods, such as electron impact ionization (EI), deal with a high energy source generating ions and an extensive range of fragmentation products from the molecule. Due to the high energy, the fragmentation spectra obtained are highly uniform between instruments and can therefore be used for creating universal databases such as

the National Institute of Standards and Technology (NIST) mass spectral library, which is used worldwide in GC-MS studies for spectra matching. In contrast, soft ionization methods, like matrix assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), treat the molecule gentler, aiming to generate quasi-molecular ions from the intact molecule. While with hard ionization methods each compound is characterized by a multitude of mass fragments, with soft ionization only a couple of mass signals are produced from each compound. The advantage of using soft ionization is its higher sensitivity in LC-MS analyses and the feasibility to conduct stepwise fragmentation of the quasi-molecular ion by collision induced dissociation (CID) or collisionally activated dissociation (CAD). At the end, either a tandem mass spectrum (MS/MS or MS²) or a hierarchical mass spectrum, also called multi-stage mass spectra (MSⁿ), is created [Sheldon *et al.*, 2009, Hooft *et al.*, 2011, van der Hooft *et al.*, 2011]. Unfortunately, MS/MS data are generally not as reproducible between labs as EI data [Bristow *et al.*, 2004, Jansen *et al.*, 2005]. However, new advances have been proposed leading to more reproducible MS/MS data. Examples are application of a tuning point protocol to standardize CID conditions prior to data acquisition [Hopley *et al.*, 2008, Champarnaud & Hopley, 2011] or usage of a fragmentation energy index for LC-MS to normalize collision energies [Palit & Mallard, 2009]. These new developments encourage researchers to start creating MS/MS spectral libraries [Horai *et al.*, 2010, Akiyama *et al.*, 2008, Smith *et al.*, 2006]. Compared to MS/MS (i.e. MS²), ion trap-based MSⁿ generates more specific and detailed information about the relation between the product ion and its direct fragments and the fragments derived from these fragments, including the spectral hierarchy. This sequential fragmentation approach increases the ability of mass spectrometrists, to structurally characterize and identify unknown metabolites detected in untargeted LC/MS-based metabolomics approaches [Wolfender *et al.*, 2000, Sheldon *et al.*, 2009, Hooft *et al.*, 2011, van der Hooft *et al.*, 2011, Scheubert *et al.*, 2011]. In addition, the relative intensities of the MSⁿ fragment ions generated, and thus the fragmentation spectra, are highly reproducible between different experiments and hardly influenced by small changes in instruments settings [Sheldon *et al.*, 2009, Hooft *et al.*, 2011, van der Hooft *et al.*, 2011]. This reproducibility indicates that there is a good potential for MSⁿ spectral tree approaches to create and search fragmentation tree libraries, like searching EI-spectra in the NIST library. The applicability for library searching has at least been shown for nominal mass MSⁿ data generated on the same instrument [Sheldon *et al.*, 2009]. By coupling Ion Trap MSⁿ fragmentation to accurate mass read-out of the fragments generated, e.g. using the LTQ-Orbitrap FTMS hybrid MS system (Thermo), elemental composition of fragments can be readily obtained, which can further help in structural elucidation of unknown compounds. However, for optimal use and implementation in metabolomics studies, the accurate mass MSⁿ spectral tree ap-

proach still lags behind, as compared to EI-spectra matching, on three main points: I) fragmentation spectra representation, II) spectra storage and III) comparison and matching of spectra. Current software to handle MSⁿ data is either commercial and not flexible enough to do dedicated follow-up data processing, e.g. MassFrontierTM (ThermoFinnigan), or not specifically developed to process MSⁿ data, e.g. XCMS [Smith *et al.*, 2006]. Appropriate processing of MSⁿ data is crucial for obtaining robust data to be stored in a reference fragmentation tree database. Recently our group developed a new freely available tool called MEF (Mass Elemental Formula) [Rojas-Chertó *et al.*, 2011] which processes and enriches high mass resolution MSⁿ data and generates fragmentation trees. The MEF tool extracts the most relevant signals from the MS spectra (representing the ions/fragments) and enriches these with the assignment of an elemental composition. Additionally, MEF generates elemental compositions for the neutral losses. To facilitate further analysis the MEF tool can export the resulting information to other formats (Chemical Markup Languages (CML) [Murray-Rust *et al.*, 2001, Holliday *et al.*, 2006, Kuhn *et al.*, 2007], portable document format (pdf) or into Comma Separated Value (CSV)). In this paper we define a fragmentation tree as a hierarchical organization of fragment ions describing the fragmentation reactions between them where the nodes refer to the fragments (either represented by their nominal mass (NM) values or elemental compositions (EC)) and the edges refer to the fragmentation reactions [Rasche *et al.*, 2011]. As soon as a library has been created, an algorithm to automatically query that library, to find similar MS spectral data, is needed. Currently, several search algorithms for comparing soft ionization MS/MS spectra of small molecules exist [Oberacher *et al.*, 2009, Wolf *et al.*, 2010, Rasche *et al.*, 2011], mostly differing in the way how the MS/MS spectra are represented and how the similarities are calculated. The main search concepts are based on the spectral-contrast-angle method [Wan *et al.*, 2002], the probability based matching (PBM) algorithm [McLafferty *et al.*, 1998], or the dot-product algorithm search [Stein & Scott, 1994]. All of them have been applied first to EI data and more recently they have also been introduced for the analyses of soft ionization MS/MS data analysis [Hansen & Smedsgaard, 2004]. Approaches capable to calculate the similarity between MSⁿ data are the recently published work of Rasche [Rasche *et al.*, 2012] which compares hypothetical fragmentation trees, and the commercial software tool MassFrontierTM (Thermo) which is based on the dot-product function.

Most algorithms mentioned above represent MS spectra by a set of equidistant bins, where each ion is encapsulated into one specific bin according to its mass-to-charge ratio (m/z). When two spectra fill up the same bins they are considered to be identical. The disadvantage of this representation is that only single spectra can be compared with each other and not the complete spectral tree, and, furthermore, that the maximum number of bins for representing a spectrum (or the bin width) is limited by the mass resolution of the data. In

the field of chemical similarity searching, in which structure databases are queried, other approaches are used. These approaches look for certain substructures being present in the structures and are called fingerprint-based algorithms. As a consequence, these algorithms are not dependent on the mass accuracy of the fragments but on the presence or absence of certain relations between the fragments and therefore more suitable for handling MS^n spectral trees generated at high mass resolution. The more commonly used algorithms in this field make use of the Tanimoto (or Jaccard), cosine or Dice coefficients, the Euclidean or Hamming distance [Willett *et al.*, 1998] or modifications of these coefficients or distances [Fligner *et al.*, 2002]. Many different studies have compared the performances of these similarity measures [Baldi & Nasr, 2010]. The Tanimoto coefficient is the most widely used coefficient for similarity-based querying, because it is easy-to-use and computationally efficient.

In this paper a new cheminformatics approach enabling the comparison of high mass resolution MS^n data and spectral trees is presented. Contrary to common comparison algorithms, in which MS data is represented as counts of certain m/z values, our method is based on binary features of specific combinations of fragments and neutral losses being present or not present in the fragmentation trees. The degree of similarity between MS^n data is calculated using the Tanimoto coefficient. By means of two different MS^n libraries, i.e. compounds from plant and from human origins, the principle and the performance of the new method are shown. Moreover, the potential of this approach to elucidate substructures of unknown compounds is demonstrated.

4.3. Material and Methods

4.3.1. Metabolite MS^n Libraries

In this work two different in-house libraries containing multi-stage mass spectra (MS^n) data have been used. The first library was created at the Division of Analytical Biosciences, Leiden University, Leiden, The Netherlands, and contains 705 MS^n spectra from 447 different human metabolites. The second library was created at Plant Research International, Wageningen-UR, Wageningen, The Netherlands, and contains 167 MS^n spectra of 118 plant metabolites belonging to the class of polyphenols. This plant library includes different series of isomers in which hydroxyl, glycosyl or methoxy groups are attached to different positions of the core flavonoid structure [Hooft *et al.*, 2011]. More information about these libraries can be found in Table 1. The main difference between the two libraries was the diversity level of the molecules (see Supplemental Text 1 in the Supporting Information). The complete set of spectral trees from both libraries were used for analyzing the perfor-

Library	Comp.	Mass range (Da)	Median (Da)	Average (Da)	Spectra	Diversity level
Human	447	59.0-1525.6	196.1	256.4	705	0.82
Plant	118	172.1-792.3	293.0	347.4	167	0.48

Table 4.1: Data Sets Used in This Study.

mance of the mass spectral tree comparison method presented here. Both libraries were generated on LTQ-Orbitrap FTMS XL instruments (Thermo Electron Corp.) using a Nano-mate injection robot (advion) with chip-based ESI nanospray. The MS^n experiments were run in both positive and negative ionization modes using a data-dependent scanning function, limited to 15 minutes acquisition time, with the criteria to select the five most intense ions detected for MS^2 and MS^3 , and the three most intense ions for the rest of the MS levels [van der Hooft *et al.*, 2011]. The data were generated in centroid mode with a FWHM resolution of 60.000.

4.3.2. Processing of MS^n Fragmentation Trees

Raw data were converted to mzXML format [Pedrioli *et al.*, 2004] using ReadW software which was provided by the Institute for Systems Biology, Seattle, United States. All MS^n data of the reference compounds in the two selected libraries were processed with an extended version of the so-called MEF tool [Rojas-Chertó *et al.*, 2011].

The parameters used to process the MS^n data are described in Supplemental Text 3 in the Supporting Information. Because the depth of the fragmentation tree, i.e. the number of fragments present, highly depends on the concentration of the compound under investigation [Hooft *et al.*, 2011], the direct comparison of MS^n spectra generated from compounds in biological samples, which are often present at low concentrations, with those stored in the library, generated at relative high concentrations, can be difficult. To cope with such differential concentrations, all MS^n spectra in the reference databases were processed at four different values for the relative intensity threshold parameter in the MEF tool. The relative intensity is defined as the intensity relative to the base peak, i.e. the most intense fragment peak, in each spectrum. Signals that had a relative intensity lower than the relative intensity threshold were not considered in the further processing. The values for the relative intensity threshold were set to 0% (default setting of the MEF tool, i.e. all signals taken into account), 5% (signals below 5% of the base peak were omitted), 10% and 20%. In this way for each reference compound a series of MS^n spectra representing a theoretical dilution series was generated in silico. Since each raw data file may contain several fragmentation

trees of the same compound, i.e. repetitions, peaks that did not appear in at least 40% of the repetitions were considered as irreproducible and were therefore omitted.

4.3.3. MS^n Fragmentation Tree Representation

The efficiency of comparing fragmentation trees in a MS^n library depends directly on the way how fragmentation trees are represented and what similarity measure is applied. Because both a fragmentation tree and a chemical structure can be represented as a graph, i.e. they both contain nodes (fragments or atoms) and edges (fragmentation reactions or bonds), the similarity measures used for molecules can also be applied to fragmentation trees. A binary fingerprint is a commonly used representation for molecules in which the presence or absence of predetermined substructural features are indicated with ones or zeros, respectively [Leach & Gillet, 2007]. In the present research this fingerprint-based representation for structures was extended to MS^n fragmentation trees. The fragmentation tree is represented by a series of zeros and ones in a linear bitmap, where each bit in the fingerprint is related to the absence or presence, respectively, of a particular feature of the fragmentation tree. Different types of features were defined in accordance with the different ways nodes were connected. All features used are shown in Supplemental Figure 1 in the Supporting Information. Next to generating a fragmentation tree, the MEF tool also extracts a neutral loss tree from raw MS^n data files. Such a tree is a hierarchical organization of the neutral losses in a graphical form. Also for these neutral losses binary fingerprints were generated and concatenated to the corresponding fragmentation tree fingerprints.

4.3.4. MS^n Fragmentation Tree Similarity Measures

For our purpose we applied the Tanimoto coefficient as a similarity measure to enable calculating the degree of similarity between fragmentation trees. Because the Tanimoto coefficient is molecule size dependent [Fligner *et al.*, 2002] and the equation has an inherent bias towards certain similarity values [Holliday *et al.*, 2003], we applied a prefiltering of the two fingerprints that are going to be compared by omitting large dissimilar features that would give rise to a series of smaller, likewise dissimilar features (see Supplemental Text 2 in the Supporting Information).

Next to having a quantitative measure describing how similar fragmentation trees are, we also visualize this similarity by showing which nodes in the compared trees overlap. For this, the concept of maximal common subgraph (MCS) was used. MCS is defined as the largest possible subgraph that two objects (structures of fragmentation trees) share in common. By calculating the MCS of the structures of the reference compounds in the database with the highest fragmentation tree similarity values, we obtained struc-

turally relevant information on substructure level for the unknown compound that is queried. In our study we generated the maximum common substructures (MCSS) for multiple molecules for a given list of InChI [Coles *et al.*, 2005] identifiers by using the CDK library [Steinbeck *et al.*, 2003, Steinbeck *et al.*, 2006].

4.4. Results and Discussions

The method developed and described in this paper is used for comparison of fragmentation trees and is based on extracting the fingerprints of both trees and then calculating their similarity using the Tanimoto coefficient. The fingerprint is build-up of representative features of the fragmentation tree, which forms different combinations between the nodes and the edges. Two different types of fragmentation trees, differing in the way how the nodes are represented, were considered in this study. One type was constituted with nodes enriched with nominal mass information, called nominal mass fragmentation tree (NMFT), while the other type consisted of nodes enriched with accurate mass-derived elemental composition information, called elemental composition fragmentation tree (ECFT). The NMFT is generated by using only the peak detection functionality of the MEF tool. The ECFT type is obtained using the fully functional MEF tool, including combinatorial rules to calculate elemental compositions. The NMFT may be useful as an alternative to the ECFT in case the mass accuracy of the data does not allow elemental composition calculation of the parent ion and its fragments. Obviously, the NMFT contains much more nodes/fragments than the ECFT generated from the same MS^n data file, because for generating a NMFT no additional chemical constraining is used, in contrast to generating a ECFT. Thus, a major drawback of the nominal mass fragmentation tree is the less precise representation of its nodes because a single nominal mass can still point to many different elemental compositions and thus possible fragments. Also, in an elemental composition fragmentation tree, chemical constraints checking elemental composition consistency are applied. As a consequence, peaks not fitting the elemental compositions of the precursor and child ions are eliminated and the fragments are more precisely represented. (See 4.17 and Supplemental Figures 5-10 in the Supporting Information, for examples of fragmentation trees showing the difference between NMFT and ECFT representations).

To identify which type of fragmentation tree representation (ECFT or NMFT) is more suitable to compare fragmentation trees using the implemented fingerprint-based search algorithm, the following experiment was carried out. Two different metabolites were analyzed at four different labs, but on a similar MS instrument and using the same acquisition

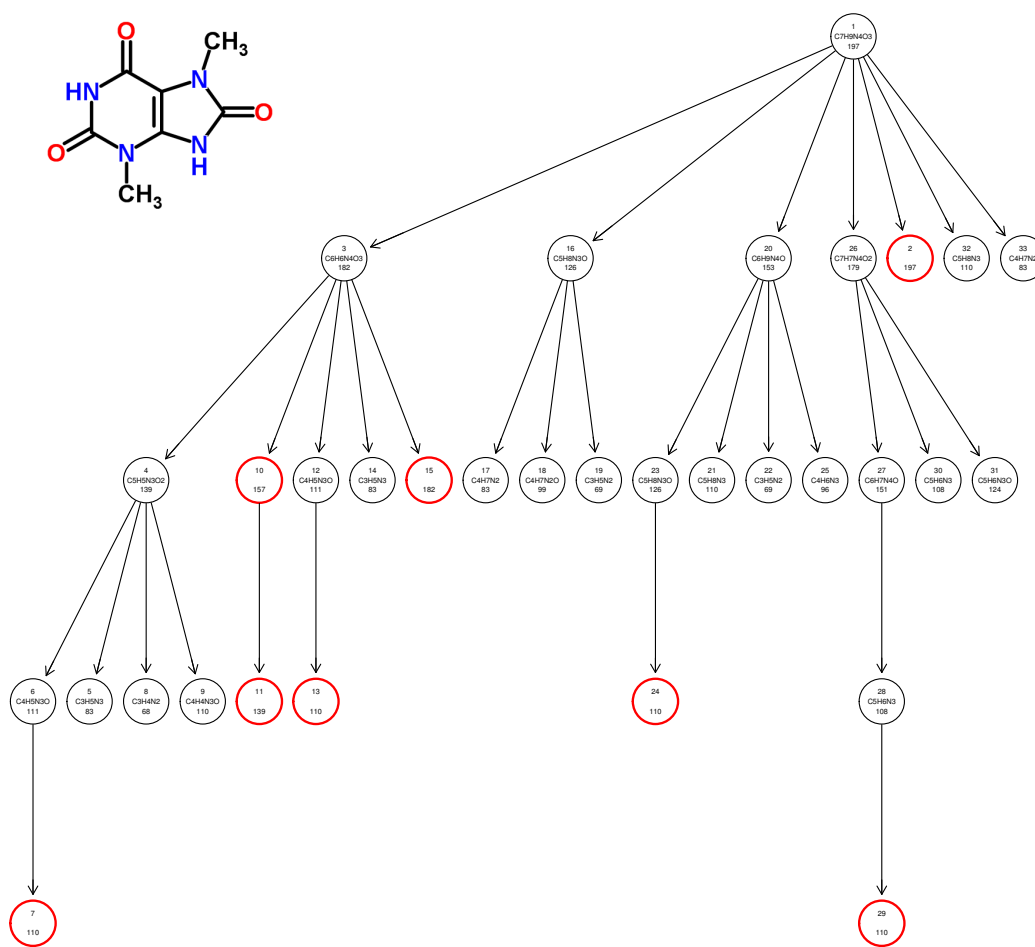


Figure 4.1: Fragmentation tree of 3,7-Dimethyluric acid [InChI=1/C7H8N4O3/c1-10-3-4(8-6(10)13)11(2)7(14)9-5(3)12/h1-2H3,(H,8,13)(H,9,12,14)]. The nodes for which the elemental composition is not calculated are drawn in red.

protocol and data processing tools. Afterwards, both the elemental composition fragmentation tree (ECFT) and the nominal mass fragmentation tree (NMFT) were generated from each acquired spectral tree. 4.2 shows the number of fragments generated for each type of fragmentation tree for each lab and the degree of similarity of the fragmentation trees between the different labs. For all labs, the NMFTs contained more nodes/fragments than the ECFTs (4.2a). This higher number of NMFT features was mainly due to peaks not related to the fragmentation pattern and peaks that were characteristic for a specific lab. As a consequence, the similarity values calculated from ECFTs are consistently higher than those calculated from NMFT (4.2b). We therefore concluded that MS^n spectra generated using nominal masses are less efficient in comparing fragmentation trees, and thus database searching, than MS^n spectra based on accurate masses enabling elemental compositions calculations.

