The handle http://hdl.handle.net/1887/28508 holds various files of this Leiden University dissertation.

**Author:** Peironcely Miguel, Julio Eduardo  
**Title:** Automated de novo metabolite identification with mass spectrometry and cheminformatics  
**Issue Date:** 2014-09-03
6 Conclusions and perspectives
Conclusions and perspectives

Metabolite identification is still one of the biggest bottlenecks in metabolomics. The de novo structure elucidation of metabolites is very time consuming and poses several challenges. However, without the identity of a peak detected with mass spectrometry-based metabolomics methods its biological role cannot be interpreted. In addition, effective integration with other ‘omics’ data such as genomics and proteomics requires the identity of metabolites. Therefore there is a huge need for more efficient strategies to identify metabolites. Multi-stage mass spectrometry (MS\textsuperscript{n}) is a promising type of mass spectrometry from which information on the fragmentation pattern of the metabolite can be obtained.

In this thesis new algorithms and methods that enable the de novo identification of metabolites have been developed. The aim was to find methods to propose candidate structures for unknown metabolites using MS\textsuperscript{n} data as starting point. Ideally this list of candidate identities should be as short as possible by using additional constraints so that an expert can easily inspect it and select some structures for further validation. The algorithms and methods, which have been developed in this thesis and a parallel project, have been integrated to into a semi-automated pipeline to identify new metabolites starting from multi-stage mass spectrometry. The focus in this thesis was on human metabolites. The discovery of new metabolites will improve our capability to understand disease via its metabolic fingerprint, to develop personalized treatments and to discover new drugs. In
addition, the cheminformatics methods presented in this thesis increase our understanding on the properties of human metabolites.

In Chapter 2 a structure generator, the Open Molecule Generator (OMG), was developed. This tool allowed generating candidate structures for unknown metabolites with a certain elemental composition and known substructure(s). While research in computer assisted structure elucidation (CASE) dates back to the 1960s, recent developments have been scarce. The most notable example was MOLGEN, an efficient, but at the same time “black box” commercial structure generator. Therefore the in Chapter 2 developed structure generator was customizable and open source: this allowed to implement methods and algorithms that were relevant to the envisioned identification pipeline as described in Chapter 5. Being open source, OMG permits other scientists to improve it further and to adapt it to future needs.

The efficiency of the generation of structure candidates for a given elemental composition (EC) and substructure(s) was demonstrated for a number of metabolites. The results showed that OMG produced all possible chemical structures for a given input (EC and/or substructure(s)). These results only contained chemically valid structures according to the valence rule, whereas MOLGEN could produce some unwanted atom types: for instance, for P valence 5, MOLGEN would generate P atoms with a triple and a double bond. OMG allowed also to use several substructures as constrain, whereas MOLGEN allows only one substructure as constrain. The use of prescribed substructures constrained significantly the number
of candidate structures obtained. They reduced for example a list of hundreds of millions of molecules (unconstrained generation) to a few thousand structures (constrained generation). While this is a significant reduction, reducing the list further using additional constraints appeared essential to turn the identification of unknowns into a tractable problem. In conclusion, a structure generator was developed which was superior to MOLGEN in several aspects. However, OMG was slow when generating many molecules. The reason was that the algorithm used in OMG was based on the canonical augmentation path. This algorithm produces all possible molecules without duplicates, but it requires the use of a canonizer for duplicate removal. This canonizer calculates the canonical representative for each intermediate (unfinished) molecule. This conversion from a molecule to its canonical representative is computationally time expensive. We concluded that a better algorithm with a better duplicate removal approach was required to speed up calculations and make the tool more attractive for practical use. Such an algorithm was developed in Chapter 5.

In Chapter 3 we studied the nature of human metabolites. The aim was two-fold: (i) to learn what characteristics differentiate metabolites from other small (< 1000 Da) non-metabolite molecules and (ii) to predict the metabolite-likeness of a chemical structure, i.e. how likely it is to be a human metabolite. In the classification of molecules, as in many machine-learning problems, one has to deal with an interpretability trade off. While easy to understand molecule descriptions (such as physicochemical properties or scaffolds) combined with easy to interpret predictive models mostly provide poor predictive results, complex descriptions (structural
fingerprints) together with difficult to interpret models provide often better predictions. In other words, if we want to accurately predict metabolite-likeliness we will probably not be able to easily understand in chemical terms what these predictions are based on.

