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

**Author:** Rojas-Chertó, Miquel  
**Title:** Towards automated identification of metabolites using mass spectral trees  
**Issue Date:** 2014-06-19
CHAPTER 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\textsuperscript{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\textsuperscript{n} spectrometry. The aim of this thesis was to develop a novel semi-automatic approach for the identification of metabolites using MS\textsuperscript{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\textsuperscript{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\textsuperscript{n} data are known as being very complex. In Chapter 2 the development of a MS\textsuperscript{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\textsuperscript{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\textsuperscript{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 uM). 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\textsuperscript{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 an 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\textsuperscript{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\textsuperscript{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 MetiT ree web application (http://MetiT ree.nl) was developed (Chapter 5). MetiT ree 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 MetiT ree 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. MetiT ree helps researchers to identify metabolites of interest by finding similar MS$^n$ data. In summary, MetiT ree 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.