Metabolite Identification Using Automated Comparison of High-Resolution Multistage Mass Spectral Trees

ARTICLE *in* ANALYTICAL CHEMISTRY · MAY 2012

Impact Factor: 5.64 · DOI: 10.1021/ac2034216 · Source: PubMed

8 AUTHORS, INCLUDING:

Julio Peironcely
Leiden University
14 PUBLICATIONS **189** CITATIONS
SEE PROFILE

Justin J. J. van der Hooft
University of Glasgow
36 PUBLICATIONS **429** CITATIONS
SEE PROFILE

Ric De Vos
Wageningen UR
215 PUBLICATIONS **8,660** CITATIONS
SEE PROFILE

Rob J. Vreeken
Janssen Pharmaceutica
95 PUBLICATIONS **1,809** CITATIONS
SEE PROFILE

analytical chemistry

Article

pubs.acs.org/ac

Metabolite Identification Using Automated Comparison of High-Resolution Multistage Mass Spectral Trees

Miquel Rojas-Cherto,*[,†,‡] Julio E. Peironcely,[†,‡,§] Piotr T. Kasper,[†,‡] Justin J. J. van der Hooft,[†,∥,⊥] Ric C. H. de Vos,[†,∥,#] Rob Vreeken,[†,‡] Thomas Hankemeier,[†,‡] and Theo Reijmers*[,†,‡]

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

[‡]Analytical Biosciences, Leiden University, Einsteinweg 55, 2300 RA Leiden, The Netherlands

[§]TNO Research Group Quality and Safety, P.O. Box 360, 3700 AJ Zeist, The Netherlands

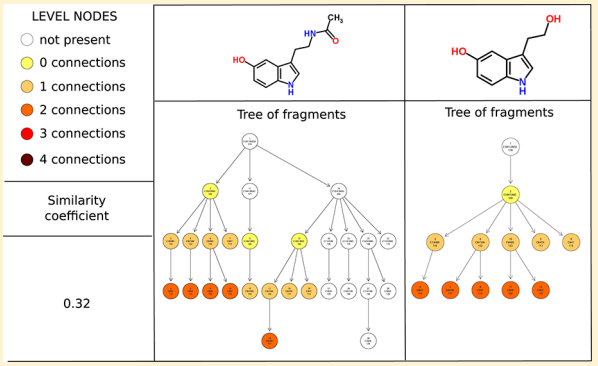
[∥]Plant Research International, Wageningen University and Research Centre, P.O. Box 16, 6700 AA Wageningen, The Netherlands

[⊥]Laboratory of Biochemistry, Wageningen University and Research Centre, Dreijenlaan 3, 6703 HA Wageningen, The Netherlands

[#]Centre for Biosystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

Supporting Information

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



Metabolomics emerges from the need to study and understand the function of the genes through their end products, so-called metabolites. Over the past years, mass spectrometry (MS) has proven itself as a powerful technology for the detection and annotation of compounds and became important for analyzing the metabolome of any organism.[1] Depending on the nature of the sample and the information scientists want to extract from it, mostly two different MS ionization techniques are used, i.e., hard and soft ionization. Both ionization techniques can be used separately[2,3] as well as combined.[4] Hard ionization methods, such as electron impact ionization (EI), deal with a high-energy source generating ions and an extensive range of fragmentation products from the molecule. Due to the high energy, the fragmentation spectra obtained are highly uniform between instruments and can therefore be used for creating universal databases such as the National Institute of Standards and Technology (NIST) mass spectral library, which is used worldwide in GC/MS studies for spectrum matching. In contrast, soft ionization methods, such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), treat the molecule more gently, aiming to generate quasi-molecular ions from the intact molecule. While with hard ionization methods each compound is characterized by a multitude of mass fragments, with soft ionization only a couple of mass signals are produced from each compound. The advantage of using soft ionization is its higher sensitivity in LC/MS analyses and the feasibility to conduct stepwise fragmentation of the quasi-molecular ion by collision-induced dissociation (CID) or collisionally activated dissociation (CAD). At the end, either a tandem mass spectrum or a hierarchical mass spectrum, also called a multistage mass spectrum, is created.[5−7] Unfortunately, tandem mass spectrometry

Received: January 12, 2012
Accepted: May 21, 2012
Published: May 21, 2012

   dx.doi.org/10.1021/ac2034216 | *Anal. Chem.* 2012, 84, 5524−5534

(MS/MS or MS^2) data are generally not as reproducible between laboratories as EI data.[8,9] However, new advances have been proposed leading to more reproducible MS/MS data. Examples are application of a tuning point protocol to standardize CID conditions prior to data acquisition[10,11] or usage of a fragmentation energy index for LC/MS to normalize collision energies.[12] These new developments encourage researchers to start creating MS/MS spectral libraries.[13−15] Compared to MS/MS (i.e., MS^2), ion-trap-based multistage mass spectrometry (MS^n) generates more specific and detailed information about the relation between the product ion and its direct fragments and the fragments derived from these fragments, including the spectral hierarchy. This sequential fragmentation approach increases the ability of mass spectrometrists to structurally characterize and identify unknown metabolites detected in untargeted LC/MS-based metabolomics approaches.[16,5,7,6,17] In addition, the relative intensities of the MS^n fragment ions generated, and thus the fragmentation spectra, are highly reproducible between different experiments and hardly influenced by small changes in instrument settings.[5−7] This reproducibility indicates that there is a good potential for MS^n spectral tree approaches to create and search fragmentation tree libraries, such as searching EI spectra in the NIST library. The applicability for library searching has at least been shown for nominal mass MS^n data generated on the same instrument.[5] By coupling ion trap MS^n fragmentation to accurate mass readout of the fragments generated, e.g., using the linear trap quadrupole (LTQ)−Orbitrap FTMS hybrid MS system (Thermo Fisher Scientific), the elemental composition of fragments can be readily obtained, which can further help in structural elucidation of unknown compounds. However, for optimal use and implementation in metabolomics studies, the accurate MS^n spectral tree approach still lags behind, as compared to EI spectrum matching, on three main points: (I) fragmentation spectrum representation, (II) spectrum storage, and (III) comparison and matching of spectra. Current software to handle MS^n data is either commercial and not flexible enough to do dedicated follow-up data processing, e.g., MassFrontier (Thermo Fisher Scientific), or not specifically developed to process MS^n data, e.g., XCMS.[15] Appropriate processing of MS^n data is crucial for obtaining robust data to be stored in a reference fragmentation tree database. Recently our group developed a new freely available tool called MEF (mass elemental formula)[18] which processes and enriches high-resolution MS^n data and generates fragmentation trees. The MEF tool extracts the most relevant signals from the MS spectra (representing the ions/fragments) and enriches these with the assignment of an elemental composition. Additionally, MEF generates elemental compositions for the neutral losses. To facilitate further analysis, the MEF tool can export the resulting information to other formats (chemical markup language (CML),[19−21] portable document format (PDF), or comma-separated value (CSV)). In this paper we define a fragmentation tree as a hierarchical organization of fragment ions describing the fragmentation reactions between them where the nodes refer to the fragments (represented by either their nominal mass (NM) values or their elemental compositions (EC)) and the edges refer to the fragmentation reactions.[22] As soon as a library has been created, an algorithm to automatically query that library to find similar MS spectral data is needed. Currently, several search algorithms for comparing soft ionization MS/MS spectra of small molecules exist,[23,24,22] mostly differing in how the MS/MS spectra are represented and how the similarities are calculated. The main search concepts are based on the spectral-contrast-angle method,[25] the probability-based matching (PBM) algorithm,[26] or the dot-product algorithm search.[27] All of them have been applied first to EI data, and more recently, they have also been introduced for the analyses of soft ionization MS/MS data analysis.[28] Approaches capable of calculating the similarity between MS^n data are the recently published work of Rasche,[29] which compares hypothetical fragmentation trees, and the commercial software tool MassFrontier (Thermo Fisher Scientific), which is based on the dot-product function.

