Multiple Mass Spectrometry Fragmentation Trees Revisited: Boosting Performance and Quality

Kerstin Scheubert, Franziska Hufsky, and Sebastian Böcker

Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, Jena, Germany,
{kerstin.scheubert,franziska.hufsky,sebastian.boecker}@uni-jena.de

This is a preprint of the article: Kerstin Scheubert, Franziska Hufsky and Sebastian Böcker. Multiple Mass Spectrometry Fragmentation Trees Revisited: Boosting Performance and Quality. In Proc. of Workshop on Algorithms in Bioinformatics (WABI 2014), volume 8701 of Lect Notes Comput Sci, pages 217-231. Springer, Berlin, 2014.
The final publication is available at springerlink.com.

Abstract. Mass spectrometry (MS) in combination with a fragmentation technique is the method of choice for analyzing small molecules in high throughput experiments. The automated interpretation of such data is highly non-trivial. Recently, fragmentation trees have been introduced for *de novo* analysis of tandem fragmentation spectra (MS^2), describing the fragmentation process of the molecule. Multiple-stage MS (MS^n) reveals additional information about the dependencies between fragments. Unfortunately, the computational analysis of MS^n data using fragmentation trees turns out to be more challenging than for tandem mass spectra.

We present an Integer Linear Program for solving the Combined Colorful Subtree problem, which is orders of magnitude faster than the currently best algorithm which is based on dynamic programming. Using the new algorithm, we show that correlation between structural similarity and fragmentation tree similarity increases when using the additional information gained from MS^n. Thus, we show for the first time that using MS^n data can improve the quality of fragmentation trees.

Keywords: metabolomics, computational mass spectrometry, multiple-stage mass spectrometry, fragmentation trees, Integer Linear Programming

1 Introduction

Studying metabolites and other small biomolecules with mass below 1000 Da, is relevant, for example, in drug design and the search for new signaling molecules and biomarkers [14]. Since such molecules cannot be predicted from the genome sequence, high-throughput *de novo* identification of metabolites is highly sought. Mass spectrometry (MS) in combination with a fragmentation technique is commonly used for this task. In liquid chromatography MS, a selected molecule can be fragmented in a second step typically using collision-induced dissociation (CID). The resulting fragment ions are recorded in tandem mass spectra (MS^2 spectra). For metabolites, the understanding of CID fragmentation is still in its infancy.

Multiple-stage MS (MS^n) allows to select the product ions of the initial fragmentation step (manually or automatically) and subject them to another fragmentation reaction. This reveals additional information about the dependencies between the fragments. The resulting fragment ions can, in turn, again be selected as precursor ions for further fragmentation. Typically, with each additional fragmentation reaction, the quality of mass spectra is reduced and measuring time increases. Thus, analysis is usually limited to a few fragmentation reactions beyond MS^2.

CID mass spectra (both MS^2 and MS^n) are limited in their reproducibility on different instruments, making spectral library search a non-trivial task [16]. Furthermore, spectral libraries are vastly incomplete. Recent approaches tend to replace searching in spectral libraries by searching in the more comprehensive molecular structure databases [1, 9–11, 26, 31]. However, many metabolites even remain uncharacterized with respect to their structure and function [17].

For the *de novo* interpretation of tandem mass spectra of small molecules, Böcker and Rasche [5] introduced fragmentation trees to identify the molecular formula of an unknown and its fragments.

Moreover, fragmentation trees are reasonable descriptions of the fragmentation process and hence can also be used to derive further information about the unknown molecule [19]. Scheubert *et al.* [23, 24] adjusted the fragmentation tree concept to MS^n data to reflect the succession of fragmentation reactions.

Adjusting the fragmentation tree concept to MS^n data, results in the NP-hard Colorful Subtree Closure problem [24] which has to be solved in conjunction with the original NP-hard Maximum Colorful Subtree problem [5], resulting in the Combined Colorful Subtree problem [24]. To solve this problem, Scheubert *et al.* [24] presented a fixed-parameter algorithm based on dynamic programming (DP) with worst-case running time depending exponentially on the number of peaks in the spectrum.