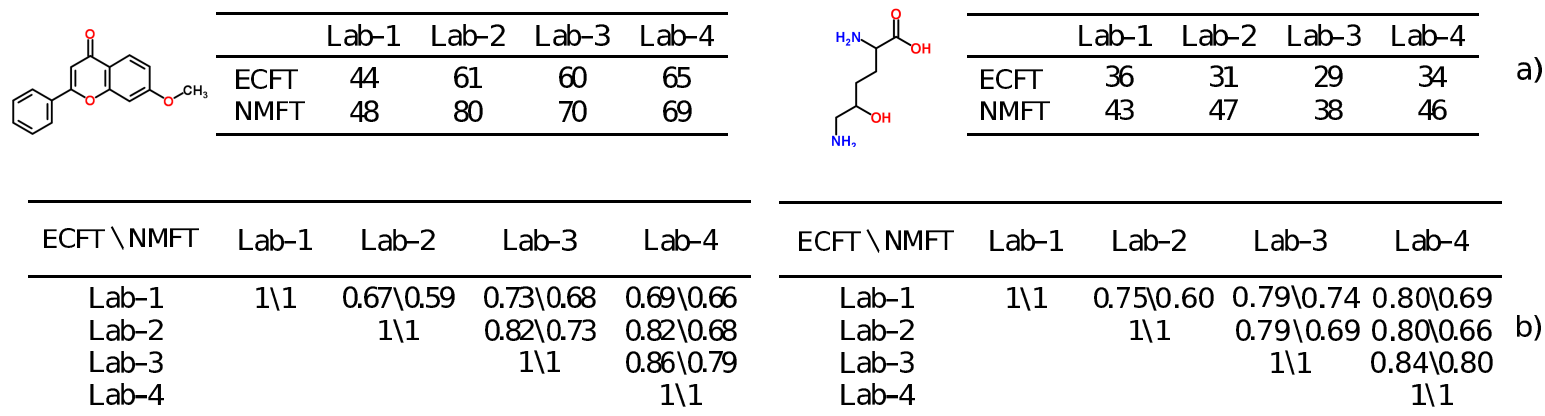


Figure 4.2: a) the number of fragments in the elemental composition (ECFT) and nominal mass (NMFT) fragmentation trees for the different labs. b) the similarity value calculated using the fingerprints generated from elemental composition fragments (ECFTs in front of slash) versus nominal mass fragments (NMFTs behind the slash). The compounds analyzed are 7-methoxy-2-phenyl-4H-chromen-4-one (left) and 5-hydroxylysine (right).

In practice, when measuring biological samples, the experimental fragmentation tree topology of a certain metabolite will be different, to a more or lesser extent, from the one stored in the reference library, because the number of acquired fragments of a certain compound may change between experiments, e.g. depending upon specific instrument sensitivity. However, the most relevant factor that influences the size (depth and width) of the fragmentation tree, i.e. the number of fragments per level, will be the concentration of the sample measured: the higher the concentration of the sample, the more molecular ions will be trapped and the larger the number of fragments generated, and therefore the larger the size of the fragmentation tree. We therefore tested whether our method is capable of assigning the correct identity to a certain metabolite while its spectral tree was not at the same depth as the reference compounds. In this study MS^n spectra of three different flavonoids were acquired at 5 different concentration levels (2500, 500, 250, 50 and 25 ng/ml). After processing the MS^n data in the elemental composition fragmentation tree mode, the fragmentation trees were searched and compared to the reference library obtained at 2500 ng/ml. In addition, in order to simulate spectra generated at different compound concentrations in their performance in library matching, the MS^n data generated at 2500 ng/ml were processed by the MEF tool at four different relative intensity threshold settings, i.e. 0, 5, 10, and 20% of relative intensity compared to the base peak, thereby simulating a concentration series *in silico*.

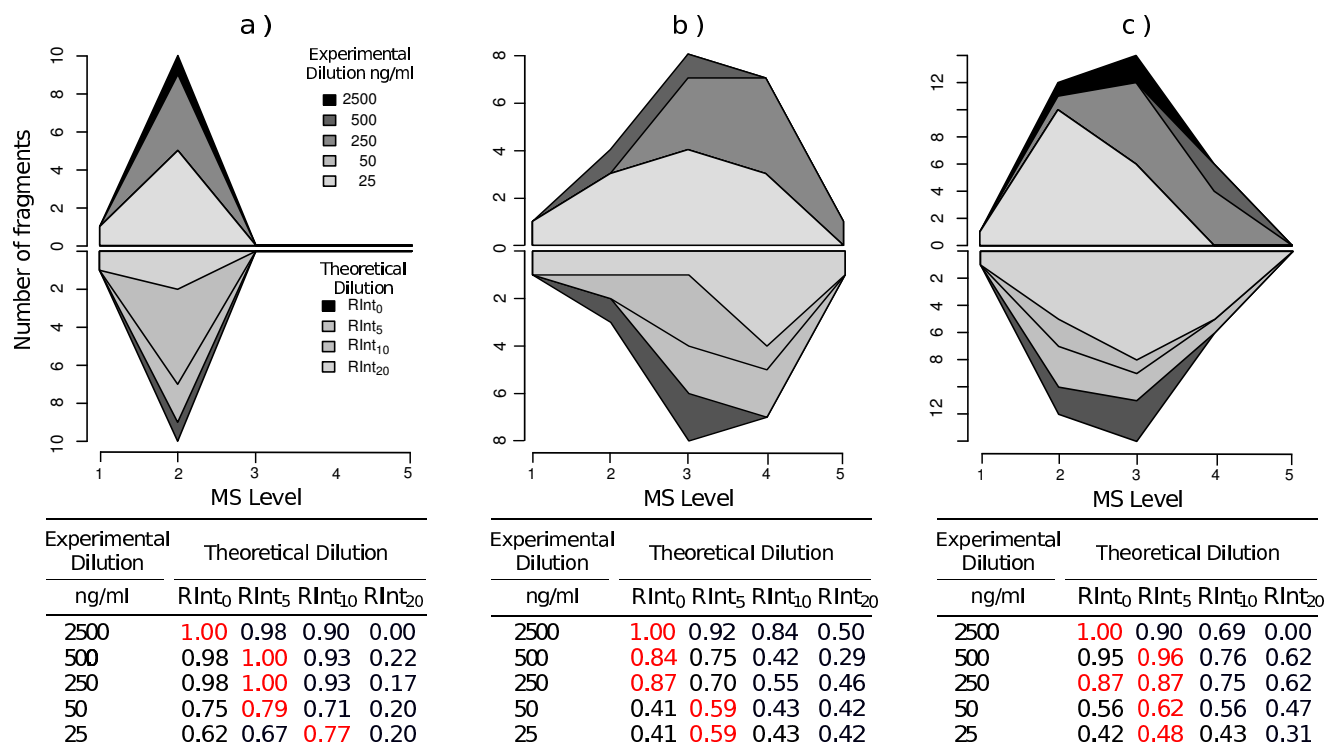


Figure 4.3: Fragmentation topology curves (number of fragments at each MS level) of fragmentation trees (ESI in positive mode) obtained from metabolites at different concentrations versus a simulated dilution series in silico extracted from the data obtained at the highest compound concentration by using different threshold settings for peak picking in each spectrum. RInt 0, 5, 10 and 20 refer to Relative Intensities thresholds of 0, 5, 10 and 20% of the base peak. The metabolites are 7-hydroxyflavone (a), 6-methoxyflavone (b) and 3,2'-dihydroxyflavone (c). The tables in the lower part of the figure show the similarity values between experimental and simulated trees. Values in red represent the highest similarity values when compared with the experimental data.

4.3 shows the resemblance between the fragmentation tree topologies of the simulated and the experimental dilutions. The effect of changing the compound concentration was comparable between simulated and experimentally obtained fragmentation trees: both types of trees showed a loss of fragments going from a high to a low concentration. However, the MS level at which shrinkage of the fragmentation trees occurs differs. Ideally (for optimal dereplication use of an MS^n library) the theoretically generated dilution series should cover as much as possible the fragmentation tree series of its experimental dilution series. In the lower part of 4.3, similarity matrices between simulated and experimental dilutions were calculated. The fragmentation trees simulated for low compound concentrations were most similar to experimentally obtained fragmentation trees at low concentrated compounds, while fragmentation trees simulated for higher compound concentrations were most similar to the more concentrated experimentally obtained trees. These results indicate that I) the reduction of fragments by lowering the relative intensity can be used to simulate spectra generated at lower compound concentrations, and II) by applying this MS^n data processing at different peak intensity thresholds the probability to correctly assign the metabolite identity is increased, even though its MS^n data were obtained at a compound concentration different from that used to populate the reference library. Although the identification probability is increased, it remains lower than if the compounds were present at similar concentrations. To ensure that the theoretical fragmentation spectrum of a library compound will simulate correctly the experimental spectrum of that compound at an unknown sample concentration, it would be optimal to simulate spectra at a large number of thresholds for relative intensity. In this study 4 different intensity threshold parameters are used but this number can easily be increased in order to more precisely compare and match experimental fragmentation spectra from compounds present at relative low concentrations in biological samples with library spectra obtained at relative high concentrations.

After defining the fragmentation tree representation (by means of elemental compositions) and the way how to deal with differences in concentrations of the obtained fragmentation trees (processing of the library trees with multiple relative intensity settings) the dereplication functionality of the library was further investigated by monitoring the identity predictions of replicate measurements of a number of metabolites which were already in the library. For 282 metabolites 765 replicate fragmentation trees were acquired and used as validation samples. Using our fingerprint-based approach, 722 (94%) of these fragmentation trees were correctly identified. Nevertheless, 43 fragmentation trees (belonging mainly to plant metabolites) were found having a higher similarity with a tree of another compound than the measured compound. In all cases the difference between the highest similarity value and the similarity value of the tree of the compound measured was relatively small. Most of these cases were due to isomers such as Epicatechin and Catechin, of which it is

known that these compounds are very difficult to discern [van der Hooft *et al.*, 2011]. In Supplemental Figure 8 in the Supporting Information the distributions are shown of all similarity values of the measured compounds together with the similarity values of the compound most similar to the measured one. For dereplication/identity search use the difference between these should be large to get clear identity assignments. For similarity search however, the similarity value of the first non-identical compound should be as high as possible to extract relevant substructural information about the unknown. The use of these libraries for similarity search is further investigated in the next section of this document.

Similarity searching for molecules frequently aims to detect molecules having similar biological activity [Stumpfe & Bajorath, 2011]. Adopting this concept, we aimed to elucidate whether similar chemical structures will result in (partially) similar fragmentation trees and vice versa, which would help in the identification of unknown compounds. For the entire set of compounds present in each library (human and plant metabolites), all pairwise similarity values (both the chemical structure similarity and the fragmentation tree similarity value) were calculated and plotted in 4.4. A perfect correlation of the chemical structure metric with the fragmentation tree metric should emerge as a diagonal line. The correlation coefficient between structural and fragmentation tree similarity for both libraries, human and plant, are ($r_{human} = 0.54, r_{plant} = 0.41$). Plots a1) and b1) reflect that mainly the region below this diagonal is occupied with similarity pairs. This means that similar fragmentation trees are typically the result of compounds having similar chemical structures, while compounds having similar chemical structures do not by definition generate similar fragments. In the latter case, only a part of both fragmentation trees will overlap leading to relatively low fragmentation similarity values. The observed phenomenon that a pair of compounds with a (relatively) high fragmentation tree similarity value also has a high chemical structure similarity value, is called neighborhood behavior [Patterson *et al.*, 1996]. The two sets of compound libraries tested here differ with respect to their variation in chemical structures. A structure diversity analysis of both libraries, see Supplemental Text 1 in Supporting Information, showed that the plant database, having a diversity value of 0.48, is structurally less diverse than the human metabolite database, having a diversity value of 0.82. This difference is due to the fact that the plant library mostly contains structurally related polyphenol structures, while the human library is mainly composed of lipids, amino acids, and sugars. The distribution plots of fragmentation tree similarities (4.4 a2 & b2), in both libraries, show at relatively low similarity value an optimum, indicating that fragmentation trees are unique and characteristic for each compound, making the process of replication more efficient.

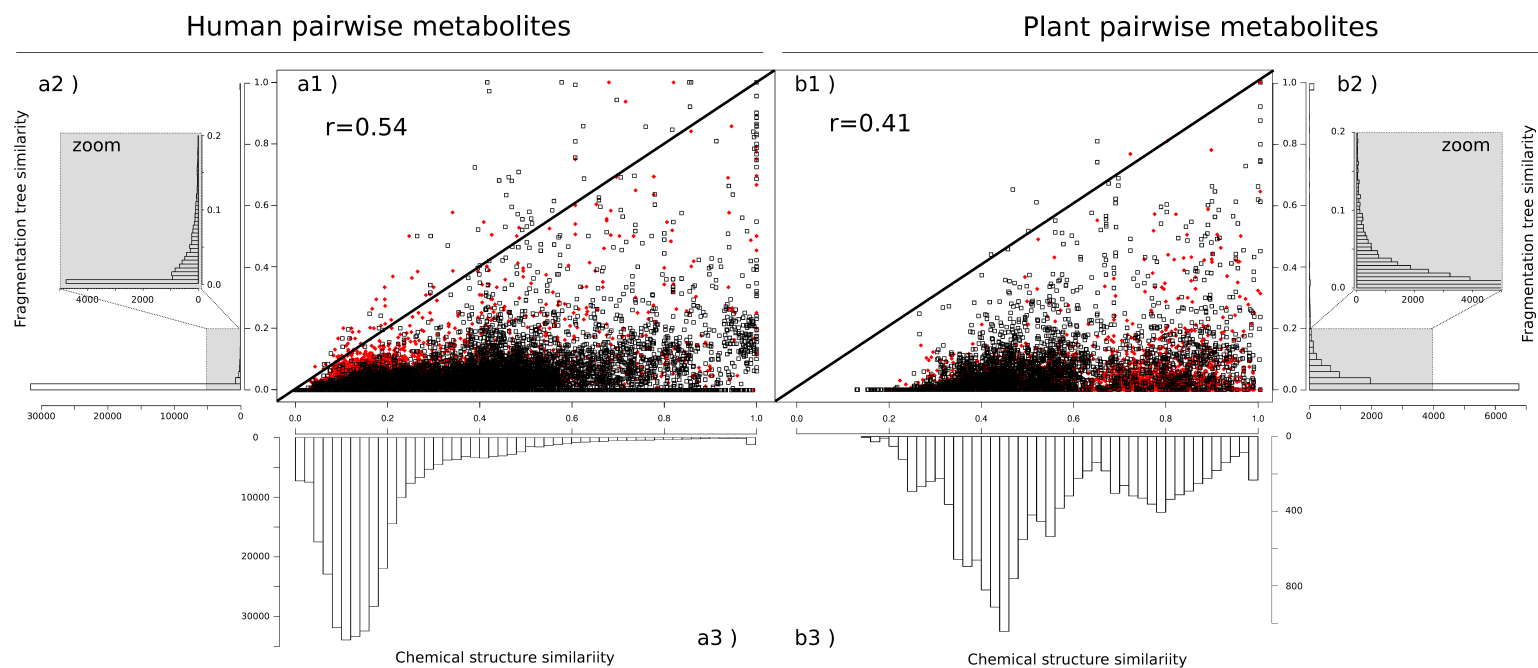


Figure 4.4: Plots a1 & b1 show the pairwise chemical structure similarity versus fragmentation tree similarity for human (a) and plant metabolites (b). Plots a2 & b2 show the fragmentation tree similarity distributions and plots a3 & b3 the chemical structure distributions.

