BIOINFORMATICS

Vol. 24 ECCB 2008, pages i49–i55
doi:10.1093/bioinformatics/btn270

Towards *de novo* identification of metabolites by analyzing tandem mass spectra

Sebastian Böcker[1,2] and Florian Rasche[1,*]

[1]Chair for Bioinformatics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena and [2]Jena Centre for Bioinformatics, Jena, Germany

ABSTRACT

Motivation: Mass spectrometry is among the most widely used technologies in proteomics and metabolomics. Being a high-throughput method, it produces large amounts of data that necessitates an automated analysis of the spectra. Clearly, database search methods for protein analysis can easily be adopted to analyze metabolite mass spectra. But for metabolites, *de novo* interpretation of spectra is even more important than for protein data, because metabolite spectra databases cover only a small fraction of naturally occurring metabolites: even the model plant *Arabidopsis thaliana* has a large number of enzymes whose substrates and products remain unknown. The field of bio-prospection searches biologically diverse areas for metabolites which might serve as pharmaceuticals. *De novo* identification of metabolite mass spectra requires new concepts and methods since, unlike proteins, metabolites possess a non-linear molecular structure.

Results: In this work, we introduce a method for fully automated *de novo* identification of metabolites from tandem mass spectra. Mass spectrometry data is usually assumed to be insufficient for identification of molecular structures, so we want to estimate the *molecular formula* of the unknown metabolite, a crucial step for its identification. The method first calculates all molecular formulas that explain the parent peak mass. Then, a graph is build where vertices correspond to molecular formulas of all peaks in the fragmentation mass spectra, whereas edges correspond to hypothetical fragmentation steps. Our algorithm afterwards calculates the maximum scoring subtree of this graph: each peak in the spectra must be scored at most once, so the subtree shall contain only one explanation per peak. Unfortunately, finding this subtree is NP-hard. We suggest three exact algorithms (including one fixed-parameter tractable algorithm) as well as two heuristics to solve the problem. Tests on real mass spectra show that the FPT algorithm and the heuristics solve the problem suitably fast and provide excellent results: for all 32 test compounds the correct solution was among the top five suggestions, for 26 compounds the first suggestion of the exact algorithm was correct.

Availability: http://www.bio.inf.uni-jena.de/tandemms

Contact: florian.rasche@minet.uni-jena.de

1 INTRODUCTION

When analyzing the metabolome of an organism, mass spectrometry in combination with liquid or gas chromatography is the most widely used high-throughput technique (von Roepenack-Lahaye *et al.*, 2004). Since the manual interpretation of mass spectra is tedious and time-consuming, methods for an automated analysis are required. For metabolite identification, most established methods rely on a database of reference mass spectra. But *de novo* identification of metabolites is highly sought: today, metabolite databases contain primary metabolites directly relevant for growth, development and reproduction of a cell or an organism. In contrast, most of the metabolites not directly involved in the aforementioned functions remain unknown. These *secondary metabolites* are especially abundant in plant signal transduction: for the model plant *Arabidopsis thaliana* 200 secondary metabolites have been identified (D'Auria and Gershenzon, 2005), but the number of genes coding for enzymes of the secondary metabolism suggests that most metabolites are still unknown (Arabidopsis Genome Initiative, 2000). For proteins, one can build a database containing all hypothetical proteins if the genome of the species under investigation has been previously sequenced. Unfortunately, there is no way to build a non-trivial database of hypothetical metabolites.

In proteomics, bioinformaticians have developed *de novo* interpretation methods capable of interpreting MS data with no need of any database (Chen *et al.*, 2001; Pitzer *et al.*, 2007). Developing such techniques for metabolite spectra is more difficult, because metabolites show more diverse structures: they neither possess a linear structure such as proteins, nor a tree-like structure such as glycans. Recently, there has been some progress on the *de novo* interpretation of metabolite mass spectra: Kind and Fiehn (2006) suggest a pipeline for analyzing high-resolution mass spectra of metabolites, and Böcker *et al.* (2006) provide an automated method for this analysis. Both methods require that masses are measured with excellent accuracy (below 2 ppm), and that the isotope distribution of the metabolite is known. Current techniques that fulfill these requirements, such as Fourier Transform Ion Cyclotron Resonance mass spectrometry, are expensive and usually not suited for high-throughput data acquisition. Zhang *et al.* (2005) use a similar approach for the identification molecular formulas of peptides, but artificially restrict the search space of molecular formulas. Heinonen *et al.* (2006) interpret tandem mass spectra by finding a best match in a database of molecular structures, modeling the fragmentation process as cuts of a molecule graph.

