

Automated Pipeline for De Novo Metabolite Identification Using Mass-Spectrometry-Based Metabolomics

Julio E. Peironcely,^{†,‡,§} Miguel Rojas-Chertó,^{‡,§} Albert Tas,[†] Rob Vreeken,^{‡,§} Theo Reijmers,^{‡,§} Leon Coulier,^{†,§} and Thomas Hankemeier^{*,‡,§}

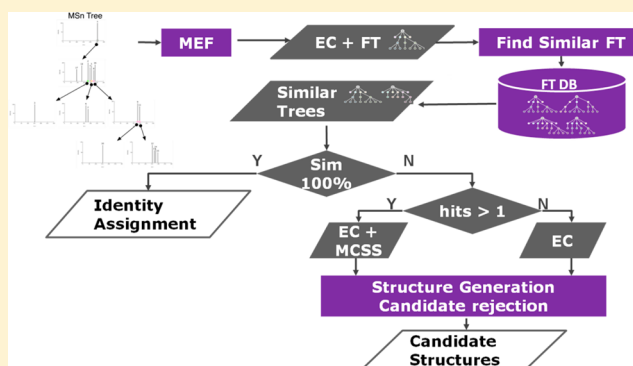
[†]TNO Research Group Quality & Safety, P.O. Box 360, NL-3700 AJ Zeist, The Netherlands

[‡]Leiden Academic Center for Drug Research, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands

[§]Netherlands Metabolomics Centre, Einsteinweg 55, 2333 CC Leiden, The Netherlands

S Supporting Information

ABSTRACT: Metabolite identification is one of the biggest bottlenecks in metabolomics. Identifying human metabolites poses experimental, analytical, and computational challenges. Here we present a pipeline of previously developed cheminformatic tools and demonstrate how it facilitates metabolite identification using solely LC/MSⁿ data. These tools process, annotate, and compare MSⁿ data, and propose candidate structures for unknown metabolites either by identity assignment of identical mass spectral trees or by de novo identification using substructures of similar trees. The working and performance of this metabolite identification pipeline is demonstrated by applying it to LC/MSⁿ data of urine samples. From human urine, 30 MSⁿ trees of unknown metabolites were acquired, processed, and compared to a reference database containing MSⁿ data of known metabolites. From these 30 unknowns, we could assign a putative identity for 10 unknowns by finding identical fragmentation trees. For 11 unknowns no similar fragmentation trees were found in the reference database. On the basis of elemental composition only, a large number of candidate structures/identities were possible, so these unknowns remained unidentified. The other 9 unknowns were also not found in the database, but metabolites with similar fragmentation trees were retrieved. Computer assisted structure elucidation was performed for these 9 unknowns: for 4 of them we could perform de novo identification and propose a limited number of candidate structures, and for the other 5 the structure generation process could not be constrained far enough to yield a small list of candidates. The novelty of this work is that it allows de novo identification of metabolites that are not present in a database by using MSⁿ data and computational tools. We expect this pipeline to be the basis for the computer-assisted identification of new metabolites in future metabolomics studies, and foresee that further additions will allow the identification of even a larger fraction of the unknown metabolites.



Metabolomics is the study and characterization of metabolites, which are the small molecules (molecular weight below 1000 Da) of an organism, biofluid, tissue, or biocompartment. Metabolites are substrates or products of metabolic processes and therefore describe accurately the phenotype of an organism.¹ Metabolite identification is frequently cited as one of the major bottlenecks in metabolomics.^{2–4} Knowing the identity of the metabolites that are relevant in studies is necessary for a proper biological interpretation of the results. This work focuses on mass spectrometry (MS) only rather than including nuclear magnetic resonance (NMR) because the former is more sensitive than the latter.

While no agreement exists on how to perform metabolite identification, some guidelines do exist that define how to report identities of metabolites.⁵ The highest reporting level is level 1, where an identity is proposed and validated using two independent and orthogonal data sources relative to an

authentic compound analyzed under identical experimental conditions, for instance accurate mass and multistage mass spectrometry (MSⁿ) spectra or retention time and m/z or MSⁿ data. Level 2 is used for putatively annotated compounds, where an identity is proposed based on MS/MS or MSⁿ spectral similarity of the unknown to the spectra of a known compound present in a database, but the identity is not validated with chemical reference standards. Level 3 includes putatively annotated compound classes, based on spectral similarity of the unknown to known compounds belonging to a certain chemical class. Level 4 includes unknown compounds that can be traced and quantified using spectral data in different experiments, but no structural information has been reported

Received: November 11, 2012

Accepted: January 31, 2013

Published: January 31, 2013



before. At the beginning of an identification project the unknown compounds can be divided into “known unknowns” and “unknown unknowns”.⁶ A known unknown is a compound that has been previously described for a certain analytical platform, for instance by a certain mass and retention time window, but that has not yet been identified in the current study. An unknown unknown is a new compound that has not been previously described or identified.