To compare two molecules based on their fragmentation spectra, Rasche *et al.* [18] introduced fragmentation tree alignments. By this, similar fragmentation cascades in the two trees are identified and scored. This allows us to use fragmentation trees in applications such as database searching, assuming that structural similarity is inherently coded in the CID spectra fragments. Improving the quality of the fragmentation trees using the additional information provided by MS^n, may improves this downstream analysis.

Here, we present a novel exact algorithm for solving the Combined Colorful Subtree problem. This Integer Linear Program (ILP) is faster than the DP algorithm. Further, we demonstrate the impact of the additional information from MS^n data for the downstream analysis: We compute fragmentation tree alignments [18] and find that correlation between the similarity score of two fragmentation trees and the structural similarity score of the corresponding molecules increases when using the additional information gained from the succession of fragments in multiple MS.

2 Constructing Fragmentation Trees

Given the molecular structure of a molecule and the measured fragmentation spectrum, an MS expert can assign peaks to fragments of the molecule and derive a "fragmentation diagram". *Fragmentation trees* are similar to experts' "fragmentation diagrams" but are extracted directly from the data, without knowledge about a molecule's structure. A fragmentation tree consists of vertices annotated with the molecular formulas of the precursor ion and fragment ions, and directed edges representing the fragmentation steps. Fragmentation trees must not be confused with *spectral trees* for multiple stage mass spectrometry [22, 25]. Spectral trees are a formal representation of the MS setup and describe the relationship between the MS^n spectra, but do not contain any additional information.

For the computation of fragmentation trees [5], a fragmentation graph is constructed (see Fig. 1): vertices represent all fragment molecular formulas with mass sufficiently close to the peak mass [3, 4]; and weighted edges represent the fragmentation steps leading to those formulas. Two vertices u, v are connected by a directed edge if the molecular formula of v is a sub-molecule of the molecular formula of u. We assume the molecular formula of the full molecule to be given (see [19] for details). The resulting graph is a directed acyclic graph (DAG) $G = (V, E)$, since fragments can only lose, never gain, weight. Vertices in the graph are colored $c : V \to \mathcal{C}$, such that vertices that explain the same peak receive the same color. Edges are weighted, reflecting that some fragmentation steps are more likely. Common fragmentation steps get a higher weight than implausible fragmentation steps. Also peak intensities and mass deviations are taken into account in these weights. The resulting fragmentation graph contains all possible fragmentation trees as subgraphs. The *weight* of an induced tree $T = (V_T, E_T)$ is defined as the sum of its edge weights: $w(T) := \sum_{(u,v)\in E_T} w\big(u, v\big)$.

The MS^n data does not only hint to direct but also to indirect successions, that is a fragment is not only scored based on its direct ancestor (its parent node), but also on indirect ancestors (grandparent node etc). Thus, we also have to score the transitive closure of the induced subtrees [24]. The *transitive closure* $G^+ = (V, E^+)$ of a DAG $G = (V, E)$ contains the edge $(u, v) \in E^+$ if and only if there is a directed path in G from u to v. As MS^n data does not differentiate between different explanations of the peaks, we score pairs of colors: $w^+ : \mathcal{C}^2 \to \mathbb{R}$. The *transitive weight* of an induced tree $T = (V_T, E_T)$ with transitive closure $T^+ = (V_T, E_T^+)$ is defined as