Most algorithms mentioned above represent MS spectra by a set of equidistant bins, where each ion is encapsulated into one specific bin according to its mass-to-charge ratio (m/z). When two spectra fill the same bins, they are considered to be identical. The disadvantage of this representation is that only single spectra can be compared with each other and not the complete spectral tree and, furthermore, that the maximum number of bins for representing a spectrum (or the bin width) is limited by the mass resolution of the data. In the field of chemical similarity searching, in which structure databases are queried, other approaches are used. These approaches look for certain substructures being present in the structures and are called fingerprint-based algorithms. As a consequence, these algorithms are not dependent on the mass accuracy of the fragments but on the presence or absence of certain relations between the fragments and therefore are more suitable for handling MS^n spectral trees generated at high mass resolution. The more commonly used algorithms in this field make use of the Tanimoto (or Jaccard), cosine, or Dice coefficients, the Euclidean or Hamming distance,[30] or modifications of these coefficients or distances.[31] Many different studies have compared the performances of these similarity measures.[32] The Tanimoto coefficient is the most widely used coefficient for similarity-based querying because it is easy to use and computationally efficient.

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

MATERIAL AND METHODS

Metabolite MS^n Libraries. In this work two different in-house libraries containing MS^n data have been used. The first library was created at the Division of Analytical Biosciences, Leiden University, Leiden, The Netherlands, and contains 705 MS^n spectra from 447 different human metabolites. The second library was created at Plant Research International, Wageningen University and Research Centre, Wageningen, The Netherlands, and contains 167 MS^n spectra of 118 plant metabolites belonging to the class of polyphenols. This plant library includes different series of isomers in which hydroxyl, glycosyl, or methoxy groups are attached to different positions of the core flavonoid structure.[6] More information about these

dx.doi.org/10.1021/ac2034216 | Anal. Chem. 2012, 84, 5524−5534

Table 1. Data Sets Used in This Study

library	no. of compds	mass range (Da)	median mass (Da)	av mass (Da)	no. of spectra	diversity level
human	447	59.0−1525.6	196.1	254.6	705	0.82
plant	118	172.1−792.3	293.0	347.4	167	0.48

libraries can be found in Table 1. The main difference between the two libraries was the diversity level of the molecules (see Supplemental Text 1 in the Supporting Information). The complete set of spectral trees from both libraries was used for analyzing the performance of the mass spectral tree comparison method presented here. Both libraries were generated on LTQ−Orbitrap FTMS XL instruments (Thermo Fisher Scientific) using a TriVersa NanoMate injection robot (Advion) with chip-based ESI nanospray. The MS^n experiments were run in both positive and negative ionization modes using a data-dependent scanning function, limited to a 15 min acquisition time, with the criterion to select the five most intense ions detected for MS^2 and MS^3 and the three most intense ions for the rest of the MS levels.[7] The data were generated in centroid mode with a full width at half-maximum (FWHM) resolution of 60.000.

Processing of MS^n Fragmentation Trees. Raw data were converted to mzXML format[33] using ReadW software which was provided by the Institute for Systems Biology, Seattle, WA. All MS^n data of the reference compounds in the two selected libraries were processed with an extended version of the so-called MEF tool.[18]

The parameters used to process the MS^n data are described in Supplemental Text 3 in the Supporting Information. Because the depth of the fragmentation tree, i.e., the number of fragments present, highly depends on the concentration of the compound under investigation,[6] the direct comparison of MS^n spectra generated from compounds in biological samples, which are often present at low concentrations, with those stored in the library, generated at relatively high concentrations, can be difficult. To cope with such differential concentrations, all MS^n spectra in the reference databases were processed at four different values for the relative intensity threshold parameter in the MEF tool. The relative intensity is defined as the intensity relative to the base peak, i.e., the most intense fragment peak, in each spectrum. Signals that had a relative intensity lower than the relative intensity threshold were not considered in the further processing. The values for the relative intensity threshold were set to 0% (default setting of the MEF tool, i.e., all signals taken into account), 5% (signals below 5% of the base peak were omitted), 10%, and 20%. In this way, for each reference compound, a series of MS^n spectra representing a theoretical dilution series were generated in silico. Since each raw data file may contain several fragmentation trees of the same compound, i.e., repetitions, peaks that did not appear in at least 40% of the repetitions were considered as irreproducible and were therefore omitted.

MS^n Fragmentation Tree Representation. The efficiency of comparing fragmentation trees in an MS^n library depends directly on how fragmentation trees are represented and what similarity measure is applied. Because both a fragmentation tree and a chemical structure can be represented as a graph, i.e., they both contain nodes (fragments or atoms) and edges (fragmentation reactions or bonds), the similarity measures used for molecules can also be applied to fragmentation trees. A binary fingerprint is a commonly used representation for molecules in which the presence or absence of predetermined substructural features is indicated with ones or zeros, respectively.[34] In the present research, this fingerprint-based representation for structures was extended to MS^n fragmentation trees. The fragmentation tree is represented by a series of zeros and ones in a linear bitmap, where each bit in the fingerprint is related to the absence or presence, respectively, of a particular feature of the fragmentation tree. Different types of features were defined in accordance with the different ways nodes were connected. All features used are shown in Supplemental Figure 1 in the Supporting Information. Next to generating a fragmentation tree, the MEF tool also extracts a neutral loss tree from raw MS^n data files. Such a tree is a hierarchical organization of the neutral losses in a graphical form. Also, for these neutral losses, binary fingerprints were generated and concatenated to the corresponding fragmentation tree fingerprints.

MS^n Fragmentation Tree Similarity Measures. For our purpose we applied the Tanimoto coefficient as a similarity measure to enable calculation of the degree of similarity between fragmentation trees. Because the Tanimoto coefficient is molecule size dependent[31] and the equation has an inherent bias toward certain similarity values,[35] we applied a prefiltering of the two fingerprints that were going to be compared by omitting large dissimilar features that would give rise to a series of smaller, likewise dissimilar features (see Supplemental Text 2 in the Supporting Information).