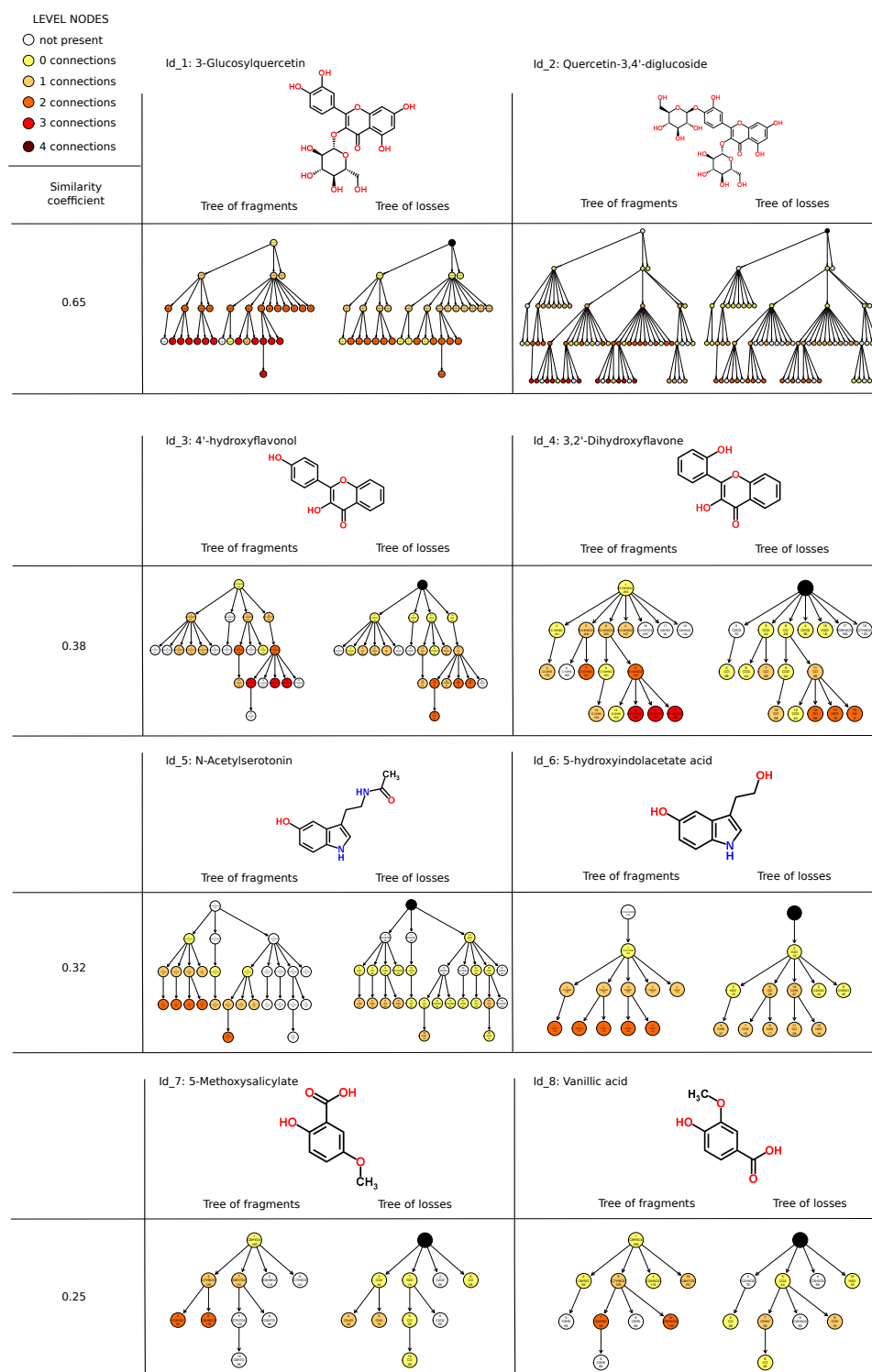


Figure 4.5: Comparison of fragmentation and neutral loss trees for two different compounds. The first column shows the fragmentation tree similarity value. In the next columns the fragmentation and neutral loss trees are plotted. White nodes indicate that they are unique in the corresponding tree. Colored nodes correspond to overlapping nodes or branches. The different colors indicate the number of upper nodes also overlapping and connected.

Knowledge about fragmentation tree similarity and which branches/building blocks are in common, is relevant for posterior interpretation of the MS^n results and annotation of the metabolite(s) under investigation. In the recently published MEF tool [Rojas-Chertó *et al.*, 2011] a new visualization feature has been implemented that highlights nodes that are common between the fragmentation trees compared (see 4.5). Clearly, the degree of similarity between fragmentation trees can be deduced from the number of colored nodes. This colored visualization of pairs of fragmentation trees can also help to discover relevant building blocks (subtrees) to focus on when interpreting fragmentation spectra of unknown metabolites. Compounds N-acetylserotonin and 5-hydroxyindoleacetic acid clearly illustrate this added value: the fragmentation tree of 5-hydroxyindoleacetic acid is almost a complete building block of the fragmentation tree of N-acetylserotonin. We also emphasize the relevance of comparing neutral loss trees for the identification of similar fragmentations. In some cases it can happen that only the neutral loss tree provides structural likeness between metabolites.

As several studies [Wolfender *et al.*, 2000, Sheldon *et al.*, 2009, van der Hooft *et al.*, 2011] have shown previously, the existence of correlation between the building blocks generated in the MS^n data and the substructure of the measured molecule makes partial structural elucidation of unknown compounds possible, provided that similar building blocks are found in a reference library. For the MS^n data in our library, structure information for each fragment or neutral loss is not (yet) available. Although we were able to find similar building blocks in the MS^n data, in the library no substructure information was returned. However, we were able to extract the maximum common substructure (MCSS) from the structures that have the most similar fragmentation trees (4.6 and Supplemental Table 4 in the Supporting Information). Thus, the extracted MCSS is a substructure that is likely part of the unknown molecule. This information can help MS experts with the identification of unknown compounds, e.g. by using it as an input together with the elemental composition of the unknown compound in a structure generator [Braun *et al.*, 2004]. To generate the MCSS we need to define the list of compounds used as input for the MCSS calculation tool. As a consequence, the obtained MCSS depends heavily on the number of compounds considered to have similar fragmentation trees. This list of compounds can be defined by setting a fragmentation similarity threshold. Upon decreasing this threshold, the number of similar fragmentation tree hits increases and therefore the MCSS gets smaller and less specific. The larger the MCSS, the more structural information is available for identifying the unknown compound. We therefore aimed to determine the smallest threshold to set while still retrieving a structural relevant MCSS.

Structures to test	Structures with most similar fragmentation trees			Maximum common substructures
8,9-EET Nr.Frag:46 Coef.Value:1	11,12-EET Nr.Frag:37 Coef.Value:0.73	bicyclo-PGE2 Nr.Frag:39 Coef.Value:0.22	9(10)-EpOME Nr.Frag:22 Coef.Value:0.18	
Rutin Nr.Frag:12 Coef.Value:1	3-Glucosyquercetin Nr.Frag:7 Coef.Value:0.42	Quercitrin 6'acetate Nr.Frag:11 Coef.Value:0.35		
4'hydroxyflavonol Nr.Frag:24 Coef.Value:1	3,2'Dihydroxyflavone Nr.Frag:31 Coef.Value:0.78	6,2'Dihydroxyflavone Nr.Frag:20 Coef.Value:0.57	Daidzein Nr.Frag:48 Coef.Value:0.43	
N-Acetylserotonin Nr.Frag:28 Coef.Value:1	5-Hydroxyindoacetic acid Nr.Frag:12 Coef.Value:0.32	2-Hydroxyphenylalanine Nr.Frag:21 Coef.Value:0.11		
6-hydroxy-m-Anisic acid Nr.Frag:10 Coef.Value:1	Vanillic acid Nr.Frag:10 Coef.Value:0.25	3-Hydroxycinnamic acid Nr.Frag:5 Coef.Value:0.17	3-Amino Salicylate Nr.Frag:5 Coef.Value:0.17	

Figure 4.6: The result of querying the MSⁿ library with fragmentation trees derived from 'unknown' (test) metabolites (first column). The 2nd, 3rd and 4th columns show structures with most similar fragmentation trees. The boxes below the structures list the number of fragments (Nr.Frag) that are characteristic of the compound's fragmentation tree and the similarity value (Coef.Value) compared to the fragmentation tree of the unknown metabolite. The last column shows the maximum common substructure (MCSS) extracted from the compounds listed in the middle columns.

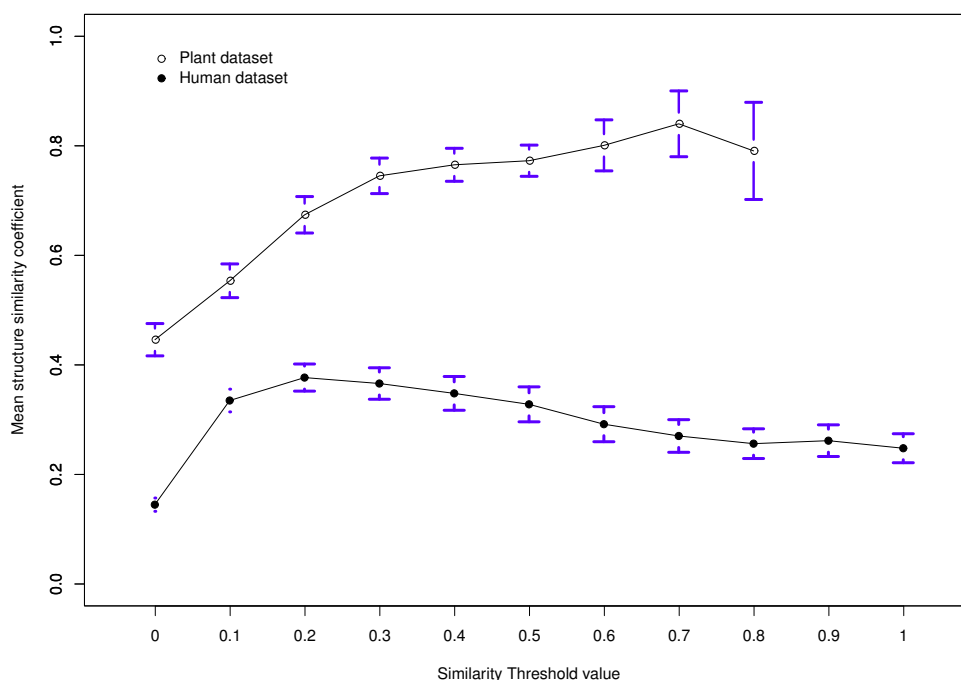


Figure 4.7: The mean structure similarity value of the maximum common substructure (MCSS) and the structure of the queried metabolite for different fragmentation tree similarity thresholds (used for generating the MCSS) for the human and plant MS^n libraries.

In 4.7 the effect of lowering the fragmentation tree similarity threshold value on the calculated MCSS is shown. Each fragmentation tree entry was compared to the rest of fragmentation trees in the library, so ultimately the average and standard deviation of the structural similarity values for all entries in the database were calculated. The MCSS obtained from the plant library has higher structural similarity values than the MCSS extracted from the human library over the whole range of fragmentation similarity threshold values. This is obviously due to the fact that the plant library contains several series of isomers and structurally related compounds. Because a higher structural MCSS similarity is correlated with the size of the MCSS, the size of the MCSS of the plant library is larger and structurally more informative than the MCSS obtained from the human library. This result underlines the importance of filling the database with as much structurally related compounds as possible, in order to obtain as much as possible information about the identity of unknown metabolites. Over the whole range of fragmentation similarity thresholds, the generated MCSS was relatively stable. Below values of 0.3 for the plant library and 0.2 for the human database, the threshold reached a value where the obtained MCSS seems to become structurally less informative, which in practical terms means we can use a fragmentation similarity threshold of about 0.25 as a good compromise to extract structural information of an unknown metabolite from a compound library. Overall, although fragmentation trees may not be very

similar, they may still be helpful in providing structure information and in partly elucidating the structure of unknown compounds.

4.5. Conclusion

This article introduces a new cheminformatical approach to calculate the similarity between mass spectral fragmentation trees, which can be helpful in the annotation of compounds detected using LC/MS-based metabolomics approaches. The new approach can be used to query multi-stage mass spectra data in MS^n libraries to define structure-spectra relationships and potentially deduce substructures within unknowns

Extracting the maximum common substructure (MCSS) from a list of structures that have the most similar fragmentation trees appears a valuable tool to obtain information about which molecular parts are also present in spectra of yet unknown compounds and can be used to structurally elucidate, at least partly, the unknown metabolite, providing that the library contains many structurally related compounds.

Our future work will focus on further populating the library with MS^n data and developing new cheminformatics tools to automatically annotate substructure information to the MS^n fragments. This will contribute to a more reliable hypothesis about the fragment structures present in unknown compounds. Furthermore, a new web-based tool called MetiTree (www.MetiTree.nl) has been built to provide the metabolomics community a platform to elucidate unknown structures using accurate mass MS^n data. Overall, we showed that our new tools can help in comparing MS^n data and in the annotation and identification of known and unknown compounds.

4.6. Acknowledgement

This project was financed by the Netherlands Metabolomics Centre (NMC) and the Centre for Biosystems Genomics (CBSG), which both are part of the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research. The authors wish to thank David Wishart for providing samples of compounds from the Human Metabolome Database (HMDB), the laboratories at the DSM Biotechnology Centre and the TNO Research Group Quality and Safety, The Netherlands, and their technicians involved in the MS^n measurements.

4.7. Supporting Information

4.7.1. Supplemental text 1 - Diversity level between the two libraries

The main difference between the human compounds and plant compounds based libraries used in this study was the diversity level between the molecules analyzed (Supplemental Text 1 in the Supporting Information). The diversity value (equation 1) of a data set A with $N(A)$ molecules was obtained by calculating the average value of all pairwise structural dissimilarities [Turner *et al.*, 1997]. A diversity value near 1 indicates that a dataset is highly diverse, while a value close to 0 indicates that the dataset contains very similar molecules or at least the structural features used to characterize the molecules make them very similar. The similarity measure $SIM(J,K)$ of any pair of molecules, J and K, is calculated by applying the Tanimoto coefficient (equation 2) using the CDK 2D-fingerprint library [Steinbeck *et al.*, 2003]. The Tanimoto coefficient $T(J,K)$ is calculated by dividing the number of features present in both molecules (z) by the number of unique features present for each molecule (x and y) minus the features present in both molecules (z).

$$DIVERSITY(A) = 1 - \frac{\sum_{J=1}^{N(A)} \sum_{K=1}^{N(A)} SIM(J, K)}{N(A)^2} \quad (4.1)$$

$$SIM(J, K) = T(J, K) = \frac{z}{x + y - z} \quad (4.2)$$

4.7.2. Supplemental text 2 - Prefiltering of the fingerprints

We applied the Tanimoto coefficient as a similarity measure to enable calculating the degree of similarity between fragmentation trees. Firstly, a prefiltering of the two fingerprints that are going to be compared is applied. This prefiltering is intended to omit large dissimilar features that would give rise to a series of smaller, likewise dissimilar features (Supplemental Figure 2). By reducing the non-overlapping features to the most basic dissimilar building blocks redundancy is avoided. On the other hand, this prefiltering step will emphasize the occurrence of overlapping fragmentation branches. The larger the overlapping branch between the fragmentation trees, the more weight is given to the similarity value calculated. For example, when comparing two fragmentation trees with the linear features A-B, B-C and A-B-C all present in only one of the trees, the filtering process will eliminate the complete feature A-B-C from the fingerprint. The same concept is applied in case of parallel features like A-(B-D), where both B and D are derived from A: if the linear features A-B and A-D are only present in one of the trees, the complete feature A-(B-D) will be eliminated. With this

prefiltering step we aimed to correct for the influence of the size of the fragmentation tree when different fragmentation trees are compared.