In this work, we present a method for the automated *de novo* identification of metabolites from quadrupole time-of-flight tandem mass spectra. Mass accuracy of this method is ∼20 ppm, one order of magnitude worse than for Fourier Transform mass spectrometry. The metabolite is fragmented using collision-induced dissociation (CID) (Wells and McLuckey, 2005), and several mass spectra are recorded for different fragmentation energies. We use this fragmentation information to identify the *molecular formula* of the metabolite. Mass spectra in our test dataset do not contain isotope peaks,

[*]To whom correspondence should be addressed.

so our method *does not* use isotopic patterns to identify the molecular formula. Such information can be easily integrated into the method and will further increase its identification accuracy.

We develop a model for the fragmentation process resulting in a graph theoretical problem called Maximum Colorful Subtree problem. Unfortunately, we can show that this problem is NP-hard. Despite this negative result, we develop several exact and heuristic algorithms for its solution. One of these exact algorithms is fixed-parameter tractable (FPT) (Niedermeier, 2006). The FPT algorithm and the heuristics show good performance in practice both with respect to identification accuracy and running times, as tests on real spectra reveal: we use a test dataset containing tandem mass spectra of 32 non-trivial metabolites, five of them with a mass over 400 Da. In all cases, the correct solution was among the top five candidates computed by our algorithms. For 26 compounds (81%), the first suggestion of the exact algorithm was correct. Unexpectedly, one heuristic shows a systematic error that even improves the results. Each algorithm needs about 1.5 min to process all mass spectra.

2 FRAGMENTATION MODEL

For proteins and glycans, experimental parameters are chosen in a way that molecules only fragment at specific chemical bonds, and fragmentation is generally well understood. However, metabolites can fragment at almost any chemical bond, and the fragmentation process is not completely understood and difficult to predict (Williams, 2002). In this work, we account for missing comprehension by allowing arbitrary fragmentation. To increase available information, several tandem mass spectra of each metabolite are measured at different collision energies. In our example data, collision energies were 15, 25, 40, 55 and 90 eV. Higher energies lead to smaller fragments, because more chemical bonds break.

To produce a set of candidate molecular formulas for the analyzed metabolite, we first decompose its parent mass using the Round Robin algorithm (Böcker and Lipták, 2007). In the next step, we rank candidates according to some score that measures agreement between the tandem mass spectra and the candidate. The first simple idea is to use a peak counting score: every peak that can be explained as a fragment of the parent molecule gets a score of 1. We rank the candidates according to the number of fragment peaks they *explain*: a molecular formula explains a fragment peak, if a sub-molecule of that formula possesses approximately the mass of the fragment peak.

But this simple scoring contradicts experimental reality: fragments of the parent molecule are fragmented for a second time when higher collision energies are applied, (Fig. 3) . This observation has been experimentally verified, and fragmentation trees can be reconstructed using ion trap mass spectrometers. To account for this multi-step fragmentation, we score candidate molecular formulas using a hypothetical fragmentation tree. This fragmentation tree is merely an aid to identify the molecular formula, although we found a high agreement with real fragmentation trees for the limited test data available.

To estimate the hypothetical fragmentation tree for a candidate molecular formula, we decompose all fragment peaks and keep only those molecular formulas which are sub-molecules of the candidate molecular formula: for every element, the molecular formula contains at most as many atoms as the candidate molecular formula. All these formulas and the candidate are considered vertices of a directed graph. Every vertex gets a color that represents the peak it explains, since one peak mass may be explained by a multitude of molecular formulas. We create a directed edge between two vertices if one molecular formula is a sub-molecule of the other molecular formula. Doing so, we represent every possible fragmentation step. Since the 'sub-molecule' relation is transitive, the constructed graph is also transitive: $(u,v),(v,w)\in E$ implies $(u,w)\in E$.

Finally, we calculate weights for all edges: each vertex is assigned to a unique sum formula and a (usually non-unique) peak, and we score the vertex using properties such as peak intensity or mass deviation. We can shift the vertex score to its incoming edge, since every vertex of a tree may only have one incoming edge. We use edge weights because we also want to score the hypothetical fragmentation step: for example, certain neutral losses are observed more frequently than others. Now, we search the resulting graph for a subtree that has maximum weight. One peak may result in many molecular formulas explaining it, therefore, to avoid peak double counting, we have to ensure that the subtree does not use any color twice. In the very rare case that two fragments share the same mass this will prevent correct double interpretation of a peak. But due to the transitivity of the graph, this will only lead to a vertex being jumped in the calculated fragmentation tree and a minor difference in the score. The score of the candidate molecular formula is set to the sum of edge weights of the fragmentation tree.

