Genomics driven metabolomics novel strategies for the discovery and identification of secondary metabolites
Ries, M.

Citation

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/1887/24303

Note: To cite this publication please use the final published version (if applicable).
The handle http://hdl.handle.net/1887/24303 holds various files of this Leiden University dissertation

**Author:** Ries, Marco  
**Title:** Genomics driven metabolomics: novel strategies for the discovery and identification of secondary metabolites  
**Issue Date:** 2014-02-25
Chapter 5

Chemoinformatics supported MS^n Comparison Pipeline (CMCP): Towards automated *de novo* structure elucidation using multiple-stage fragmentation tree comparison

*Based on*

*Chemoinformatics supported MS^n Comparison Pipeline (CMCP): Towards automated de novo structure elucidation using multiple-stage fragmentation tree comparison*

*In preparation for publication*
Abstract

Here we present a novel multiple stage fragmentation tree based structure elucidation pipeline for the de novo structure identification of small molecules, coined CMCP (Chemoinformatics supported MS^n Comparison Pipeline) which combines sample extraction, fragmentation tree acquisition, database screening and data interpretation. Based on high-resolution multiple-stage fragmentation mass spectra (MS^n), fragment and neutral-loss trees are generated and compared to a MS^n database to extract compounds with similar fragmentation behavior. By identifying their shared fragmentation mechanisms, chemical structures of unknowns can be deduced without them needing to be present in the database. As the structure elucidation presented here is based on structural identification of similar fragments and neutral-losses rather than solely their comparison, confident identification of unknown molecular structures is possible beyond classifications into compound classes. Most important conceptual outcomes of a database query using fragment and the newly introduced neutral-loss tree comparison are discussed showing how this information can ultimately be used to deduce complete chemical (sub)structures of unknown compounds. For their demonstration, the de novo identification of the complex metabolites roquefortine C, dehydrohistidyltryptophanyldiketopiperazine (DHTD), roquefortine F and roquefortine D extracted from cultures of Penicillium chrysogenum is described in detail.
Introduction

With the prevalent application of metabolomics, several methods and techniques for the simultaneous analysis of a large number of metabolites have been developed which provide detailed information about the metabolome of complex biological samples. However, their structural identification, necessary to describe and interpret cellular processes, is still a major bottleneck in metabolomics (Kind and Fiehn, 2006).

Mass spectrometry (MS) in combination with chromatographic separation techniques is widely used for the analysis of metabolites due to its high sensitivity, low demand on sample purity and rapid analysis time (Villas-Boas, et al., 2005). Especially liquid chromatography MS (LC-MS), which is increasingly used in recent years, allows to analyze relative polar and thermally unstable compounds without prior derivatization. Different types of MS instruments are available to fragment ionized molecules and detect their charged fragments providing structural information. Using these instruments, several MS/MS databases and software tools have been generated in order to facilitate metabolite identification (Horai, et al., 2010; Wishart, et al., 2009). However, most approaches are limited to identity searches, in which the unknown compound needs to be present in the database or to the assignment of unknowns into compound classes based on similar fragmentation (Rasche, et al., 2012; Sheldon, et al., 2009). This leaves the complete structure elucidation of unknowns still a major challenge.

Compared to MS/MS approaches in which fragments can be immediately fragmented further after generation, MS\(^n\) methods provide a more detailed fragmentation pathway. As only precursor ions are exited, fragments cool after formation, thus avoiding further fragmentation giving deeper insight into relationships between fragments (Stein, 2012). With the use of such a multidimensional mass spectrometry approach spectral trees can be generated containing precursor-product ion relationships (Sheldon, et al., 2009). Based on correlations between specific fragments in the MS\(^n\) spectra and the substructure of the measured molecule, structural elucidation of unknown compounds can be performed (Chen, et al., 2002; Fandino, et al., 2002; Wang, et al., 1999). However, as extensive databases and software tools are scarce, MS\(^n\) data handling and interpretation is mainly still limited to the manual identification of characteristic fragments from defined classes of compounds (Cui, et al., 1999; Kang, et al., 2007; Rochfort, et al., 2008), leaving the full power of multiple stage fragmentation mass spectra untapped. As a consequence our group recently extended the MEF (mass elemental formula) tool (Rojas-Cherto, et al., 2011; Rojas-Cherto, et al., 2012) for processing and comparing MS\(^n\) data which extracts ion-signals from MS\(^n\) spectra, followed by assignment of their corresponding elemental composition.