MS experiments yield the m/z of the compound, from which the mass can be derived. For each mass, one or multiple elemental compositions are possible, and the more accurate the mass is determined the fewer candidate elemental compositions are obtained. The mass accuracy depends on the instrument employed, and even for high accuracies, such as in the low part per million (ppm) or subppm range, unique elemental compositions cannot always be obtained.⁷ The number of possible elemental compositions can be reduced by incorporating information on the other molecules present in the sample and the possible biotransformations that could have occurred.⁸ Additionally, a database of ionization products and frequent neutral losses when MS/MS data are available can be used to annotate the elemental compositions of metabolites.⁹ In the case when a unique elemental composition is available, multiple molecules can still be found with that composition. Additional information of the compound can be obtained by performing MS/MS experiments, where the compound is fragmented and the m/z of the resulting fragments can be measured. These spectra can then be matched with existing spectra databases for identity assignment or similarity search.¹⁰

As an alternative, MSⁿ data can be used to characterize a compound in more detail by fragmenting it, detecting its fragments, isolating them, and fragmenting them multiple times.¹¹ The resulting information is a mass spectral tree of fragments connected hierarchically to the original parent ion,^{11,12} which contains more structural information of the unknown compound than regular MS and MS/MS data. MSⁿ data can be processed and enriched with open source tools like the multistage elemental formula (MEF),¹³ which creates a fragmentation tree where the parent ion and each fragment ion are annotated with their elemental composition, instead of the mass and a tree of neutral losses representing the fragmentation pattern of the compound. Actually, this tool can be used to exclude many possible elemental compositions for a given MSⁿ tree, so that often only one elemental composition for a spectral tree is obtained.

Different approaches have been recently presented to query and compare spectral data, most of them relying on concepts of fingerprint similarity. A fingerprint from the fragmentation tree of an unknown compound is an array of features like the elemental compositions of the fragments and the different branches, and it is used to query a database of known compounds, for which a fragmentation tree fingerprint has been previously computed. The assumption for using fingerprint similarity is that similar fragmentation trees are produced by similar compounds.¹⁴ Hypothetical fragmentation trees have been derived using a probabilistic model not from MSⁿ data, but from HPLC/MS/MS¹⁵ or GC/TOF-MS,¹⁶ and used to build a fingerprint comparison method¹⁷ that could assign the class of unknown compounds and in some cases the identity. A different approach¹⁸ involved building a spectral fingerprint directly from the MS/MS spectrum and relate it to a fingerprint containing structural information of the molecule. Recently, Rojas-Chertó et al.¹⁴ developed a similar approach using MSⁿ

data to build fingerprints and use them to query experimental MSⁿ data, where the hierarchical relations between fragments in the fragmentation tree were measured experimentally instead of computationally simulated. These fingerprints were implemented in the web application Metitere to process, handle, store, and analyze MSⁿ spectra.¹⁹

In the best case, querying fragmentation trees using fingerprint similarity can return a perfect match if the unknown was present in the database, which would be a level 1 identification of a “known unknown” (if the unknown and the standard were measured in the same conditions). In a less favorable case, the unknown is not in the database, and it is necessary to propose candidate structures via computer assisted structure elucidation (CASE) like our open source structure generator OMG.²⁰ In such situations, Rojas-Chertó et al.¹⁴ suggested to use the chemical structures of the similar trees in the fragmentation tree database to create the maximum common substructure (MCSS) under the assumption that the unknown metabolite, which belongs to the same class, will possess the same moiety. This MCSS together with the elemental composition of the unknown could be the input for a structure generator that would produce all the possible molecules complying with these criteria. CASE has been used to identify pollutants and toxic compounds in environmental samples by generating candidates with a structure generator like MOLGEN and filter or rank them using specific criteria related to the problem at hand.²¹ In a similar fashion, Schymanski et al.²² initially used gas chromatography coupled with electron-ionization mass spectrometry (GC/EI-MS), and possible filtering criteria were the prediction of spectra using tools like MetFrag,²³ retention index prediction and steric energy calculation. In a posterior study²⁴ a consensus score combining these criteria was used to rank candidate molecules.

MSⁿ data and software tools to process and evaluate these data have been presented as the key factors of success for the identification of small molecules.²⁵ Many cheminformatics tools that contribute to the elucidation of compounds have been developed, but in the field of metabolite identification, they require the unknown metabolite to be present in a database like PubChem.^{26–28} Furthermore, no combination of tools in a pipeline has been used for de novo metabolite identification as it was done for environmental pollutants, which used MS/MS data. Previous studies²⁹ used MSⁿ to identify plant metabolites, but required manual intervention and concluded that there is a need for pipelines of chemoinformatics to improve metabolite identification. In the work presented here, we combine different tools in a pipeline that enables, for the first time, de novo identification of metabolites from MSⁿ data as well as identity assignment in an automated fashion. In order to demonstrate the use of such an identification pipeline, we acquired 30 mass spectral trees of metabolites present in human urine and attempted to identify them with this pipeline.