We denote the number of vertices of the graph n, the number of edges m and the number of colors k. Combining the requirement that every color may only occur once and the constraint that the result graph must be a tree, we define a colorful subtree as follows:

Definition 1. A *colorful subtree* $T=(V_T,E_T)$ of a vertex-colored directed acyclic graph G is a subtree of G which uses every color in C at most once:

$$\text{for all } c\in C: \; |\{v\in V_T \mid \text{color}(v)=c\}| \leq 1$$

As we are interested in the fragmentation tree with the highest score, it is necessary to look for the colorful subtree with maximum weight. Therefore, we define the following computational problem:

Definition 2. (Maximum Colorful Subtree)
Input: A vertex-colored edge-weighted directed acyclic graph G
Task: Find the colorful subtree of G that has maximal weight.

Unfortunately, this problem is computationally hard, as we show in the next section. But our experiments show that limiting ourselves to colorful subtrees is inevitable.

3 HARDNESS OF THE PROBLEM

Theorem 1. Maximum Colorful Subtree *is NP-hard, even if G is a tree with unit edge weights.*

Proof. We prove NP-hardness by reduction from the SAT problem that is known to be NP-complete (Garey and Johnson, 1979). An algorithm solves the SAT problem if it can decide whether a given Boolean expression in conjunctive normal form (CNF) is satisfiable. This proof is analogous to the proof that the Graph Motif problem on vertex colored graphs is NP-hard (Fellows *et al.*, 2007).

Given an instance of SAT as a CNF formula $\Phi=c_1\wedge\cdots\wedge c_s$ over variables $x_1,\ldots,x_t$, we construct an instance of Maximum Colorful Subtree as follows: we shall construct a tree with $2t+st+1$ vertices

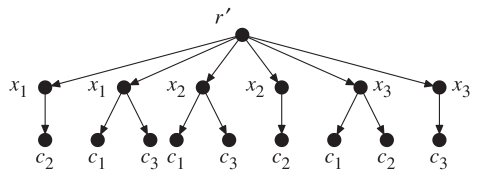

Fig. 1. Example for the construction of G. Vertices are labeled with their color. The Boolean expression consists of the three clauses: $c_1 = (\bar{x}_1 \vee x_2 \vee x_3), c_2 = (x_1 \vee \bar{x}_2 \vee x_3)$ *and* $c_3 = (\bar{x}_1 \vee x_2 \vee \bar{x}_3)$.

and $s+t+1$ distinct colors denoted $r', x_1, \ldots, x_t, c_1, \ldots, c_s$. We color the root-vertex r of the tree G by r' and connect $2t$ children to it. Then we assign every color x_i to exactly two of these children. The two vertices with color x_i represent both truth assignments for x_i. If a truth assignment to x_i satisfies clause c_j we connect a new vertex colored c_j to that vertex colored x_i, that corresponds to this truth assignment. This assignment is done for all literals in all clauses. To complete the construction we assign a score of 1 to every edge of G. An example for the construction can be found in Figure 1.

As the constructed tree has as many leaves as there are literals in Φ, the construction is polynomial. We claim that Φ is satisfiable if and only if a maximum colorful subtree of G has a score equal to the number of clauses s plus the number of variables t of Φ. To prove the forward direction, assume a truth assignment ϕ that satisfies Φ. Define $A \subset V(G)$ to be the subset of vertices in the colors x_i that correspond to the assignment ϕ. Then, for every j there exists at least one vertex colored c_j in the neighborhood of A. Add an arbitrary representative of these vertices colored c_j to the set $B \subset V(G)$. The union of these sets $A \cup B \cup \{r\}$ forms a colorful subtree T of G. It has a score of $s+t$, as for each clause and for each variable there is one edge in T. No colorful subtree with a higher score can exist; therefore T is a maximum colorful subtree of G.

To prove the backward direction assume there is a maximum colorful subtree T of G with score $s+t$. The tree T then contains all colors of G. Therefore, the truth assignment corresponding to the vertices of T colored x_i satisfies Φ, as for all j at least one vertex colored c_j is connected to these vertices, otherwise T would not contain all colors.

4 ALGORITHMS

We now describe algorithms to solve the Maximum Colorful Subtree problem. Despite the complexity of the problem, especially the FPT algorithm performs sufficiently fast in practice even for large molecules. We also present two heuristics which further reduce running times.

4.1 Brute force

The brute force (BF) approach iterates over all combinations of vertices that can form a colorful subtree. For every such combination the algorithm calculates the Maximum Spanning Tree (MST) using Kruskal's algorithm (Kruskal, 1956) with some adjustments to ensure that it results in a directed tree. The MST with the highest weight is then returned as result.

4.2 Branch and bound

For this algorithm, we use the classical branch and bound approach to test all combinations of edges that could define a maximum colorful subtree. Depth first search is applied to find a good solution early during the search, which can then be used for bounding. As relaxation during the search for an upper bound, edges may violate the tree property and render the tree not colorful in this phase of the algorithm. Since this algorithm turned out to perform badly in applications, we omit further details on this concept.