Next to having a quantitative measure describing how similar fragmentation trees are, we also visualize this similarity by showing which nodes in the compared trees overlap. For this, the concept of a maximal common subgraph (MCS) was used. MCS is defined as the largest possible subgraph that two objects (structures of fragmentation trees) share in common. By calculating the MCS of the structures of the reference compounds in the database with the highest fragmentation tree similarity values, we obtained structurally relevant information on substructure level for the unknown compound that is queried. In our study we generated the maximum common substructures (MCSSs) for multiple molecules for a given list of IUPAC international chemical identifiers (InChIs)[36] by using the chemistry development kit (CDK) library.[37,38]

RESULTS AND DISCUSSION

The method developed and described in this paper is used for comparison of fragmentation trees and is based on extracting the fingerprints of both trees and then calculating their similarity using the Tanimoto coefficient. The fingerprint is built up of representative features of the fragmentation tree, which forms different combinations between the nodes and the edges. Two different types of fragmentation trees, differing in how the nodes are represented, were considered in this study. One type was constituted with nodes enriched with nominal mass information, called the nominal mass fragmentation tree (NMFT), while the other type consisted of nodes enriched with accurate mass-derived elemental composition information, called the elemental composition fragmentation tree (ECFT). The NMFT is generated by using only the peak detection functionality of the MEF tool. The ECFT type is obtained using the fully functional MEF tool, including combinatorial

dx.doi.org/10.1021/ac2034216 | Anal. Chem. 2012, 84, 5524−5534

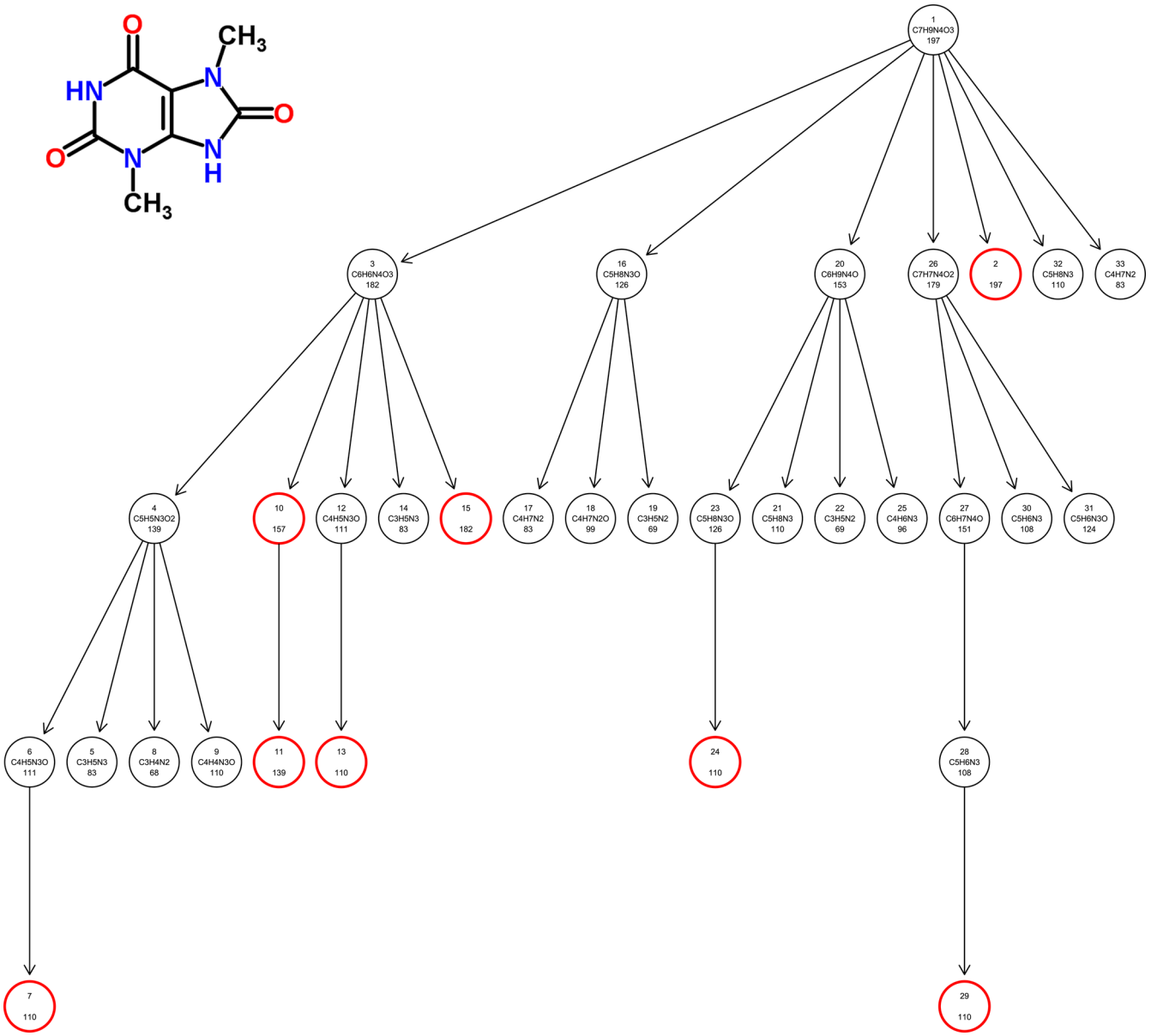
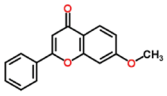


Figure 1. Fragmentation tree of 3,7-dimethyluric acid [InChI = 1/C7H8N4O3/c1-10-3-4(8-6(10)13)11(2)7(14)9-5(3)12/h1-2H3,(H,8,13)-(H,9,12,14)]. The nodes for which the elemental composition is not calculated are drawn in red.

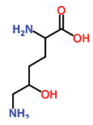
rules to calculate elemental compositions. The NMFT may be useful as an alternative to the ECFT in case the mass accuracy of the data does not allow elemental composition calculation of the parent ion and its fragments. Obviously, the NMFT contains many more nodes/fragments than the ECFT generated from the same MS^n data file, because for generating an NMFT no additional chemical constraining is used, in contrast to generating an ECFT. Thus, a major drawback of the nominal mass fragmentation tree is the less precise representation of its nodes because a single nominal mass can still point to many different elemental compositions and thus possible fragments. Also, in an elemental composition fragmentation tree, chemical constraints checking elemental composition consistency are applied. As a consequence, peaks not fitting the elemental compositions of the precursor and child ions are eliminated and the fragments are more precisely represented. (See Figure 1 and Supplemental Figures 5−10 in the Supporting Information, for examples of fragmentation trees showing the difference between NMFT and ECFT representations.)

