Computing Fragmentation Trees from Tandem Mass Spectrometry Data

Florian Rasche,† Aleš Svatoš,‡ Ravi Kumar Maddula,‡ Christoph Böttcher,§ and Sebastian Böcker*†

†Chair for Bioinformatics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
‡Research Group Mass Spectrometry, Max Planck Institute for Chemical Ecology, Hans-Knöll-Straße 8, D-07745 Jena, Germany
§Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, D-06120 Halle, Germany

Supporting Information

ABSTRACT: The structural elucidation of organic compounds in complex biofluids and tissues remains a significant analytical challenge. For mass spectrometry, the manual interpretation of collision-induced dissociation (CID) mass spectra is cumbersome and requires expert knowledge, as the fragmentation mechanisms of ions formed from small molecules are not completely understood. The automated identification of compounds is generally limited to searching in spectral libraries. Here, we present a method for interpreting the CID spectra of the organic compound’s protonated ions by computing fragmentation trees that establish not only the molecular formula of the compound and all fragment ions but also the dependencies between fragment ions. This is an important step toward the automated identification of unknowns from the CID spectra of compounds that are not in any database.

The rapid identification of small compounds from limited amounts of substance is of interest in many areas of biology and medicine such as metabolomics, bioprospecting, biomarker discovery, diagnostics, pharmaceutical chemistry, or the discovery of signaling molecules. In the field of metabolomics, where the goal is to comprehensively detect and quantify the low-molecular-weight metabolites in any given higher eukaryote are largely unknown: The automated identification of compounds is generally limited to searching in spectral libraries. Here, we present a method for interpreting the CID spectra of the organic compound’s protonated ions by computing fragmentation trees that establish not only the molecular formula of the compound and all fragment ions but also the dependencies between fragment ions. This is an important step toward the automated identification of unknowns from the CID spectra of compounds that are not in any database.

Today, mass spectrometry (MS) is a key technology for the identification of small molecules.1,2 Compared to nuclear magnetic resonance (NMR), MS offers sensitivity that is orders of magnitude higher. Various analytical setups have been developed, most notably gas chromatography MS (GC/MS) and liquid chromatography MS (LC-MS). GC/MS spectra can be interpreted automatically via database search, since reference spectra were collected over many years and the fragmentation mechanisms are reproducible on different instruments.3 However, GC/MS requires the metabolite to be thermally stable. This is not the case for several biologically important compound classes, like polar lipids, sugar-containing metabolites, phenolics, peptides, or nucleotides. These compounds are analyzed by LC-MS. Here, the adduct ion of the studied compound (e.g., [M + H]⁺) is mass selected and fragmented (collision-induced dissociation, CID) in an inert gas-filled collision cell, and mass-to-charge ratios (m/z) of the resulting fragments are recorded as tandem mass spectra. The computational analysis of such data is in its infancy, and this is presumed to be one of the major technological hurdles in metabolomics today.4–6 Even the time-consuming manual analysis of these data is nontrivial, as the fragmentation of small molecules under varying fragmentation energies is not completely understood.7 Due to the limited reproducibility of CID mass spectra on different instruments, even searching in spectral libraries is a serious problem.8–10 Apart from a few pioneering studies,11–14 there has been little progress toward an automated analysis of CID data.

In this study, we present a method for the automated and swift analysis of tandem MS data from small molecules (Figure 1). The method does not require a compound structure or mass spectral database for its analysis. The result is a hypothetical fragmentation tree in which nodes are annotated with molecular formulas of the fragments and arcs represent fragmentation events. These trees can, for example, be compared with each other to identify compound classes of unknowns. We evaluate our results on three different levels: First, we will show that hypothetical fragmentation trees agree well with expert annotations. Second, we will compare our tree topologies to multiple-stage MS data. Third, we will use commercial software (Mass Frontier) to simulate fragmentation patterns and compare these to our results. Both experts and the commercial software use the molecular structure of the sample compound in order to annotate the tandem MS data.

Received: July 19, 2010
Accepted: November 30, 2010
data. In contrast, our method operates without this knowledge. We find excellent agreement between our results and the annotations by the experts and the commercial software. We argue that this is an important step toward “identifying the unknowns” using tandem MS: Our method allows for the automated and high-throughput analysis of small-compound MS data beyond elemental composition, without relying on databases of reference spectra or compounds.