4.3 FPT dynamic programming

Scott *et al.* (2006) propose an algorithm to find colorful subtrees in so-called protein-protein interaction networks which they colored randomly. The basis for this parametrized algorithm is dynamic programming (DP) over the vertices and all possible subsets of the color set C. Let $W(v,S)$ denote the maximal score of a colorful subtree with root v and using colors $S \subseteq C$. We derive the following recurrence:

$$W(v,S) = \max \begin{cases} \displaystyle\max_{u:c(u)\in S\setminus\{c(v)\}} W(u, S\setminus\{c(v)\}) + w(v,u) \\ \displaystyle\max_{\substack{(S_1,S_2):S_1\cap S_2=\{c(v)\}\\ S_1\cup S_2=S}} W(v,S_1)+W(v,S_2) \end{cases}$$

with initial condition $W(v,\{c(v)\})=0$ and the weight of non- existent edges set to $-\infty$. The first line extends a tree by just introducing v as new root, and adding the score of the edge (v,u) to the score of the tree. In the program this is done by iterating over all outgoing edges of v. The second line merges two trees, which have nothing in common but their root. This is the expensive calculation, although in practice the implementation iterates over defined values only. Not all entries of W are defined, as there does not necessarily exist a subtree of the input rooted in v using exactly the colors in S. The worst case running time for this algorithm is $O(3^k km)$ and the necessary space is $O(2^k n)$. The algorithm needs 3^k steps to calculate the second line of the recurrence, since the k colors are divided into three groups: those not contained in S, element of S_1 or element of S_2. There are 3^k possibilities to perform this division. This yields an fixed-parameter algorithm as the number of colors k, representing the number of peaks in the input spectra, restricts the exponential growth.

The major disadvantage of this method is its memory consumption. To perform backtracking and thus construct the fragmentation tree it is necessary to store the order the colors were added to the sets. But if the user is not interested in the fragmentation tree, it is possible to implement the color sets S as bit strings, minimizing the necessary space. Although this optimization only decreases memory demands by a constant factor, this often makes the difference between finding a solution and running out of memory. If the user is only interested in the fragmentation trees of the best f decompositions, we can optimize space demands as follows: first the best scoring decompositions of the parent ion are determined using bit strings. Afterwards a graph is constructed containing only the best f parent mass decompositions and their children.

4.4 Combination of dynamic programming and brute force

If the vertices-to-color ratio n/k is small, the brute force algorithm outperforms the dynamic programming approach, whereas the dynamic programming algorithm is faster for larger n/k.

Therefore, we combine both techniques and use the one more suitable for the current input. Experiments show that a n/k ratio of 4 is a good change-over point to optimize running times. The ratio n/k may also be read as explanations per peak. As there are few explanations for small molecules, this concept corresponds to processing small molecules with the brute force algorithm, whereas we use the dynamic programming to analyze larger molecules.

4.5 Heuristics

The two heuristic algorithms described here usually do not compute the optimal score, but are nevertheless useful for metabolite identification since their error appears to be systematic.

The *greedy heuristic* is a simplification of the branch and bound approach. We sort edges according to their scores in descending order. We then pick the first edge in this list. Afterwards the next edge from the list that does neither violate the tree property nor make the result tree not colorful, is selected. The algorithm continues until $k-1$ edges have been selected. This corresponds to the 'first guess' of the branch and bound approach.

The *top–down heuristic* is also a greedy concept, but here the algorithm always tries to find paths beginning at the root. The algorithm starts at the root and follows the best scoring outgoing edge. To follow an edge means to add it to the solution set and continue from the vertex at its end. At the next vertex, it again follows the best scoring outgoing edge that can be added without rendering the result not colorful or violating the tree property. If no such edge exists, the algorithm moves back to the root. It terminates if no edge at the root can be selected. This way, all colors are present in the resulting tree because the input graph is transitive.

5 SCORING

A peak counting score is not sufficient to obtain good results, since many molecular formulas can explain all peaks in a spectrum. Therefore, we introduce a scoring concept which takes the following properties into account: peak intensities, mass deviation between explanation and peak, chemical properties of the molecular formula, collision energies of the spectra and the neutral loss arising from the fragmentation step. Since scores represent the likelihood that a certain fragmentation step occurs we use log scores as edge weights to obtain an additive scoring scheme. Recall that scores assigned to vertices, such as the intensities and the mass deviation, are applied to all incoming edges of the respective vertex.

5.1 Merging peaks

In this work, we analyze series of spectra with different collision energies, but we shall treat the data as if it were only a single spectrum. Hence it is necessary to merge peaks into one spectrum before scoring can take place. This is done by applying a threshold, merging peaks from different spectra whose mass difference is smaller than the threshold. We also require that peaks have to be in adjacent spectra. For example, if peaks with similar mass occurred in the spectrum with 15 and 35 eV, but not in the spectrum with 25 eV, the program would not merge them, as they most likely have different explanations but incidentally the same mass: otherwise, a peak at 25 eV must also exist.