To identify which type of fragmentation tree representation (ECFT or NMFT) is more suitable to compare fragmentation trees using the implemented fingerprint-based search algorithm, the following experiment was carried out. Two different metabolites were analyzed at four different laboratories, but on a similar MS instrument and using the same acquisition protocol and data processing tools. Afterward, both the ECFT and NMFT were generated from each acquired spectral tree. Figure 2 shows the number of fragments generated for each type of fragmentation tree for each laboratory and the degree of similarity of the fragmentation trees between the different laboratories. For all laboratories, the NMFTs contained more nodes/fragments than the ECFTs (Figure 2a). This higher number of NMFT features was mainly due to peaks not related to the fragmentation pattern and peaks that were characteristic for a specific laboratory. As a consequence, the similarity values calculated from ECFTs are consistently higher than those calculated from NMFTs (Figure 2b). We therefore concluded that MS^n spectra generated

dx.doi.org/10.1021/ac2034216 | *Anal. Chem.* 2012, 84, 5524−5534



	Lab–1	Lab–2	Lab–3	Lab–4
ECFT	44	61	60	65
NMFT	48	80	70	69



	Lab–1	Lab–2	Lab–3	Lab–4
ECFT	36	31	29	34
NMFT	43	47	38	46

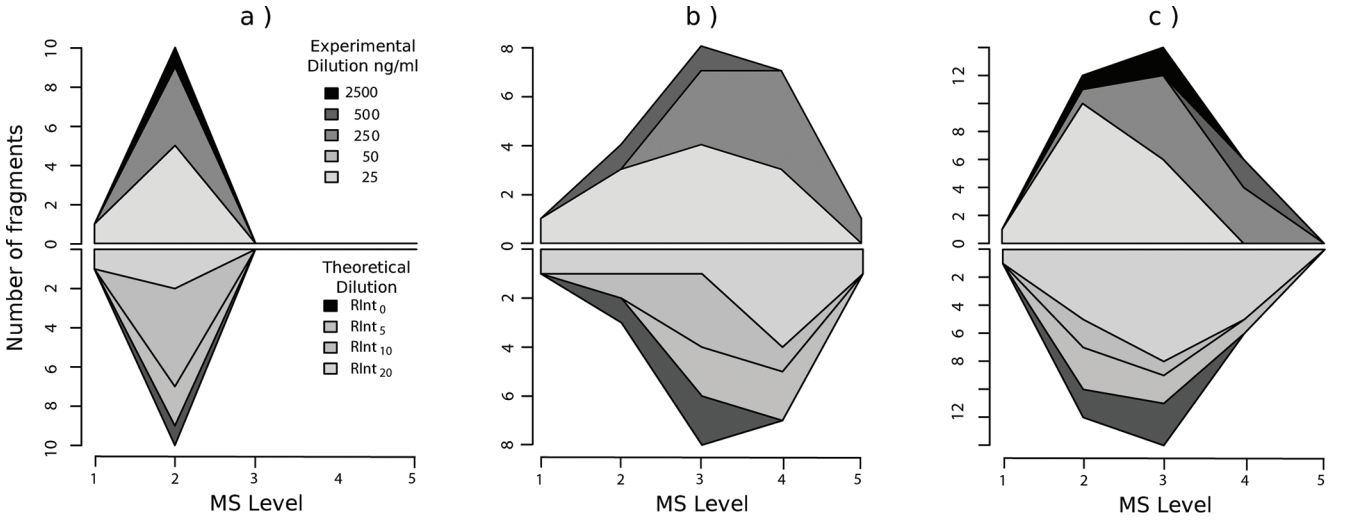
a)

ECFT \ NMFT	Lab–1	Lab–2	Lab–3	Lab–4
Lab–1	1\1	0.67\0.59	0.73\0.68	0.69\0.66
Lab–2		1\1	0.82\0.73	0.82\0.68
Lab–3			1\1	0.86\0.79
Lab–4				1\1

ECFT \ NMFT	Lab–1	Lab–2	Lab–3	Lab–4
Lab–1	1\1	0.75\0.60	0.79\0.74	0.80\0.69
Lab–2		1\1	0.79\0.69	0.80\0.66
Lab–3			1\1	0.84\0.80
Lab–4				1\1

b)

Figure 2. (a) Number of fragments in the ECFTs and NMFTs for the different laboratories. (b) Similarity value calculated using the fingerprints generated from ECFTs (in front of slash) versus NMFTs (behind the slash). The compounds analyzed are 7-methoxy-2-phenyl-4*H*-chromen-4-one (left) and 5-hydroxylysine (right).



Experimental Dilution	Theoretical Dilution			
ng/ml	$RInt_0$	$RInt_5$	$RInt_{10}$	$RInt_{20}$
2500	1.00	0.98	0.90	0.00
500	0.98	1.00	0.93	0.22
250	0.98	1.00	0.93	0.17
50	0.75	0.79	0.71	0.20
25	0.62	0.67	0.77	0.20

Experimental Dilution	Theoretical Dilution			
ng/ml	$RInt_0$	$RInt_5$	$RInt_{10}$	$RInt_{20}$
2500	1.00	0.92	0.84	0.50
500	0.84	0.75	0.42	0.29
250	0.87	0.70	0.55	0.46
50	0.41	0.59	0.43	0.42
25	0.41	0.59	0.43	0.42

Experimental Dilution	Theoretical Dilution			
ng/ml	$RInt_0$	$RInt_5$	$RInt_{10}$	$RInt_2$
2500	1.00	0.90	0.69	0.00
500	0.95	0.96	0.76	0.62
250	0.87	0.87	0.75	0.62
50	0.56	0.62	0.56	0.47
25	0.42	0.48	0.43	0.31

Figure 3. Fragmentation topology curves (number of fragments at each MS level) of fragmentation trees (ESI in positive mode) obtained from metabolites at different concentrations versus a simulated dilution series in silico extracted from the data obtained at the highest compound concentration by using different threshold settings for peak picking in each spectrum. $RInt_0$, $RInt_5$, $RInt_{10}$, and $RInt_{20}$ refer to relative intensity thresholds of 0%, 5%, 10%, and 20% of the base peak. The metabolites are 7-hydroxyflavone (a), 6-methoxyflavone (b), and 3,2′-dihydroxyflavone (c). The tables in the lower part of the figure show the similarity values between experimental and simulated trees. Values in red represent the highest similarity values when compared with the experimental data.

using nominal masses are less efficient in comparing fragmentation trees, and thus database searching, than MS^n spectra based on accurate masses enabling elemental composition calculations.