In previous work, the functionality of the MEF tool has been demonstrated for the determination of similarities between database entries (Rojas-Cherto, et al., 2012) and the tentative identification of small molecules (Peironcely, et al., 2013). Howe-
ver, the conserved structural information of the analyte, reflected by specific fragmentation mechanisms and the chemical structure of fragments and neutral-losses, was not exploited.

Here, we present a chemoinformatics guided pipeline for the structural identification of unknown compounds coined CMCP (Chemoinformatics supported MS\(^n\) Comparison Pipeline). By combining small scale sample extraction, fragmentation tree acquisition and the extended MEF tool with MS expert knowledge, the chemical structure of low concentrated complex molecules can be identified using solely mass spectrometry (Figure 1). We discuss the most important conceptual outcomes of a database query using CMCP and demonstrate how this information can ultimately be used to deduce complete chemical structures for unknown compounds.

![Figure 1. Schematic view of the CMCP pipeline for the identification of unknown compounds in a biological sample and structures of identified compounds 1 – 4.](image)

**Experimental procedures**

**Strain and culture conditions**

*Penicillium chrysogenum* strain DS54555, which lacks the *Ku70* and penicillin cluster genes was supplied by DSM Anti-infective (Delft, The Netherlands). Cells were grown on SMP medium (glucose, 5.0 g/L; lactose, 75 g/L; urea, 4.0 g/L; Na\(_2\)SO\(_4\), 4.0 g/L; CH\(_3\)COONH\(_4\), 5.0 g/L; K\(_2\)HPO\(_4\), 2.12 g/L; KH\(_2\)PO\(_4\), 5.1 g/L) using a shaking incubator at 200 rpm for 168 hours at 25°C.

**Small scale isolation of analytes**

Low amounts of the four structurally complex metabolites 1 – 4 (Figure 1), were ex-
tracted from liquid cultures of *P. chrysogenum* using the LC-UV-MS method described elsewhere (Ali, et al., 2013). Briefly, 230 µL of methanol was added to 50 µL of *P. chrysogenum* culture filtrate for protein precipitation. The sample was vortexed, centrifuged and 100 µL supernatant transferred to an Eppendorf vial which was subsequently evaporated in a Thermo-Speedvac (Thermo Scientific, San Jose, CA). The dried sample was re-dissolved in 100 µL water containing 2 % acetonitrile, vortexed and transferred to an autosampler vial. For separation, an Agilent 1200 Capillary pump (Agilent, Santa Clara, CA) coupled to a Surveyor PDA detector (Thermo Scientific, San Jose, CA) and LTQ-FT Ultra mass spectrometer (Thermo Scientific, San Jose, CA) was used. A sample of 5 µL was injected onto a Waters Atlantis T3 column (2.1 x 100 mm, 3 µm) (Waters, Milford, MA). The elution was performed with a linear gradient from 98 % of solvent A (1 % acetonitrile and 0.1 % formic acid in water) and 2 % solvent B (1 % water and 0.1 % formic acid in acetonitrile) to 100 % of Solvent B in total of 25 minutes (first 1.5 minutes isocratic at 98% A, then to 40% of solvent B at 22 minutes and 10% solvent B at 25 minutes) at a flow rate of 300 µL/min. The column was flushed for 10 minutes at 100% B followed by equilibration for 8 minutes at 100% A. Fractions of 1, 2, 3 and 4 were collected manually and subsequently evaporated to dryness in a Speedvac.