■ MATERIALS AND METHODS

Mass spectral trees were acquired for the features measured in human urine samples. Details on analytical methods are provided in Supporting Information. More than 450 metabolite features representing most probably metabolites were detected with deconvolution (using the software Dissect, Bruker Daltonics, Bremen, Germany) in urine. Mass spectral trees were acquired for the 30 most abundant peaks (Table S-1) and processed with the metabolite identification pipeline presented. The identities of these 30 features and their trees were

unknown upon selection. Our approach did not attempt to provide a comprehensive analytical coverage of urine metabolites. The aim of this study was to illustrate how the software pipeline can improve the identification of “known unknowns” and “unknown unknowns” in metabolomics.

Mass Spectral Tree Processing and MSⁿ Database. The first step in the pipeline (Figure 1) is to process and annotate

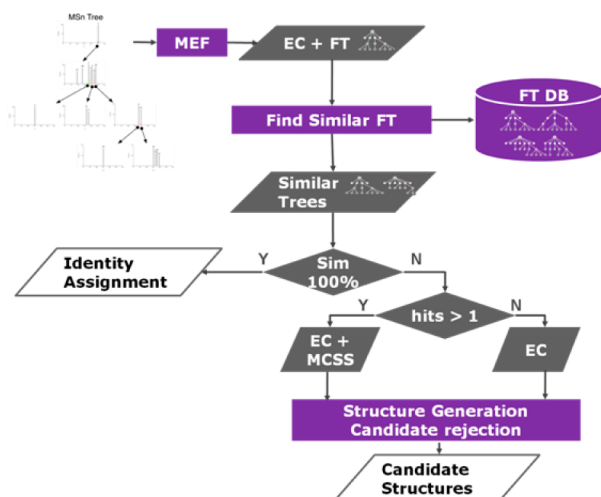


Figure 1. Metabolite identification pipeline. Abbreviations: MEF, multistage elemental formula; EC, elemental composition; FT, fragmentation tree; MCSS, maximum common substructure; Sim, similarity.

the mass spectral trees into fragmentation trees. Mass spectral trees were processed using the MEF tool,¹³ which resolves a unique elemental composition for each parent and fragment ion, as well as for the neutral losses. The result of using the MEF tool to process a mass spectral tree is a fragmentation tree and a neutral loss tree with elemental compositions assigned to the nodes of the tree. An in-house library of MSⁿ data of reference metabolites was used as described by Rojas-Chertó et al.¹⁴ This database contains fragmentation trees and neutral loss trees of 447 human metabolites and 118 plant polyphenolic metabolites. All MSⁿ spectra in the library were processed with MEF tool.

Data Comparison and Fragmentation Tree Similarity Search. The 30 unknown metabolites were compared to the known metabolites stored in the MSⁿ database using the fragmentation tree fingerprint and similarity calculation presented by Rojas-Chertó et al.¹⁴ A 10% similarity or more was considered to be relevant for identification purposes by educated guess.¹⁴ In the case that an unknown compound has 100% similarity with a metabolite in the database, we assign the identity to the “known unknown”, which in our case is level 1 identification. When no metabolite is found with 100% similarity, we are facing the identification of an “unknown unknown”. In such a case, multiple metabolites can be found with a certain degree of similarity, which is class assignment (level 3 identification) if these metabolites belong to the same class. Additionally, we used these similar compounds to calculate the maximum common substructure (MCSS) they shared and assumed it to be present in the structure of the unknown metabolite.

Candidate Structure Generation. We used the structure generator Open Molecule Generator (OMG)²⁰ to in silico

generate all possible candidate structures for the unknowns (Figure 2), taking as an input the elemental composition of the

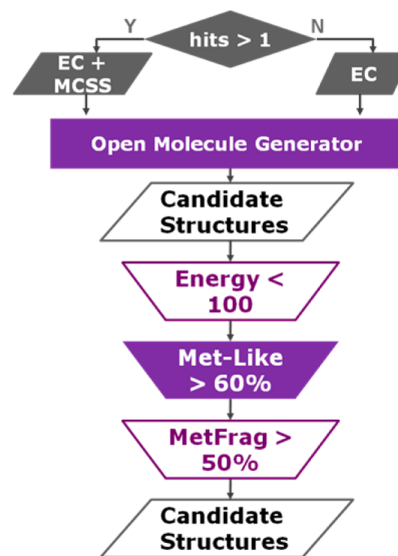
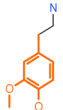
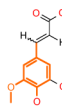
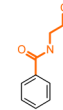
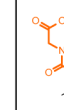
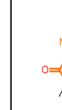
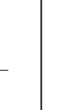
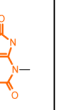
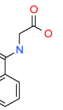
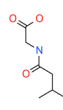
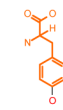
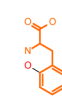







Figure 2. Structure generation and candidate rejection. Abbreviations: EC, elemental composition; MCSS, maximum common substructure.

unknown. OMG generates all the possible chemical structures containing exactly those atoms. This list of candidates, even for small elemental compositions, tends to contain millions or billions of possible molecules. Optionally, one can force the output molecules to contain one or multiple nonoverlapping prescribed substructures, which reduces drastically the number of candidate structures generated. The bigger the substructure or the more substructures used, the fewer candidates are produced. In this work, we used the MCSS found in the similarity search to be present in the generated structures.