5.2 Filtering

As mentioned before, we want to avoid strict filters, because this allows us to find solutions with unexpected properties. If unlikely properties of a molecule are encountered, we only decrease its score. In our experiments, it turned out that we had to apply one filter: this was done to reduce the number of candidate molecular formulas to a point that could be handled by our algorithms with reasonable time and space. We filter on the basis of Senior's (1951) third theorem which states that the sum of valences has to be greater than or equal to twice the number of atoms minus one. Molecules violating Senior's third theorem are rare, especially in natural compounds: these molecules often have an unusual elemental composition such as a high amount of fluorine. Kind and Fiehn (2007) find 64 substances violating the rule in the 45 000 entries of the Wiley mass spectral database (McLafferty, 2005). Thus the filter has a sensitivity of 99.86%. If the user is interested in molecules violating Senior's theorem, the filter can easily be disabled. This results in longer running times and slightly worse identification results.

5.3 Peak intensities

Clearly, a solution explaining more intense peaks should receive a higher score. Therefore, the intensity of the peak has to influence scoring. There is one problem in this concept, however. Peak intensities are usually normalized within a single spectrum: all peaks are scaled relatively to the most intense peak. Therefore intensities of two spectra are not comparable. We propose two possibilities to overcome this restriction.

One is to re-calculate the raw intensities using the total ion current (TIC) as a base value. The TIC is usually stored as a property of the spectrum. The other idea is to rank the spectra and to normalize these ranks between 0 and 1: thus the highest peak of a spectrum receives a score of 1. The advantages of the latter concept are similar to those of the Spearman rank order correlation. We use re-calculation of raw intensities, because it yields slightly better results on the test data.

5.4 Mass deviation

When interpreting mass spectrometry data, a scoring has to take into account the deviation between the calculated decomposition mass and the measured peak. Since mass spectrometrists assume that the mass error of a device roughly is normal distributed, we evaluate the logarithmized Gaussian probability density function at the measuring error value, and add it to the score. As SD σ we use $1/3$ or $1/2$ of the relative mass error, assuming that 99.8% or 95% of all measured peaks, respectively, have a mass error smaller than the given value. In our experiments, we apply $\sigma = 20/3$ Da.

5.5 Hetero atom to carbon ratio

In organic chemistry, all atoms not being carbon and hydrogen are called hetero atoms. The hetero to carbon ratio can be used to test if a molecular formula corresponds to a real molecule, as this ratio is typically between 0.25 and 1 in biologically relevant molecules. We find the hetero to carbon ratio of the molecules in the KEGG database (Kanehisa *et al.*, 2006) to be normally distributed. The distribution has a mean of 0.59 and a SD of 0.56. We use the density function of this distribution for scoring. But the hetero atom to carbon ratio has the disadvantage of being hereditary. If a molecule has a high ratio, its fragments are also likely to have a high hetero atom to carbon ratio. Thus the high ratio will be penalized in every fragmentation step, which is not desirable. We avoid the problem by the following procedure: first, hetero-ratio scores of both decompositions incident to an edge are calculated. If the score of a fragment is better than the score of its predecessor, the hetero-ratio score of this edge remains unchanged. Otherwise the difference between the score of the fragment and the score of the predecessor is subtracted from the score of the edge. This decreases the score only if the fragment has a more unexpected hetero to carbon ratio than its predecessor.

5.6 Collision energies

Because mass spectra are measured using different collision energies, we can deduce that some peaks cannot represent direct fragments of other peaks: this is the case if they appeared at a lower energy than the predecessor peak or at a high energy where the predecessor peak has long disappeared. Ideally, there is a collision energy where both peaks appear. This will be given full score.

It is highly unlikely that the fragment peak appears before the predecessor peak, therefore $\log(\alpha)$, $\alpha \ll 1$ is added to the score. It is possible though, that the mass spectrometer did not detect the predecessor peak in the relevant spectrum, thus the edge is not deleted.

Another possibility is that the predecessor peak ceases, there is one spectrum where it cannot be found, and then at the next higher collision energy, the fragment peak starts to emerge. This fragment did most likely not directly emerge from the predecessor, so this is as well given a score of $\log(\alpha)$.

If there is no spectrum in which both peaks can be found, but neither a spectrum containing none of the peaks in question, it is possible that the molecules are direct fragments, but there might as well exist another fragment between them. Therefore in this situation $\log(\beta)$, $\alpha < \beta < 1$ is added to the score. We use $\alpha = 0.1$ and $\beta = 0.8$ as initial estimates that appear to reasonably capture the experimental set-up. To avoid overfitting, we did not optimize these parameters for high identification rates. We shall adjust these values in a training process when a larger training set is available.

