Comparing Fragmentation Trees from Electron Impact Mass Spectra with Annotated Fragmentation Pathways

Franziska Hufsky[1,2] and Sebastian Böcker[1]

1 **Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany,** {franziska.hufsky,sebastian.boecker}@uni-jena.de
2 **Max Planck Institute for Chemical Ecology, Beutenberg Campus, Jena, Germany**

Abstract

Electron impact ionization (EI) is the most common form of ionization for GC-MS analysis of small molecules. This ionization method results in a mass spectrum not necessarily containing the molecular ion peak. The fragmentation of small compounds during EI is well understood, but manual interpretation of mass spectra is tedious and time-consuming. Methods for automated analysis are highly sought, but currently limited to database searching and rule-based approaches. With the computation of hypothetical fragmentation trees from high mass GC-MS data the high-throughput interpretation of such spectra may become feasible. We compare these trees with annotated fragmentation pathways. We find that fragmentation trees explain the origin of the ions found in the mass spectra in accordance to the literature. No peak is annotated with an incorrect fragment formula and 78.7 % of the fragmentation processes are correctly reconstructed.

1998 ACM Subject Classification J.2 Physical Sciences and Engineering (Chemistry)

Keywords and phrases metabolomics, GC-MS, computational mass spectrometry, fragmentation trees

Digital Object Identifier 10.4230/OASIcs.GCB.2012.12

1 Introduction

Metabolomics, also called "metabonomics" or "metabolic profiling", is a rapidly developing field of 'omics' research, dealing with the detection, identification and quantification of low molecular-weight compounds (typically below 1000 Da) in cells, organs or organisms. The analysis and identification of small molecules is important in many areas of biology and medicine such as biomarker discovery, diagnostics, pharmaceutical chemistry and functional genomics [2, 9, 36]. The metabolome consists of various compounds that belong to a wide array of compound classes, including sugars, acids, bases, lipids, hormonal steroids, and many others [5, 18]. The structural diversity of metabolites is extraordinarily large despite of their small size [21]. Unlike biopolymers such as proteins and glycans, metabolites are not made up of repeated building blocks. The genome sequence does not reveal information about metabolite structure, as it does for protein structure. The number of metabolites in any higher eukaryote is estimated between 4000 and 20 000 [7]. Unfortunately, an astounding number of these metabolites remain uncharacterized with respect to their structure and function [26].

At the moment there is no single instrumental platform that can analyze all metabolites [5, 21]. Mass spectrometry (MS), typically coupled with a chromatographic separation

German Conference on Bioinformatics 2012 (GCB'12).
Editors: S. Böcker, F. Hufsky, K. Scheubert, J. Schleicher, S. Schuster; pp. 12–22

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

technology, is one of the key technologies for the identification of small molecules. It has excellent compound specificity and high sensitivity. In particular, MS sensitivity is orders of magnitude higher than that of nuclear magnetic resonance (NMR) [21, 29]. Several kinds of analytical apparatus have been developed and most of these combine chromatography with a fragmentation technique to increase compound specificity. Gas chromatography coupled to mass spectrometry (GC-MS) is one of the most frequent tools for profiling metabolites and it was in existence decades before liquid chromatography MS (LC-MS) [8, 14]. The amount of data produced during metabolomic analysis is hard to process and analyze manually [18].

The most common ionization technique in GC-MS is electron impact ionization (EI). The resulting fragment-rich mass spectra are in general consistent and specific for each molecule [20, 21] and fragmentation mechanisms are already well described [23]. Reference spectra were collected over many years, allowing for automated interpretation via database search [24]. Where the compound is unknown, comparing the spectrum obtained to a spectral library will result in imprecise or incorrect hits, or no hits at all [8, 18, 20]. To cover a wider range of compounds in silico fragmentation is used to predict spectra of compounds with known structure [11, 12, 19, 37]. A first step towards the structural elucidation of fully unknown compounds is feature-based identification of the compound class [17, 19, 34, 35]. See Kind and Fiehn [20] for a comprehensive review of computational techniques for small molecule mass spectrometry.

Böcker and Rasche [3] introduced fragmentation trees for the *de novo* interpretation of metabolite fragmentation data. The fragmentation tree concept helps to identify the molecular formulas of the compound and to interpret the fragmentation process. Nodes are annotated with the molecular formulas of the fragments, and edges represent fragmentation events, that is, neutral or radical losses. Computing fragmentation trees does not require databases of compound structures or mass spectra or expert knowledge of fragmentation. The trees can be compared to each other to identify compound classes of unknowns [28]. Expert evaluation suggests that the fragmentation trees from LC-MS^2 data are of very good quality [29]. Fragmentation trees can also be computed from LC-MS^n data [31]. Recently, Hufsky *et al.* [16] presented a computational method for the *de novo* interpretation of EI fragmentation data, based on fragmentation tree construction, and applied it to real world data. Besides a list of common losses, this method does not use any chemical expert knowledge, but does require high mass accuracy of the measurements [16].