**Preparation of Working Solutions**

Working solutions were prepared by re-dissolving the dried fraction of each metabolite in 20 µL water containing 50% acetonitrile and 0.1% formic acid. Right before acquisition, working solutions were mixed with 5 µL isopropanol in a 384 well plate to increase nano-electrospray stability (Eppendorf, Hamburg, Germany).

**Fragmentation tree acquisition, processing and database comparison**

Fragmentation tree acquisition was carried out on a LTQ-Orbitrap-XL mass spectrometer (Thermo Fisher Scientific, Waltham, MA) equipped with an Advion Nanomate nano-electrospray source (Advion, Ithaca, NY) (1.5 kV source voltage, 120°C capillary temperature, 5 V capillary voltage, 70 V tube lens voltage and 35 % normalized collision energy for CID fragmentation) according a modified protocol described elsewhere (Kasper, et al., 2012). The scan rate of the mass spectrometer was set to 6 micro scans. At least three complete fragmentation trees were acquired per compound. The ten most abundant fragment ions in MS², the nine highest in MS³ and three highest in MS⁴ were isolated and further fragmented resulting in 371 (1 MS², 10 MS³, 90 MS⁴ and 270 MS⁵) possible fragmentations. Isolation width was set to 1.5 m/z and minimal precursor intensity required for further fragmentation was fixed to 50,000 counts. Acquired Thermo Xcalibur files were converted into mzXML format using ReadW software (Pedrioli, et al., 2004). Chemical formulas were assigned to all fragments of the acquired tree using the MEF tool (Rojas-Cherto, et al., 2011). MEF processing parameters were as following: Signal-to-noise ratio 1, mass accuracy 10 without applying the nitrogen rule or degree of unsaturation (Pellegrin, 1983). Finally, all processed trees were compared to an in-house metabolite MS⁰ database, containing more than different 400 entries, to find similar fragments and neutral-loss patterns (Rojas-Cherto, et al., 2012).
Chapter 5

Results

Fragmentation tree generation and similarity search
To demonstrate the different conceptual outcomes of a database query and to show how this information can be used to deduce structural information, the *de novo* identification of the four unknown metabolites roquefortine C (1), dehydrohistidyl-tryptophanyl diketopiperazine (DHTD) (2), roquefortine F (3) and roquefortine D (4) is described in detail (Figure 1). Compounds 1 - 4 were obtained, next to six additional compounds, from comparative metabolite profiling of host and various deletion strains of *P. chrysogenum*, using high resolution mass spectrometry (HPLC-HR-MS) (Ali, et al., 2013). Their mass was determined as 390.192 (1), 322.129 (2), 420.203 (3) and 392.208 (4) dalton, corresponding to the protonated molecules with formula \( C_{22}H_{24}N_5O_2 \), \( C_{17}H_{16}N_5O_2 \), \( C_{23}H_{26}N_5O_3 \) and \( C_{22}H_{26}N_5O_2 \). As none of the ions had a corresponding entry in a public MS database like MassBank (Horai, et al., 2010), HMDB (Wishart, et al., 2009) or PRIMe (Akiyama, et al., 2008; Sakurai, et al., 2013) and because the biosynthetic mechanism of their producing gene cluster was not known, additional structural information was not available leaving their structure completely unknown. In order to acquire multiple stage fragmentation mass spectra, small scale fraction collection was performed to extract low amounts of pure analytes from *P. chrysogenum* culture broth. Subsequently, fragmentation trees were acquired and processed according the description in the method section. The resulting fragment and neutral-loss trees were compared to an in-house metabolite MS* database, which was searched for similar mass fragments and neutral-losses present in database entries. Database entries were ranked according their degree of similarity, represented by the Tanimoto coefficient (Flower, 1998), yielding the database entry of protonated roquefortine C with formula \( C_{22}H_{24}N_5O_2 \) as most similar compound to all four metabolites.