Candidate Structure Filtering. We used three filters (Figure 2) to remove unlikely candidate chemical structures: steric energy, metabolite-likeness, and fragmentation prediction. While OMG produces candidate structures that are valid according to the valence rule, many are unstable and, therefore, unlikely to be found in a biological system. First, we used the component “Molecular Energy” from Pipeline Pilot Student Edition 6.1³⁰ to calculate the internal energy of the generated structures and those with an energy value of 100 or above were removed. This threshold value was selected after observing that all the metabolites present in the Human Metabolome Database (HMDB)³¹ have energy values below 100 when calculated with the same component. In order to use the energy score for further candidate ranking, we scaled energy values to unit range between 0% (for a candidate with energy value of 100) and 100% (for the candidate with the lowest energy value). Second, we used a predictive model of Metabolite-Likeness³² to remove candidate structures that are unlikely to structurally resemble human metabolites. We reported that almost all known human metabolites obtain a Metabolite-Likeness of 50% or more. Therefore, we set a conservative minimum threshold of 60% Metabolite-Likeness to consider structures for further identification. Third, we used the spectra prediction tool MetFrag²³ to remove candidates that cannot explain many of the peaks observed in the experimental spectra. MetFrag uses as an input a list of molecules and a list of the experimental spectral peaks, defined by the *m/z* and intensities.

Table 1. De Novo Metabolite Identification of “Unknown Unknown” Metabolites Not Present in the MSⁿ Database but Having a Degree of Fragmentation Tree Similarity with One or More Fragmentation Trees of Known Metabolites in the Database^a

Unknown	28		9		17		15		27		
Candidate Structures	6.8M		150M		Billions		Billions		Billions		
EC Hits	C ₉ H ₁₀ O ₂	2	C ₇ H ₇ NO ₄	2	C ₆ H ₆ N ₄ O ₃	2	C ₉ H ₉ NO ₄	2	C ₉ H ₁₃ NO ₄ P ₂	3	
Similar Structures											
Similarity	25%	11%	19%	11%	18%	10%	24%	12%	32%	30%	13%
MCSS											
Candidate Structures MCSS	82		65,445		4		8		281		
Candidate structures filtering	8		2,312		4		5		182 (40)		

^aCandidate structures are generated with Open Molecule Generator using the EC and MCSS and filtered using energy, Metabolite-Likeness, and MetFrag.

By cleavage of bonds, MetFrag fragments the molecules and computes for each one how many of the provided spectral peaks can be explained by the fragments. With this information, a score is built describing how well each candidate molecule can describe the experimental spectra. We used the settings of [M + H]⁺ mode, positive charge, 0.01 Mzabs, and 10 Mzppm. We rejected candidate structures that did not obtain at least 50% MetFrag score. Lastly, we combined the three scores in a unique consensus score, as proposed by Schymanski et al.²⁴ to rank the remaining structures in order and prioritize them for further manual identification by an expert.

RESULTS

MSⁿ spectral trees of 30 unknown metabolites acquired in human urine were analyzed with the metabolite identification pipeline described in the Materials and Methods section. The fragmentation trees of the unknowns were used to query the MSⁿ database for identical or similar fragmentation trees. From the 30 unknown metabolites, 10 obtained a 100% fragmentation tree similarity match, 9 found one or more similar trees (10% < similarity value < 100%), and 11 did not obtain a single hit in the database. At this stage, for these 11 unknowns we could only derive the elemental composition from the data. Using OMG to generate candidate structures for them would return billions of structures, and therefore, these unknowns were not studied further and remained unidentified. This indicates that the MSⁿ database used in this study should be enriched with more and varied metabolites.

Identity Search. The database query returned a 100% similarity match for 10 fragmentation trees (Table S-2). This is the highest possible similarity score and implies that both the fragmentation tree and neutral loss tree are identical for the unknown metabolite and the standard compound in the database. These 10 identified metabolites are creatinine, acetaminophen, phenylalanine, 7-methylxanthine, uric acid,

hippuric acid, paraxanthine, o-tyrosine, l-acetylcarnitine, and tryptophan.

Both the authentic standards present in the database and the unknown metabolites from urine were acquired using high resolution MS and MSⁿ in the same lab using the same equipment (as described in Materials and Methods). Hence, we are confident to have achieved a full level 1 identification as proposed in the MSI⁵ for these 10 unknown compounds. The authentic standards were acquired by direct infusion, in order to obtain deep and wide mass spectral trees, containing as much structural information as possible. Therefore, they miss an associated retention time. Ideally, these standards should be measured in the same HPLC system as the unknowns in order to have an extra analytical technique to support the full identification. It is interesting to mention that, despite being characteristic of the chemical structure, a mass spectral tree could theoretically not be unique for a given molecule; i.e. two isomeric structures with the same elemental composition but different structure could produce the same mass spectral tree. Hence, there is a need for complementary analytical methods, like NMR, to validate the identification of metabolites.