In this study, we evaluate the quality of fragmentation trees computed from EI fragmentation data [16]. To evaluate the potential of fragmentation trees to reconstruct fragmentation processes we compare them to annotated fragmentation pathways of 22 compounds from the literature. The constructed fragmentation trees were not supposed to depict the actual fragmentation reactions. They however agree well with the annotated pathways explaining the origin of the respective ions found in the mass spectra. No peak was annotated with an incorrect fragment formula and 78.7 % of the fragmentation processes were correctly reconstructed. For the annotation of the fragmentation processes in the literature the molecular structures of the compounds were used. In contrast, the computation of fragmentation trees works without this knowledge. The assignment of molecular formulas to all fragments and explanation of relevant fragmentation reactions independent of existing library knowledge, supports the structural elucidation of unknown compounds. Combined with a method for the automated comparison of fragmentation trees [15, 28] it will enable the automated analysis of metabolites that are not included in common libraries.

2 Methods

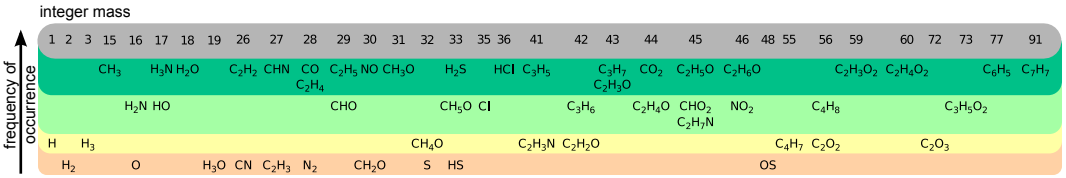
For the interpretation of the GC-MS spectra we use fragmentation trees as introduced by Böcker and Rasche [3] for LC-MS^2 spectra. A hypothetical fragmentation tree models fragmentation cascades by annotating nodes with the molecular formulas of fragments, and edges with fragmentation events, that is, neutral or radical losses. The root of the fragmentation tree is labeled with the molecular formula of the unfragmented ion. For LC-MS^2 data the molecular ion mass is known. EI results in a mass spectrum not necessarily containing the molecular ion peak. Hufsky *et al.* [16] proposed a method for computing fragmentation trees from such data.

To compute a fragmentation tree from an EI fragmentation spectrum a fragmentation graph is constructed. All candidate molecular formulas within the mass accuracy of the instrument are computed for each peak. The fragmentation graph contains a node for each decomposition. The nodes are colored, such that all explanations of the same peak receive the same color. Nodes are weighted using mass deviation and peak intensity [3]. Two nodes are connected by an edge (corresponding to a loss) if the second molecular formula is a subformula of the first. Edges are weighted according to their plausibility as real fragmentation steps considering the mass of the loss, the ratio between carbon and hetero atoms, and common losses for EI fragmentation (see Table 1). See [16] for a detailed description of the scoring.

The colorful subtree with maximum sum of edge weights is the explanation of the observed fragments, that fits best with the given conditions. Considering trees every fragment is explained by a unique fragmentation pathway, see [29]. Considering only colorful trees every peak is explained by a single fragment. Several fragments resulting in a single peak is an extremely rare event in practice.

By demanding that each fragment in the fragmentation spectrum is generated by a single fragmentation pathway we slightly oversimplify the problem. Our optimization algorithm will choose the mainly occurring pathway to compute a fragmentation tree. There are two exceptions to this reasoning: (1) In the resulting fragmentation tree, assume that some fragment f_3 is cleaved from f_2, which is in turn cleaved from f_1. Solely from the EI fragmentation pattern and without additional structural information, it cannot be ruled out that fragment f_3 is in truth cleaved directly from f_1. Both interpretations are implicitly encoded in the fragmentation tree: the fragmentation may occur from the fragment's direct parent in the tree or from any of its parents (see Figure 1(a)). (2) In the resulting fragmentation tree, assume that some fragment f_2 is cleaved from a fragment f_1 by loosing l_1 and another fragment f_3 is cleaved from f_1 by loosing l_2. Further, another fragment f_4 is cleaved from f_2 by loosing l_2. Solely from the data, it cannot be ruled out that fragment f_4 is in truth cleaved from f_3 by loosing l_1. Again, both interpretations are implicitly encoded

Table 1 List of common losses over the alphabet used throughout this study (CHNOPSCl) for scoring EI fragmentation reactions [13]. The losses are sorted by integer mass and their probability of occurrence in a GC-MS spectrum [16]. Losses in the first row (dark green) are very common and thus score high, while losses in the last row (orange) are not-that-common and thus score comparatively low.