Structure elucidation based on identical fragment and neutral-loss trees
With the precursor and almost all fragments and neutral-losses of 1 being present and similarly arranged as in the fragment and neutral-loss tree of roquefortine C, an identical chemical structure of 1 and roquefortine C was indicated (Figure 2). This was supported by a comparison of their multiple stage fragmentation spectra which showed fragments with almost identical mass-over-charge ratios and relative intensities over multiple fragmentation stages (data not shown). Based on the known structure of the database entry, the structure of 1 was tentatively assigned as roquefortine C which was ultimately confirmed using HPLC-MS/MS comparing the retention time of 1 and a roquefortine C standard.

Structure elucidation based on similar fragment and neutral-loss trees
For compound 2, the database query returned partial similarity to its most similar database entry roquefortine C with almost all fragments and neutral-losses of compound 2 present and similarly arranged as in the database entry (Figure 3, Supplemental Figure 1 and 2). However, compared to the identification of compound 1, in which compound 1 and the database entry showed the same ion in MS1, here different ions were observed, respectively \( m/z \) 322 (ID 1, MS1, \( C_{17}H_{16}N_5O_2 \)) in compound 2 and \( m/z \) 390 (ID 1, MS1, \( C_{22}H_{24}N_5O_2 \)) in roquefortine C.
For further structure elucidation, the two highly similar subtrees of fragment m/z 322 (ID 28, MS², C₁₇H₁₆N₅O₂) in roquefortine C and m/z 322 (ID 1, MS¹, C₁₇H₁₆N₅O₂) in compound 2 were used. Their similarly arranged fragments and neutral-losses, containing an identical elemental composition, indicated that the fragment with m/z 322 originating from the precursor ion m/z 390 (ID 1, MS¹, C₂₂H₂₄N₇O₂) in roquefortine C has an identical structure as the ion m/z 322 of 2, which represents the protonated molecule in MS¹. This was further supported by comparing the multiple stage fragmentation spectra of their common subtrees which showed fragments with almost identical mass-over-charge ratios and relative intensity over multiple fragmentation stages (Figure 4). By identifying the chemical structure of the fragment m/z 322 in roquefortine C, formed by a loss of isoprene, and transferring the information to 2, the structure of protonated 2 could be determined as dehydrohistidyltryptophanyldiketopiperazine (DHTD) (Figure 5). Its tentative structure was ultimately confirmed by NMR experiments (Ali, et al., 2013).

An example in which the fragmentation tree of a database entry is a complete subtree of an unknown compound is the identification of compound 3, which shares a common subtree with its most similar database entry roquefortine C, initiated by
Chapter 5

Figure 3. Fragment tree comparison of compound \( m/z \ 108 \) and its most similar database entry roquefortine C (right) revealed that the entire tree of C is a subtree of roquefortine library ID 88. C. The shared fragmentation path of the most abundant fragment per section is highlighted by colored arrows and shown in more detail in the middle of the figure.
Chemoinformatics supported MS\textsuperscript{n} Comparison Pipeline

Chapter 5

ion m/z 390 (ID 1, MS\textsuperscript{1}, C\textsubscript{22}H\textsubscript{24}N\textsubscript{5}O\textsubscript{2}) in roquefortine C and m/z 390 (ID 100, MS\textsuperscript{2}, C\textsubscript{22}H\textsubscript{24}N\textsubscript{5}O\textsubscript{2}) in 3 (Supplemental Figure 3 and 4). As the complete tree of roquefortine C is present in the tree of 3, including the precursor ion in MS\textsuperscript{1}, an identical structure for the ion m/z 390 in both compounds is concluded. This is supported by identical fragmentation spectra over multiple fragmentation stages initiated by a fragmentation of the ions at m/z 390 in both compounds (Supplemental figure 5). Compared to roquefortine C, compound 3 contains one additional carbon, two additional hydrogens and one additional oxygen which were cleaved off during first fragmentation yielding ion m/z 390 (ID 100, MS\textsuperscript{2}, C\textsubscript{22}H\textsubscript{24}N\textsubscript{5}O\textsubscript{2}) with an identical structure as protonated roquefortine C. Supported by the competing losses of CH\textsubscript{3}• and CH\textsubscript{4}O from a fragmentation of m/z 420 (ID 1, MS\textsuperscript{1}, C\textsubscript{23}H\textsubscript{26}N\textsubscript{5}O\textsubscript{2}) yielding the fragments m/z 389 (ID 46, MS\textsuperscript{2}, C\textsubscript{22}H\textsubscript{23}N\textsubscript{5}O\textsubscript{2}) and m/z 388 (ID 29, MS\textsuperscript{2}, C\textsubscript{22}H\textsubscript{23}N\textsubscript{5}O\textsubscript{2}), the structure of the loss producing m/z 390 could be determined as formaldehyde originating from the cleavage of a methoxygroup (Stevigny, et al., 2004). Therefore, a roquefortine C structure is concluded for compound 3 with a methoxygroup attached. Its position in the structure of 3 is unclear, as the methoxygroup is cleaved off in the first