4.7.3. Supplemental text 3 - Library to process and compare MS^n data

The library of Java procedures to process and compare MS^n data is free available as an open source project in Sourceforge ([Samsn, 2012]). The README file contains a tutorial explaining the features and the command lines to use.

All MSⁿ data was processed with the above tool using the following command line:

```
> java -jar sams.jar -occurr=0.4 -sn=1 -mzgap=0.5 -rint=0 -acc=15 -  
ec=[MY_ELEMENTS] -rules=[RDBER] -imzXML namefile.mzXML -ocml  
namefile.cml process
```

The resulting output-file is an CML-file which contains a description of the enriched fragment peaks with the elemental composition information. The parameters used are:

```
sn = 1 . Signal to noise threshold  
mzgap = 0.5 . Minimal distance between adjacent peaks  
rint = 0.0, 0.05, 0.1, 0.2 . Relative intensity threshold. We process with 4  
different values to simulate a dilution series.  
acc = 15 . Maximal accuracy range set in ppm  
rules = RDBE . (Ring Double Bond Equivalents) Constraint rules applied to  
the formula occur = 0.4 . Minimum occurrence to appear in all repetitions  
within one file to be accepted as a fragment.  
ec = C0..10,H1..20. Elements to be included, together with the upper-  
/lower-limit of the number of atoms. They will depend on the compound  
to be analyzed. E.g. ec=C0..10,H1..20 means that the range of the carbon  
atom is set between 0 and 10 and the range of the hydrogen atom is set  
between 1 and 20.
```

To visualize the results the following command line can be used:

```
> sams -i1cml NAME_FILE_1.cml -i2cml NAME_FILE_2.cml compare
```

4.7.4. Figures

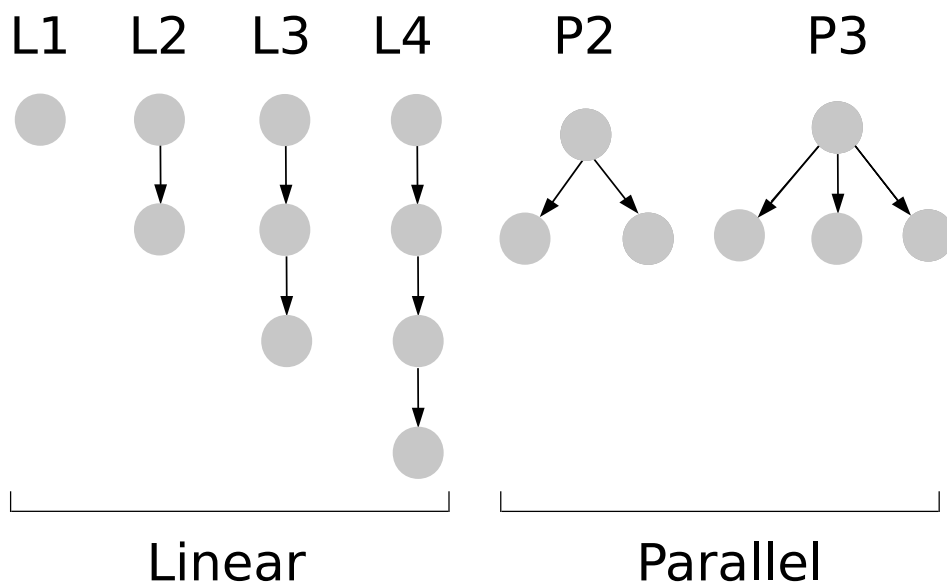


Figure 4.8: Supplemental Figure 1: Features template used to compare fragmentation trees. Each node contains elemental compositions (EC) or nominal masses (NM). We distinguish between linear and parallel connections of features, in combination with the number of nodes involved. These features are extracted both for the fragments and for the neutral losses.

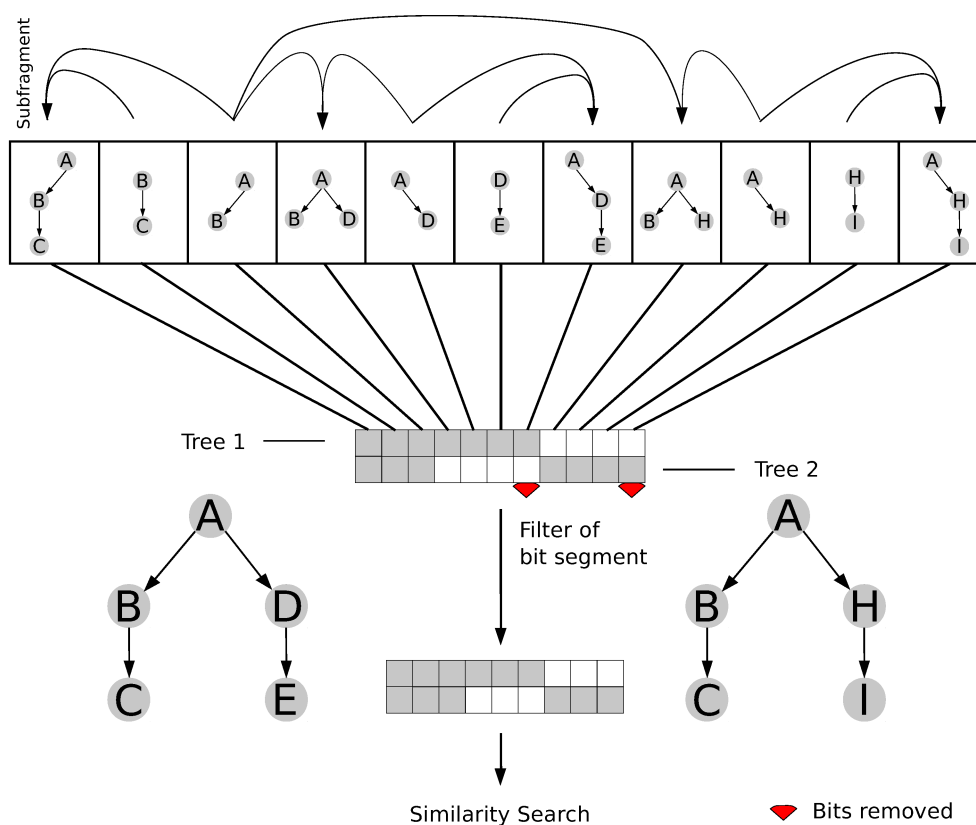


Figure 4.9: Supplemental Figure 2: Example of encoding two simple fragmentation trees and posteriorly fingerprint filtering. Bit positions are set 'on' (gray) if the corresponding feature is present in the fragmentation tree. Otherwise they are set 'off' (white). The filter process removes those bits, where all its subfragment bits are not present in one of the fragmentation trees.

4.7.5. Examples of Fragmentation trees

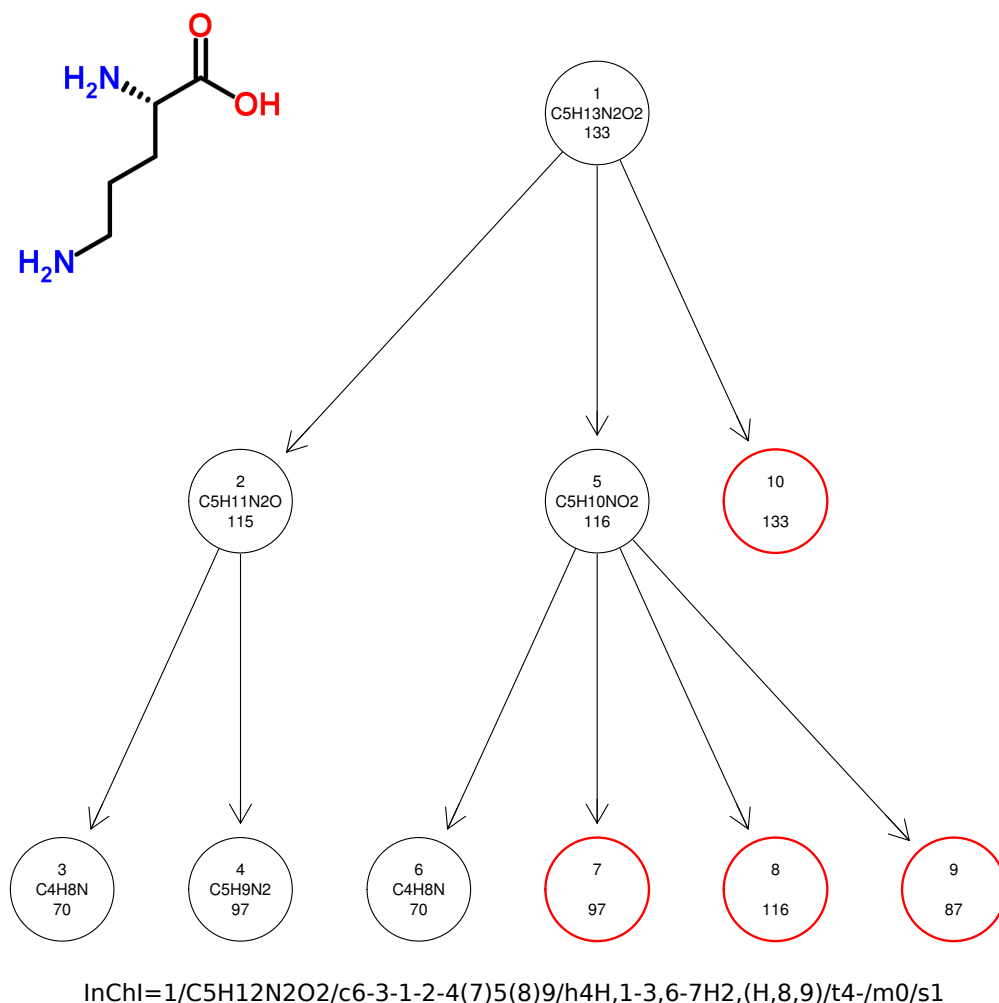
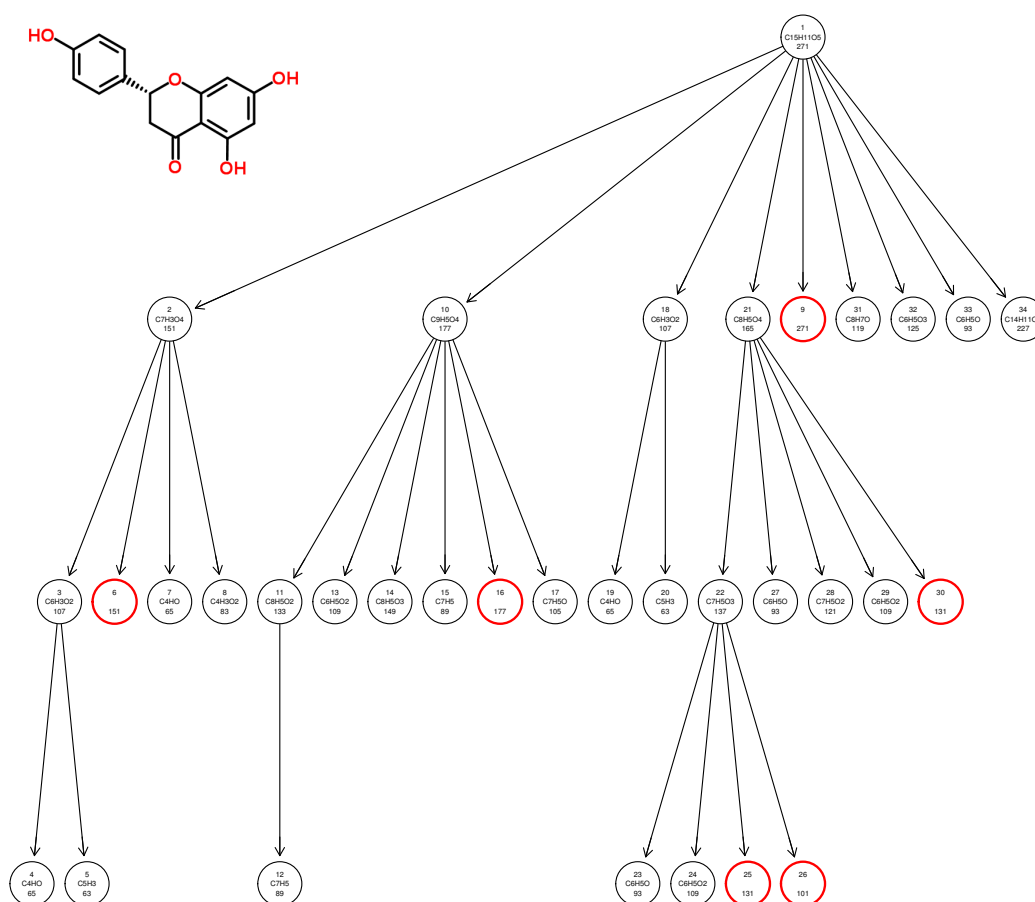
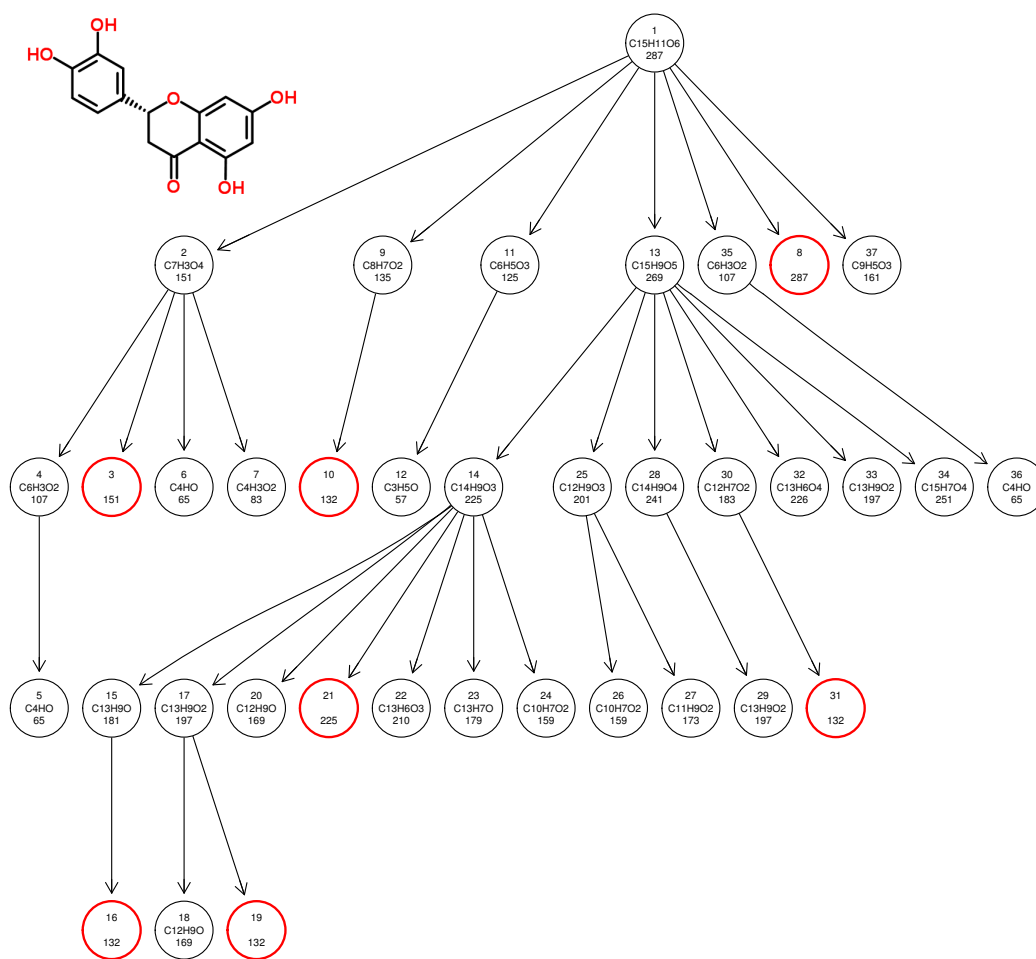


Figure 4.10: Supplemental Figure 3: Fragmentation tree of 5-Amino-L-norvaline [InChI=1/C5H12N2O2/c6-3-1-2-4(7)5(8)9/h4H,1-3,6-7H2,(H,8,9)/t4-/m0/s1]. The nodes for which the elemental composition is not calculated are drawn in red.



InChI=1S/C15H12O5/c16-9-3-1-8(2-4-9)13-7-12(19)15-11(18)5-10(17)6-14(15)20-13/h1-6,13,16-18H,7H2/t13-/m1/s1

Figure 4.11: Supplemental Figure 4: Fragmentation tree of (R)-naringenin [InChI=1S/C15H12O5/c16-9-3-1-8(2-4-9)13-7-12(19)15-11(18)5-10(17)6-14(15)20-13/h1-6,13,16-18H,7H2/t13-/m1/s1]. The nodes for which the elemental composition is not calculated are drawn in red.