We have recently developed a method for the automated comparison of fragmentation trees. In conjunction, these two methods will enable the automated analysis of large MS data sets for identifying unknown compounds: For example, we can analyze drug degradation products from pharmacokinetic and drug-stability studies by LC-MS analysis and subsequently perform automated computation as well as compare fragmentation trees. In so doing, we may identify pathways of drug degradation, on the most likely fragmentation tree is calculated, containing the fragmentation reactions that presumably took place. For validation, these trees were checked by mass spectrometry experts, compared to fragmentation patterns derived from analysis of MS spectra, and evaluated against Mass Frontier calculations. (6) We exemplarily show that comparing the computed trees can reveal information about unknown compounds.

### EXPERIMENTAL SECTION

To show the applicability of our method to diverse types of MS data, we use three data sets in this study (see Table 1). The first data set consists of 37 compounds, mostly representing plant secondary metabolites, measured on an Orbitrap mass spectrometer. The second contained 42 compounds measured on an API QSTAR. The third data set with 102 compounds was measured on a Micromass QTOF instrument.

**Chemicals.** Supplementary Tables S1, S2, and S3 (Supporting Information) list the chemicals of the data sets. Substances were from our laboratory stocks and were either previously purchased or isolated from natural sources. The samples were dissolved in methanol (ca. 1 mg/1 mL) and individually measured by infusion into electrospray sources or mixed and LC/MS or LC/MS/MS experiments were performed.

**Table 1. Data Sets Used in This Study**

<table>
<thead>
<tr>
<th>instrument</th>
<th>ppm</th>
<th>CID (eV)</th>
<th>IP</th>
<th>compounds</th>
<th>used</th>
<th>mass range</th>
<th>median</th>
<th>average</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orbitrap</td>
<td>5</td>
<td>35,45,55,70</td>
<td>yes</td>
<td>37</td>
<td>37</td>
<td>152.0–822.4</td>
<td>298.1</td>
<td>345.2</td>
<td>Table S-1, Supporting Information</td>
</tr>
<tr>
<td>API QSTAR</td>
<td>20</td>
<td>15,25,45,55,90</td>
<td>yes</td>
<td>42</td>
<td>42</td>
<td>89.0–441.2</td>
<td>174.6</td>
<td>207.5</td>
<td>Table S-2, Supporting Information</td>
</tr>
<tr>
<td>Micromass QTOF</td>
<td>20</td>
<td>10,20,30,40,50</td>
<td>no</td>
<td>102</td>
<td>100</td>
<td>137.1–609.3</td>
<td>357.7</td>
<td>372.5</td>
<td>Table S-3, Supporting Information</td>
</tr>
</tbody>
</table>

a Mass accuracy of the measurement. b Isotope patterns available. c Number of compounds. d Number of compounds after exclusion. e Three to five distinct collision energies were used for each compound. See text for details.