Figure 5. Proposed fragmentation mechanism of database entry roquefortine C yielding compound 2 in situ in the mass spectrometer. After an initial protonation, the bond between C2 and N14 is cleaved in roquefortine C followed by a 1,4-hydride shift leading to the loss of isoprene, yielding m/z 322.
Figure 6. Neural-loss tree comparison of compound 4 (left) and regorafenib (right).showing highly similar neural-loss cascades present in both trees. The shaded neural-loss path used for the denotation of 4 is shown in the middle in more detail. Detailed neutral-loss trees of regorafenib C and 4 obtained from neutral-loss tree comparison can be found in the supplement.

Database entry not present.

compound 4

databases entry

Level nodes
Chemoinformatics supported MS\(^n\) Comparison Pipeline

Chapter 5

fragmentation stage. Using NMR experiments, the structure of 3 was determined as roquefortine F, containing the structure of roquefortine C with a methoxygroup on the nitrogen in the indol part of the molecule which is in good agreement with the results obtained from CMCP (Chapter 3).

It should be pointed out that although the protonated form of 2 and roquefortine C share the identical ion \(m/z\) 322, their subtrees differ significantly in depth and width. Therefore, not all fragments of \(m/z\) 322 (ID1, MS\(^1\)) in 2 could be found in the subtree of \(m/z\) 322 (ID28, MS\(^2\)) in roquefortine C. This is due to the fact that next to different observed absolute intensities of the ions, two ions originating from different fragmentation stages are compared. While for 2, ion \(m/z\) 322 was found in MS\(^1\), for roquefortine C it was found in MS\(^2\). Due to the limitations of the acquisition protocol, the further fragmentation of an ion in MS\(^1\) can result in maximal 371 fragmentations (1 MS\(^3\), 10 MS\(^4\), 90 MS\(^5\), 270 MS\(^6\)) whereas the further fragmentation for an ion in MS\(^2\) is limited to maximal 37 fragmentations (1 MS\(^3\), 9 MS\(^4\), 27 MS\(^5\)). As a consequence, a much richer fragmentation tree was acquired for the ion \(m/z\) 322 in compound 2. Similar results were obtained for the identical ion \(m/z\) 390 present in MS\(^1\) of roquefortine C and MS\(^2\) in 3 for which a much a richer fragmentation was observed in roquefortine C.

**Structure elucidation based on similar neutral-loss cascades**

For compound 4, the database query returned partial similarity to its most similar database entry roquefortine C, showing similar neutral-loss cascades over various fragmentation stages with only few shared unconnected fragments (Figure 6, Figure 7, Supplemental Figure 6 and 7). For further structure identification, the neutral-loss cascade \(C_3H_8 - C_9H_7N - C_3H_3NO_2\) shared by roquefortine C (ID 28, ID 62, ID 66) and compound 4 (ID 20, ID 107, ID 108) was used. As this relative specific neutral-loss cascade is shared by both compounds, it can be concluded that 4 might have similar