InChI=1S/C₁₅H₁₂O₆/c16-8-4-11(19)15-12(20)6-13(21-14(15)5-8)7-1-2-9(17)10(18)3-7/h1-5,13,16-19H,6H2/t13-/m1/s1

Figure 4.12: Supplemental Figure 5: Fragmentation tree of Eriodictyol [InChI=1S/C₁₅H₁₂O₆/c16-8-4-11(19)15-12(20)6-13(21-14(15)5-8)7-1-2-9(17)10(18)3-7/h1-5,13,16-19H,6H2/t13-/m1/s1]. The nodes for which the elemental composition is not calculated are drawn in red.

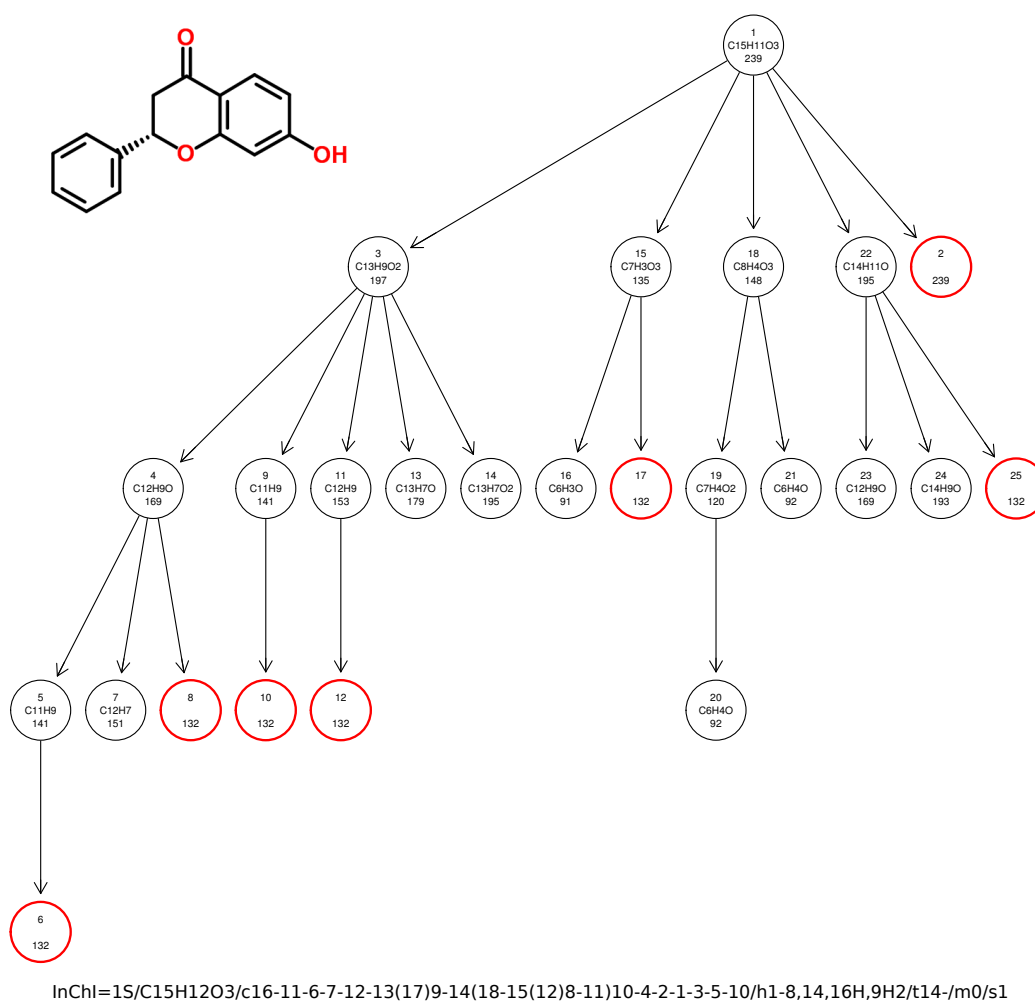


Figure 4.13: Supplemental Figure 6: Fragmentation tree of (2S)-7-hydroxyflavanone [InChI=1S/C₁₅H₁₂O₃/c16-11-6-7-12-13(17)9-14(18-15(12)8-11)10-4-2-1-3-5-10/h1-8,14,16H,9H2/t14-/m0/s1]. Drawn in red those nodes which the elemental composition is not generated.

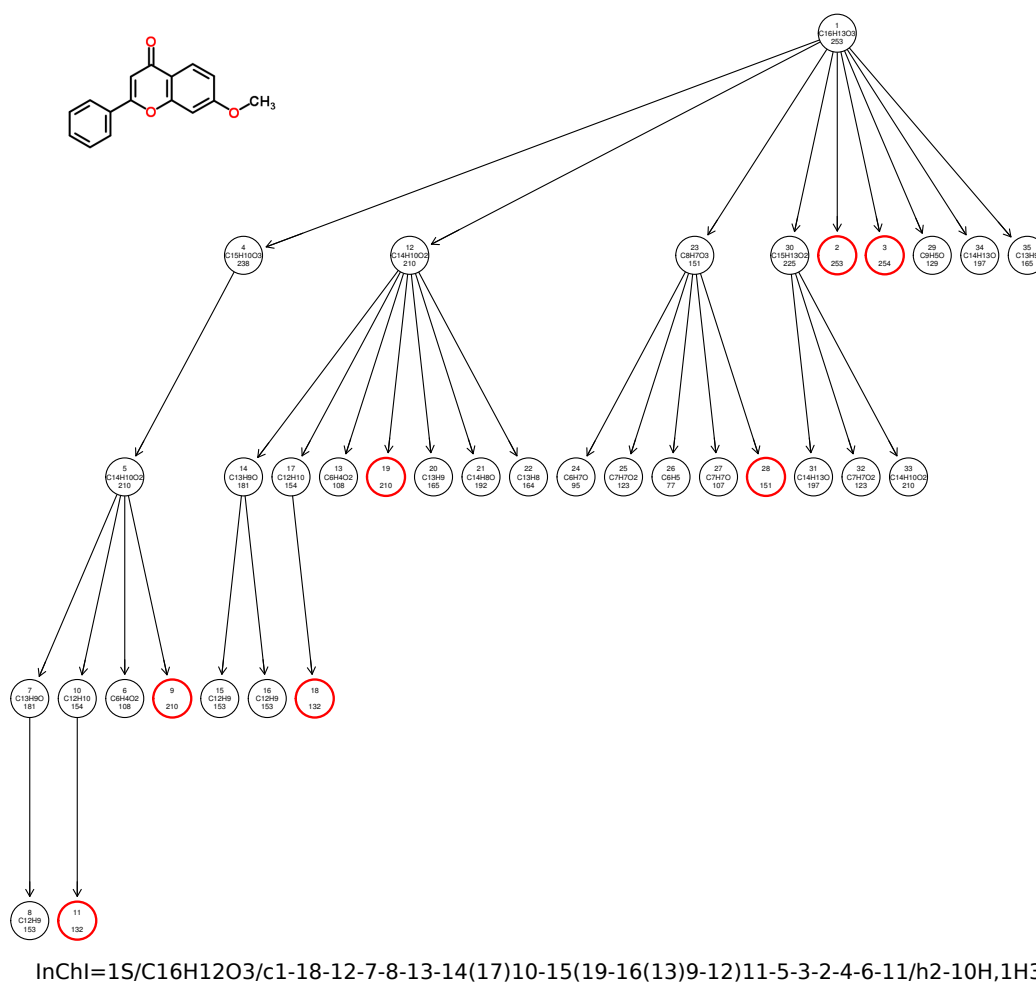


Figure 4.14: Supplemental Figure 7: Fragmentation tree of 6-Methoxyflavone [InChI=1S/C₁₆H₁₂O₃/c1-18-12-7-8-13-14(17)10-15(19-16(13)9-12)11-5-3-2-4-6-11/h2-10H,1H3]. The nodes for which the elemental composition is not calculated are drawn in red.

4.7.6. Supplemental text 4 - Evaluation of the fingerprint-based approach

For validation of the fingerprint-based approach, extra 765 fragmentation trees acquired either in positive mode or negative mode, corresponding to 282 different reference compounds, were matched to an established library. The library contains at least a fragmentation tree of the same reference compound. The library is composed of fragmentation trees of 454 reference compound acquired in positive mode and fragmentation trees of 422 reference compounds acquired in negative mode. Using the fingerprint-based approach 94% of the fragmentation trees were correctly identified. 43 fragmentation trees were assigned to another reference compound. These failing cases were due to isomers like Epicatechin and Catechin, which are very difficult to discern.

In Supplemental Figure 8 we show the distribution of the similarity score values obtained when the extra fragmentation trees are compared to the fragmentation trees in the library containing the same reference compound (v1). The figure is also showing the similarity score values obtained when it is compared with the most similar fragmentation tree not from the same reference compound (v2). We can observe that there is a significant difference of the similarity score between the first and the second most similar fragmentation trees. This confirms the high specificity of MS^n data to discern fragmentation trees.

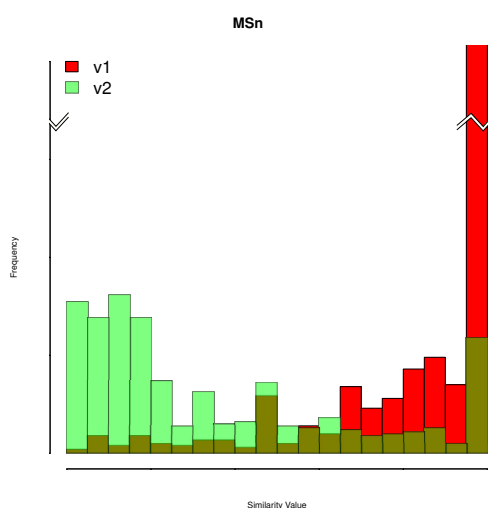


Figure 4.15: Supplemental Figure 8: Figure showing the distribution of the similarity score values obtained when extra 765 (MS^n) fragmentation trees are compared to the fragmentation trees (MS^n) in the library. v1 refers to the similarities with the same reference compound. v2 reflexes the distribution of the similarity score values obtained when it is compared with the most similar fragmentation tree in the library which is not the same reference compound.

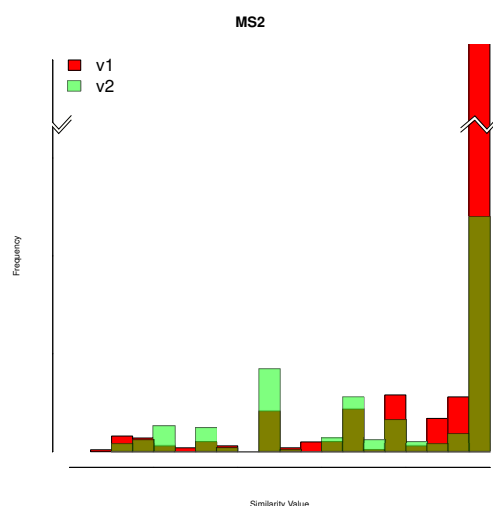


Figure 4.16: Supplemental Figure 9: Figure showing the distribution of the similarity score values obtained when extra 765 (MS^2) fragmentation trees are compared to the fragmentation trees (MS^2) in the library. v1 refers to the similarities with the same reference compound. v2 reflexes the distribution of the similarity score values obtained when it is compared with the most similar fragmentation tree in the library which is not the same reference compound.

It is also relevant to know the advantage of using MS^n instead of MS^2 . As we did not acquired MS^2 data, we recreated it in silico from MS^n using only the fragments present in MS level 2. We repeated with these data the validation tests. The first observation is that the fingerprints representing the fragmentation trees are smaller. 96% of the fragmentation trees was correctly identified. This is slightly better than the results obtained with MS^n data (94%). The main reason is that the reproducibility of the MS^2 data is better than the MS^n data. In several MS^n cases fragments beyond MS level 2 were not acquired. This ultimately leads to a slightly higher probability to find the correct compound using MS^2 data. This exercise also showed that querying the MS^2 library in several cases lead to equally scoring fragmentation trees and thus multiple possible identities meaning that MS^2 data is less specific. For MS^2 data we also observed that 63% of the fragmentation trees were only giving a similarity value for the true reference compound and not matched to any other reference compound. Whereas comparing MS^n data, only 20% of the fragmentation trees were not matched to any other reference compound. This shows that MS^n data is better for the extraction of a list of structures with similar fragmentation trees (similarity search/finding substructures).

4.7.7. Supplemental text 5 - Maximum common substructure analysis

In Figure 7, the effect of lowering the fragmentation tree similarity threshold value on the calculated maximum common substructure (MCSS) is shown. The Y axis describes the similarity value of the maximum common substructure (MCSS) with the structure of the queried metabolite. What we observe is that 0.25 fragmentation tree similarity value is a good compromise to extract still relevant structural information of an unknown compound. A similar analysis can be done by looking at the relative MCSS size (number of atoms of the MCSS divided by the number of the atoms of the queried metabolite). More precisely in Supplemental Figure 10, we analyse the effect of the fragmentation similarity threshold on the MCSS relative size obtained. Like previously the larger the fragmentation tree similarity threshold, the smaller the size of the MCSS generated.

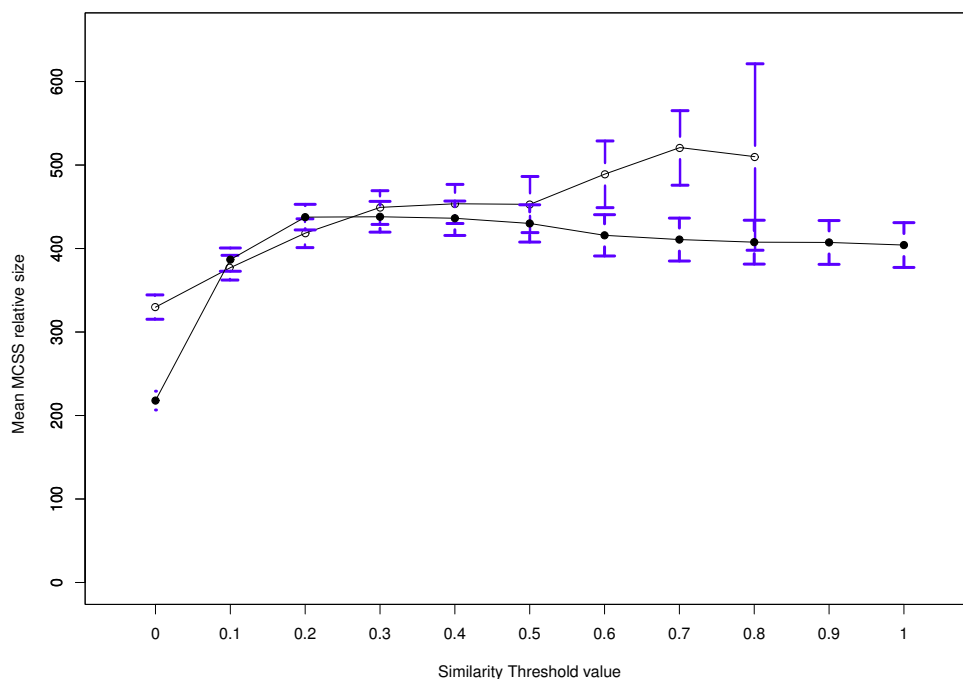


Figure 4.17: Supplemental Figure 10: The relative size (number of atoms) of the maximum common substructure (MCSS) for different values of fragmentation tree similarity (used for generating the MCSS) for the human and plant MSⁿ libraries.

Bibliography

- [Akiyama *et al.*, 2008] Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K., Hirai, M. Y., Sakurai, T., Kikuchi, J. & Saito, K. (2008) PRIME: a Web site that assembles tools for metabolomics and transcriptomics. *In Silico Biology*, **8** (3-4), 339–345.
- [Baldi & Nasr, 2010] Baldi, P. & Nasr, R. (2010) When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *Journal of chemical information and modeling*, **50** (7), 1205–22.
- [Braun *et al.*, 2004] Braun, J., Gugisch, R., Kerber, A., Laue, R., Meringer, M. & Rücker, C. (2004) MOLGEN-CID—A canonizer for molecules and graphs accessible through the Internet. *Journal of chemical information and computer sciences*, **44** (2), 542–8.
- [Bristow *et al.*, 2004] Bristow, A. W. T., Webb, K. S., Lubben, A. T. & Halket, J. (2004) Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries. *Rapid communications in mass spectrometry : RCM*, **18** (13), 1447–54.
- [Champarnaud & Hopley, 2011] Champarnaud, E. & Hopley, C. (2011) Evaluation of the comparability of spectra generated using a tuning point protocol on twelve electrospray ionisation tandem-in-space mass spectrometers. *Rapid communications in mass spectrometry : RCM*, **25** (8), 1001–7.
- [Coles *et al.*, 2005] Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S. & Zhang, Y. (2005) Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic & biomolecular chemistry*, **3** (10), 1832–4.