We observed that metabolites are a heterogeneous family of compounds and compared to non-metabolites, metabolites have simpler structures, less ring systems, more hydrophilic groups, less nitrogen and sulfur atoms, and more oxygen and phosphor atoms. These easy to interpret features were not complex enough to be used in a model that would discern between metabolite and non-metabolite molecules. Therefore, we developed and validated a metabolite-likeliness predicting model, which used the molecular descriptor MDL Public Keys and the Random Forest classification algorithm to assign a metabolite score to a molecule. This model achieved the highest classification accuracy at the expense of low interpretability. The model was effectively used in Chapter 4 to reject non-metabolite candidate structures for unknown metabolites, demonstrating that metabolite-likeliness prediction is one of the tools that can be routinely used in metabolite identification studies.

In Chapter 4 the tools developed in Chapters 2 and 3 of this thesis and developed in a parallel project within the Netherlands Metabolomics Centre [69–71] were integrated into a pipeline to identify metabolites. This pipeline was composed of four modules. The first module annotated MS^n data to obtain a fragmentation tree comprising fragment ions and neutral losses of known elemental composition using the
Multistage Elemental Formula (MEF) tool [69]. The second module compared this fragmentation tree of an unknown with the fragmentation trees of known metabolites in a home-build database. If a fragmentation tree with 100% similarity was found, the identity of the unknown was provisionally assigned. If more than one fragmentation tree was found with less than 100% similarity, a substructure of that unknown metabolite was calculated via the maximum common substructure (MCSS) from the metabolites being similar to the fragmentation tree. The third module was the OMG structure generator, using as input the elemental composition and, if found, a substructure of that metabolites. The fourth module used three filters to reduce the list of candidates generated by OMG: (i) a Metabolite-likeness filter, which kept only molecules resembling human metabolites, (ii) a internal energy filter, which kept only energetically stable molecules and (iii); the MetFrag filter, which predicted the mass spectral fragmentation of molecules and kept only those candidate molecules that could explain half of the fragments observed experimentally.

This metabolite identification pipeline was tested for the identification of 30 different MSn spectra obtained from unknown metabolites in a human urine sample. For 10 of the 30 unknowns a perfect match of the acquired fragmentation tree and a fragmentation tree of a known metabolite in the database was found, and therefore, these metabolites were provisionally identified. For 9 unknown metabolites two or more similar metabolites were found in the MS^n database, which allowed the calculation of a maximum common substructure as input to constrain the number of structures obtained by the OMG structure generator. For 3 out of these 9 unknowns,
OMG and the different filters provided a short list of 8 or less candidate structures. For 6 out of these 9 unknowns, the list of candidates was excessively large, i.e. larger than 40 candidates. Lastly, 11 unknowns remained unidentified because no similar metabolite was found in the database and therefore OMG using only the elemental composition provided more than one million of candidate structures.

In summary, the developed identification pipeline proved to be useful at identifying unknown metabolites using only MS^n data. If a metabolite is already in a MS^n database, the metabolites can be well provisionally identified based only on the MS^n spectrum; obviously, in this manner only metabolites already known can be identified. In order to identify truly new metabolites, i.e. de novo identification, one needs to produce a short list of candidate structures for a given fragmentation tree, and that is only possible if sufficient substructure information is available and powerful filters are used. This actually depends heavily on the availability of comprehensive MS^n database of known metabolites from which similar metabolites can be found and a maximum common substructure can be derived. It would be very beneficial if annotated subtrees in the database with their corresponding substructures would be available. This would allow finding multiple annotated subtrees with their substructures for a MS^n tree of an unknown metabolites in such a database. This would provide multiple prescribed substructures as input for the structure generator. Such multiple substructures as input are only possible using manual interpretation of a MS^n tree. The three filters used (Metabolite-Likeness, internal energy and fragmentation prediction) proved to be useful at reducing the number of candidates to an amount that could be inspected by an expert for further
validation. Obviously this identification pipeline for known and fully unknown molecules (i.e. not reported in any database) does not only apply to human metabolites, but can be applied also to plant metabolites or other types of molecules.

In Chapter 5, we implemented and tested the Parallel Molecule Generator (PMG). This structure generator addresses some of the lessons learned in Chapter 2 (OMG) and Chapter 4 (metabolite identification pipeline). Firstly in both chapters we observed that using multiple substructures as constraints could make many identification problems feasible using the in Chapter 4 developed identification pipeline. Secondly, OMG should be improved by using a faster algorithm, by reducing the need for a full canonizer, i.e. use a less computationally demanding method to remove duplicate structures, and by parallel execution of the algorithm. Lastly, in Chapter 4 we have learned that filters like removing energetically unfavorable structures are important and they should be already incorporated in the structure generating process. The rationale for implementing such filters in PMG is that providing a short list of candidates produced slowly is more desirable than a long list of candidates produced quickly, since a shorter list of candidates brings us closer to the identification of the unknown metabolite. We achieved a reduction in the number of possible candidate structures for a given elemental composition by including several constraints, i.e. using prescribed substructures, good and bad lists, exclusion of energetically unfavorable bad rings. In addition, we reduced the impact on the computational time of including these constraints in the method by implementing two new algorithms. These run in parallel, therefore they produce
Conclusions and perspectives

results faster, they can accommodate more constraints and their results are as complete as those the original OMG algorithm.