---

**Figure 4.** MS\(^1\) fragmentation spectrum of compound 2 (A) and MS\(^1\) fragmentation spectrum of the database entry roquefortine C (B) showing an identical fragmentation pattern.
structural features as roquefortine C, resulting in similar fragmentation mechanisms with similar neutral-losses upon excitation. Identifying these common fragmentation mechanisms, based on the known structure of the database entry, resulted in a partial identification of 4, containing a dimethylallyl (C₅H₈), indol (C₉H₇N) and diketopiperazine (C₅H₈NO₂) moiety similarly arranged (Figure 8). Striking is to see that the difference in the elemental composition between both compounds, two additional hydrogens in the structure of 4, are conserved throughout the entire fragmentation process and still being present in the final fragment with m/z 110 (ID 108, MS⁴, C₅H₈N₃). Although the structure of corresponding fragment of roquefortine C can be deduced based on the known structure of roquefortine C, structural information about the final fragment of 4 is not available, as a further fragmentation was not successful. However, in combination with gene deletion experiments (Ali, et al., 2013), concluding that compounds 1 – 4 originate from the same biosynthetic gene cluster, the structure of 4 was tentatively identified as roquefortine D which was subsequently confirmed by NMR experiments (Ali, et al., 2013).

![Figure 8. Proposed fragmentation of database entry roquefortine C (a) and compound 4 (b). After protonation, isoprene, represented by the loss of C₅H₈, is cleaved off in the first fragmentation step yielding the fragment m/z 322 in roquefortine C and m/z 324 for compound 4. Due to the subsequent loss of the indole part of the structure, indicated by a loss of C₉H₇N, ions m/z 193 in roquefortine C and m/z 195 in 4 are formed. In the final fragmentation step the ring of the diketopiperazine structure opens, leading to a loss of C₃H₃NO₂ and the formation of m/z 108 in roquefortine C and m/z 110 in 4.](image)

Discussion

The results presented here demonstrate how the metabolite identification pipeline CMCP can be used to obtain detailed structural information of structurally complex molecules. Using fragment and neutral-loss tree comparison, structural similar compounds were extracted from a MSⁿ database. As structural information is coded in the fragments, neutral-losses and fragmentation mechanisms, structural similarities between the unknown compound and the database entry are revealed. Identifying these shared mechanisms, linked to substructure information stored in the database or using MS expert knowledge, and applying them to the unknown molecule allows complete (sub)structures identification even without the unknown being present in the database, as shown for the de novo identification of compounds 2 - 4.

Depending on the structural overlap between the analyte and a database entry, high similarity can be a result of similar fragments, similar neutral-losses or a combination of both. Particularly the comparison of neutral loss trees, which is used here for the first time to obtain detailed structural information, represents a valu-
able structure elucidation approach as it uses complementary information to fragment trees for the determination of similarity. As almost no fragments are shared between the database entry roquefortine C and compound 4, a comparison solely based on similar fragments would not have been as successful. However, it should be pointed out that there can be a large gap between the degree of similarity obtained from fragment tree comparison and neutral-loss tree comparison. This is due to the fact that similarity based on fragments can be verified as fragments can be further fragmented generating a fragmentation subtree which contains structural information about the particular fragment. In case the obtained multiple stage fragmentation spectra are identical, it can be assumed that their precursor possess an identical chemical structure. In contrast, loss tree comparison is based exclusively on uncharged neutral-losses which can, by definition, not further be fragmented. Although their structures can be deduced from the known structure of the database entry and applied to the unknown analyte, as shown for the identification of compound 4, it is not possible to absolutely confirm the identical identity of shared losses in both compounds. Certainty that the exact same losses with the same fragmentation mechanisms occur in both compounds is depending on the length of the shared loss path and the uniqueness of the losses. The loss path C₅H₈ – C₉H₇N – C₃H₃NO₂ used for the identification of 4 contains relatively unique fragments corresponding to dimethylallyl, tryptophan and diketopiperazine substructures which are only shared with the database entry roquefortine C.