In practice, when measuring biological samples, the experimental fragmentation tree topology of a certain metabolite will be different, to a more or lesser extent, from the one stored in the reference library, because the number of acquired fragments of a certain compound may change between experiments, e.g., depending upon specific instrument sensitivity. However, the most relevant factor that influences the size (depth and width) of the fragmentation tree, i.e., the number of fragments per level, will be the concentration of the sample measured: the higher the concentration of the sample, the greater the number of molecular ions trapped and the larger the number of fragments generated, and therefore the larger the size of the fragmentation tree. We therefore tested whether our method is capable of assigning the correct identity to a certain metabolite while its spectral tree was not at the same depth as the reference compounds. In this study MS^n spectra of three different flavonoids were acquired at five different concentration levels (2500, 500, 250, 50, and 25 ng/mL). After the MS^n data were processed in the elemental composition fragmentation tree mode, the fragmentation trees were searched and compared to the reference library obtained at 2500 ng/mL. In addition, the MS^n data generated at 2500 ng/mL were processed

dx.doi.org/10.1021/ac2034216 | *Anal. Chem.* 2012, 84, 5524–5534

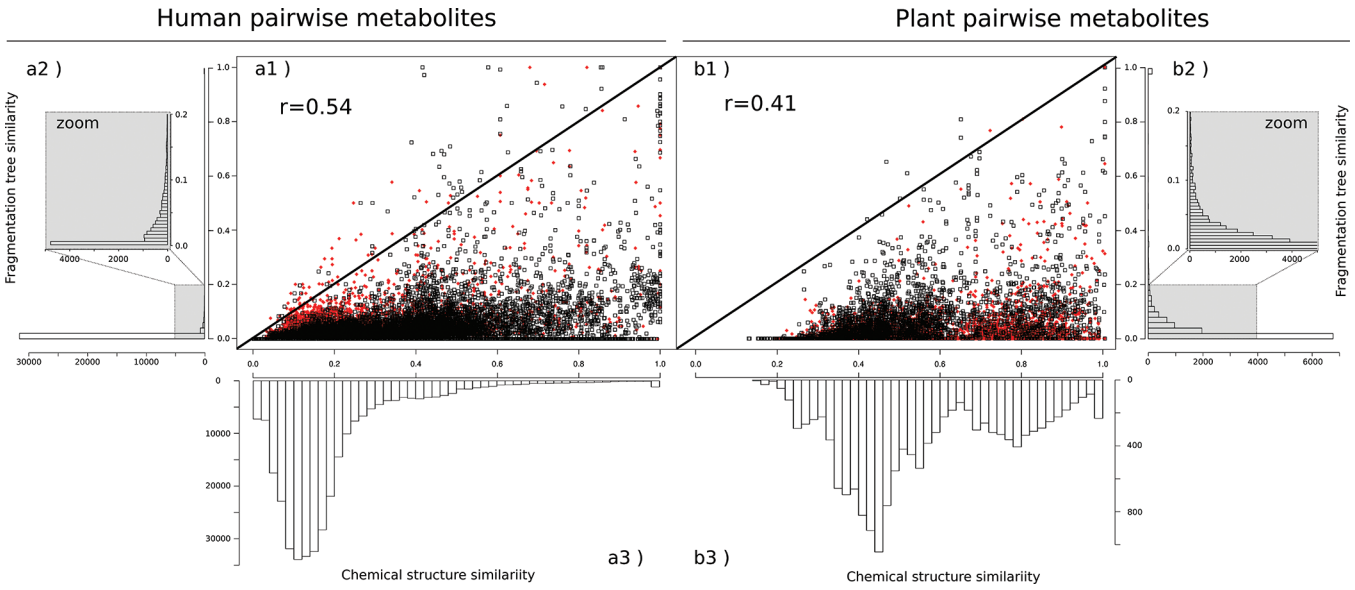


Figure 4. (a1, b1) Pairwise chemical structure similarity versus fragmentation tree similarity for human (a) and plant (b) metabolites. (a2, b2) Fragmentation tree similarity distributions. (a3, b3) Chemical structure similarity distributions.

by the MEF tool at four different relative intensity threshold settings, i.e., 0%, 5%, 10%, and 20% relative intensity compared to the base peak, thereby simulating a concentration series in silico.

Figure 3 shows the resemblance between the fragmentation tree topologies of the simulated and the experimental dilutions. The effect of changing the compound concentration was comparable between simulated and experimentally obtained fragmentation trees: both types of trees showed a loss of fragments going from a high to a low concentration. However, the MS level at which shrinkage of the fragmentation trees occurs differs. Ideally (for optimal dereplication use of an MS^n library) the theoretically generated dilution series should cover as much as possible the fragmentation tree series of its experimental dilution series. In the lower part of Figure 3, similarity matrices between simulated and experimental dilutions were calculated. The fragmentation trees simulated for low compound concentrations were most similar to experimentally obtained fragmentation trees at low concentrations, while fragmentation trees simulated for higher compound concentrations were most similar to the more concentrated experimentally obtained trees. These results indicate that (I) the reduction of fragments by lowering the relative intensity can be used to simulate spectra generated at lower compound concentrations and (II) by applying this MS^n data processing at different peak intensity thresholds, the probability to correctly assign the metabolite identity is increased, even though its MS^n data were obtained at a compound concentration different from that used to populate the reference library. Although the identification probability is increased, it remains lower than if the compounds were present at similar concentrations. To ensure that the theoretical fragmentation spectrum of a library compound will simulate correctly the experimental spectrum of that compound at an unknown sample concentration, it would be optimal to simulate spectra at a large number of thresholds for relative intensity. In this study four different intensity threshold parameters were used, but this number can easily be increased to more precisely compare and match experimental fragmentation spectra from compounds present at relatively low concentrations in biological samples with library spectra obtained at relatively high concentrations.

After definition of the fragmentation tree representation (by means of elemental compositions) and how to deal with differences in concentrations of the obtained fragmentation trees (processing of the library trees with multiple relative intensity settings), the dereplication functionality of the library was further investigated by monitoring the identity predictions of replicate measurements of a number of metabolites which were already in the library. For 282 metabolites, 765 replicate fragmentation trees were acquired and used as validation samples. Using our fingerprint-based approach, 722 (94%) of these fragmentation trees were correctly identified. Nevertheless, 43 fragmentation trees (belonging mainly to plant metabolites) were found having a higher similarity with a tree of another compound than the measured compound. In all cases the difference between the highest similarity value and the similarity value of the tree of the compound measured was relatively small. Most of these cases were due to isomers such as epicatechin and catechin; it is known that these compounds are very difficult to discern.[7] In Supplemental Figure 8 in the Supporting Information, the distributions are shown of all similarity values of the measured compounds together with the similarity values of the compound most similar to the measured one. For dereplication/identity search use, the difference between these should be large to get clear identity assignments. For similarity search, however, the similarity value of the first nonidentical compound should be as high as possible to extract relevant substructural information about the unknown. The use of these libraries for similarity search is further investigated in the next section of this paper.

Similarity searching for molecules frequently aims to detect molecules having similar biological activity.[39] Adopting this concept, we aimed to elucidate whether similar chemical structures will result in (partially) similar fragmentation trees and vice versa, which would help in the identification of unknown compounds. For the entire set of compounds present in each library (human and plant metabolites), all pairwise similarity values (both the chemical structure similarity and the fragmentation tree similarity values) were calculated and plotted in