- [Fligner *et al.*, 2002] Fligner, M. A., Verducci, J. S. & Blower, P. E. (2002) A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics*, **44** (2), 10.
- [Grange *et al.*, 2002] Grange, A. H., Genicola, F. A. & Sovocool, G. W. (2002) Utility of three types of mass spectrometers for determining elemental compositions of ions formed from chromatographically separated compounds. *Rapid communications in mass spectrometry RCM*, **16** (24), 2356–2369.
- [Hansen & Smedsgaard, 2004] Hansen, M. E. & Smedsgaard, J. r. (2004) A new matching algorithm for high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, **15** (8), 1173–80.
- [Hernández *et al.*, 2011] Hernández, F., Portolés, T., Pitarch, E. & López, F. J. (2011) Gas chromatography coupled to high-resolution time-of-flight mass spectrometry to analyze trace-level organic compounds in the environment, food safety and toxicology. *TrAC Trends in Analytical Chemistry*, **30** (2), 388–400.
- [Holliday *et al.*, 2006] Holliday, G. L., Murray-Rust, P. & Rzepa, H. S. (2006) Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions. *Journal of chemical information and modeling*, **46** (1), 145–57.
- [Holliday *et al.*, 2003] Holliday, J. D., Salim, N., Whittle, M. & Willett, P. (2003) Analysis and display of the size dependence of chemical similarity coefficients. *Journal of chemical information and computer sciences*, **43** (3), 819–28.
- [Hooft *et al.*, 2011] Hooft, J. J. J., Vervoort, J., Bino, R. J. & Vos, R. C. H. (2011) Spectral trees as a robust annotation tool in LC-MS based metabolomics. *Metabolomics*, **8** (4), 691–703.
- [Hopley *et al.*, 2008] Hopley, C., Bristow, T., Lubben, A., Simpson, A., Bull, E., Klagkou, K., Herniman, J. & Langley, J. (2008) Towards a universal product ion mass spectral library - reproducibility of product ion spectra across eleven different mass spectrometers. *Rapid communications in mass spectrometry RCM*, **22** (12), 1779–1786.
- [Horai *et al.*, 2010] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K. & Nishioka, T. (2010) Mass-Bank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS*, **45** (7), 703–14.
-

- [Jansen *et al.*, 2005] Jansen, R., Lachatre, G. & Marquet, P. (2005) LC-MS/MS systematic toxicological analysis: comparison of MS/MS spectra obtained with different instruments and settings. *Clinical Biochemistry*, **38** (4), 362–372.
- [Kind & Fiehn, 2010] Kind, T. & Fiehn, O. (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews*, **2** (1-4), 23–60.
- [Kuhn *et al.*, 2007] Kuhn, S., Helmus, T., Lancashire, R. J., Murray-Rust, P., Rzepa, H. S., Steinbeck, C. & Willighagen, E. L. (2007) Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. *Journal of chemical information and modeling*, **47** (6), 2015–34.
- [Leach & Gillet, 2007] Leach, A. R. & Gillet, V. J. (2007) *An Introduction to Chemoinformatics: Revised Edition*. Springer.
- [McLafferty *et al.*, 1998] McLafferty, F. W., Zhang, M. Y., Stauffer, D. B. & Loh, S. Y. (1998) Comparison of algorithms and databases for matching unknown mass spectra. *Journal of the American Society for Mass Spectrometry*, **9** (1), 92–5.
- [Murray-Rust *et al.*, 2001] Murray-Rust, P., Rzepa, H. S. & Wright, M. (2001) Development of chemical markup language (CML) as a system for handling complex chemical content. *New Journal of Chemistry*, **25** (4), 618–634.
- [Oberacher *et al.*, 2009] Oberacher, H., Pavlic, M., Libiseller, K., Schubert, B., Sulyok, M., Schuhmacher, R., Csaszar, E. & Köfeler, H. C. (2009) On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm. *Journal of mass spectrometry : JMS*, **44** (4), 494–502.
- [Palit & Mallard, 2009] Palit, M. & Mallard, G. (2009) Fragmentation energy index for universalization of fragmentation energy in ion trap mass spectrometers for the analysis of chemical weapon convention related chemicals by atmospheric pressure ionization-tandem mass spectrometry analysis. *Analytical Chemistry*, **81** (7), 2477–2485.
- [Patterson *et al.*, 1996] Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. & Weinberger, L. E. (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *Journal of Medicinal Chemistry*, **39** (16), 3049–3059.
- [Pedrioli *et al.*, 2004] 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. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, **22** (11), 1459–66.
- [Portolés *et al.*, 2011] Portolés, T., Pitarch, E., López, F. J., Hernández, F. & Niessen, W. M. A. (2011) Use of soft and hard ionization techniques for elucidation of unknown compounds by gas chromatography/time-of-flight mass spectrometry. *Rapid communications in mass spectrometry : RCM*, **25** (11), 1589–99.
- [Rasche *et al.*, 2012] Rasche, F., Scheubert, K., Hufsky, F., Zichner, T., Kai, M., Svatos, A. & Böcker, S. (2012) Identifying the unknowns by aligning fragmentation trees. *Analytical chemistry*, **84** (7), 3417–3426.
- [Rasche *et al.*, 2011] Rasche, F., Svatosl̂ň, A., Maddula, R. R. K., Bořlttcher, C. & Bořlcker, S. (2011) Computing Fragmentation Trees from Tandem Mass Spectrometry Data. *Analytical Chemistry*, **83** (4), 1243–1251.
- [Rojas-Chertó *et al.*, 2011] Rojas-Chertó, M., Kasper, P. T., Willighagen, E. L., Vreeken, R., Hankemeier, T. & Reijmers, T. (2011) Elemental Composition determination based on MSn. *Bioinformatics*, **27** (17), 2376–2383.
- [Samsn, 2012] Samsn (2012). samsn.
- [Scheubert *et al.*, 2011] Scheubert, K., Hufsky, F., Rasche, F. & Böcker, S. (2011) Computing Fragmentation Trees from Metabolite Multiple Mass Spectrometry Data. *Journal of computational biology*, **18** (11), 377–391.
- [Sheldon *et al.*, 2009] Sheldon, M. T., Mistrik, R. & Croley, T. R. (2009) Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *Journal of the American Society for Mass Spectrometry*, **20** (3), 370–6.
- [Smith *et al.*, 2006] Smith, C. A., O'Maille, G., Want, E. J., Abagyan, R. & Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using non-linear peak alignment, matching, and identification. *Analytical chemistry*, **78** (3), 779–87.
- [Stein & Scott, 1994] Stein, S. E. & Scott, D. R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, **5** (9), 859–866.
- [Steinbeck *et al.*, 2003] Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, **43** (2), 493–500.
-

- [Steinbeck *et al.*, 2006] Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R. & Willighagen, E. L. (2006) Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Current Pharmaceutical Design*, **12** (17), 2111–2120.
- [Stumpfe & Bajorath, 2011] Stumpfe, D. & Bajorath, J. (2011) Similarity searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **1** (2), 260–282.
- [Turner *et al.*, 1997] Turner, D. B., Tyrrell, S. M. & Willett, P. (1997) Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *Journal of Chemical Information and Modeling*, **37** (1), 18–22.
- [van der Hooft *et al.*, 2011] van der Hooft, J. J. J., Mihaleva, V., de Vos, R. C. H., Bino, R. J. & Vervoort, J. (2011) A strategy for fast structural elucidation of metabolites in small volume plant extracts using automated MS-guided LC-MS-SPE-NMR. *Magnetic resonance in chemistry : MRC*, **49 Suppl 1**, S55–60.
- [Wan *et al.*, 2002] Wan, K. X., Vidavsky, I. & Gross, M. L. (2002) Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, **13** (1), 85–8.
- [Willett *et al.*, 1998] Willett, P., Barnard, J. & Downs, G. (1998) Chemical Similarity Searching. *Journal of Chemical Information and Modeling*, **38** (6), 983–996.
- [Wolf *et al.*, 2010] Wolf, S., Schmidt, S., Muller-Hannemann, M. & Neumann, S. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, **11** (1), 148.
- [Wolfender *et al.*, 2000] Wolfender, J., Waridel, P., Ndjoko, K., Hobby, K. R., Major, H. J. & Hostettmann, K. (2000) Evaluation of Q-TOF-MS/MS and multiple stage IT-MS_n for the dereplication of flavonoids and related compounds in crude plant extracts. *Analusis*, **28** (10), 895 – 906.
-

CHAPTER 5

Metitree: A web-application to organize and process multi-stage mass spectrometry data

Miguel Rojas-Chertó ^{1,2}, Michael van Vliet,^{1,3} Julio E. Peironcely ^{1,2,4}, Thomas Hanke-meier ^{1,2} and Theo Reijmers ^{1,2}

Bioinformatics:2012:28(20):2707-9

¹Netherlands Metabolomics Centre, Leiden, The Netherlands

²Division of Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands

³NBIC, The Netherlands

⁴TNO, Zeist, The Netherlands

5.1. Abstract

Identification of metabolites using high resolution multi-stage mass spectrometry (MS^n) data is a significant challenge demanding access to all sorts of computational infrastructures. MetiTree is a user-friendly, web application dedicated to organize, process, share, visualize, and compare MS^n data. It integrates several features to export and visualize complex MS^n data, facilitating the exploration and interpretation of metabolomics experiments. A dedicated spectral tree viewer allows the simultaneous presentation of three related types of MS^n data, namely, the spectral data, the fragmentation tree, and the fragmentation reactions. MetiTree stores the data in an internal database to enable searching for similar fragmentation trees and matching against other MS^n data. As such MetiTree contains much functionality that will make the difficult task of identifying unknown metabolites much easier.

5.2. Availability:

MetiTree is accessible at <http://www.MetiTree.nl>.

The source code is available at

<https://github.com/NetherlandsMetabolomicsCentre/metitree>.

5.3. Introduction

Metabolite identification is a challenging but essential step for the interpretation and understanding of many biological processes for an increasing number of applications such as biomarker discovery, drug discovery, or nutritional studies. The feasibility of using multi-stage mass spectrometry (MS^n) for identification of metabolites has been shown before [Sheldon *et al.*, 2009]. The complexity of the data generated demands new computational infrastructures to organize the data and to extract relevant information. Recently, databases have been set up for storing fragmentation spectra such as MS^n data (e.g. MassBank [Horai *et al.*, 2010]), and tools have been developed to process MS^n data (e.g. the MEF tool [Rojas-Chertó *et al.*, 2011]), or to compare MS^n data (e.g. Mass Frontier (Thermo Fisher Scientific))). In this paper we present a web-application called MetiTree with the novelty that it combines the processing of high resolution MS^n data with a personal local library to organize the fragmentation data. Furthermore, it allows the comparison of MS^n data to help the researcher with the identification of metabolites. MetiTree is available at <http://www.MetiTree.nl> together with some test MS^n data.

5.4. METHODS

Web Application: MetiTree (Metabolite Identification Tree) is a web application intended to aid in the metabolite identification process. Currently, MetiTree offers the possibility to organize, process, share, visualize, and search for similar high resolution multi-stage mass spectrometry (MS^n) data. MetiTree's web interface is accessed through a web browser and it was created using the Grails (<http://grails.org>) frame-work.

Data Processing And Comparison: In order to process MS^n data, MetiTree integrates the MEF tool ([Rojas-Chertó *et al.*, 2011]), which extracts chemical information from the fragments assigning the elemental composition to the ions and neutral losses. MetiTree also allows the comparison of newly acquired MS^n data to data that is already stored in an internal library ([Rojas-Cherto *et al.*, 2012]).

Data Visualization: MetiTree incorporates a JavaScript spectral tree viewer developed to visualize MS^n data (<https://trac.nbic.nl/brsp201017/>), in order to facilitate the exploration, interpretation, and validation of the results. This viewer interconnects three MS^n items: the spectrum, which contains mass peaks, the fragmentation tree, which contains fragment nodes/elemental compositions, and the fragmentation reactions, which contain structures.

5.5. USAGE EXAMPLE

MS^n data previously published by our group [Rojas-Chertó *et al.*, 2011, Rojas-Cherto *et al.*, 2012] are used to demonstrate how metabolites can be identified using the MetiTree web application. These data are freely accessible as test data in MetiTree.

Data processing: The required input to process MS^n data is mzXML files and the settings of the processing parameters. Processing parameters are grouped into those to extract the mass spectrometry information (m/z , intensity, and retention time) and those to enrich the MS data with chemical information (elements and number of atoms). MetiTree allows individual file as well as batch processing. Furthermore, the same mzXML file can be processed several times with different sets of parameters. Results and parameters information are stored to allow for posterior revision.

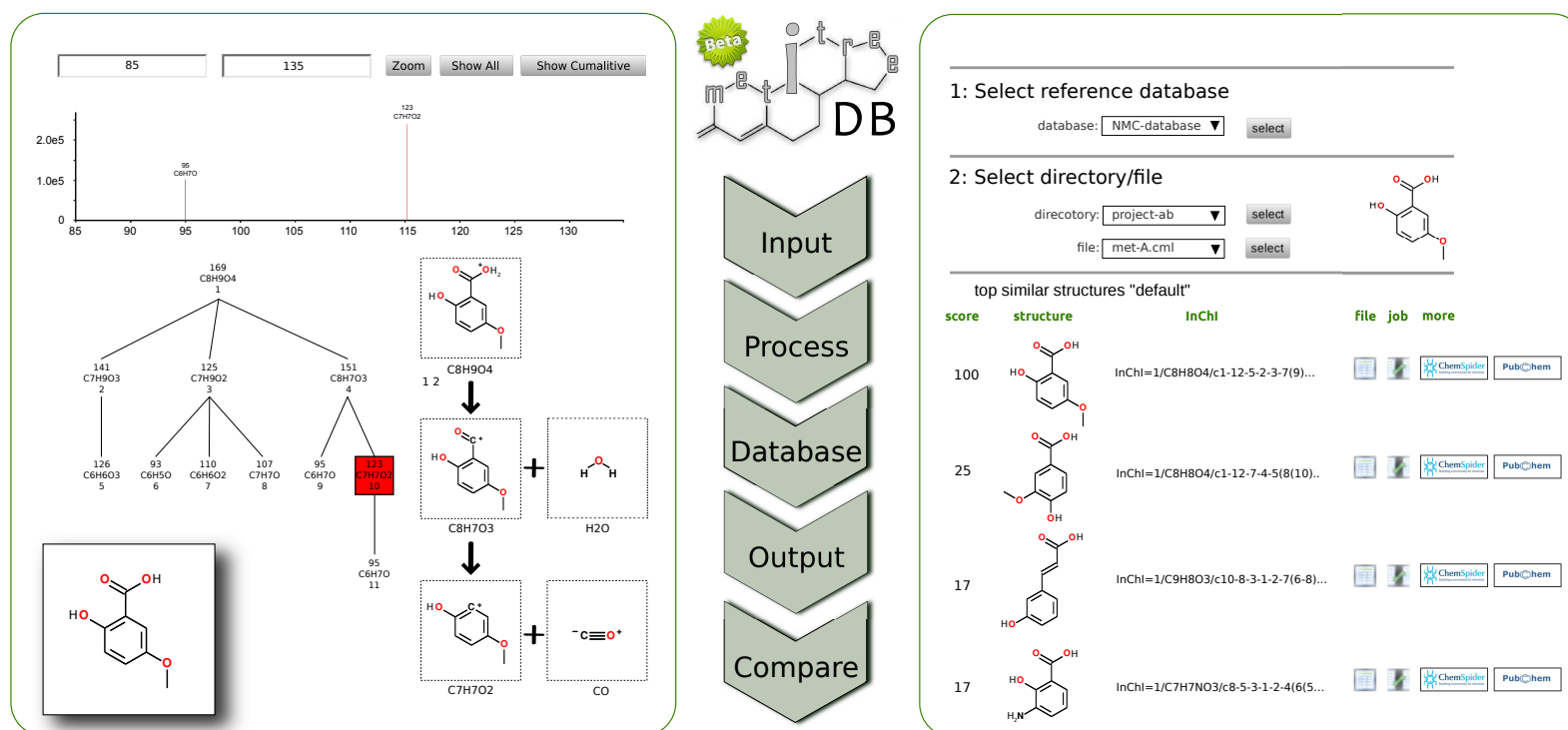


Figure 5.1: Overview of the MetiTree process flow. After the user submits MSⁿ spectra MetiTree will process these according to a set of parameters. Afterwards, the processed data can be stored in an internal library and labeled as a reference compound using the InChI identifier. The results are presented in different formats and viewers, which facilitates the exporting of text and figures for their use in reports and publications (A). Finally, MSⁿ data can be queried to find similar MSⁿ data in the library (B) and the query results are presented in a list.

5.6. USAGE EXAMPLE

MSⁿ data previously published by our group [Rojas-Chertó *et al.*, 2011, Rojas-Cherto *et al.*, 2012] are used to demonstrate how metabolites can be identified using the MetiTree web application. These data are freely accessible as test data in MetiTree.