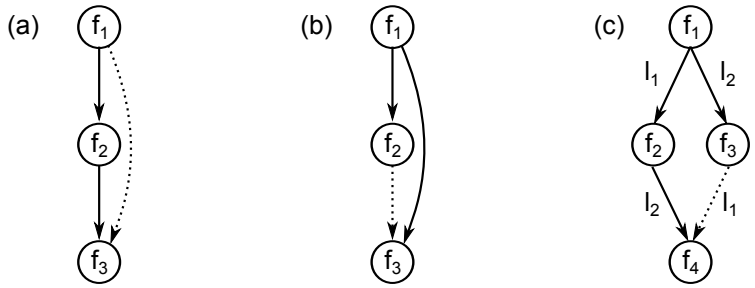


Figure 1 Comparing fragmentation trees (solid edges) with annotated pathways from the literature (dashed edges). (a) In the fragmentation tree, fragment f_3 is cleaved from f_2, which is in turn cleaved from f_1. It cannot be ruled out that fragment f_3 is in truth cleaved directly from f_1 (dotted edge). Both interpretations are implicitly encoded in the fragmentation tree. We evaluate these edges as *correct*. (b) In the fragmentation tree, fragment f_3 is cleaved directly from f_1, while in truth it is cleaved from f_2 (dotted edge), which is in turn cleaved from f_1. We evaluate these edges as inserted *to high*. (c) *Parallelogram:* Fragment f_4 is cleaved from f_2 by loosing l_2, which is in turn cleaved from f_1 by loosing l_1. In truth fragment f_4 is cleaved by loosing l_2 first and l_1 afterwards (dotted edge). Both interpretations are implicitly encoded in the fragmentation tree. We evaluate these edges as *correct*.

in the fragmentation tree: the fragmentation may occur by loosing l_1 first and l_2 afterwards, or the other way (see Figure 1(c)). In the following, we call this constellation *parallelogram*.

EI is a hard ionization technique often resulting in missing or low intensity molecular ion peaks [20, 21]. Different from LC-MS^2 analysis the mass of the molecular ion is not known. All nodes in the graph are possible roots of the fragmentation tree. Molecular formulas explaining each peak cannot be restricted to sub-molecular formulas as proposed in [29]. Therefore the identification of the molecular ion and formula and the computation of the complete fragmentation tree is done in two separate steps.

We first identify the molecular ion and molecular formula using only a set of peaks that appear to be most relevant for the compound. These peaks are selected using three different criteria. We choose the k_1 most intense peaks, the k_2 peaks with highest *score*, and the k_3 peaks with highest *score* in the *upper m/z range*. The *score* is a combination of m/z value and relative intensity $m/z \cdot \ln(100 \cdot int_{rel})$ and the *upper m/z range* is the m/z region from $0.9\tilde{M}$ to $\tilde{M}$ where $\tilde{M}$ is the highest m/z of a peak detected in the spectrum. In this step fragmentation trees are computed using Dynamic Programming [3]. Afterwards we compute a fragmentation tree for the complete spectrum assuming that we know the correct molecular ion and molecular formula of the compound. The resulting fragmentation tree is rooted in this molecular formula. In this step, fragmentation trees are computed using Integer Linear Programming [30].

3 Data

The EI induced fragmentation of small molecules is well described in the literature. To evaluate the potential of fragmentation trees to reconstruct fragmentation processes we extract annotated fragmentation pathways for 22 compounds from different compound classes (see Table 2). In [1] Acheson *et al.* describe the fragmentation of alkyl acridines. We choose two simple alkylacridines and two reduced acridines containing chlorine. Further, we choose seven compounds from a study on alkyl isocyanides and methyl branched alkyl cyanides [10]. From [25] we select four dihydro-1,4-oxathiines with fragmentation paths additionally invest-

Table 2 Overview of the 22 reference compounds with fragmentation pathways annotated in the literature. If more than one compound of the same class is used, we denote the class name and give the mass range and average mass.

		mass	
compound (class)	**#**	**range**	**average**
alkyl acridines [1]	4	207.1 - 399.2 Da	276.1 Da
alkyl isocyanides & α-branched alkyl cyanides [10]	7	41.0 - 83.1 Da	69.1 Da
dihydro-1,4-oxathiines [25]	4	146.0 - 235.1 Da	178.8 Da
gossypol [27]	1	518.2 Da	
ephedrine [33]	1	165.1 Da	
2,1-benzisothiazoline 2,2-dioxide nitro derivatives [4]	5	214.0 - 242.0 Da	228.0 Da
all	22	41.0 - 518.2 Da	187.6 Da

igated with qualitative collisionally induced dissociation (CID) measurements. From [27] we extract the fragmentation pathway of gossypol and from [33] the one from ephedrine. Further, we choose five 2,1-benzisothiazoline 2,2-dioxide nitro derivatives from [4].