Uncovering relationships between neutral-losses and providing extensive structural information about individual fragments by further fragmenting them, is the main advantage of MSⁿ methods compared to MS/MS techniques, as it offers crucial information beyond elemental compositions and intensities. However, the structural information conserved in the fragments also needs to be exploited. Especially for smaller databases which do not provide enough positive hits to determine the structure of a fragment or neutral-loss solely on the structural overlap of a common substructure of multiple similar database entries, fragments and losses need to be manually assigned to a particular part of the molecule. Without using this structural information, the identification of 2 - 4 would have been limited to a structural classification resulting in a somehow relationship to roquefortine C. With the advent of novel software tools for the processing of multiple stage fragmentation data, tandem mass spectra databases may be replaced by MSⁿ databases due to the additional structural information available within.

Traditionally, the number of analytes which can be identified using mass spectrometry depends on the comprehensiveness of the database, as identity searches require each analyte to be represented by an identical database entry. Thus, large databases with numerous entries are necessary to cover a large chemical space. By searching for fragmentation similarities between multiple stage fragmentation trees as presented here, rather than complete identities, less comprehensive databases are required. For the identification of compounds 1 – 4, only one database entry was necessary to represent the fragmentation of a specific structural group of compounds, as shown for roquefortine C which was used to identify the metabolites 1 – 4.

In addition, CMCP paves the way towards a fully automated structure identification using solely mass spectrometry. With the ongoing storage of structural information
of fragments and neutral-losses into the MS\textsuperscript{n} database, detailed structural information of nodes in database entries is extended. Using this information, structures of similar nodes present in the unknown target compound can be automatically assignment enabling an immediate identification of fragments and neutral-losses without much intervention of MS experts. By automatically combining the structural overlap of various subtrees or losses even from different similar database entries, complete (sub)structures could be deduced allowing a completely automated structure identification.

**Conclusion**

Here, we present a structure elucidation pipeline that allows the complete structural identification of unknown molecules beyond structural classification without the need of the compound to be present in the database. It could be shown that fragment and neutral-loss tree comparison yield complementary information which makes neutral-loss tree comparison a valuable addition for the identification of unknowns. In combination with the ongoing structural identification of fragments and neutral-losses, automatic substructure assignment is possible, paving the way for a completely automated structure elucidation.

Concluding, fragmentation tree comparison is an important component of our metabolite identification pipeline which combines comparative metabolite profiling and structure elucidation using CMCP and NMR analysis. With this pipeline various (novel) metabolites of *P. chrysogenum* were identified and their biosynthesis could be determined. Future work on direct-infusion experiments and data processing tools aim to improve the throughput of this pipeline by overcoming the relative resource-intensive small scale extraction of the unknown compounds.

**Acknowledgements**

This project was supported by the Perspective Genbiotics program subsidized by the Stichting voor de Technische Wetenschappen (STW) and (co)financed by the Netherlands Metabolomics Centre (NMC) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. The authors are grateful to Prof. Dr. Nico M. Nibbering for his help and guidance during the elucidation of fragmentation mechanisms.

**References**


Cui, M., Sun, W., Song, F., Liu, Z., and Liu, S. (1999). Multi-stage mass spectrometric studies of triterpenoid saponins...


Supplemental Information

Supplemental Figure 1. Fragment tree of compound 2 compared to the fragment tree of database entry roquefortine C as indicated by colored nodes.
Supplemental Figure 2. Fragment tree of the database entry roquefortine C compared to the fragment tree of compound 2.
Supplemental Figure 3. Fragment tree of compound 3 compared to the fragment tree of database entry roquefortine C.
Supplemental Figure 4. Fragment tree of the database entry roquefortine C compared to the fragment tree of compound 3.

Supplemental Figure 5. MS² fragmentation spectrum of compound 3 (A) and MS³ fragmentation spectrum of the database entry roquefortine C (B) showing an identical fragmentation pattern.
Supplemental Figure 6. Neural-loss tree of compound 4 compared to the neutral-loss tree of the database entry roquefortine C.
Supplemental Figure 7. Neutral-loss tree of database entry roquefortine C compared to neutral-loss tree of compound 4.