**Data Collection.** The first data set was measured on an Orbitrap XL instrument (Thermo Fisher Scientific, Bremen, Germany) and a Hewlett-Packard HP-1100 HPLC (Agilent, Waldbronn, Germany) system consisting of a binary pump, an autosample, and a column heater. HPLC separations were performed using a Symmetry C18 Column, (100 × 2.1 mm, 3.0 μm, Waters, Milford, MA, USA) at a constant flow rate of 300 μL/min using a binary solvent system: solvent A, water with 0.1% formic acid, and solvent B, acetonitrile with 0.1% formic acid. The HPLC gradient system started with 5% B, linearly increased to 95% B in 24 min, was held for 4 min, and then brought back to the 5% B initial condition within 0.1 min and re-equilibration 4 min prior to the next injection. Full-scan mass spectra were generated using 7500 resolution power. This low resolution enabled us to measure CID spectra of eluting peaks at several collision energies in the LC/MS/MS workflow. Source parameters were S0, 10, and 5 arb units for the sheath, auxiliary, and sweep gas (nitrogen) flow rates, respectively. The spray voltage was set at 4500 V with a tube lens voltage at 140 V. The automatic gain control (AGC) was set at $5.0 \times 10^5$ for FTMS full scan and $3.0 \times 10^3$ for FTMS MS$^3$ scans. The capillary temperature and voltage were set at 275 °C and 39 V, respectively. For MS$^2$ experiments, the total run time was divided into seven or eight segments; one m/z is selected in each segment, and six different runs were performed to complete the MS$^2$ experiments of all compounds. For individual compounds, the acquisition window comprised individual CID experiments at 35%, 45%, 55%, and 70% normalized collision energies. MS$^3$ spectra were acquired under operator control with the 7500 resolution settings, and the samples were introduced using a built-in infusion pump with a flow rate of 5 μL/min. The activation time was set at 30 ms with the activation parameter q = 0.25. For MS$^2$ and MS$^3$ analysis, an isolation window of 1.5 mass units was used to isolate the precursor ions. The mass accuracy was better than 5 ppm for MS and MS$^3$ experiments.
Our second data set, from an API QSTAR QTOF instrument with mass accuracy 20 ppm, contains 42 compounds, including biogenic amines, amino acids, and phenolic choline esters. In this case, the LC-MS system consisted of a capillary LC system (Ultimate from Dionex, Idstein, Germany) and an API QSTAR Pulsar Hybrid QTOF-MS (Applied Biosystems/MDS Sciex, Mississauga, Ontario, Canada) equipped with an IONSPRAY electrospray ion source. Capillary LC was performed using a modified C18 column (GROMSIL ODS 4 HE, 150 × 0.3 mm, particle size of 3 μm, pore size of 120 Å, guard column of 10 × 0.3 mm (Alltech Grom)) at a flow rate of 6 μL/min, applying a similar binary solvent system as described for Orbitrap analysis. The following gradient program was used 0–5 min, isocratic 5% B; 5–45 min, linear from 5–95% B; 45–52 min, isocratic 95% B; 52–60 min, isocratic 5% B. Eluted compounds were detected in positive ion mode using the following instrument settings: ion spray voltage, 5.5 kV; DP1, 50 V; DP2, 15 V; FP, 220 V; nebulizer gas (nitrogen), 25 arb units; curtain gas (nitrogen), 20 arb units; collision gas (nitrogen), 4 arb units; pulser frequency, 9.986 kHz; accumulation time, 2 s. Ions were recorded in “enhance all” mode from m/z 75 to 1000 within four Q1 transmission windows: m/z 55–113, 10% scan time; m/z 113–225, 20% scan time; m/z 225–450, 35% scan time; m/z 450–1000, 35% scan time. The mass spectrometer was operated under Analyst QS 1.0. Mass calibration was performed using protonated ALILTILVS (m/z 829.5393) and a fragment ion (m/z 149.0233) originating from phthalate-type plasticizers. Mass resolution for [M + H]+ of the calibration peptide was R_{cal} = 8500. Product ion spectra were acquired with Q3 operating at unit resolution and applying collision energies in the range of 15–90 eV. Nitrogen was used as a collision gas at a pressure of 4 arb units. Product ions were recorded in “enhance all” mode using q3 transmission windows and pulser frequencies as suggested by Analyst QS 1.0. To improve the mass accuracy of product ion, spectra internal calibration was applied using precursor and fragment ions with unequivocally identified elemental compositions. Acquired spectra were centroided and exported as peak lists. Each data set was measured at three to five distinct collision energies using CID fragmentation. Peak picking and calibration was done using vendor software. For these two data sets, we also recorded isotope patterns of precursor ions in profile mode. The third data set, consisting of 102 compounds, was measured on a Micromass Q-TOF II instrument (Beverly, MA) with mass accuracy 20 ppm. We excluded two compounds from this data set, since precursor peaks had mass accuracy worse than 50 ppm. Experimental details for the Micromass Q-TOF data set can be found in ref 11. In each data set, MS spectra were measured at up to five distinct collision energies.

Peak lists related to product ion spectra acquired from the same precursor at different collision energies were merged into a single peak list. For that, peaks with less than 50 mDa distance were considered to represent the same fragment ion and combined to a single one by calculating their signal intensity weighted mean. This relatively large mass window was found to improve the mass accuracy of the data. Intensities were not scaled, since this would compromise comparisons between the peaks between spectra taken at different collision energies.

### COMPUTATION METHODS AND PROGRAMS

#### Molecular Formula Identification

Numerous methods to determine the molecular formula of small molecules without any user interaction have been published recently.\textsuperscript{11,16–19} We use the following approach to identify molecular formulas: First, we analyze isotope patterns using the SIRIUS software.\textsuperscript{18} This method compares the isotope pattern of the unfragmented molecule compared to theoretical isotope patterns of all molecules with monoisotopic mass sufficiently close to the measured monoisotopic mass. Candidates are evaluated using Bayesian statistics,\textsuperscript{20} and both intensities and masses of isotope peaks are taken into consideration:

\[
P(M|\mathcal{D}, \mathcal{R}) = \frac{P(M|\mathcal{D} \cap \mathcal{R}) \cdot P(\mathcal{D} | \mathcal{R})}{\sum_i P(M_i | \mathcal{D}) \cdot P(\mathcal{D} | \mathcal{R})}
\]

where \(\mathcal{D}\) is the data (the measured spectrum), \(M\) are the models (the candidate molecules), and \(\mathcal{R}\) stands for any prior background information. As background information, we set the prior to zero for all molecules but the decompositions of the monoisotopic mass and for molecular formulas that cannot correspond to a molecule because of chemical considerations.\textsuperscript{21} Finally, we set

\[
P(\mathcal{D} | M, \mathcal{R}) = \prod_f P(M | m_f) \prod_j P(f | p_{j})
\]

where \(P(M | m_f)\) is the probability to observe peak \(j\) at mass \(M\) when its true mass is \(m_f\) and \(P(f | p_{j})\) is the probability to observe peak \(j\) with intensity \(f\) when its true intensity is \(p_j\). The method takes into account that mass accuracy and intensity accuracy worsen with decreasing intensity. Thus, there are parameters for the accuracies at maximum intensity and at zero intensity. The actual expected accuracies of a peak are determined by linear interpolation using the peaks intensity. To correct errors originating from baseline subtraction, a correction offset is added to the intensities. The mass accuracy at full intensity \(\alpha_f\) is set to the instruments mass accuracy (see Table 1), all other parameters are kept in default setting: \(\alpha_0 = 1.5 \alpha_f\) (mass accuracy at zero intensity), \(\beta_1 = 10\) (intensity accuracy at full intensity), \(\beta_0 = 90\) (intensity accuracy at zero intensity), and \(\delta f = +0.02\) (intensity correction offset). The alphabet of elements is CHNOPS. See ref 18 for more details. We then analyze the fragmentation patterns of the compounds, as described in ref 16. This method works similarly to the fragmentation tree calculation described below but calculates a fragmentation tree for every molecular formula within the mass accuracy of the precursor peak and ranks the formulas according to the overall score of their fragmentation tree. The scoring was the same as for the fragmentation tree calculation, which is described in detail in the Supporting Information. Finally, we combine results of the two identification methods: This combined score is \(S \log P_{iso} + s_{frag}\) where \(P_{iso}\) is the likelihood from the Bayesian analysis of isotope patterns and \(s_{frag}\) is the score of the fragmentation pattern analysis. The constant is chosen to make the scores comparable.

#### Computing Fragmentation Trees

Using tandem MS, fragments often result from a subsequent series of fragmentation events. We model these fragmentation cascades using fragmentation trees (Figure 2): The nodes of this tree are labeled with the molecular formulas of the molecule and its fragments, whereas the arcs (directed edges) correspond to neutral losses (NLs). The root of the fragmentation tree is labeled with the unfragmented ion. Apart from being easily representable in a computer, the fragmentation tree concept helps us to visualize the sometimes complicated fragmentation process; see Figure 3. Our task is to compute such a hypothetical fragmentation tree solely from the tandem MS data.
For every peak of the fragmentation spectrum, we compute all molecular formulas that are within the mass accuracy of the instrument and that are subformulas of the compound molecular formula. We use these molecular formulas as the vertices of a fragmentation graph; see Figure 2 for an example. Vertices are colored so that two molecular formulas corresponding to the same peak also receive the same color. Two vertices are connected by a directed edge if the second molecular formula is a subformula of the first. We weight the fragmentation graph using log likelihoods or log odds. This enables a statistical interpretation of the outcome (i.e., maximum likelihood). For vertices, we use log odds to differentiate between the model (the peak is truly a fragment with the proposed molecular formula) and the background (the peak is noise). Under the model, we use the mass difference between the measured peak and the molecular formula and assume mass differences to be Normal distributed.22,23 For the background model, we use the peak intensity for this purpose and assume that noise peak intensities are exponentially distributed; for example, see Figure 4 in ref 24. Finally, we incorporate prior probabilities adding a constant to each vertex score. For weighting edges, we consider common neutral losses, unlikely neutral losses containing only one atom type, the mass of the loss, collision energies, and the ratio between carbon and hetero atoms.