As the measured spectra are not available to us, we simulate spectra from the pathways. From the molecular formulas in the fragmentation pathway, we compute exact peak masses, and simulate "measured" spectra by adding a normal distributed error of 10 ppm on the mass of the fragment formula (without considering ionization). Peak intensities of the fragment peaks are taken from the literature. They are either given as actual number or estimated from the plotted spectrum. In addition, we add 70 % noise peaks with uniformly distributed masses smaller than the parent mass, and pareto distributed intensities.

4 Results

4.1 Molecular Ion Peak and Formula Identification

Fragmentation trees enable the identification of the molecular ion and the molecular formula of a metabolite if the molecular ion is present in the spectrum. EI is a hard ionization technique resulting in missing molecular ion peaks in about 30 % of the spectra [22]. For two compounds in our dataset, namely gossypol and ephedrine, the relative intensity of the molecular ion peak given in the literature was 0 %. We test the identification of the molecular ion and the molecular formula on the remaining 20 spectra containing a molecular ion peak.

To identify the molecular ion peak and its formula an alphabet of potential elements must be provided to the method. For all compounds, we use the six elements most abundant in metabolites, namely carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S) [18]. When analyzing the two compounds in our dataset containing chlorine (Cl), we also add this element to the alphabet. Information on whether a compound contains chlorine can be usually obtained from isotope pattern analysis.

Computing the molecular ion peak and molecular formula requires an average of 4.6 s for each compound. This time includes peak decomposition and graph construction. We discard peaks with no decomposition. We then choose the subset of peaks that appear to be most relevant for the compound as described above. We choose $k_1 = 10$ and $k_2 = k_3 = 5$, resulting in at most 20 peaks if the sets are not overlapping.

For all compounds, the method correctly detect the molecular ion peak. This is not surprising, since the molecular ion peaks have the highest m/z values in all spectra, based on the generation of noise peaks as described above. For 17 of the 20 compounds (85 %), the highest scoring suggestion for both the molecular ion peak and its molecular formula is

Table 3 Results of the tree evaluation. (a) Peak explanations in the annotated pathways compared to the computed fragmentation trees. [1]Percent of the explanations in the annotated pathways. [2]Percent of the explanations in the computed fragmentation trees. (b) Evaluation of the fragmentation events annotated in the fragmentation trees. For 5 of the 277 correct peak explanations, the fragmentation process leading to this fragment is not given in the literature. (c) Evaluation of the frequency of parallelograms in the annotated pathways. A parallelogram is *closed* if both fragmentation ways are annotated, and *"open"* otherwise. [3]Percent of the "open" parallelograms.

(a) fragments	pathway total	tree total	correct	missing	additional
peak explanations	296	284	277	19	7
precentage			93.6 %[1]	6.4 %[1]	2.5 %[2]

(b) losses	total	*correct*	correct, but *to deep*	correct, but *reverse order*	*to high*	*wrong*
losses	272	214	31	8	19	39
percentage		78.7 %	11.4 %	2.9 %	7.0 %	14.3 %

(c) parallelograms	total	*closed*	*"open"*	different in tree
parallelograms	99	29	70	8
percentage		29.3 %	70.7 %	11.4 %[3]

correct. For the remaining three compounds, the correct molecular formula is the second suggestion.

4.2 Fragmentation Tree Quality

We compute a hypothetical fragmentation for every compound, assuming that we know the correct molecular ion and molecular formula of the compound. In this step, all peaks of the spectrum are used for computation. Computation, including decomposition and graph construction, requires 1.5 s on average and a maximum of 18.5 s for the largest compound, namely gossypol. For this compound with mass of 518.2 Da, decomposition of all peaks requires 17.9 s (97 % of the total running time).

We compare the computed fragmentation trees with annotated fragmentation patterns from the literature. The fragmentation trees annotate 284 peaks in total (see Table 3(a)). Only seven of this explanations (2.5 %) are false positives, that is explanations of noise peaks as fragments. The remaining 277 peaks are annotated with the correct fragment formula. From all 296 fragments described in the pathways from literature 19 (6.4 %) could not be explained. There are different reasonings for a peak not being explained in the tree. For some peaks, the mass deviation of the measured peak mass to the exact mass is to high. This effect is getting stronger for smaller peaks, since mass deviation penalty is dependent of the peak intensity [16]. For other peaks, the fragmentation step resulting in this fragment gets a bad score. For example, the loss C_2H_2N that was annotated in the literature as a first fragmentation step for three of the alkyl isocyanides is not included in the list of common losses for EI fragmentation and is not even a combination of these (see Table 1 and [16]). Therefore the fragments resulting from this step could not be identified. Nevertheless, the method is capable of identifying losses that are very specific for a single compound or compound class and therefore not listed as a common loss (see [16]).