5.7 Mass of the neutral loss

Since the base of the scoring is still the peak counting concept, the algorithm would often suggest a star as the fragmentation tree. As stated above, that is not the desired situation, as we want to model the multiple steps of fragmentation. Therefore, we penalize large neutral losses although this is not justified chemically. The neutral loss corresponding to an edge can easily be calculated by subtracting the molecular formulas of the incident vertices from each other. We add the logarithm of the complement of the ratio between the mass of the neutral loss represented by the current edge and the parent mass, $\log(1 - \frac{\text{mass(neutral loss)}}{\text{parent mass}})$, to the score.

5.8 Common neutral losses

Certain neutral losses occur often during fragmentation, especially in biological compounds. Chemists even rely on those to classify analytes. An expert provided us with a short list of common neutral losses. This list can be found in Table 1. To account for this fact we increase an edge score by $\log(\gamma), \gamma > 1$ if a neutral loss is among this list, or is a combination of at most three list entries. Additionally, radicalic neutral losses are rare. Therefore we decrease the score of an radicalic neutral loss by $\log(\delta), \delta < 1$. To determine whether the neutral molecule with a given formula is radicalic, we check whether its double-bond equivalent is non-integer. We use $\gamma = 2$ and $\delta = 0.25$, but again need to optimize these heuristic values by a training process, when more data is available.

Table 1. The common neutral losses the program uses

Name	Formula	Name	Formula
Methyl	CH_3	Isobutene	C_4H_8
Methane	CH_4	Isopentene	C_5H_8
Oxy	O	Formic acid	CH_2O_2
Hydroxyl	H_2O	Malonic acid	$C_3H_2O_3$
Carbon monoxide	CO	Xylose	$C_5H_8O_4$
Nitrogen	N_2	Rhamnose	$C_6H_{10}O_4$
Ammonia	NH_3	Hexose	$C_6H_{10}O_5$
Ethyl	C_2H_4	Glucuronic acid	$C_6H_8O_6$
Formaldehyde	CH_2O		

If an entry of this table occurs in a fragmentation step, the score of the step is significantly increased.

6 EMPIRICAL RESULTS

We implemented our algorithms in Java 1.5. Running times were measured on an Intel Pentium IV, 1.8 GHz with 512 MB memory. As test data, we used 150 tandem mass spectra of 32 metabolites (unpublished). These metabolites were either commercially available reference compounds or extracted from the seed of *A.thaliana* plants. The test set contained the biogenic amino acids and many complex choline derivatives. Separation was done using a capillary HPLC system. The spectra were measured on an API QSTAR Pulsar Hybrid Quadrupole TOF instrument by Applied Biosystems. Raw data were preprocessed using the AnalystQS software supplied with the instrument. A more detailed description of the experimental setup can be found in von Roepenack-Lahaye *et al.* (2004). The test set was analyzed with the following options: masses were decomposed using a relative mass error of 20 ppm over the standard CHNOPS-alphabet containing the six elements most abundant in living organisms. The original intensities were re-calculated using the TICs, and afterwards peaks closer than 0.1 Da have been merged. Scoring was performed as described in the previous section, using parameters $\sigma = 20/3$ Da, $\alpha = 0.1$, $\beta = 0.8$, $\gamma = 2$ and $\delta = 0.25$.

Identification results can be found in Table 2. The exact algorithms excellently identify metabolite molecular formulas. For the majority of compounds the correct molecular formula is ranked first, even for such large compounds as 4-hexosylvanilloyl choline (416 Da). All correct formulas can be found among the first five solutions, enabling researchers to restrict further analysis to the top five candidates. For comparison, we dropped the requirement that our fragmentation tree has to be colorful: identification accuracy degrades dramatically if colors are ignored, and for none of the metabolites above 400 Da the correct sum formula was among the first 18 solutions. This clearly suggests that we must force our algorithms to select at most one explanation per peak.

Table 2. The identification rates of the exact algorithm, the greedy heuristic and the top down heuristic

Mass range	Number of compounds	Exact and greedy heuristic			Top–down heuristic		
		Top 1(%)	Top 2(%)	Top 5(%)	Top 1(%)	Top 2(%)	Top 5(%)
100–200 Da	15	100	100	100	100	100	100
200–300 Da	10	70	80	100	80	90	100
300–400 Da	2	50	100	100	50	100	100
400–500 Da	5	60	80	100	100	100	100