Different fragmentation pathways may lead to fragments with identical molecular formula or even identical structure. This implies that the highest scoring fragmentation process will contain all fragmentation reactions, but we are interested in the major fragmentation events that mainly occurred. Hence, we slightly oversimplify the problem: We demand that each fragment in the fragmentation spectrum is generated by a single fragmentation pathway. With a scoring that considers criteria for “mainly occurring fragmentations”, our optimization algorithm will then choose the mainly occurring pathway. This simplification means that we are searching for a fragmentation tree inside the fragmentation graph. There is one exception to the above reasoning: Assume that some fragment $f_3$ is cleaved from fragment $f_2$ and that $f_2$ is in turn cleaved from $f_1$. Solely from the tandem MS data and without additional structural information, we cannot rule out that $f_3$ is directly cleaved from $f_1$. However, this information is implicitly encoded in a fragmentation tree: the fragmentation may occur from the fragment’s direct parent in the tree or from any of its parents. This was recognized often during the expert evaluation process and MSn fragmentation experiments, and we refer to it as “pull-up”, since we “pulled-up” fragment $f_3$ and attach it to some node closer to the root.

Also, several fragments may result in a single peak in the fragmentation spectrum, but we argue that this is an extremely rare event in practice. To this end, we demand that our fragmentation tree is colorful: Each vertex color and, hence, each peak in the fragmentation spectrum is scored at most once. Now, we search for a colorful tree inside the fragmentation graph that has a maximal sum of edge weights. Several heuristics have been proposed for this,16 but experts evaluated the fragmentation trees computed by heuristics as inaccurate. We use dynamic programming over the vertices to find the maximum
colorful subtree in the graph. We encode the colors used so far as part of the dynamic programming matrix. Optimal solutions can be computed by combining optimal solutions of subproblems. Memory usage increases exponentially with the number of peaks: maximum memory usage for 10 peaks was 200 MByte and for 15 peaks was 900 MByte, and for the 86 compounds with 20 or more peaks, memory usage was projected to be around 23 GByte. Therefore, we only use the 10 most intensive peaks to calculate an exact tree to save memory and heuristically attach the remaining peaks to this skeleton. Using the 15 most intensive peaks for exact calculation did not change the trees, so using only the 10 most intense peaks appears to be completely sufficient in application. Algorithmic and scoring details can be found in the Supporting Information.

RESULTS

Molecular Formula Identification. We analyzed the Orbitrap data set using both isotope pattern analysis and fragmentation pattern analysis. Results for the isotope patterns are given in Table S1 (Supporting Information): For 30 out of 37 compounds, the correct molecular formula was the TOP 1 hit. In all 37 cases, the correct interpretation was found within the three most significant hits (TOP 3) of the output. To improve identifications, we add fragmentation pattern information: This method is similar in spirit to that described in the Experimental Section for the computation of fragmentation trees. Here, 19 out of the 38 compounds are identified at the top position, and in all but six cases, the correct molecular formula was found in the TOP 5 of the output. Finally, we combined this somewhat orthogonal information, as described in the Experimental Section: We correctly identify all compounds as the TOP 1 hit of our output; see again Table S1 (Supporting Information). We stress that no biochemical background information, except the elements CHNOPS and Table 2 for common neutral losses, are used, and no user interaction is required. All parameters were kept at default settings. When adding chlorine and bromine to the alphabet, we correctly identify the 34 compounds with mass below 500 Da. For all compounds, the correct molecular formula was in the TOP 3.

We also determined molecular formulas for the QSTAR data set using CHNOPS, Cl, and Br as elements. For 34 out of 42 compounds, the correct molecular formula was ranked at the top position of the isotope pattern analysis (Table S2, Supporting Information). For 37 compounds, the correct molecular formula was at the top position for the fragmentation pattern analysis. Combining these results, the correct molecular formula was identified at the top position for all 42 compounds.

This validates that we may consider the molecular formula of the unknown compound to be known during fragmentation tree calculation.

Computing Fragmentation Trees. For each compound in all three data sets, we use the above method to compute a hypothetical fragmentation tree; see Figures 2 and 3 for examples and the Supporting Information for a complete list. Our algorithm processes each compound in 1.5 s on average and 14.4 s in the worst case, measured on a laptop computer with a dual-core Intel Core duo processor T2400 at 1.83 GHz and 1 GB RAM, running a 32-bit Linux system, and a Java virtual machine version 1.6.0. Clearly, analysis time depends both on the size of the compound and on the number of peaks in the spectrum, but even for large compounds with mass above 600 Da and 15 or more

---

**Figure 3.** Hypothetical fragmentation tree of chelidonine (C$_{20}$H$_{19}$NO$_5$) computed by our method using Orbitrap data. Nodes (blue) correspond to peaks in the tandem mass spectra and their annotated molecular formula; arcs (red) correspond to hypothetical neutral losses.
spectral MS peaks, the average running time was only 9.2 s. To find the correct molecular formula using the tandem MS data (see above), it is necessary to compute fragmentation trees for all molecular formulas within the mass accuracy of the parent peak. Doing so requires 1.8 s per compound on average on a single processor.