dx.doi.org/10.1021/ac2034216 | Anal. Chem. 2012, 84, 5524−5534

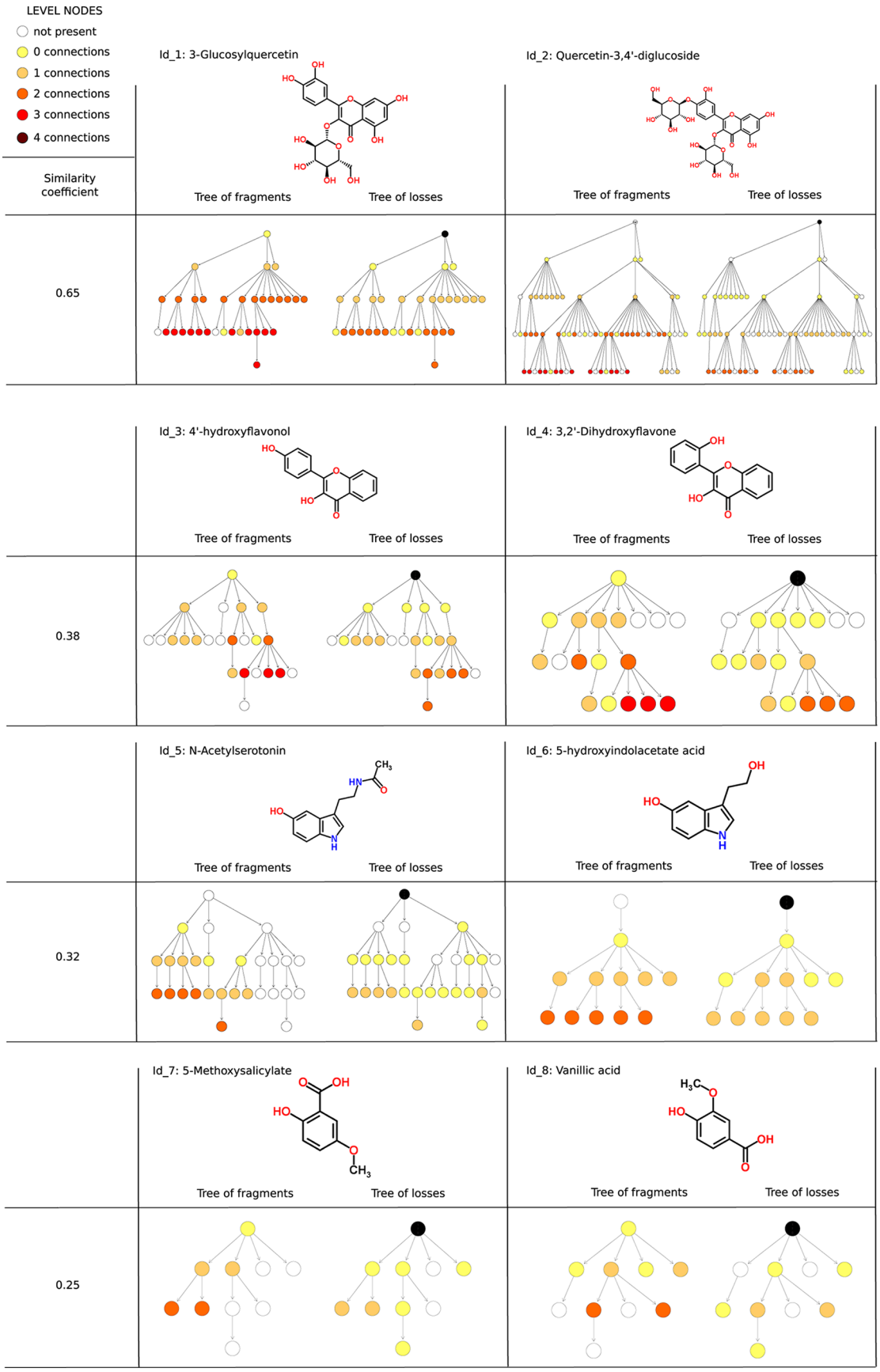


Figure 5. Comparison of fragmentation and neutral loss trees for two different compounds. The first column shows the fragmentation tree similarity value. In the next columns the fragmentation and neutral loss trees are plotted. White nodes indicate that they are unique in the corresponding tree. Colored nodes correspond to overlapping nodes or branches. The different colors indicate the number of upper nodes also overlapping and connected.

dx.doi.org/10.1021/ac2034216 | *Anal. Chem.* 2012, 84, 5524−5534

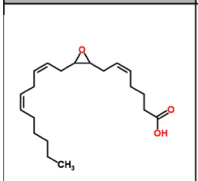
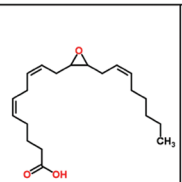
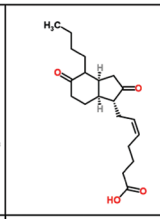
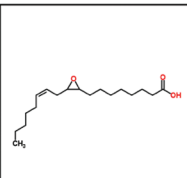
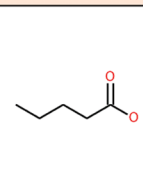
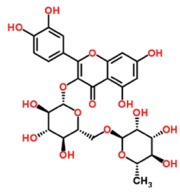
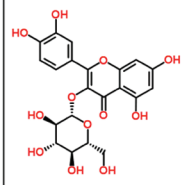
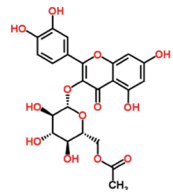
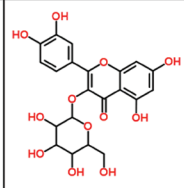
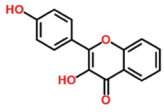
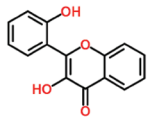
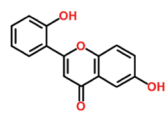
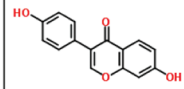
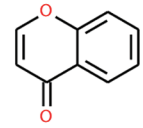
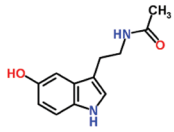
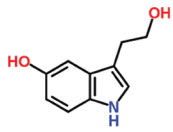
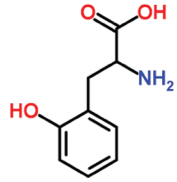
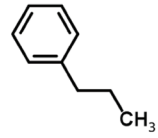
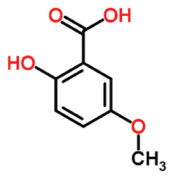
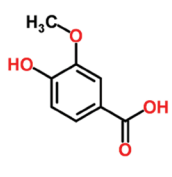
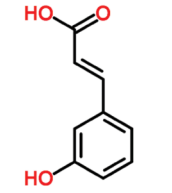
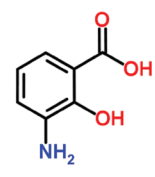
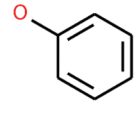
Structures to test	Structures with most similar fragmentation trees			Maximum common substructures
				
8,9-EET Nr.Frag:46 Coef.Value:1	11,12-EET Nr.Frag:37 Coef.Value:0.73	bicyclo-PGE2 Nr.Frag:39 Coef.Value:0.22	9(10)-EpOME Nr.Frag:22 Coef.Value:0.18	
				
Rutin Nr.Frag:12 Coef.Value:1	3-Glucosyquercetin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

Figure 6. Result of querying the MS^n library with fragmentation trees derived from "unknown" (test) metabolites (first column). The second, third, and fourth columns show structures with the most similar fragmentation trees. The boxes below the structures list the number of fragments (Nr.Frag) that are characteristic of the compound's fragmentation tree and the similarity value (Coef.Value) compared to the fragmentation tree of the unknown metabolite. The last column shows the MCSSs extracted from the compounds listed in the middle columns.