$$w^+(T) := \sum\nolimits_{(u,v)\in E_T^+} w^+\big(c(u), c(v)\big) \tag{1}$$

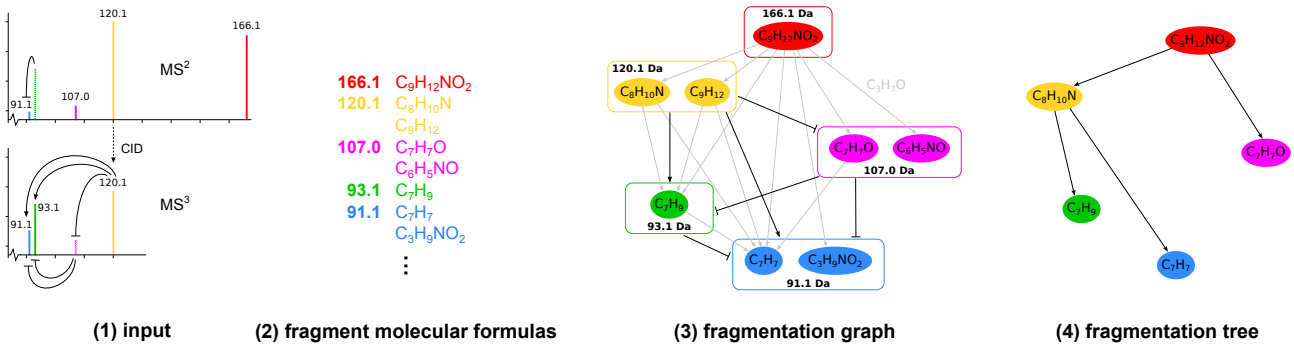

Fig. 1. (1) As input we use MS^n spectra that contain additional information on the succession of fragments. (2) For each peak, we compute all fragment molecular formulas with mass sufficiently close to the peak mass. (3) A fragmentation graph is constructed with vertices for all fragment molecular formulas and edges (grey) for all possible fragmentation steps. Explanations of the same peak receive the same color. The transitive closure of the graph is scored based on pairs of colors. To simplify the drawing, we only show non zero edges of the transitive closure (black). (4) The colorful subtree with maximum combined weight of the edges and the transitive closure is the best explanation of the observed fragments.

Scheubert *et al.* [24] introduced three parameters σ_1, σ_2 and σ_3 to score the transitive closure. Parameter σ_1 rewards fragments of an MS^n spectrum that are successors of its parent fragment ($\sigma_1 \geq 0$). Parameter σ_2 penalizes fragments that are successors of a parent fragment of an MS^n spectrum although the corresponding peak is not contained in this spectrum ($\sigma_2 \leq 0$). Parameter σ_3 penalizes direct and indirect fragmentation steps that occur at high collision energy but not at low collision energy ($\sigma_3 \leq 0$). For a more detailed description of the parameters see [24].

Now, each subtree of the fragmentation graph corresponds to a possible fragmentation tree. Considering trees, every fragment is explained by a unique fragmentation pathway. To avoid the case that one peak is explained by more than one molecular formula, we limit our search to *colorful* trees, where each color is used at most once. In practice, it is very rare that a peak is indeed created by two different fragments. Searching for a colorful subtree of maximum sum of edge weights is known as the Maximum Colorful Subtree problem, which is NP-hard [5,8]. Searching for a colorful subtree of maximum weight of the transitive closure is known as the Colorful Subtree Closure problem, which is again NP-hard (even for unit weights) [24]. In addition, both problems are even hard to approximate [6, 24, 27]. The problem we are interested in combines the two above problems, that is searching for a colorful subtree of maximum combined weight of the edges and the transitive closure, which is the best explanation of the observed fragments [24]:

Combined Colorful Subtree problem. Given a vertex-colored DAG $G = (V, E)$ with colors $\mathcal{C}$, edge weights $w : E \to \mathbb{R}$, and transitive weights $w^+ : \mathcal{C}^2 \to \mathbb{R}$. Find the induced colorful subtree T of G of maximum weight $w^*(T) = w(T) + w^+(T)$.