Similarity Search. For 9 of our 20 remaining unknown metabolites, we found in the database metabolites with similar fragmentation trees. Three metabolites only found one similar metabolite in the database. In such cases, we were neither able to propose the class of the unknown nor to extract a maximum common substructure, since we would need at least two similar metabolites of the same class. The only possible course of action according to our pipeline was to generate candidate structures using OMG for the elemental composition. Additionally, candidate molecules that are structurally dissimilar to the metabolite in the database could be removed using a chemical similarity filter, but this was out of the scope of the current work, because the resulting list of candidates would be too large. Unknown 16 returned 21 similar metabolites, which produced a very small MCSS (C-C). Such a MCSS, when used

in OMG, would not constrain the generation process and return billions of candidate structures. Therefore we did not proceed with the identification of this unknown.

Identification of Unknowns. Five unknowns returned two or three similar metabolites in the database. All the similar metabolites are found in urine according to HMDB. We calculated an MCSS from these metabolites, generated structures using OMG, and filtered the candidates using the three filtering criteria. For unknown 28, similarity search returned two similar metabolites, with 25% to 3-methoxytyramine and 11% to sinapic acid using fragmentation tree similarity (Table 1). The MCSS used as prescribed substructure in the candidate generation process with OMG returned only 82 molecules, instead of 6.8 million molecules if only the elemental composition was used. This list of candidate structures was reduced by using energy, metabolite-likeness and MetFrag filters. This resulted in 8 candidate structures, which are presented in Table 2 sorted by a consensus score (CS). For unknown 9, the database query returned two similar metabolites, with fragmentation tree similarity of 19% to hippuric acid and 11% to isovalerylglycine (Table 1). OMG generates more than 150 million molecules for the elemental composition of this unknown, and using the MCSS derived from the similar metabolites, this list is reduced to 65445 compounds and filtered further to 2312 candidate structures, of which 1279 obtained a CS of 90% or higher. This made a selection of smaller list of candidate structures not feasible. For unknown 17 two metabolites with similar fragmentation trees, 18% similarity to 1,3-dimethyluric acid and 10% to 1,3,7-trimethyluric, were found in the database (Table 1). OMG generated a much smaller list of candidate structures, only 4, using the MCSS as constraint (Table 3).





Similarity search returned 2 metabolites similar to unknown 15, with fragmentation tree similarity of 24% to hippuric acid and 12% to isovalerylglycine. At this step we observed three things: (i) the elemental composition of the unknown was the elemental composition of hippuric acid with an extra oxygen atom; (ii) the fragmentation tree similarity of 24% was due to the neutral loss tree, which were identical for the unknown and the compound in the database, indicating that both compounds had a similar structure and fragmentation pattern; (iii) the fragmentation tree measured for the unknown was almost identical to the one in the database, except for an additional oxygen atom in each of the fragment ions. This indicated that the chemical structure of the unknown was the structure of hippuric acid, which we used as MCSS (Table 1), with an additional oxygen atom. OMG generated 8 candidate molecules using the MCSS as constraint (Table 4), which results from adding one oxygen atom in all possible ways to hippuric acid. Further filtering using our three criteria removed candidates 6, 7, and 8, which despite having favorable values of energy score and metabolite-likeness, were not able to explain any of the experimental fragments and therefore MetFrag assigned them a 0% score. A close examination of the *in silico* fragments proposed by MetFrag for the experimental fragment revealed that all of them contained a phenol group, a feature that is not present in the three rejected candidates. Therefore, we propose that unknown 15 has the same structure as a compound with an oxygen atom attached to the benzene ring. The position of the oxygen in the phenol group remains unknown. Additionally, NMR measurements of standards could be used to elucidate the position of the oxygen in the molecule and confirm the identity of this unknown.

Table 2. Candidate Structures for Unknown 28

Candidate Structure		Energy	Met likeness	MetFrag	Consensus Score
		HMDB / InChIKey			
1		-1.35 (100%)	81.0%	99.5%	93.5%
		None / XNKMBCYYBMXTPO-UHFFFAOYSA-N			
2		-1.24 (99.9%)	80.8%	94.4%	91.7%
		HMDB13744 / YOMSJEATGXXYPX-UHFFFAOYSA-N			
3		12.42 (86.41%)	68.6%	93.4%	82.8%
		None / QTPWGUHKASDDHO-UHFFFAOYSA-N			
4		57.76 (41.86%)	75.4%	93.6%	70.2%
		None / HUVRKDVFCZQOM-UHFFFAOYSA-N			
5		64.07 (35.45%)	72.2%	90.2%	66.0%
		None / WNOKBMFZDXCSRNUHFFFAOYSA-N			
6		61.54 (37.95%)	61.8%	93.3%	64.4%
		None / YOYNEOCOGAQSSV-UHFFFAOYSA-N			
7		83.35 (16.43%)	66.6%	93.6%	58.9%
		None / QVXRGADGDBVPIE-UHFFFAOYSA-N			
8		94.26 (5.66%)	77.2%	93.4%	58.9%
		None / SYCBIYPZGRLQD-UHFFFAOYSA-N			