Data processing: The required input to process MSⁿ data is mzXML files and the settings of the processing parameters. Processing parameters are grouped into those to extract the mass spectrometry information (m/z, intensity, and retention time) and those to enrich the MS data with chemical information (elements and number of atoms). MetiTree allows individual file as well as batch processing. Furthermore, the same mzXML file can be processed several times with different sets of parameters. Results and parameters information are stored to allow for posterior revision.

Data visualization: Once the data is processed, it can be displayed using the spectral tree viewer (Figure 1A) When the node (a fragment) is selected the corresponding spectrum is displayed, together with the concatenated reactions that connect the parent ion with the selected fragment. The structure of the fragment can only be displayed if it has been previously assigned. The results generated by MetiTree can be exported to different formats (CSV, CML [Murray-Rust & Rzepa, 1999], and PDF) for further analysis or for presenting results in reports and publications.

Library storage: MetiTree creates directories for grouping mzXML files, assisting with the organization of the data according projects or topics. Processed MSⁿ data can be stored in one or multiple internal databases (Figure 1B). Because the users are organized in groups, they can share files and libraries with other group members. All MSⁿ data can be labeled with an InChI identifier of the compound, which is automatically cross referenced with PubChem and ChemSpider databases.

Data search: MetiTree integrates the functionality to query for similar MSⁿ data [Rojas-Cherto *et al.*, 2012] stored in the library. The results are presented in a list showing the chemical structures of the most similar MSⁿ data and the corresponding similarity values. A value near 100 indicates that MSⁿ data are highly similar; while a value close to 0 illustrates that they are very different. If a fragmentation tree of the same compound is present in the library, complete identification is possible (identity search). If similar fragmentation data is found (similarity search), this substructure information can be used to generate candidate structures of the unknown compound (partial identification).

Future: In the near future this new web-application will accept also the uploading of other types of MSⁿ files (e.g. cml, mzML) and manual annotation of MSⁿ data with chemical structural information (assigning substructures to the nodes of the fragmentation trees) will

also be possible.

5.7. Conclusion

The growing interest in metabolite identification has increased the need to create computational and visual tools for MS^n analysis. MetiTree which gathers several in-house developed tools is an easy-to-use web-application that combines processing, sharing, visualizing, and querying MS^n data to help researches to identify metabolites of interest and decrease the time-consuming task of identifying metabolites.

5.8. Acknowledgement

This project was financed by the research programme of the Netherlands Metabolomics Centre (NMC), which is a part of The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. This work was part of the BioAssist programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI). The authors are also grateful to Prof. Dr. N.M.M. Nibbering, Dr. A.C. Tas, P.T. Kasper, Dr. R. Vreeken, J.J.J. van der Hooft, and M. Ries for their input.

Bibliography

- [Horai *et al.*, 2010] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K. & Nishioka, T. (2010) Mass-Bank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS*, **45** (7), 703–14.
- [Murray-Rust & Rzepa, 1999] Murray-Rust, P. & Rzepa, H. S. (1999) Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *Journal of Chemical Information and Modeling*, **39** (6), 928–942.
- [Rojas-Chertó *et al.*, 2011] Rojas-Chertó, M., Kasper, P. T., Willighagen, E. L., Vreeken, R., Hankemeier, T. & Reijmers, T. (2011) Elemental Composition determination based on MSn. *Bioinformatics*, **27** (17), 2376–2383.
- [Rojas-Cherto *et al.*, 2012] Rojas-Cherto, M., Peironcely, J. E., Kasper, P. T., van der Hooft, J. J. J., de Vos, R. C. H., Vreeken, R., Hankemeier, T. & Reijmers, T. (2012) Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. *Analytical chemistry*, **84** (13), 5524–34.
- [Sheldon *et al.*, 2009] Sheldon, M. T., Mistrik, R. & Croley, T. R. (2009) Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *Journal of the American Society for Mass Spectrometry*, **20** (3), 370–6.

CHAPTER 6

Summary

The detailed description of the chemical compounds present in organisms, organs/tissues, biofluids and cells is the key to understand the complexity of biological systems. The small molecules (metabolites) are known to be very diverse in structure and function, and they can act as intermediates or end products in all sorts of reactions occurring in a biological system. However, the identification of the chemical structure of metabolites is one of the major bottlenecks in metabolomics research. Once this is achieved further interpretation of their biochemical role in biological systems can follow. Hence, the annotation and the structure elucidation of the metabolites are essential to understand the biological system under study. Actually, no single analytical platform exists that can measure and identify all existing metabolites. In current metabolomics research different analytical platforms are being used covering different classes of metabolites which ultimately allow profiling of the metabolome as complete as possible. Multistage mass spectrometry (MS^n) is a powerful analytical technique that helps identifying all these metabolites. This technique provides detailed structural information of the unknown metabolite by fragmenting the metabolite and its fragments recursively. However, at the moment only computational tools can provide a fast and straightforward analysis of the large amount of complex data that is generated by using MS^n spectrometry. The aim of this thesis was to develop a novel semi-automatic approach for the identification of metabolites using MS^n data. Furthermore, these tools were to be integrated into a pipeline to assign identities to unknown metabolites present in databases but especially to unknown metabolites not present in a database. The tools were to be released as open-source tools to make sure that other scientists can also profit of this approach. The research in this thesis focusses on to the identification of mainly human and, also, plant metabolites.

Electrospray ionisation multistage mass spectrometry (ESI- MS^n) is a very useful technique used with ion trap mass spectrometers, especially when coupled to a high resolution mass spectrometer such as an Orbitrap or a Fourier transform ion cyclotron resonance mass spectrometer. The data generated for each metabolite are a batch of spectra related to each other in a hierarchical manner, and these MS^n data are known as being very complex. In **Chapter 2** the development of a MS^n method is described and its potential demonstrated for the identification of metabolites. To reduce the complexity of the generated data it was necessary to represent the fragment ions by nominal m/z values or elemental compositions. The last manner can in principle distinguish unambiguously those ions not related to the fragmentation process. In **Chapter 2** the elemental formula path (EFP) concept is proposed to characterize and represent fragment ions in MS^n spectra. EFP is a linear string of concatenated elemental compositions describing the path from the top precursor ion till the fragment ion of interest. It is demonstrated that representing MS^n spectra by means of a collection of EFP's facilitates the comparison between fragmentation data obtained from

the same metabolite or from different metabolites. Using this concept to compare spectra, the influence of the concentration of a metabolite on the reproducibility and robustness of the obtained fragmentation tree was studied. The results show that the extracted EFP's are reproducible across the whole range of glutathione concentrations (from 1 to 1000 μM). However, the number of observed EFP's decrease at lower concentrations, suggesting that for metabolites at very low concentrations not enough information will be acquired using MS^n spectra for assignment or elucidation of its structure. The concept of EFP's allowed to distinguish two isomeric prostaglandins (PGE and PDD), which are structurally very similar but have different biological functions. It was observed that the number of unique characteristic features (peaks) for each prostaglandin increased with depth of the MS^n experiment, i.e. going to the MS5 and MS6 level. The study also shows that the isolation width parameter and the collision energy have to be carefully chosen as these two parameters influenced mainly the relative intensity of the fragment ion peaks and therefore the complete EFP set per metabolite.

The first step in the elucidation of a structure of an unknown metabolite (or compound) is the determination of its elemental composition. In **Chapter 3** the development of the Multistage Elemental Formula (MEF) tool is described. The MEF tool processes MS^n data to obtain clean fragmentation trees and to enable the correct assignment of the elemental composition to molecular ions, their fragment ions, and neutral losses. The MEF tool reduces efficiently the list of possible elemental composition candidates by constraining the elemental compositions of each ion by its parent (precursor ion) and descendants (fragments). A correlation has been found between the mass tolerance/mass error chosen for data acquisition (i.e. the isolation width of ions in the ion trap) and the topology (depth and width) of the fragmentation tree. It is demonstrated that usage of MEF and MS^n requires a lower mass accuracy than when using only MS/MS spectra. It was demonstrated that the incorporation of additional MS^n levels improves the determination of the elemental composition. Including more fragments in the fragmentation tree provide more dependencies between elemental formulas lists of the different fragments, leading to stronger constraints and revealing the correct elemental composition with MS data obtained with less accuracy. The effect is stronger for metabolites with a low molecular weight or containing fragments with low mass/charge ratios. This allows the use of a less expensive mass spectrometer with less high resolution power for the determination of the elemental composition. Acquiring mass spectral tree data is time consuming and for this reason it is necessary to find out which nodes of the fragmentation tree need to be acquired preferably. It was shown that most information can be retrieved when acquiring fragmentation trees as deep as possible with low mass fragments to be preferred. The MEF tool was validated to identify situations in which the tool may not deliver correct elemental assignments. Unreliable results were

obtained when (i) the mass tolerance applied to the mass/charge ratios was smaller than the experimental accuracy, when (ii) an EC is assigned to a mass peak that was an artifact and not due to the compound of interest and when (iii) certain mass peaks are removed because no EC is found but that belonged to the compound of interest. However, for all these situations proper solutions were found, and the approach allows the automated and reliable assignment of elemental compositions to all fragment ions of a spectral trees and at the same time allows to remove artifacts, i.e. mass peaks that are not due to the compound of interest.

Once for a compound the (correct) elemental compositions are assigned to all fragment ions, a search of identical or similar MS^n data in a (our) MS library is followed. Two different approaches to identify whether the spectral tree of an unknown compound is already present in any MS library have been followed: an identity search or a similarity search. Whereas an identity search demands that all features are present in both spectral trees being compared, a similarity search quantifies the number of features present or absent in both spectral trees.

In **Chapter 4**, a new method to compare MS^n data has been introduced. The method compares the presence or absence of certain features in both fragmentation tree. The features are defined in accordance with the different ways fragments and neutral losses are connected. To demonstrate the performance of the method we used two libraries containing 867 MS^n spectra from 549 different plant and human metabolites. In our study we found that there is a unidirectional correlation between the chemical structure of a compound and the fragmentation tree: metabolites with similar fragmentation trees have similar chemical structures, and dissimilar fragmentation trees are from metabolites with dissimilar structure. This correlation is in one direction only because similar chemical structures not always result in similar fragmentation trees. Another issue encountered is that for compounds present at lower concentrations the fragmentation tree is not complete, and an uncomplete fragmentation tree is compared with more complete fragmentation trees in databases acquired at higher concentrations. This challenge could be addressed by applying different intensity thresholds as a parameter when preprocessing the mass spectral tree using the MEF tool fragmentation trees so that simulated fragmentation trees acquired at different compound concentrations are available in a database for comparison when the fragmentation tree in the database was acquired at higher concentrations, what is usually the case. Furthermore, we developed a method calculating the maximum common substructure (MCSS) from a list of structures that have similar fragmentation trees so that structural information can be extracted from the database entries although the unknown metabolite is not in the library present. For this strategy it is very important that a database is available with as much compounds structurally comparable to the unknown metabolites of interest, in order

to obtain as much as possible reliable information about the common structure between the unknown metabolite and the compounds in the databases. Ultimately, a database in which to all fragment ions the substructure is assigned is of course the most suitable for metabolites identification.

To address the growing interest in metabolite identification and the need for easy-to-use computational tools the MetiTree web application (<http://MetiTree.nl>) was developed (**Chapter 5**). Metitree integrates, in an easy-to-use way, all tools developed in (**Chapter 3** and **Chapter 4**) and provides access to these tools from any computer through a web browser. This web application helps to overcome several challenges like the processing of the MS^n data to obtain fragmentation trees. Fragmentation tree data are complex data which should be visualized in a simplified manner so that these data can be interpreted and compared in an intuitive manner. The developed fragmentation tree viewer in MetiTree offers a simple and straightforward way to visualize a fragmentation tree and to analyze all the fragments, its precursor ion and children fragments. In summary you can study what reaction are happening in each level. It provides a valuable tool for interpretation, since MS^n data show the fundamentals of fragmentation reactions in a mass spectrometer, and teaching purposes, since it can be used in schools to demonstrate the reactions happening in the spectrometer. The developed method also supports the validation of the new acquired data. The comparison functionality allows comparison of your data with MS^n data previously analyzed to determine if data are correctly acquired. MetiTree helps researchers to identify metabolites of interest by finding similar MS^n data. In summary, MetiTree offers the functionalities to organize, process, share, visualize, and compare MS^n data. In general it speeds-up the process of the de-novo identification.

In summary, the in this thesis developed concepts and methods facilitate the extraction of relevant information from MS^n data and helps posteriorly identification of the chemical structure of unknown compounds. The developed platform integrating the methods developed allow to identify unknown metabolites in a faster, more precise, and more automated way. Together with other recent developments such as a structure generator allowing to use as input several substructures of a molecule, constraints such as the energy, prediction of the fragmentation, octanol-water coefficient a highly automated identification pipeline is feasible providing a short list of possible candidates, reducing the time required for identification of unknown significantly. In those cases, where too many candidates are obtained, or where several possible structures are obtained that a difficult to differentiate by mass spectrometry, nuclear magnetic resonance spectroscopy coupled to LC can be used to obtain additional information in a targeted manner, due to recent progress in the sensivity of LC-NMR due to efficient coupling using an solid phase extraction or a hanging droplet evaporation interface. In addition, the research presented in this thesis is also useful to

other topics outside of metabolomics,. Such as proteomics or the identification of organic molecules in general.

Samenvatting

Om de complexiteit van biologische systemen beter te begrijpen is het essentieel een zo gedetailleerd mogelijk beeld te krijgen van alle chemische verbindingen aanwezig in het organisme, de organen/weefseltypen, de lichaamsvloeistoffen en de cellen die bestudeerd worden. Die kleine moleculen (ook wel metabolieten genoemd) staan erom bekend heel divers qua structuur en functie te zijn. Metabolieten functioneren dan ook vaak als intermediair of eindproduct bij allerlei soorten chemische reacties in het biologisch systeem. Eén van de grootste knelpunten binnen metabolomics onderzoek is het identificeren van de metabolieten (het toekennen van een chemische structuur aan een metaboliet). Wanneer dit eenmaal is gedaan, is verdere interpretatie van hun biochemische rol in biologische systemen mogelijk. De toekenning van een structuur aan metabolieten is zeer essentieel om het bestudeerde biologisch systeem beter te begrijpen. Op dit moment bestaat er geen enkel analytisch meetplatform dat alle bestaande metabolieten kan meten en identificeren. Binnen het huidige metabolomics onderzoek worden daarom vaak verschillende analytische platforms naast elkaar gebruikt. Elke platform is gekoppeld aan een of meerdere metabolietklasse zodat uiteindelijk een zo compleet mogelijk geheel metaboloom gemeten kan worden. Gedetailleerde massa spectrometrie fragmentatie (MS^n) is een krachtige analytische techniek die het mogelijk maakt al deze metabolieten te identificeren. Deze techniek levert gedetailleerde structuur informatie op van de onbekende metaboliet d.m.v. herhaaldelijke fragmentatie van de metaboliet en de resulterende fragmenten. Op dit moment bestaan er alleen computationele hulpmiddelen voor een snelle en simpele, maar oppervlakkige, analyse van deze grote hoeveelheid van complexe MS^n data. Het doel van deze proefschrift is om een nieuwe, semi-automatische aanpak te ontwikkelen voor de identificatie van metabolieten op basis van MS^n data. Deze hulpmiddelen zouden met broncode opgeleverd worden om ervoor te zorgen dat andere wetenschappers ook kunnen profiteren van deze aanpak. Het onderzoek in deze proefschrift concentreert zich op de identificatie van hoofdzakelijk metabolieten aanwezig in de mens maar ook in de plant.