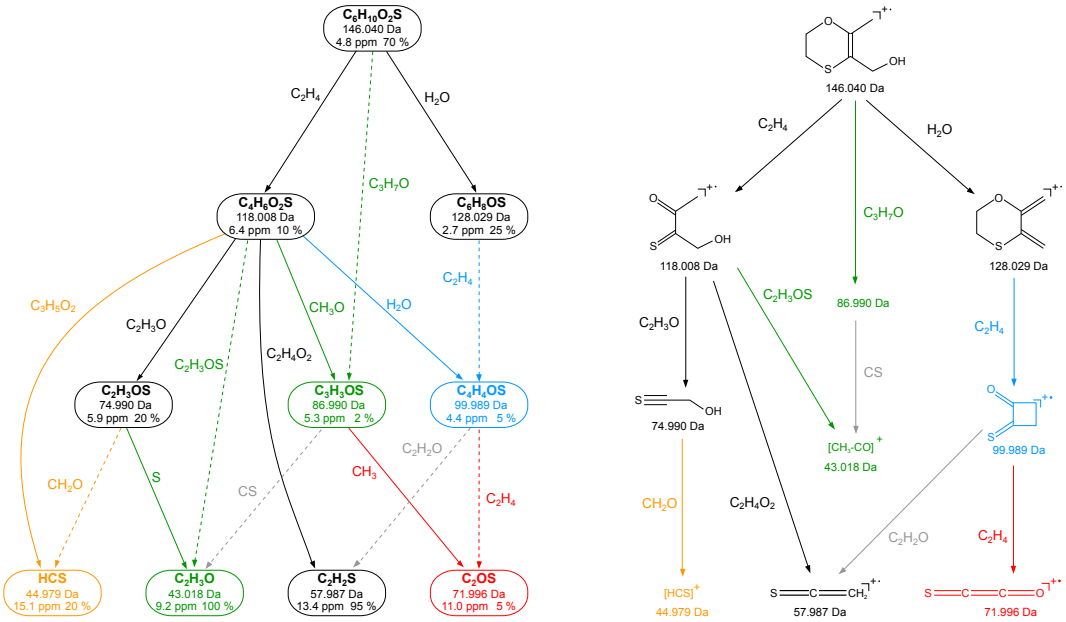


Figure 2 Computed fragmentation tree (solid edges) of 5,6-hydro-3-hyroxymethyl-2-methyl-1,4-oxathiine (left) compared to the annotated pathways [25] from the literature (right). This compound is a worst-case example to visualize all the things that can go wrong. All fragments are annotated with the correct molecular formula. Dashed edges in the tree are losses from the annotated pathways. Black edges in the fragmentation tree agree with the annotated pathways. Grey dashed edges are additional pathways that could not be computed since the tree property would have been violated. The blue fragment was actually cleaved in *reverse order* from the molecular ion. The green fragments were inserted *to deep*, and the orange fragment was inserted *to high* in the fragmentation tree. The red fragment was inserted into a completely different pathway. Note that mass errors of more than 10 ppm occur as we added the simulated mass error on the mass of the fragment formula (without considering ionization).

Individual edges from the fragmentation tree were compared to those in the annotated pathways, and matching losses were assigned as *correct*. In some cases, consecutive edges of the fragmentation tree can be combined to give the molecular formula of a single fragmentation step in the annotated fragmentation pathways (see Figure 1(a)). In some other cases two consecutive losses in the fragmentation tree are described in reverse order in the annotated fragmentation pathways (see Figure 1(c)). We evaluate those fragments that were inserted *to deep* or in *reverse order* in the fragmentation trees as *correct*, since without a given structural formula and solely from the EI fragmentation data, the correct case cannot be distinguished from our method's suggestion. If the fragmentation step in the resulting fragmentation tree is explained by several consecutive steps in the annotated pathway, the fragment was inserted *to high* (see Figure 1(b)). If the fragment was inserted into a completely different pathway the edge is assigned as *wrong*.