Evaluation against Expert Knowledge and Multiple MS. The fragmentation trees of all 79 compounds in the Orbitrap and API QSTAR data set were manually evaluated by mass spectrometry experts with broad experience in the structural elucidation of natural products. Fragmentation trees were compared with expected fragmentation patterns that the experts deduced from the provided chemical structures and merged CID spectra. Fragmentations were rationalized by a sequence of charge-assisted, charge-remote fragmentation, elimination, or retro-cyclo-reversion reactions and related skeletal rearrangement events.3,26–28 First, the theoretical fragmentation pathway was formulated on the basis of the fragmentation rules for protonated even-numbered electron ions. Individual arcs in the pathway were compared to those in the fragmentation tree, and matching losses were assigned as “correct”. Numerous agreements between NLs listed in Table 2 and manually assigned fragmentation steps were found. In some cases, several consecutive arcs of the fragmentation tree can be combined to give the molecular formula of a NL in the experts’ fragmentation pathways. We evaluate those summed “pull-up” arcs as “correct”, since without a given structural formula and sole MS² data, the “correct” case cannot be distinguished from our method’s suggestion. Some NLs assigned by the automated method cannot be ruled out, but experts were unable to rationalize them in a fragmentation pathway; these NLs are annotated as “unsure”. Arcs which give molecular fragments with questionable stability under experimental conditions and those that cannot be explained via a “pull-up” were assigned as “wrong”. Whenever possible, the literature was used to support the assignment; however, not all references provided useful data, due to the paucity of well-evaluated CID fragment spectra of metabolites in the literature.

For protonated (−)-epicatechine, we now describe in more detail how we evaluated the fragmentation tree; see Figure S-1 (Supporting Information): The fragmentation pathway was based on the CID fragmentation of structurally related kaempferol29 as no reliable literature on (−)-epicatechin exists. Obvious water
and C₇H₈O₃, o-chinone neutral losses followed by another water loss (nodes 291, 273, 165, 151) were found in the calculated tree and annotated as “correct” (Table S1, Supporting Information). A loss of CO from node 151 is possible, but the abundant m/z 123.045 is more likely formed by an rDA reaction from protopontine and annotated as "unsure". In all evaluated trees, similar reasoning processes were used to evaluate the hypothetical fragmentation trees. We find that the NLs of O, C, N, \( \cdot \)CH₂, C₆O, C₅O, C₆H₅, C₇H₈, and C₇H₂ were the arcs most frequently annotated as "wrong".

For the Orbitrap data set, 352 of 458 NLs (76.9%) were assigned as "correct", 57 (12.4%) as "unsure", and 49 (10.7%) as "wrong". In cases of methoxylated aromatic compounds, well-pronounced radical losses, namely, \( \cdot \)CH₃, \( \cdot \)HO, and CH₃O, were not presented in the calculated trees of compounds such as berberine or emitine although the corresponding peaks were found in the spectra. It should be understood that this is solely a problem of the objective function used, not of the general approach: We will include radical losses in a subsequent program version. For the QSTAR data set, 286 of 350 NLs (81.7%) were assigned as "correct", 51 (14.5%) as "unsure", and only 13 (3.7%) as "wrong". For 15 fragmentation trees in the Orbitrap data set and 22 trees in the QSTAR data set, all NLs in the tree were annotated as "correct". See Tables S1 and S2 (Supporting Information) for details. In general, the calculated trees are very close to the experts’ assignment, which is remarkable if we consider the comparatively simple optimization objective the automated assignment is based on, compared to years of experience on the human side. We stress that, unlike the experts during evaluation, our method has no information about the molecular structure.

The results of the heuristics, in contrast, were not promising. There are two ways to evaluate the quality of a heuristic solution to an optimization problem: either compare the scores of the heuristic with that of an exact solution or compare the actual solutions computed by heuristic and exact methods. In both respects, quality of the heuristics proposed in ref 16 are unsatisfactory for computing fragmentation trees. The scores calculated by the heuristics ranged between 30% and 100% of the score calculated by the exact algorithm, but for 97 (53.6%) of the 181 compounds, the heuristic scores are less than half the optimal score. Note that this does not necessarily mean that fragmentation trees computed by the heuristic are of low quality.