Similarity search for unknown 27 returned three metabolites with fragmentation tree similarity of 32% to L-tyrosine, 30% to o-tyrosine, and 13% to DL-dopa (Table 1). OMG generated a list of 281 candidate structures using the MCSS, which was reduced to 182 after filtering. We observed that two of the similar metabolites in the database had a phenol (benzene ring

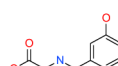
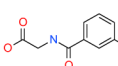
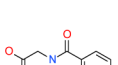
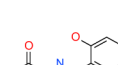
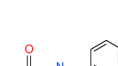
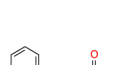


Table 3. Candidate Structures for Unknown 17

Candidate Structure		Energy	Met likeness	MetFrag	Consensus Score
		HMDB / InChIKey			
1		18.91 (100%)	98.4%	100%	99.4%
		HMDB01970 / ODCYDGXXCHTFIR-UHFFFAOYSA-N			
2		21.09 (97.31%)	97.6%	100%	98.3%
		HMDB01973 / XJEJWDFDVPDMAS-UHFFFAOYSA-N			
3		21.49 (96.82%)	96.2%	100%	97.7%
		HMDB11107 / YHNNPKUFPWLTOP-UHFFFAOYSA-N			
4		18.98 (99.91%)	98.4%	89.7%	96.0%
		HMDB03099 / QFDRTOONISXGJA-UHFFFAOYSA-N			

with an attached oxygen atom) and the third one a catechol (benzene ring with two attached oxygen atoms). Hence, we assumed that our unknown also had at least one oxygen atom attached to the benzene ring. We selected from the 182 candidates those that contained a phenol, which resulted in a final list of 40 candidate structures. A P–P bond was present in all the candidates, which despite being a rarity among known metabolites did not penalize the scores obtained by the molecules. This P–P moiety would immediately raise an alarm flag for any metabolite identification expert. It could be caused by poor experimental acquisition of the mass spectral tree or by an incorrect assignment of the elemental composition by MEF. Inspection by an expert determined that for a m/z of 262.03809 MEF should produce an elemental composition like $C_9H_{13}NO_6P$, which belongs to phosphotyrosine, instead of $C_9H_{14}NO_4P_2$. Therefore, we confirmed that the analytical conditions were identical for this unknown as for the other compounds and that all the elemental compositions generated and forced MEF to use the elemental composition $C_9H_{13}NO_6P$ for the parent ion, but it failed to annotate the elemental compositions of the fragments. In other words, this elemental composition could not explain the fragment ions measured experimentally. As a result, we considered the experimental data and the elemental composition $C_9H_{14}NO_4P_2$ to be valid and all 40 candidates to be possible. Ideally, authentic standards of them should be measured and compared with the spectral data of the unknown.

Tentative Validation of MCSS Assignment. We assessed whether the use of the MCSS and the filtering can lead to a wrong identification or to miss the good molecule in the list of candidate structures. We applied the structure generation and

Table 4. Candidate Structures for Unknown 15

Candidate Structure		Energy	Met likeness	MetFrag	Consensus Score
		HMDB / InChIKey			
1		-1.51 (100%)	96.8%	100%	98.9%
		HMDB06116 / XDOFWFNMYJRHEW-UHFFFAOYSA-N			
2		-1.51 (100%)	96.8%	100%	98.9%
		HMDB06116 / XDOFWFNMYJRHEW-UHFFFAOYSA-N			
3		-1.49 (99.99%)	96.8%	100%	98.9%
		HMDB13678 / ZMHLUFWWWPBTIU-UHFFFAOYSA-N			
4		-1.39 (99.88%)	94.6%	100%	98.1%
		HMDB00840 / ONJSZLXSECQROL-UHFFFAOYSA-N			
5		-1.39 (99.88%)	94.6%	100%	98.1%
		HMDB00840 / ONJSZLXSECQROL-UHFFFAOYSA-N			
6		-1.11 (99.61%)	96.2%	0%	65.2%
		None / NVVKRZSRWCCEAU-UHFFFAOYSA-N			
7		-0.53 (99.03%)	98.2%	0%	65.7%
		HMDB02404 / GCWCVCCEIQXUQU-MRVPVSSYSA-N			
8		1.04 (97.49%)	87%	0%	61.5%
		None / FMYVYJPEMYKYRE-UHFFFAOYSA-N			

filtering strategy to the 10 identified metabolites. For only four of these metabolites similar trees were found in the database and a MCSS could be generated (Table S-3). We observed in each of these four cases that the MCSS found is a substructure of the metabolite, with which OMG generated among others the good structure. The filtered list of candidate structures

always contained the good molecule, which was ranked high according to the consensus score in three of the four cases. In previous work,¹⁴ it was observed that with a tree similarity below 20% the MCSS obtained was not very informative. In these four examples, the MCSS used were informative enough when obtained from metabolites with at least 12% tree similarity. When including metabolites with tree similarity between 12% and 10% the MCSS was a carboxylic acid for unknowns 22, 12, and 18. For unknown 28 it was a benzene ring. These MCSS belong to the metabolite, but OMG would return millions of candidates; therefore, we did not use them for further confirmation. From this we conclude that the use of the pipeline can provide good candidate structures. Further validation should be performed to understand whether there are cases for which the pipeline could lead to incorrect results.