For 5,6-hydro-3-hyroxymethyl-2-methyl-1,4-oxathiine [25], we now describe in more detail how we evaluate the edges of the fragmentation tree (see Figure 2). We choose this compound as worst-case example to visualize all the things that can go wrong. The loss of ethene from the molecular ion (146-118) followed by a loss of C_2H_3O (118-75) as well as a loss of $C_2H_4O_2$ (118-58) are annotated as *correct*, as they can be found in the annotated pathways. The water loss from the molecular ion (146-128) is also annotated in the literature. In the fragmentation

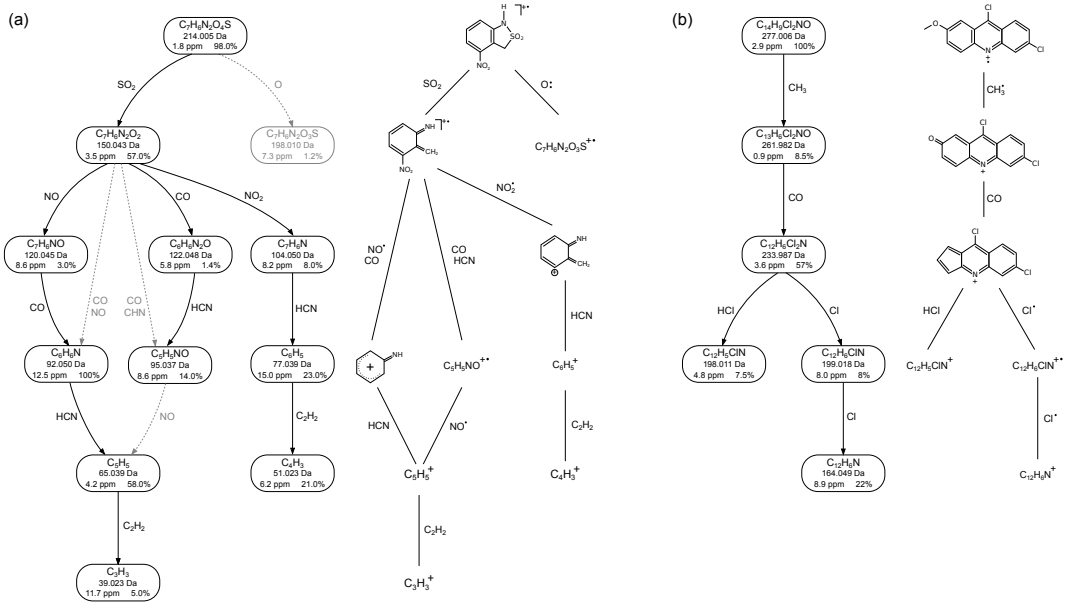


Figure 3 Fragmentation trees compared to annotated pathways from the literature. (a) Fragmentation tree (left) and annotated pathway (right) of a 2,1-benzisothiazoline 2,2-dioxide nitro derivative (compound 6 from [4]). The grey fragment is not explained in the fragmentation tree as it has very low intensity and results from a rather uncommon loss (see Table 1). Dashed edges in the tree are additional losses from the annotated pathways that cannot be explained by our method since the tree property would be violated. In the literature the edge (150-92) combines two fragmentation steps (150-120-92), since the 120 Da peak is very small. In truth, it is very likely, that this fragmentation always proceeds in two steps, but that the lifetime of the intermediate ions is too short [4]. The same applies to edge (150-95) combining the two fragmentation steps (150-122-95). (b) The fragmentation tree (left) and the annotated pathway (right) of 6,9-dichloro-2-methoxyacridine [1] match completely. Note that mass errors of more than 10 ppm occur as we added the simulated mass error on the mass of the fragment formula (without considering ionization).

tree ethene gets lost first and water afterwards (146-118-100), while in the annotated pathway these losses are cleaved in *reverse order*. Edges between nodes 118-75-43 can be combined to the expected loss of C_2H_3OS so the loss of sulfur is considered as *correct*. Pulling up the edges between nodes 146-118-87 results in a total loss of C_3H_7O, so the CH_3O loss was inserted *to deep* and is considered as *correct* by pull-up. Cleaving fragment 45 directly from 118 is considered as *to high*. Fragment 72 was cleaved by loosing ethene from fragment 100 in the annotated pathway. Therefore the methyl loss (87-72) in the fragmentation tree is annotated as *wrong*.

We use similar reasoning processes to evaluate all hypothetical fragmentation trees (see Figure 3 for two examples and Table 3(b) for an overview). For 5 of the 277 correct peak explanations, the fragmentation process leading to this fragment is not given in the literature. From the remaining 272 losses in our data set, 214 losses (78.7 %) are assigned as *correct*. From these, 31 fragments (11.4 %) are inserted *to deep* and 8 fragments (2.9 %) are actually cleaved in *reverse order*. Further, we find that 19 fragments (7.0 %) are inserted *to high* and 39 edges (14.3 %) are annotated as *wrong*. We stress that, unlike for the annotation of the fragmentation processes in the literature, our method has no information about the molecular structure of the compounds.