3 Integer Linear Programming for Fragmentation Trees

For the computation of fragmentation trees from tandem MS data, several exact and heuristic algorithms to solve the Maximum Colorful Subtree problem have been proposed and evaluated [5, 19,20], inter alia a fixed-parameter algorithm using dynamic programming (DP) over vertices and color subsets [5,7], and an Integer Linear Program (ILP) [20] (see below) – both computing an exact solution. For multiple MS data, Scheubert *et al.* [24] presented an exact DP algorithm for the Combined Colorful Subtree problem, which is parameterized by the number of colors k in the graph. Here, we present an ILP for solving the Combined Colorful Subtree problem. ILPs are a classical approach for finding exact solutions of computationally hard problems.

3.1 ILP for Tandem MS

We first repeat the ILP introduced by Rauf *et al.* [20] for tandem MS data. By mapping all peaks into a single "pseudo tandem MS" spectrum we can also use this ILP to find a fragmentation tree

for multiple MS data. However, by doing so, we ignore the additional information gained from the succession of fragments in multiple MS.

Let $G = (V, E)$ be the input graph, and let $\mathcal{C} : V \to C$ denote the vertex coloring of G. We assume that G has a unique source r that will be the root of the subtree. For each color $c \in C$ let $V(c)$ be the set of all vertices in G which are colored with c. We introduce binary variables x_{uv} for each edge $uv \in E$, where $x_{uv} = 1$ if and only if uv is part of the subtree.

$$\max \sum_{uv \in E} w(u, v) \cdot x_{uv} \tag{2}$$

$$\text{s.t.} \sum_{u \text{ with } uv \in E} x_{uv} \leq 1 \qquad \text{for all } v \in V \setminus \{r\}, \tag{3}$$

$$x_{vw} \leq \sum_{u \text{ with } uv \in E} x_{uv} \qquad \text{for all } vw \in E \text{ with } v \neq r, \tag{4}$$

$$\sum_{uv \in E \text{ with } v \in V(c)} x_{uv} \leq 1 \qquad \text{for all } c \in C, \tag{5}$$

$$x_{uv} \in \{0, 1\} \qquad \text{for all } uv \in E. \tag{6}$$

Constraints (3) ensure that the feasible solution is a tree, whereas constraints (5) make sure that there is at most one vertex of each color present in the solution. Finally, constraints (4) require the solution to be connected. Note that in general graphs, we would have to ensure for every cut of the graph to be connected to some parent vertex. That would require an exponential number of constraints [15]. But since our graph is directed and acyclic, a linear number of constraints suffice. White *et al.* [30] pointed out that constraints (3) are redundant due to constraints (5). However, in the following we will refer to the original ILP from [20].

3.2 ILP for Multiple MS Allowing Transitivity Penalties Only

A rather simple ILP for solving the Combined Colorful Subtree problem extends the ILP from Rauf *et al.* [20] by adding constraints similar to [2] to capture the transitivity of the closure. To this end, we will introduce additional variables that capture the edges of the transitive closure of the tree. Unfortunately, this simple approach is only working for negative weights for all edges of the transitive closure and cannot be generalized to arbitrary transitivity scores.

Let $G^+ = (V, E^+)$ be the transitive closure of the input graph G. We assume that $w^+(c(u), c(v)) \leq 0$ holds for all edges uv of the transitive closure. Let us define binary variables x_{uv} for each edge $uv \in E$, and z_{uv} for each edge $uv \in E^+$. We assume $x_{uv} = 1$ if and only if uv is part of the subtree; and $z_{uv} = 1$ if uv is part of the closure of the subtree. We can formulate the following ILP:

$$\max \sum_{uv \in E} w(u, v) \cdot x_{uv} + \sum_{uv \in E^+} w^+(c(u), c(v)) \cdot z_{uv} \tag{7}$$

satisfying constraints (3), (4), (5) and, in addition:

$$x_{uv} \leq z_{uv} \qquad \text{for all } uv \in E, \tag{8}$$

$$z_{uv} + z_{vw} - z_{uw} \leq 1 \qquad \text{for all } uv, vw \in E^+, \tag{9}$$

$$x_{uv} \in \{0, 1\} \qquad \text{for all } uv \in E, \tag{10}$$

$$z_{uv} \in \{0, 1\} \qquad \text{for all } uv \in E^+. \tag{11}$$

As $w^+(c(u), c(v)) \leq 0$ for all $uv \in E^+$ we may assume that $z_{uv} = 0$ holds unless required otherwise by (8) or (9). Constraint (8) requires that all edges of the subtree are also edges of the closure; constraint (9) results in the transitivity of the closure.

Unfortunately, the above ILP cannot be generalized to arbitrary transitivity scores, demonstrated by the example that $z_{uv} = 1$ for all $uv \in E^+$ satisfies both constraints (8) and (9), independently of the actual assignment of variables x_{uv}.

3.3 ILP for Multiple MS Using General Transitivity Scores

Here, we present an ILP for solving the Combined Colorful Subtree problem using general transitivity scores. Let $G = (V, E)$ be the input graph, and let $\mathcal{C} : V \to C$ denote the vertex coloring of G. For each color $c \in C$ let $V(c)$ be the set of all vertices in G which are colored with c. Let $H = (U, F)$ be the color version of G with

$$U := \mathcal{C}(V) \quad \text{and} \quad F := \{\mathcal{C}(u)\mathcal{C}(v) \,:\, uv \in E\}.$$

We may assume $U = C$, but for the sake of clarity we will use U whenever we refer to the vertices of the color graph H.

Let us define binary variables x_{uv} for each edge $uv \in E$, and z_{ab} and y_{ab} for each edge $ab \in F$. We assume $x_{uv} = 1$ if and only if uv is part of the subtree, and $y_{ab} = 1$ if there exist $u \in V(a)$ and $v \in V(b)$ such that uv is part of the subtree, that is, $x_{uv} = 1$. Variables y_{ab} are merely helper variables that map the subtree to the color space. Finally, we assume $z_{ab} = 1$ if ab is part of the closure of the subtree in color space. The following ILP captures the maximum colorful subtree problem as well as the Colorful Subtree Closure problem using arbitrary transitivity scores:

$$\max \sum_{uv \in E} w(u, v) \cdot x_{uv} + \sum_{ab \in F} w^+(a, b) \cdot z_{ab} \tag{12}$$

satisfying constraints (3), (4), (5) and, in addition:

$$x_{uv} \leq y_{C(u)C(v)} \qquad \text{for all } uv \in E, \tag{13}$$

$$y_{ab} \leq \sum_{u \in V(a), v \in V(b)} x_{uv} \qquad \text{for all } ab \in F, \tag{14}$$

$$y_{ab} \leq z_{ab} \qquad \text{for all } ab \in F, \tag{15}$$

$$z_{ab} + y_{bc} - 1 \leq z_{ac} \qquad \text{for all } bc \in F,\ a \in U, \tag{16}$$

$$z_{ab} - y_{bc} + 1 \geq z_{ac} \qquad \text{for all } bc \in F,\ a \in U, \tag{17}$$

$$z_{ac} \leq \sum_{\substack{b \in U \text{ with} \\ bc \in F}} y_{bc} \qquad \text{for all } ac \in F, \tag{18}$$

$$x_{uv} \in \{0, 1\} \qquad \text{for all } uv \in E, \tag{19}$$

$$y_{ab}, z_{ab} \in \{0, 1\} \qquad \text{for all } ab \in F. \tag{20}$$

Constraints (13) and (14) ensure that there is an edge in the color version of the tree if and only if there is an edge between vertices of the corresponding colors. Constraints (15) guarantee that for each edge that is part of the solution, also its transitive edge is part of the solution. Constraints (16) and (17) ensure the transitivity of the transitive closure of the solution: For a given edge y_{bc} in the color version of the tree and an arbitrary color a, a is either an ancestor of b (and thus also of c), or not. The first case implies that there must be transitive edges from a to b as well as from a to c. In the second case, transitive edges from a to b as well as from a to c are prohibited. Constraints (18) guarantee that only the transitive closure of the solution tree is part of the solution, and not the transitive closure of other subgraphs.