Figure 4. A perfect correlation of the chemical structure metric with the fragmentation tree metric should emerge as a diagonal line. The correlation coefficients between structural and fragmentation tree similarity for both libraries, human and plant, are $r_{human} = 0.54$ and $r_{plant} = 0.41$. Plots a1 and b1 reflect that mainly the region below this diagonal is occupied with

dx.doi.org/10.1021/ac2034216 | *Anal. Chem.* 2012, 84, 5524−5534

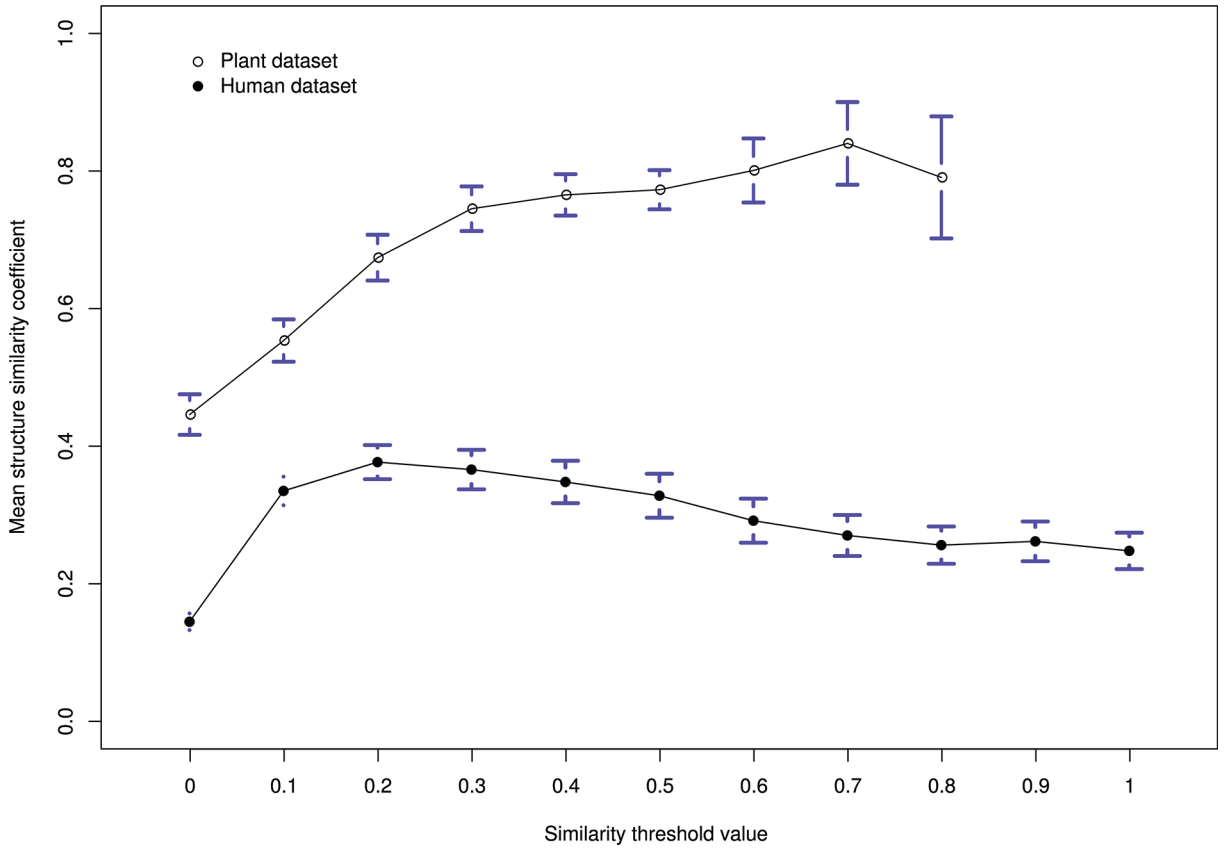


Figure 7. Mean structure similarity value of the MCSS and the structure of the queried metabolite for different fragmentation tree similarity thresholds (used for generating the MCSS) for the human and plant MS^n libraries.

similarity pairs. This means that similar fragmentation trees are typically the result of compounds having similar chemical structures, while compounds having similar chemical structures do not by definition generate similar fragments. In the latter case, only a part of both fragmentation trees will overlap, leading to relatively low fragmentation similarity values. The observed phenomenon that a pair of compounds with a (relatively) high fragmentation tree similarity value also has a high chemical structure similarity value is called "neighborhood behavior".[40] The two sets of compound libraries tested here differ with respect to their variation in chemical structures. A structure diversity analysis of both libraries, see Supplemental Text 1 in the Supporting Information, showed that the plant database, having a diversity value of 0.48, is structurally less diverse than the human metabolite database, having a diversity value of 0.82. This difference is due to the fact that the plant library mostly contains structurally related polyphenol structures, while the human library is mainly composed of lipids, amino acids, and sugars. The distribution plots of fragmentation tree similarities (Figure 4a2,b2), in both libraries, show at relatively low similarity value an optimum, indicating that fragmentation trees are unique and characteristic for each compound, making the process of replication more efficient.

Knowledge about fragmentation tree similarity and which branches/building blocks are in common is relevant for posterior interpretation of the MS^n results and annotation of the metabolite(s) under investigation. In the recently published MEF tool,[18] a new visualization feature has been implemented that highlights nodes that are common between the fragmentation trees compared (see Figure 5). Clearly, the degree of similarity between fragmentation trees can be deduced from the number of colored nodes. This colored visualization of pairs of fragmentation trees can also help to discover relevant building blocks (subtrees) to focus on when interpreting fragmentation spectra of unknown metabolites. The compounds *N*-acetylserotonin and 5-hydroxyindoleacetic acid clearly illustrate this added value: the fragmentation tree of 5-hydroxyindoleacetic acid is almost a complete building block of the fragmentation tree of *N*-acetylserotonin. We also emphasize the relevance of comparing neutral loss trees for the identification of similar fragmentations. In some cases it can happen that only the neutral loss tree provides structural likeness between metabolites.

As several studies[16,5,7] have shown previously, the existence of correlation between the building blocks generated in the MS^n data and the substructure of the measured molecule makes partial structural elucidation of unknown compounds possible, provided that similar building blocks are found in a reference library. For the MS^n data in our library, structure information for each fragment or neutral loss is not (yet) available. Although we were able to find similar building blocks in the MS^n data, in the library no substructure information was returned. However, we were able to extract the MCSS from the structures that have the most similar fragmentation trees (Figure 6 and Supplemental Table 4 in the Supporting Information). Thus, the extracted MCSS is a substructure that is likely part of the unknown molecule. This information can help MS experts with the identification of unknown compounds, e.g., by using it as an input together with the elemental composition of the unknown compound in a structure generator.[41] To generate the MCSS, we need to define the list of

dx.doi.org/10.1021/ac2034216 | *Anal. Chem.* 2012, 84, 5524−5534

compounds used as input for the MCSS calculation tool. As a consequence, the obtained MCSS depends heavily on the number of compounds considered to have similar fragmentation trees. This list of compounds can be defined by setting a fragmentation similarity threshold. Upon decreasing this threshold, the number of similar fragmentation tree hits increases and therefore the MCSS gets smaller and less specific. The larger the MCSS, the greater the amount of structural information available for identifying the unknown compound. We therefore aimed to determine the smallest threshold to set while still retrieving a structurally relevant MCSS.