4.3 Parallelograms

We evaluate the frequency of *parallelograms* in the annotated pathways from the literature (see Table 3(c)). As mentioned above, these are constellations where it cannot be decided solely from the data, whether a fragment results from cleaving loss l_1 first and l_2 afterwards or the other way round (see Figure 1), since both intermediate fragment ions are present in the spectrum. In total, we find 99 parallelograms in all but three compounds. 29 of these are *closed*, that is both fragmentation ways are annotated. This is possible since pathways from the literature not necessarily have to be trees. In contrast, our method has to choose one of these fragmentation pathways. For the remaining 70 parallelograms, either the one or the other way is annotated. From these 70 parallelograms, our method selects the other (possibly wrong) pathway in only 8 (11.4 %) cases.

5 Conclusion

We show that hypothetical fragmentation trees agree in their general information very well with annotated EI fragmentation patterns. We stress that for the computation of the trees no information about the molecular structure of the compounds is used. It is important to note that fragmentation trees are not a tool to reflect the specific mechanisms of EI fragmentation. We find that often the combination or inversion of edges results in pathways that correspond to the true fragmentation. This is not a major set-back since the relevant fragmentation can be constructed based on the trees. Fragmentation trees are a basis for the further interpretation of EI mass spectra. The information obtained, such as fragment formulas, can be used within other methods, for example to simplify in silico fragmentation and presumably improve its results.

Many available methods for analyzing fragmentation spectra of metabolites are rule-based. Mass spectral features are used for classifying compounds [35], Scott [32] uses rules for different classes to estimate the molecular mass of the compound, and rules are used to predict the fragmentation pattern of compounds not included in spectral libraries [19]. Completely unknown compounds may not necessarily follow these known rules for classification or fragmentation. In contrast, the computation and alignment of fragmentation trees is a fully automated and "rule-free" approach that is not limited to known compound classes [28]. It allows to find similar, not necessarily identical, compounds in a library search and unlike other methods it can report the significance of these hits using a decoy database. Consensus substructures of these hits may be key structural elements of the unknown compound and can be used within molecular isomer generators to enumerate all structural isomers containing these substructures [6]. This pipeline will suggest only a few molecular structures and thus can greatly reduce manual analysis time.

Fragmentation tree alignment already accounts for the combination of two consecutive edges [15]. In addition, we find that for some consecutive fragmentation steps the respective ions do not allow to determine the correct fragmentation order solely from the EI data. These constellations occur in 86 % of the compounds. Our method cannot discern the correct fragmentation order solely from the data and will select, based on the scoring properties, the more common and smaller loss twice. In the future, both fragmentation ways should be considered in the fragmentation tree alignment.

Acknowledgments F. Hufsky was supported by the International Max Planck Research School Jena. We thank Martin Rempt for pointing out EI fragmentation pathways annotated in the literature.

References

1 R. M. Acheson, R. T. Aplin, and R. G. Bolton. Electron impact induced alkyl-group fragmentation on the acridine nucleus. *Organic Mass Spectrometry*, 12:518–530, 1977.

2 R. J. Bino, R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. J. Nikolau, P. Mendes, U. Roessner-Tunali, M. H. Beale, R. N. Trethewey, B. M. Lange, E. S. Wurtele, and L. W. Sumner. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci*, 9(9):418–425, 2004.

3 S. Böcker and F. Rasche. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24:I49–I55, 2008. Proc. of *European Conference on Computational Biology* (ECCB 2008).

4 W. Danikiewicz, K. Wojciechowski, R. H. Fokkens, and N. M. M. Nibbering. Electron impact-induced fragmentation of 2,1-benzisothiazoline 2,2-dioxide. *Org Mass Spectrom*, 28:853–859, 1993.

5 K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrom Rev*, 26(1):51–78, 2007.

6 J.-L. Faulon, D. P. Visco, and D. Roe. Enumerating molecules. In K. B. Lipkowitz, R. Larter, and T. R. Cundari, editors, *Reviews in Computational Chemistry*, volume 21, chapter 3, pages 209–286. John Wiley & Sons, Inc., 2005.

7 A. R. Fernie, R. N. Trethewey, A. J. Krotzky, and L. Willmitzer. Metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol Cell Biol*, 5(9):763–769, 2004.

8 O. Fiehn. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trends Analyt Chem*, 27(3):261–269, 2008.

9 O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nat Biotechnol*, 18(11):1157–1161, 2000.

10 W. Heerma and J. J. D. Ridder. The electron-impact-induced fragmentation of some alkyl isocyanides and α-branched alkyl cyanides. *Org Mass Spectrom*, 3:1439–1456, 1970.