Electrospray ionizatie massa spectrometrie fragmentatie (ESI- MS^n) is een veel, in combinatie met ion trap massa spectrometers, gebruikte techniek, vooral als dit gekoppeld is aan hoge resolutie massa spectrometers zoals een Orbitrap of een Fourier transform ion cyclotron resonance massa spectrometer. De voor elke metaboliet gegenereerde data bestaan uit een aantal spectra die hiërarchisch aan elkaar gerelateerd zijn en deze zogenaamde MS^n data zijn daardoor uitermate complex. De ontwikkeling van een methode die analyse van MS^n data mogelijk maakt, inclusief de toepassing daarvan bij het identificeren van metabolieten, staat beschreven in Hoofdstuk 2. Om de complexiteit van de gegenereerde data te reduceren was het nodig om de fragmentatie ionen te representeren m.b.v. nominale massa waarden of elementformules (atoomsamenstellingen). Deze laatste representatie zou de ionen die niet gerelateerd zijn aan het fragmentatie proces er

in principe ondubbelzinnig uit kunnen halen. In Hoofdstuk 2 wordt het elementformule pad (EFP) concept voorgesteld die uitermate karakteristiek en representatief is voor de fragment ionen aanwezig in de MS^n spectra. Het EFP is een lineaire koord van aan elkaar gekoppelde elementformules die het fragmentatie proces beschrijven van een bepaald fragment ion inclusief de fragment ionen waaruit deze ontstaan is. Gedemonstreerd wordt dat door MS^n spectra te representeren als een collectie van EFP's, fragmentatie data verkregen voor hetzelfde metaboliet of verschillende metabolieten met elkaar vergeleken kunnen worden. Dit concept om spectra te vergelijken is gebruikt om te onderzoeken wat de invloed van de concentratie van een metaboliet is op de reproduceerbaarheid en robuustheid van de verkregen fragmentatie boom. De resultaten laten zien dat de verkregen EFP's reproduceerbaar zijn over een hele reeks van glutathion concentraties (van 1 tot 1000 μM). Het aantal waargenomen EFP's neemt echter wel af bij lagere concentraties. Dit heeft tot gevolg dat voor metabolieten aanwezig in lage concentraties er niet genoeg informatie aanwezig zal zijn in de MS^n spectra om éénduidig een structuur toe te kennen. Het EFP concept maakte het mogelijk om twee isomerische prostaglandinen (PGE en PDD), met gelijkende structuren maar verschillende biologische functies, van elkaar te onderscheiden. Het aantal unieke karakteristieke kenmerken (massa pieken) voor elke prostaglandine neemt toe als functie van de diepte van de MS^n boom, bijv. gaande van MS nivo 5 naar MS nivo 6. Deze studie laat ook zien dat de isolatie wijdte en de botsingsenergie zorgvuldig gekozen moeten worden. Beide acquisitie parameters beïnvloeden hoofdzakelijk de relatieve intensiteit van de fragment ion pieken en als zodanig ook de volledige set van EFP's per metaboliet.

De eerste stap in de opheldering van de structuur van een onbekende metaboliet (of verbinding) is de bepaling van de atoomsamenstelling. In Hoofdstuk 3 wordt de ontwikkeling van de 'Multistage Elemental Formula' (MEF) tool beschreven. De MEF tool bewerkt MS^n data zodat schone fragmentatie bomen worden verkregen waardoor er, op een korrekte wijze, atoomsamenstellingen toegekend kunnen worden aan de moleculaire ionen, hun fragment ionen en de 'neutral losses'. De MEF tool reduceert hiervoor op een efficiënte manier de lijst met alle theoretisch mogelijke atoomsamenstelling kandidaten door per ion beperkingen aan te brengen op basis van atoomsamenstellingen van de voorouder ('precursor ion') en de afstammelingen (de fragmenten). Een samenhang werd gevonden tussen de toegestane fout in de massa tijdens acquisitie van de data (bijv. de isolatie wijdte van de ionen in de ion trap) en de topologie (diepte en breedte) van de fragmentatie boom. Gedemonstreerd wordt dat gebruik van MEF in combinatie met MS^n data een lagere massa accuraatheid vereist dan wanneer er MS/MS spectra gebruikt zou worden. Toevoeging van additionele MS^n nivo's verbetert de uiteindelijke bepaling van de atoomsamenstelling. Deze uitbreiding van de fragmentatie boom creëert meer afhankelijkheden tussen elementformule lijsten van de verschillende fragmenten wat weer leidt tot sterkere beperkingen en uitein-

delijk resulteert in opheldering van de korrekte atoomsamenstelling voor MS data verkregen bij een lagere nauwkeurigheid. Dit effect is sterker voor metabolieten met een laag molecuulgewicht of wiens fragmenten lage massa waarden hebben. Dit maakt het ook mogelijk om minder dure massa spectrometers met een minder hoog oplossend vermogen te gebruiken voor de bepaling van de elementformule. De acquisitie van massa fragmentatie spectra is tijdrovend vandaar dat er ook gekeken is van welke fragmenten in de fragmentatie boom bij voorkeur data verzameld moet worden. Aangetoond werd dat de meest informatieve fragmenten op lage fragmentatie nivo's zitten waarbij de voorkeur uitgaat naar fragmenten met een lage massa. De MEF is tevens gevalideerd op situaties waarbij geen correcte elementformule het resultaat was. Onbetrouwbare resultaten werden verkregen wanneer, (i) de in MEF toegestane massa afwijking van de massa kleiner was dan de experimentele massa afwijking, (ii) een elementformule werd toegekend aan een artefact in het massa spectrum i.p.v. een massa piek behorende bij een relevante verbinding, (iii) bepaalde massa pieken verwijderd worden omdat er geen elementformule erbij gevonden wordt maar deze pieken toch behoren bij een relevante verbinding. Voor al deze situaties werden verschillende passende oplossingen gevonden zodat de aanpak geautomatiseerde en betrouwbare toekenning van elementformules toestaat aan alle fragment ionen die onderdeel zijn van een massa fragmentatie boom maar tegelijkertijd artefacten verwijderd (massa pieken die niet behoren bij de relevante verbinding).

Wanneer voor een verbinding de correcte elementformule is toegekend aan alle fragment ionen kan binnen onze MS^n library gezocht worden naar identieke of gelijkende MS^n data. Daarvoor kan dus gebruik worden gemaakt van twee verschillende zoekroutines: een identiteit of een similariteitzoektocht. Terwijl bij een identiteitzoektocht alle kenmerken tussen de te vergelijken spectrale bomen gelijk behoren te zijn, wordt bij een similariteitzoektocht het aantal overeenkomende en afwijkende kenmerken gekwantificeerd.

Hoofdstuk 4 introduceert een nieuwe methode om MS^n data te vergelijken. De methode vergelijkt de aanwezigheid of afwezigheid van bepaalde kenmerken in beide te vergelijken fragmentatie bomen. De kenmerken zijn gedefinieerd overeenkomstig de verschillende manieren waarop fragmenten en neutral losses met elkaar verbonden zijn. Het presteren van de methode wordt gedemonstreerd aan de hand van twee libraries met daarin 867 MS^n spectra van 549 verschillende plant en humane metabolieten. In onze studie vinden wij een éénrichtingsassociatie tussen de chemische structuur van een verbinding en de fragmentatie boom: metabolieten met gelijkende fragmentatie bomen hebben gelijkende chemische structuren terwijl ongelijke fragmentatie bomen komen van metabolieten met niet gelijkende structuur. Deze correlatie is in één richting omdat gelijkende chemische structuren niet altijd lijden tot gelijkende fragmentatie bomen. Een ander aandachtspunt is dat verbindingen die aanwezig zijn in lage concentraties incomplete fragmentatie bomen opleveren en dit prob-

lemen geeft wanneer deze incomplete boom met een database met complete bomen wordt vergeleken verkregen uit hoge concentratie bepalingen. Een oplossing voor dit probleem was het gebruiken van verschillende intensiteitsgrenswaarden tijdens het processen met de MEF tool van de massa spectrale bomen in de database. Op deze wijze komen naast de complete fragmentatie bomen ook gesimuleerde, lage concentratie, incomplete fragmentatie bomen van referentie metabolieten voor in de database. Additioneel werd er een methode ontwikkeld die de maximale overlappende substructuur (Maximum Common Sub-Structure) berekend, gegeven een lijst met structuren die gelijkende fragmentatie bomen vertonen. Op deze wijze kan structuur informatie uit de database gehaald worden zonder dat de onbekende metaboliet in de database zit. Om deze strategie te laten slagen is het belangrijk een database te gebruiken met daarin zoveel mogelijk gelijkende verbindingen als de verbinding die geïdentificeerd dient te worden. In de meest ideale situatie zou gewerkt moeten worden met een database waar aan alle fragment ionen substructuren toegekend zijn.

Vanwege de groeiende interesse in metaboliet identificatie en de behoefte aan makkelijk te gebruiken computationele tools, werd de web-applicatie MetiTree (www.metitree.nl) ontwikkeld (Hoofdstuk 5). MetiTree integreert, op een gebruiksvriendelijke manier, alle tools die ontwikkeld zijn in Hoofdstuk 3 en Hoofdstuk 4 en geeft toegang tot deze tools vanaf elke willekeurige computer via een web browser. Deze web-applicatie helpt bij het processen van MS^n data zodat schone fragmentatie bomen verkregen worden. Fragmentatie boom data zijn complexe data die op een simpele wijze gevisualiseerd dienen te worden zodat de data op een intuïtieve wijze geïnterpreteerd en vergeleken kunnen worden. De ontwikkelde fragmentatie boom afbeelder binnen MetiTree maakt het mogelijk om fragmentatie bomen op een simpele en eenduidige wijze te visualiseren en tegelijkertijd fragmenten te selecteren zodat de chemische structuren van de precursor ion en de fragmenten geanalyseerd kunnen worden. Op elk MS nivo kan de daar optredende reactie bekeken worden. Het is tevens een waardevolle interpretatie tool omdat MS^n data de fundamenteën van fragmentatie reacties die plaatsvinden in een massa spectrometer laat zien. Dit maakt het ook geschikt voor allerlei leerdoeleinden. De ontwikkelde methode ondersteunt ook validatie van nieuw verkregen data. De vergelijkfunctie maakt het mogelijk om eigen MS^n data te vergelijken met eerder geanalyseerde MS^n data om te zien of acquisitie naar behoren is verlopen. Verder helpt MetiTree onderzoekers om metabolieten te identificeren door naar structuren te zoeken met gelijkende MS^n data. Resumerend, MetiTree bevat functionaliteiten die het mogelijk maken om MS^n data te organiseren, bewerken, delen, visualiseren en te vergelijken. In het algemeen zal MetiTree het proces van de-novo identificatie versnellen.

De in deze proefschrift ontwikkelde concepten en methoden faciliteren de extractie van relevante informatie uit MS^n data en helpen bij het identificeren van chemische structuren

van onbekende verbindingen. Het ontwikkelde platform, die alle ontwikkelde methoden integreert, maakt het mogelijk om automatisch onbekende metabolieten sneller en preciezer te identificeren. Samen met andere recente ontwikkelingen zoals de structuur generator (met de mogelijkheid om meerdere substructuren van een verbinding als invoer te gebruiken), interne energy berekeningen, voorspelling van fragmentatie reacties en logP, levert dit een hoog-geautomatiseerde identificatie pipeline welke uiteindelijk een beperkte lijst oplevert met identiteiten. In de gevallen waar teveel identiteiten worden verkregen of waar verschillende mogelijke structuren als uitkomst gegeven worden die moeilijk met te differentiëren zijn met massa spectrometrie daar zou nucleair magnetische resonantie spectroscopie (NMR) gekoppeld aan vloeistof chromatografie (LC) gebruikt kunnen worden om op een gerichte wijze, additionele structuur informatie te krijgen. Met name de recente technische ontwikkelingen op het vlak van LC-NMR (efficiënte koppeling door gebruik solid phase extractie of een druppel verdampingsinterface) hebben de gevoeligheid zichtbaar verbeterd. Het in deze thesis gepresenteerde onderzoek is ook toepasbaar in onderzoeksgebieden buiten metabolomics zoals proteomics of de identificatie van organische moleculen in het algemeen.

Curriculum vitae

Miguel Rojas Chertó was born on 28th of February 1978 in Tortosa, Spain. In 1996 he completed his secondary high education at High school Ramon Berenguer IV in Amposta, Spain. In the subsequent two years he started his chemistry education at The Universidad Nacional de Educación a Distancia. In 1998 he moved to Tarragona (Spain) to continue full time with his studies of chemistry at the Universitat Rovira I Virgili. During his studies he conducted a 1-year ERASMUS internship with the group of Dr. M. Belén Ruiz, in the department of Theoretical Chemistry & Pharmacy at the Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany. He implemented routines to calculate several states of boron by employing a Hyleraas-CI wave function approach. He obtained the MSc of chemistry in 2004. Subsequently he worked as a computational chemistry scientific programmer in several cheminformatics projects in the group of Dr. Christoph Steinbeck at the Cologne University Bioinformatics Center (Cologne, Germany). In 2007 he started his PhD research in the Division of Analytical Biosciences group at the Leiden University, The Netherlands. Under the supervision of Prof. dr. Thomas Hankemeier, he worked on a project to develop a novel semi-automatic strategy for the identification of human metabolites in body fluids and tissues using multi-stage mass spectra. This was a collaborative project within the Netherlands Metabolomics Centre. Currently he is working as an international country project manager at CULTIDELTA, a company specialized on native and low-maintenance plants worldwide.

List of publications

First authorship:

Rojas-Chertó, M., Kasper, P. T., Willighagen, E. L., Vreeken, R., Hankemeier, T., & Reijmers, T. (2011). Elemental Composition determination based on MSn. *Bioinformatics*, 27(17), 2376-2383.

Rojas-Chertó, M., Peironcely, J. E., Kasper, P. T., Van der Hooft, J. J. J., De Vos, R. C. H., Vreeken, R., Hankemeier, T., et al. (2012). Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. *Analytical chemistry*, 84(13), 5524-34.

Rojas-Chertó, M., Van Vliet, M., Peironcely, J. E., Van Doorn, R., Kooyman, M., Te Beek, T., Van Driel, M. A., et al. (2012). MetiTree: a web application to organize and process high-resolution multi-stage mass spectrometry metabolomics data. *Bioinformatics (Oxford, England)*, 28(20), 2707-9.

Rojas-Chertó, M., Kasper, P., Julio E, P., Reijmers, T., Vreeken, R., & Hankemeier, T. (2010). The pipelined metabolite identification based on MS fragmentation. *Journal of Cheminformatics*, 2(Suppl 1), P53.

Co-authorship:

Peironcely, J. E., Rojas-Chertó, M., Tas, A., Vreeken, R., Reijmers, T., Coulier, L., & Hankemeier, T. (2013). Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics. *Analytical chemistry*, 85(7), 3576-83.

Peironcely, J. E., Rojas-Chertó, M., Fichera, D., Reijmers, T., Coulier, L., Faulon, J.-L., & Hankemeier, T. (2012). OMG: open molecule generator. *Journal of Cheminformatics*, 4(1), 21.

Kasper, P. T., Rojas-Chertó, M., Mistrik, R., Reijmers, T., Hankemeier, T., & Vreeken, R. J. (2012). Fragmentation trees for the structural characterisation of metabolites. *Rapid communications in mass spectrometry : RCM*, 26(19), 2275-86.

Kloet, F. M., Tempels, F. W. A., Ismail, N., Heijden, R., Kasper, P. T., Rojas-Chertó, M., Doorn, R., et al. (2012). Discovery of early-stage biomarkers for diabetic kidney disease using ms-based metabolomics (FinnDiane study). *Metabolomics*, 8(1), 109-119.

Peironcely, J. E., Bender, A., Rojas-Chertó, M., Reijmers, T., Coulier, L., & Hankemeier, T. (2009). Expanding and Understanding Metabolite Space. *Journal of Cheminformatics*, 2(Suppl 1), P39.

Acknowledgments

It is a challenge to mention all the people who, directly or indirectly, contributed to the realization of this thesis.

My interest in the research conducted for this thesis started actually in Germany. I am grateful to Dr. Christoph Steinbeck for introducing me to the field of cheminformatics. Together with all persons from the CUBIC team, it was fun to work when our projects were coming together.

I am very grateful to many colleges at Analytical Biosciences who have taught me about how important and nice it is to work in a team. They have all maintained a very nice environment at work. I would like to mention the super-lab-team: Bas, Jan-Willem, Jos, Kjeld, Peter, Marek, Katrin, Marco, Shanna, Gerwin, Jorne, Herman, and the staff members Rob and Heiko. I am also grateful to Piotr for working side by side on the identification of metabolites: we achieved a lot if we look back what was there when we started. Thanks also to Maya; they both have been a great help answering patiently each of my questions and doubts about mass spectrometry and its data. We three were the starters of the NMC group in Leiden and together we formed a nice team. Furthermore, I am especially grateful to Michael for the long hackings together. And of course Loes for taking the effort to arrange every bureaucratic paperwork I needed. I also enjoyed the interesting collaboration with Justin and Ric from Wageningen. It is interesting to see how two groups with different specialism (human and plant) can tackle different challenges with the same project.

Theo, without your ideas and your patience neither my project nor our manuscripts would have seen the light. Thomas, your vision that identification could be improved encouraged me to work on this also when I encountered many challenges.

One thousands thanks to two impressive Catalan guys, Julio and Jordi. They made my time in Leiden enjoyable and we developed a nice friendship. Specially Julio, whom after the long philosophical discussion to improve this world and the group, we achieved small changes.

During this time I discovered the loveliest surprise of my life, although separated by some distance. Always supporting and pushing me to finish this thesis. Lidia, t'estimo dos munts.

Finally, I would like to thank my parents. They encouraged me every time to try to achieve all my dreams and they supported me when the times are difficult. They taught me that with effort, dedication, and believe I can achieve it. That is the reason I am today here. Gràcies Papa i Mama, per estar sempre amb mi. Encara que hem estan a més de 2000km de distancia en dels altres, us porto sempre al meu cor. Us estimo moltíssim.