DISCUSSION

The results presented here demonstrate how this metabolite identification pipeline can be used to identify metabolites using MSⁿ data from human urine samples. This workflow could be adapted to work with MS/MS data, although data processing and similarity search of spectra should be then modified. Here we only used MSⁿ data and applied it to those features for which a similar fragmentation tree was present in the MSⁿ database. Such MSⁿ database can be used locally, Metitree,¹⁹ or online, Massbank³³ and MzCloud. The number of metabolites that can be identified in this way depends on how comprehensive the database of MSⁿ is. Furthermore, we showed for the first time how metabolites not present in a database could be identified.

Having substructure information is crucial to identify unknown metabolites. In our case, we observed that using a large MCSS (or alternative multiple prescribed substructures) reduced significantly the number of candidate structures; therefore, future work should focus on developing more reliable ways of generating more or larger MCSS. In the case of unknown 9, the MCSS found was linear, which allowed for the formation of many rings and, therefore, a list of more than 2000 candidate structures. For the same unknown and for unknown 28 we observed that the filtering using energy, metabolite-likeness, and MetFrag yielded a 10-fold reduction in the number of candidates, proving the value of incorporating these criteria. In those cases where the MCSS described most of the structure of the unknown, OMG produced a short list of candidates, and this was not significantly reduced with the filters, since most of the structures were acceptable. Additionally, more filters could be added in the future depending on the data available, like retention time prediction.^{34–36} Fragmentation prediction by MetFrag proved to be useful at rejecting candidates, like for unknown 15, that did not have an oxygen atom attached to a ring but to a chain.

The use of mass spectral trees was crucial to assign identities and to derive structural information of the unknown metabolite from similar metabolites. We observed that very similar metabolites could have low fragmentation tree similarity, because their fragmentation trees were different. Fortunately, the structural resemblance was captured in the neutral loss trees, which in some cases were identical between the unknown and a similar metabolite, despite having different fragmentation trees. This shows the importance of including neutral loss information in the fragmentation tree fingerprint approach and encourages future research on how to better combine

fragmentation tree and neutral loss tree information for similarity search.

CONCLUSION

In this work we have presented a pipeline that enables metabolite identification using MSⁿ data and that can be used in metabolomics studies involving experimental data. Starting from the experimental MSⁿ data of unknown metabolites, this pipeline processes, annotates, and compares MSⁿ data, and assigns the identity or provides a few putative identities for de novo identification of unknown metabolites.

By means of fragmentation tree similarity, this pipeline can assign the identity to an unknown metabolite, provided its MSⁿ spectra have been previously measured and stored in a database. In the case this metabolite is not in the database, this pipeline is capable of doing de novo metabolite identification by extracting common moieties in similar compounds and using structure generation to propose candidate structures. De novo identification is in itself the biggest contribution of this work to the field of metabolomics as the pipeline does not require the unknown metabolite to be present in any database to propose a handful of possible structures.

While the unknown is not required to be in a database to be identified, the number of the candidate structures returned will be fewer, provided substructures of the unknown can be discovered. Ideally, these substructures could be found by matching subtrees of the unknown with a database of annotated MSⁿ trees, i.e., where a structure has been assigned to the fragment ions. Unfortunately, these annotated databases are not yet available for MSⁿ data, and therefore, we searched for similar metabolites to the unknown in the MSⁿ database and generated the MCSS. On the one hand, it appears necessary to enrich MSⁿ databases with experimental data of more and varied metabolites to increase the chances of finding similar metabolites. On the other hand, finding too many compounds with similar fragmentation trees can produce a small MCSS if the chemical structures are different, which will not constrain enough the generation of candidate molecules. Therefore, it is interesting to study better ways to find similar compounds, like an initial clustering of the known metabolites and a posterior MCSS calculation within each cluster could benefit de novo identification. Additionally, the similarity threshold of fragmentation trees could be modified in order to obtain less similar compounds and as consequence a larger MCSS, provided we have a rich and comprehensive database.

To the best of our knowledge this is the first implementation of a metabolite identification pipeline that enables identity assignment and de novo metabolite identification and that makes use solely of LC/MSⁿ data, and we foresee that further additions such as the ones proposed above will allow us to identify even a larger fraction of the unknown metabolites.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: hankemeier@lacdr.leidenuniv.nl.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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.