11 M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola, and J. Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom*, 22(19):3043–3052, 2008.

12 M. Heinonen, A. Rantanen, T. Mielikäinen, E. Pitkänen, J. Kokkonen, and J. Rousu. Ab initio prediction of molecular fragments from tandem mass spectrometry data. In *Proc. of German Conference on Bioinformatics (GCB 2006)*, volume P-83 of *Lecture Notes in Informatics*, pages 40–53, 2006.

13 M. Hesse, B. Zeeh, and H. Meier. *Spectroscopic Methods in Organic Chemistry*. Thieme Medical Pub, 1997.

14 E. C. Horning and M. G. Horning. Metabolic profiles: gas-phase methods for analysis of metabolites. *Clin Chem*, 17(8):802–809, 1971.

15 F. Hufsky, K. Dührkop, F. Rasche, M. Chimani, and S. Böcker. Fast alignment of fragmentation trees. *Bioinformatics*, 28:i265–i273, 2012. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2012).

16 F. Hufsky, M. Rempt, F. Rasche, G. Pohnert, and S. Böcker. De novo analysis of electron impact mass spectra using fragmentation trees. *Anal Chim Acta*, 739:67–76, 2012.

17 J. Hummel, N. Strehmel, J. Selbig, D. Walther, and J. Kopka. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*, 6(2):322–333, 2010.

18 H. J. Issaq, Q. N. Van, T. J. Waybright, G. M. Muschik, and T. D. Veenstra. Analytical and statistical approaches to metabolomics research. *J Sep Sci*, 32(13):2183–2199, 2009.

19 A. Kerber, R. Laue, M. Meringer, and K. Varmuza. MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation. *Adv Mass Spectrom*, 15:939–940, 2001.

20 T. Kind and O. Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev*, 2(1-4):23–60, 2010.

21 T. Kind, G. Wohlgemuth, D. Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, and O. Fiehn. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal Chem*, 81(24):10038–10048, 2009.

22 S. J. Lehotay, K. Mastovska, A. Amirav, A. B. Fialkov, T. Alon, P. A. Martos, A. de Kok, and A. R. Fernández-Alba. Identification and confirmation of chemical residues in food by chromatography-mass spectrometry and other techniques. *Trends Analyt Chem*, 27(11):10170–1090, 2008.

23 F. W. McLafferty and F. Tureček. *Interpretation of Mass Spectra*. University Science Books, Mill valley, California, fourth edition, 1993.

24 S. Neumann and S. Böcker. Computational mass spectrometry for metabolomics – a review. *Anal Bioanal Chem*, 398(7):2779–2788, 2010.

25 V. Nevalainen and P. Vainiotalo. Electron impact induced fragmentation of dihydro-1,4-oxathiines. 1. 2,3-substituted 5,6-dihydro-1,4-oxathiines. *Org Mass Spectrom*, 21(9):543–548, 1986.

26 G. J. Patti, O. Yanes, and G. Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, 2012.

27 P. Przybylski, T. Pospieszny, A. Huczyński, and B. Brzezinski. EI MS and ESI MS studies of the bisesquiterpene from cotton seeds: Gossypol and its Aza-derivatives. *J Mass Spectrom*, 43(5):680–686, 2008.

28 F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, and S. Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal Chem*, 84(7):3417–3426, 2012.

29 F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher, and S. Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*, 83:1243–1251, 2011.

30 I. Rauf, F. Rasche, F. Nicolas, and S. Böcker. Finding maximum colorful subtrees in practice. In *Proc. of Research in Computational Molecular Biology (RECOMB 2012)*, volume 7262 of *Lect Notes Comput Sci*, pages 213–223. Springer, Berlin, 2012.

31 K. Scheubert, F. Hufsky, F. Rasche, and S. Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. *J Comput Biol*, 18(11):1383–1397, 2011.

32 D. R. Scott. Rapid and accurate method for estimating molecular weights of organic compounds from low resolution mass spectra. *Chemometr Intell Lab*, 16(3):193–202, 1992.

33 M. Thevis and W. Schänzer. Mass spectrometry in sports drug testing: Structure characterization and analytical assays. *Mass Spectrom Rev*, 26(1):79–107, 2007.

34 H. Tsugawa, Y. Tsujimoto, M. Arita, T. Bamba, and E. Fukusaki. GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC Bioinformatics*, 12:131, 2011.

35 K. Varmuza and W. Werther. Mass spectral classifiers for supporting systematic structure elucidation. *J Chem Inf Comp Sci*, 36(2):323–333, 1996.

36 D. S. Wishart. Current progress in computational metabolomics. *Brief Bioinform*, 8(5):279–293, 2007.

37 S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.