The increase in speed using PMG compared to the OMG was evaluated for the same elemental compositions and associated substructures as we used for the validation of the OMG in Chapter 2. In terms of generating molecules using a single core computing, PMG provided all possible structures about 40-fold faster compared to OMG with elemental compositions as only input. The speed increases was even up to 100-fold for PMG compared to OMG when also prescribed substructures were provided as input. The time to generate molecules could be further reduced by executing PMG in multiple cores, which OMG does not allow, represents an almost linear speed up increase as more cores are added.

The efficiency in reducing the number of structures obtained with the additional constrains was tested for the unknown metabolites found in human urine of Chapter 4. An expert annotated manually 30 MS^n spectral trees, of unknown metabolites found in human urine, which provided substructures for the good list. For the unknowns for which the elemental composition allowed structures with rings, the use of all additional constraints as introduced in PMG removed up to 82% of the candidate structures. In conclusion, PMG is a further improvement in the efficient generation of candidate structures for a given elemental composition, substructures and using additional constraints. We expect that the open source nature of PMG allows further improvements by other researchers, especially when more knowledge over the type of molecules to be identified is known.
Metabolomics is a growing field that still suffers from some limitations specific for metabolites, and some limitations that are also observed for other ‘omics’ areas such as proteomics. So far no generic procedure is available that allows the identification and quantification of all metabolites present in a sample compared to sequencing of a full genome, which is possible for currently just a few thousands of euros. One challenge is that databases containing metabolites and experimental data such as MS, MS^n, and NMR spectra of metabolites have been for many decades kept in-house of companies or research groups, and the available databases are containing only a fraction of the metabolites being expected. Fortunately, more international consortia are being established to tackle the challenges in metabolomics.

The research described in this thesis has shown that the success of de novo metabolite identification relies on the synergy between analytical chemistry methods (i.e. LC-MS^n) and cheminformatics tools. It can be expected that the analytical instrumentation and methods will further develop and faster methods will require less amount of sample and will detect more metabolites, for which masses and fragment ions will be detected with mass spectrometry with more accuracy and better reproducibility. The key factor for the success of MS as a standard technique for metabolite identification is its ability to produce substructure information for many analytes. Important will be also to obtain MS^n spectra on-the-flight, i.e. without the need of fractionation prior to direct infusion into the MS.
Conclusions and perspectives

In this thesis it has been presented that knowing the elemental composition and certain substructures of an unknown metabolite allows to limit the number of possible structures of that unknown. In this thesis the concept of the maximum common substructure (MCSS) from similar $M^S_n$ spectra was used to derive a substructure for the unknown metabolites. However, better alternative to determine substructures in an unknown metabolite should be developed as there are cases that not the correct substructure was obtained, or the substructure was too small. A better alternative could be to relate shared branches or subtrees among metabolites with shared substructures, i.e. the building block principle. Or, even better, an $M^S_n$ database of known metabolites should have annotated the branches and subtrees of the metabolites with their corresponding substructures. In such an approach, matching a subtree of the unknown metabolite with and annotated subtrees in the database will provide a higher confidence that the unknown contains that substructure(s). And as shown in this thesis the more substructures are used as constraints for the structure generator the fewer candidate structures are obtained. Having an $M^S_n$ database with tens of thousands of annotated metabolites would allow the identification of metabolites in a semi high-throughput fashion.

Generally high benefits can be expected from the computerization of human expertise, as was for example demonstrated for chess playing supercomputers and artificial intelligence software. However, for metabolite identification in the current situation also for the in this thesis developed pipeline the input of a human expert is still required. Human expertise is necessary at different steps of the metabolite identification process. Humans are required to evaluate the correctness of the
analytical data acquired from biological samples. It does not matter how good a software pipeline is, if the initial input is of bad quality, the output will be of bad quality. Software helps us at performing repetitive tasks more efficiently. Where software underperforms humans is at detecting anomalies in a pipeline. An expert can use tools like common sense and intuition to detect that results are overly optimistic, on the one hand, or to focus on those candidates that have more chances of succeeding. However, it can be expected that with the further development of the identification pipeline the required input from a human expert will become less and less over time.