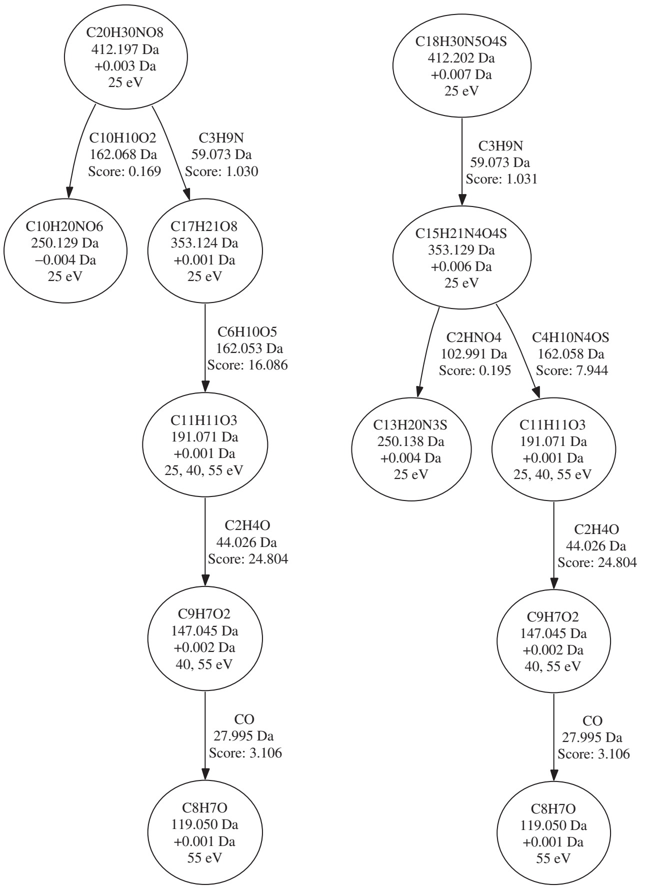

Fig. 2. Two fragmentation trees calculated from the spectra of hexosyloxycinnamoyl choline. The correct solution was ranked first of 113 candidates. Note the identical fragmentation cascades in the lower right of the trees. Choline ions have an intrinsic positive charge. Left: fragmentation tree of the correct molecular formula ranked at first position, score 97.93. Right: fragmentation tree of an incorrect molecular formula ranked at seventh position, score 81.83.

Figure 2 shows two hypothetical fragmentation trees calculated from the spectra of hexosyloxycinnamoyl choline. The tree on the left uses the correct sum formula as root, whereas the right tree is based on a wrong candidate. They exhibit a non-linear branching of the fragmentation process which we find in most of the analyzed compounds. This suggests that it is indeed necessary to search for trees and not only linear structures. The trees also illustrate that higher scoring candidates often share fragmentation cascades: two fragmentation steps at the lower right are completely identical for both candidates. The reason for the correct candidate to receive a significantly higher score is that hexose ($C_6H_{10}O_5$) is separated in the fragmentation process of the correct molecular formula. Hexose is an entry of Table 1 and therefore the corresponding fragmentation receives a high score.

The results of the greedy heuristic are identical to those of the exact algorithm. The scores calculated by the heuristics are suboptimal, but they produce a systematic error resulting in the same ranks. As an example, consider hexosyloxybenzoyl choline depicted in Figure 3: the exact algorithms find a solution with score 103.9 for the correct molecular formula and 102.65 for the runner-up, whereas the top–down heuristic scores the correct solution with 51.67 and the runner-up receives a score of 45.85. Unexpectedly, using the top–down heuristic even improves the results. Further tests on other data need to show whether this is generally the case. We cannot yet provide an explanation for this finding.

Running times of the different approaches can be found in Table 3. The speed of both heuristics and the DP + BF exact algorithm is sufficient to analyze data on the fly. It takes around 3 sec to identify one compound on a standard PC, which is significantly faster than measuring the spectra. We stress that the brute force algorithm significantly slows down for metabolites above 400 Da, which severely limits its use for even larger molecules.

Although focus of our research is the identification of metabolites, the maximum colorful subtree of the correct parent mass decomposition is also a prediction of the fragmentation process. The true fragmentation tree of one metabolite, hexosyloxybenzoyl choline, was manually constructed using ion trap mass spectrometry. In this case, the predicted fragmentation tree shown in Figure 3 exactly matched the manually constructed one.

7 CONCLUSION

We have presented a concept for the automated *de novo* identification of metabolites using tandem mass spectra. Our fragmentation model leads to the Maximum Colorful Subtree problem, which is NP-hard. We propose an exact FPT algorithm as well as two heuristics to solve the problem, and we introduce a scoring scheme based on properties of mass spectra and peak decompositions. Experiments on real mass spectra show that our method achieves very good identification results in application.