To this end, we also manually evaluated fragmentation trees calculated by the heuristic. Manual inspection of the trees revealed that the proposed fragmentation events do not agree with those described in the literature. For example, in Figure S-4 (Supporting Information), we have depicted, for compound chinchonine from the Orbitrap data set, both the optimal fragmentation tree calculated by the exact algorithm and the fragmentation tree computed by the greedy heuristic;16 results for other heuristics are similar. Through expert evaluation, we find that the number of NLs annotated as "correct" drops from 19 in the optimal tree to 9 in the tree from the heuristic; "unsure" increases from 1 to 4, and "wrong" increases from 3 to 10. In total, there were 23 NLs in the spectrum. In total, evaluation results for this compound are much worse for the heuristic fragmentation tree than for the optimal fragmentation tree. Similar results were obtained for bicuculline ("correct" 24 to 18, "unsure" 3 to 7, "wrong" 7 to 9, total 34) and chelidonine ("correct" 14 to 9, "unsure" 2 to 3, "wrong" 3 to 7, total 19). This is a clear indication that the heuristic algorithms from ref 16 cannot be used for computing fragmentation trees. Therefore, we excluded the heuristic algorithms from our study.

Expert evaluation of fragmentation trees was refined by MS² and MS³ spectra of ions of high abundance in the CID spectra and of those forming branching nodes of calculated trees.
For (S,R)-noscapine, MS$^3$ of m/z 414→396 and 414→220 transitions were recorded using a linear trap for a precursor ion preparation/selection and an orbitrap analyzer for ion detection. Additionally, MS$^4$ of m/z 414→396→378 was obtained; see Figure S-5 (Supporting Information). Transition 414→396 supported the direct formation of m/z 378 and 365 fragment ions from m/z 396. A strong peak at m/z 381, not included in the calculated tree and corresponding to radical loss (CH$_3$), was noticed. Transition 414→396→378 did not support the direct edge between m/z 378 and fragment ions 248 and 220. Those are likely formed directly from protonated noscapine. The lineage of ions 179 from m/z 220 was confirmed by transition 414→220; however, the most intense is a methyl radical loss providing m/z 205. When comparing the pathway, this data suggests, with the tree calculated from MS$^2$ data, five edges are “correct” and two are pull-ups. No wrong assignment was made; see Figure S-6 (Supporting Information). Altogether, this showed the close similarity of calculated trees and MS$^n$-spectra-derived fragmentation pathways and supports the annotation of “correct”, “unclear”, and “wrong” NLs in experts’ evaluation; see Tables S-1 and S-2 (Supporting Information).

**Evaluation against Mass Frontier.** For the final evaluation of our method, we compare the molecular formulas our method assigns to the peaks, with the predictions of the Mass Frontier software. Here, we use the Micromass QTOF data set, where predictions have previously been carried out in a different experimental context. Given the molecular structure of the compound, Mass Frontier predicts tandem mass spectra, which we match to the observed data. Regarding the accuracy of the method, we annotate more than four times as many peaks as Mass Frontier (70.3% vs 16.8%). Only 19 out of 6508 peaks were annotated by Mass Frontier but not by our software. For the 1072 peaks that both tools annotate, the same molecular formula is assigned in 97.3% of the cases (1043 peaks). This is an excellent agreement, taking into account the completely different paradigms of the two tools: Mass Frontier knows the molecular structure but not the experimental MS data, whereas our tool knows the experimental MS data but not the molecular structure. The probability that such an agreement can happen by chance (significance) is below $10^{-16}. This is the probability that, by uniformly drawing a molecular formula at 50 ppm for each peak, we reach the observed number of 1043 matching molecular formulas or an even higher number.