In Figure 7 the effect of lowering the fragmentation tree similarity threshold value on the calculated MCSS is shown. Each fragmentation tree entry was compared to the rest of the fragmentation trees in the library, so ultimately the average and standard deviation of the structural similarity values for all entries in the database were calculated. The MCSS obtained from the plant library has higher structural similarity values than the MCSS extracted from the human library over the whole range of fragmentation similarity threshold values. This is obviously due to the fact that the plant library contains several series of isomers and structurally related compounds. Because a higher structural MCSS similarity is correlated with the size of the MCSS, the size of the MCSS of the plant library is larger and structurally more informative than the MCSS obtained from the human library. This result underlines the importance of filling the database with as many structurally related compounds as possible to obtain as much information as possible about the identity of unknown metabolites. Over the whole range of fragmentation similarity thresholds, the generated MCCS was relatively stable. Below values of 0.3 for the plant library and 0.2 for the human database, the threshold reached a value where the obtained MCSS seems to become structurally less informative, which in practical terms means we can use a fragmentation similarity threshold of about 0.25 as a good compromise to extract structural information of an unknown metabolite from a compound library.

Overall, although fragmentation trees may not be very similar, they may still be helpful in providing structure information and in partly elucidating the structure of unknown compounds.

CONCLUSION

In this paper we introduce a new cheminformatical approach to calculate the similarity between mass spectral fragmentation trees, which can be helpful in the annotation of compounds detected using LC/MS-based metabolomics approaches. The new approach can be used to query multistage mass spectral data in MS^n libraries to define structure−spectrum relationships and potentially deduce substructures within unknowns.

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

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

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: m.rojas@lacdr.leidenuniv.nl. Phone: +31 71 527 4220. Fax: +31 71 527 4565. E-mail: t.reijmers@lacdr.leidenuniv.nl. Phone: +31 71 527 4320.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

(1) Kind, T.; Fiehn, O. *Bioanal. Rev.* **2010**, *2*, 23−60.
(2) Hernández, F.; Portolés, T.; Pitarch, E.; López, F. J. *TrAC, Trends Anal. Chem.* **2011**, *30*, 388−400.
(3) Grange, A. H.; Genicola, F. A.; Sovocool, G. W. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2356−2369.
(4) Portolés, T.; Pitarch, E.; López, F. J.; Hernández, F.; Niessen, W. M. A. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 1589−99.
(5) Sheldon, M. T.; Mistrik, R.; Croley, T. R. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 370−6.
(6) van der Hooft, J. J. J.; Vervoort, J.; Bino, R. J.; De Vos, R. C. H. *Metabolomics* **2011**, 1−13.
(7) van der Hooft, J. J. J.; Vervoort, J.; Bino, R. J.; Beekwilder, J.; de Vos, R. C. H. *Anal. Chem.* **2011**, *83*, 409−16.
(8) Bristow, A. W. T.; Webb, K. S.; Lubben, A. T.; Halket, J. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 1447−54.
(9) Jansen, R.; Lachatre, G.; Marquet, P. *Clin. Biochem.* **2005**, *38*, 362−372.
(10) Hopley, C.; Bristow, T.; Lubben, A.; Simpson, A.; Bull, E.; Klagkou, K.; Herniman, J.; Langley, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 1779−1786.
(11) Champarnaud, E.; Hopley, C. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 1001−7.
(12) Palit, M.; Mallard, G. *Anal. Chem.* **2009**, *81*, 2477−2485.
(13) 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*, 703−14.
(14) Akiyama, K.; Chikayama, E.; Yuasa, H.; Shimada, Y.; Tohge, T.; Shinozaki, K.; Hirai, M. Y.; Sakurai, T.; Kikuchi, J.; Saito, K. *In Silico Biol.* **2008**, *8*, 339−345.
(15) Smith, C. A.; O'Maille, G.; Want, E. J.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−87.

(16) Wolfender, J.; Waridel, P.; Ndjoko, K.; Hobby, K. R.; Major, H. J.; Hostettmann, K. *Analusis* **2000**, *28*, 895−906.
(17) Scheubert, K.; Hufsky, F.; Rasche, F.; Böcker, S. *J. Comput. Biol.* **2011**, *18*, 377−391.
(18) Rojas-Chertó, M.; Kasper, P. T.; Willighagen, E. L.; Vreeken, R.; Hankemeier, T.; Reijmers, T. *Bioinformatics* **2011**, *27*, 2376−2383.
(19) Murray-Rust, P.; Rzepa, H. S.; Wright, M. *New J. Chem.* **2001**, *25*, 618−634.
(20) Holliday, G. L.; Murray-Rust, P.; Rzepa, H. S. *J. Chem. Inf. Model.* **2006**, *46*, 145−57.
(21) Kuhn, S.; Helmus, T.; Lancashire, R. J.; Murray-rust, P.; Rzepa, H. S.; Steinbeck, C.; Willighagen, E. L. *Structure* **2007**, 2015−2034.
(22) Rasche, F.; Svatos, A.; Maddula, R. K. R. K.; Böttcher, C.; Böcker, S. *Anal. Chem.* **2011**, *83*, 1243−1251.
(23) Oberacher, H.; Pavlic, M.; Libiseller, K.; Schubert, B.; Sulyok, M.; Schuhmacher, R.; Csaszar, E.; Köfeler, H. C. *J. Mass Spectrom.* **2009**, *44*, 494−502.
(24) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
(25) Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85−8.
(26) McLafferty, F. W.; Zhang, M. Y.; Stauffer, D. B.; Loh, S. Y. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 92−5.
(27) Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859−866.
(28) Hansen, M. E.; Smedsgaard, J. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 1173−80.
(29) Rasche, F.; Scheubert, K.; Hufsky, F.; Zichner, T.; Kai, M.; Svatos, A.; Böcker, S. *Anal. Chem.* **2012**, *84*, 3417−3426.
(30) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.
(31) Fligner, M. A.; Verducci, J. S.; Blower, P. E. *Technometrics* **2002**, *44*, 10.
(32) Baldi, P.; Nasr, R. *J. Chem. Inf. Model.* **2010**, *50*, 1205−22.
(33) Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. *Nat. Biotechnol.* **2004**, *22*, 1459−66.
(34) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, revised ed.; Springer: New York, 2007.
(35) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−28.
(36) Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. *Org. Biomol. Chem.* **2005**, *3*, 1832−4.
(37) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(38) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. *Curr. Pharm. Des.* **2006**, *12*, 2111−2120.
(39) Stumpfe, D.; Bajorath, J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260−282.
(40) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(41) Braun, J.; Gugisch, R.; Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 542−8.