Zhang *et al.* (2005) present identification results almost as good as ours, but the authors limit their search to certain molecular formulas thought relevant for peptides, and use isotopic patterns to identify the fragments. As soon as metabolite tandem mass

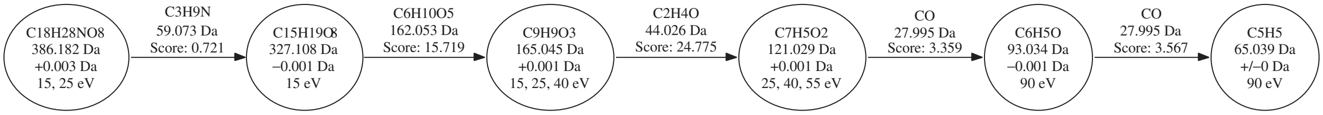

Fig. 3. The calculated and manually confirmed fragmentation tree of hexosyloxybenzoyl choline. The correct tree was ranked first of 87 candidates.

Table 3. The total running times of the algorithms

Algorithm	Running time (min)
Branch and bound	1560
Brute force	5.2
Dynamic programming	72.6
Combination DP+ BF	1.5
Greedy heuristic	1.5
Top–down heuristic	1.2

spectra containing isotopic patterns are available to us, we will integrate this data into our analysis: since modifications of the experimental setup are minor, we hope to receive such data soon. Experiments by Kind and Fiehn (2006) indicate that this will boost the discriminative power of our approach, and allow for the exact identification of large metabolites. It is easy to extend the alphabet in our approach. Therefore, we will perform tests on metabolites containing fluorine or chlorine as soon as we have corresponding data available. Our method can also be used for cross-validation in an fully automated mass spectrometry pipeline, if relying solely on database matches is to inexact and would otherwise need manual validation.

In the future, we want to use maximum colorful subtrees for the scoring of molecular structures, adopting and improving concepts of Heinonen *et al.* (2006). When tandem mass spectrometers with higher resolution become available, it might even be possible to elucidate molecular structures from the fragment spectra: the predicted fragment formulas could be used to restrict the output of a molecular formula generator.

ACKNOWLEDGEMENTS

We thank the department of Stress and Developmental Biology at the Leibniz Institute of Plant Biochemistry in Halle, Germany for supplying us with the test data and Christoph Böttcher for the manual identification of the fragmentation tree. Jürgen Schmidt provided the list of common neutral losses.

Conflict of Interest: none declared.

REFERENCES

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Böcker,S. and Lipták,Zs. (2007) A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, **48**, 413–432.

Böcker,S. *et al*. (2006) Decomposing metabolomic isotope patterns. In *Proceedings of Workshop on Algorithms in Bioinformatics (WABI 2006)*, Vol. 4175 of *Lecture Notes Computer Science*. Springer Verlag, Berlin Heidelberg, pp. 12–23.

Chen,T. *et al*. (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **8**, 325–337.

D'Auria,J.C. and Gershenzon,J. (2005) The secondary metabolism of Arabidopsis thaliana: growing like a weed. *Curr. Opin. Plant Biol.*, **8**, 308–316.

Fellows,M. *et al*. (2007) Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *International Colloquium on Automata, Languages and Programming (ICALP 2007)*, Vol. 4596 of *Lecture Notes In Computer Science*. Springer Verlag, Berlin Heidelberg, pp. 340–351.

Garey,M.R. and Johnson,D.S. (1979) *Computers and Intractability (A Guide to Theory of NP-Completeness)*. Freeman, New York.

Heinonen,M. *et al*. (2006) Ab initio prediction of molecular fragments from tandem mass spectrometry data. In *Proceedinds of German Conference on Bioinformatics (GCB 2006), Lecture Notes in Informatics*, P-83, Springer Verlag, Berlin Heidelberg, pp. 40–53.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic. Acid Res.*, **34**, D354–D357.

Kind,T. and Fiehn,O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, **7**, 234.

Kind,T. and Fiehn,O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

Kruskal,J. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.*, **7**, 48–50.

McLafferty,F.W. (2005) *Wiley Registry of Mass Spectral Data*. 7th edition with NIST 2005 spectral data edition. John Wiley & Sons, Hoboken, NJ.

Niedermeier,R. (2006) *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, Oxford.

Pitzer,E. *et al*. (2007) Assessing peptide de novo sequencing algorithms performance on large and diverse data sets. *Proteomics*, **7**, 3051–3054.

Scott,J. *et al*. (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol.*, **13**, 133–144.

Senior,J. (1951) Partitions and their representative graphs. *Am. J. Math.*, **73**, 663–689.

von Roepenack-Lahaye,E. *et al*. (2004). Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.*, **134**, 548–559.

Wells,J.M. and McLuckey,S.A. (2005) Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.*, **402**, 148–185.

Williams,A. (2002) Applications of computer software for the interpretation and management of mass spectrometry data in pharmaceutical science. *Curr. Top. Med. Chem.*, **2**, 99–107.

Zhang,J. *et al*. (2005) Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 217–230.