**Comparing Fragmentation Trees.** Now that we have evaluated the validity of our fragmentation trees (79.0% of the neutral losses were correct): How can we establish the structural similarity of molecules using the trees? Let us assume that, for an unknown compound and a set of reference compounds, we have computed fragmentation trees. By manual inspection, we compare the NLs occurring in both sample tree and reference tree and try to find a subtree that occurs in both trees, comparable to Maximum Agreement Subtrees (MAST) used in phylogenetics. The larger the common subtree, the more similar the fragmentation pattern and the more similar the molecular structures. See Figure 4 for an example, where the similarity between the compounds is hard to identify when interpreting the tandem mass spectra directly: Only by transforming these data into fragmentation trees do similarities become obvious. Using such similarities between the trees, we may also classify unknown compounds: It is well-known such classifications can be achieved using certain characteristic neutral losses, such as H$_2$O (18.011 Da), NH$_3$ (17.027 Da), and CO (27.995 Da) for amino acids, NL of formic acid, water, and carbon monoxide for carboxylic acids, or methanol, carbon monoxide, ethene, and formaldehyde for phenolic compounds to mention some additional examples. However, even when such characteristic NLs are missing, our method allows us to classify a compound based on the overall similarity of the fragmentation pattern. See Figures S-7 and S-8 (Supporting Information) for an example: The fragmentation tree of histidine is missing the characteristic NH$_3$ loss, but its fragmentation tree shows high agreement with that of asparagine and only negligible agreement with that of cafeoyl choline and other cholines.

**DISCUSSION**

Many attempts have been made to automate the interpretation of MS data from small molecules. Here, we have proposed a method for the computational analysis of tandem MS data that goes beyond database searching or assigning molecular formulas. We have introduced fragmentation trees as a method of annotating tandem mass spectra from organic compounds. We have demonstrated that the fragmentation trees computed by our method agree well with expert knowledge, rule-based fragmentation prediction, and experimental data from multiple MS experiments. Our techniques permit a comprehensive analysis of the observed data well beyond what has previously been possible. Our method can be used in a fully automated analysis pipeline without user interaction.

We can, of course, not detect compounds that are too unstable for MS analysis, but many stable compounds, e.g., the stable portion of the secondary metabolites, and their molecular formulas are also not found in any database. Our method treats all possible molecular formulas on a par and requires no database to compute fragmentation trees. We argue that this allows us to overcome limitations in the “known universe of organic chemistry”. When numerous combinations of the different scoring parameters were tested, it would probably be possible to find a parameter combination that improves the results, but when the amount of training data is small compared to the number of parameters/degrees of freedom in the optimization, using such a parameter optimization may lead to too strong adaptation to the available data. This so-called “overfitting” leads to poor generalization results of the method. Thus, we refrained from parameter optimization due to the relatively small number of compounds. The parameters are set to default values or values that are chosen ad hoc. Given larger data sets, parameter optimization will most likely further improve the results of our method.

Fragmentation trees can help MS experts elucidate the structure of unknowns: The fragmentation tree annotates fragment masses with molecular formulas and shows dependencies between the fragments through NLs. It is rather involved to derive such information from the “pure” tandem MS data. Obviously, agreement between fragmentation trees and expert opinion does not mean that our predictions are correct, but our evaluation indicates that automated analysis reaches a quality of prediction that is almost on par with a manual analysis by experts. Thus, fragmentation trees can serve as a valuable tool for compound and structural identification, even though these fragmentation trees are only a possible explanation of the observed data and not “the truth”.

We have demonstrated that, by comparing fragmentation trees, we can derive valuable structural information about an
unknown fragmentation trees. Unfortunately, the manual comparison of fragmentation trees is also laborious and time-consuming. We have recently developed a method to automate this step of the analysis, and the first results are very promising. Our results can be combined with other sources of information about the sample compound, such as characteristic neutral losses, retention time, or the collision energy resulting in a fixed survival yield (CE$_{SO}$). To simplify data collection, our method can also analyze ramp spectra, where collision energy is varied within a single tandem spectrum.

The automated analysis of large-scale compound screens will likely increase our understanding of secondary metabolism and may help us to search for "interesting" organic compounds for botanical therapeutics, biomarker discovery, or the discovery of signaling molecules. Simplification of these tasks has long been sought after.

**ASSOCIATED CONTENT**

Supporting Information. Detailed methods and additional figures as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

Corresponding Author

*E-mail: sebastian.boecker@uni-jena.de.*

Notes

The authors declare that they have competing financial interests.

**ACKNOWLEDGMENT**

The authors would like to thank David Grant and his group, especially Dennis Hill, for providing their data for this study and Emily Wheeler for editorial assistance. The Max Planck Society supported this work financially. The method is implemented in the SIRIUS STARBURST program available online (http://bioinformatic.uni-jena.de/starburst/).

**REFERENCES**

(20) Zhang, W.; Chait, B. T. Anal. Chem. 2000, 72, 2482–2489.

**NOTE ADDED AFTER ASAP PUBLICATION**

This paper was published on the Web on December 23, 2010. Additional information has been added to Table S-3 in the Supporting Information. The corrected version was reposted on January 14, 2011.