■ REFERENCES

- (1) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–71.
- (2) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, in press.
- (3) Johnson, C. H.; Gonzalez, F. J. *J. Cell. Physiol.* **2012**, *227*, 2975–81.
- (4) Scalbert, A.; Brennan, L.; Fiehn, O.; Hankemeier, T.; Kristal, B. S.; Ommen, B.; van; Pujos-Guillot, E.; Verheij, E.; Wishart, D.; Wopereis, S. *Metabolomics* **2009**, *5*, 435–458.
- (5) 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.; Reily, M. D.; Thaden, J. J.; Viant, M. R. *Metabolomics* **2007**, *3*, 211–221.
- (6) Wishart, D. S. *Bioanalysis* **2009**, *1*, 1579–1596.
- (7) Kind, T.; Fiehn, O. *BMC Bioinf.* **2006**, *7*, 234.
- (8) Rogers, S.; Scheltema, R. a.; Girolami, M.; Breitling, R. *Bioinformatics* **2009**, *25*, 512–8.
- (9) Draper, J.; Enot, D. P.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H. *BMC Bioinf.* **2009**, *10*, 227.
- (10) Roux, A.; Xu, Y.; Heilier, J.-F.; Olivier, M.-F.; Ezan, E.; Tabet, J.-C.; Junot, C. *Anal. Chem.* **2012**, *84*, 6429–6437.
- (11) Kasper, P. T.; Rojas-Chertó, M.; Mistrik, R.; Reijmers, T.; Hankemeier, T.; Vreeken, R. J. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 2275–86.
- (12) Sheldon, M. T.; Mistrik, R.; Croley, T. R. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 370–6.
- (13) Rojas-Chertó, M.; Kasper, P. T.; Willighagen, E. L.; Vreeken, R. J.; Hankemeier, T.; Reijmers, T. H. *Bioinformatics* **2011**, *27*, 2376–83.
- (14) Rojas-Chertó, M.; Peironcelly, J. E.; Kasper, P. T.; van der Hooft, J. J. J.; de Vos, R. C. H.; Vreeken, R.; Hankemeier, T.; Reijmers, T. *Anal. Chem.* **2012**, *84*, 5524–5534.
- (15) Rasche, F.; Svatos, A.; Maddula, R. K.; Böttcher, C.; Böcker, S. *Anal. Chem.* **2011**, *83*, 1243–51.
- (16) Hufsky, F.; Rempt, M.; Rasche, F.; Pohnert, G.; Böcker, S. *Anal. Chim. Acta* **2012**, *739*, 67–76.
- (17) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatos, A.; Böcker, S. *Anal. Chem.* **2012**, *84*, 3417–3426.
- (18) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics* **2012**, *28*, 2333–2341.
- (19) Rojas-Chertó, M.; van Vliet, M.; Peironcelly, J. E.; van Doorn, R.; Kooyman, M.; Te Beek, T.; van Driel, M. a.; Hankemeier, T.; Reijmers, T. *Bioinformatics* **2012**, *28*, 2707–2709.
- (20) Peironcelly, J. E.; Rojas-chertó, M.; Fichera, D.; Reijmers, T.; Coulier, L.; Faulon, J.-L.; Hankemeier, T. *J. Cheminf.* **2012**, *4*, 21.
- (21) Schymanski, E. L.; Bataineh, M.; Goss, K.-U.; Brack, W. *Trends Anal. Chem.* **2009**, *28*, 550–561.
- (22) Schymanski, E. L.; Meringer, M.; Brack, W. *Anal. Chem.* **2011**, *903*–912.
- (23) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
- (24) Schymanski, E. L.; Gallampois, C. M. J.; Krauss, M.; Meringer, M.; Neumann, S.; Schulze, T.; Wolf, S.; Brack, W. *Anal. Chem.* **2012**, *84*, 3287–3295.
- (25) Kind, T.; Fiehn, O. *Bioanal. Rev.* **2010**, *2*, 23–60.
- (26) Zhou, B.; Wang, J.; Ransom, H. W. *PLoS One* **2012**, *7*, e40096.
- (27) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, a.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B. *Analyst* **2009**, *134*, 1322–32.
- (28) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Lai, S.; Wilder, J.; Grant, D. F. *Anal. Chem.* **2012**, *84*, 9388–94.
- (29) Hooft, J. J. J.; Vervoort, J.; Bino, R. J.; Vos, R. C. H. *Metabolomics* **2011**, *8*, 691–703.
- (30) *Accelrys Pipeline Pilot, version 6.1.5*; Accelrys Inc.: San Diego, CA, 2010.
- (31) 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.; Souza, A. De; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazzyrova, A.; Shaykhtudinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. *Nucleic Acids Res.* **2009**, *37*, D603–610.
- (32) Peironcelly, J. E.; Reijmers, T.; Coulier, L.; Bender, A.; Hankemeier, T. *PLoS ONE* **2011**, *6*, e28966.
- (33) 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. *J. Mass Spectrom.* **2010**, *45*, 703–714.
- (34) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr., A* **2011**, *1218*, 6732–41.
- (35) Hall, L. M.; Hall, L. H.; Kertesz, T. M.; Hill, D. W.; Sharp, T. R.; Oblak, E. Z.; Dong, Y. W.; Wishart, D. S.; Chen, M.-H.; Grant, D. F. *J. Chem. Inf. Model.* **2012**, *52*, 1222–37.
- (36) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. *V Anal. Chem.* **2011**, *8703*–8710.