4 Correlation with Structural Similarity

Rasche *et al.* [18] presented the comparison of fragmentation trees using fragmentation tree alignments. One important application of this approach is searching in a database for molecules that are similar to the measured unknown molecule. Two structurally similar molecules have similar fragmentation trees and vice versa [18]. Hence, the similarity of high quality fragmentation trees correlates with the structural similarity of the corresponding molecules. We will use the correlation coefficient to optimize the parameters of the transitivity score and to evaluate the benefit of MS^n data compared to MS^2 data.

Fragmentation tree similarity is defined via edges, representing fragmentation steps, and vertices, representing fragments. A local fragmentation tree alignment contains those parts of the two trees where similar fragmentation cascades occurred [18]. To compute fragmentation tree alignments we use the sparse DP introduced by Hufsky *et al.* [12] which is very fast in practice.

For the comparison of molecular structures, many different similarity scores have been developed [13]. Molecular structures can be represented as binary fingerprints. Here, we use two of those fingerprint representations, that is the fingerprints from PubChem database [29] accessed via the Chemistry Development Toolkit version 1.3.37 [28] [1], and Molecular ACCess System (MACCS) fingerprints implemented in OpenBabel [2]. We use Tanimoto similarity scores (Jaccard indices) [21] to compare those binary vectors.

To assess the correlation between fragmentation tree similarity and structural similarity, we use the well-known Pearson correlation coefficient r which measures the linear dependence of two variables, as well as the Spearman's rank correlation coefficient ρ that is the Pearson correlation coefficient between the ranked variables. The coefficient of determination, r^2, measures how well a model explains and predicts future outcomes. Fragmentation tree alignment scores and structural similarity scores are two measures where one would not expect a linear dependence. This being said, we argue that any Pearson correlation coefficients $r > 0.5$ ($r^2 > 0.25$) can be regarded as strong correlation.

5 Results

To evaluate our work, we analyze spectra from a dataset introduced in [24]. It contains 185 mass spectra of 45 molecules, mainly representing plant secondary metabolites. All spectra were measured on a Thermo Scientific Orbitrap XL instrument using direct infusion. For more details of the dataset see [24].

For the construction of the fragmentation graph, we use a relative mass error of 20 ppm and the standard alphabet – that is carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur – to compute the fragment molecular formulas. For weighting the fragmentation graph, we use the scoring parameters from [19]. For scoring the transitive closure, we evaluate the influence of parameters σ_1, σ_2 and σ_3 on the quality of fragmentation trees. We assume the molecular formula of the unfragmented molecule to be given (for details, see [18, 19, 24]).

For the computation of fragmentation trees from tandem MS data, we use the DP algorithm from [5] (called DP-MS^2 in the following) and the ILP from [20] (ILP-MS^2). Recall, that we can convert MS^n data to "pseudo MS^2" data by mapping all peaks into a single spectrum and ignoring the additional information gained from the succession of fragments in MS^n. For the computation of fragmentation trees from multiple MS data, we use the DP algorithm from [24] (DP-MS^n) as well as our novel ILP (ILP-MS^n). Both DP algorithms are restricted by memory and time consumptions. Thus, exact calculations are limited to the k' most intense peaks. The remaining peaks are added in descending intensity order by a greedy heuristic (see the tree completion heuristic from [20, 24]). For solving the ILPs we use Gurobi 5.6[3]. The experiments were run on a cluster with four nodes each containing 2x Intel XEON 6 Cores E5645 at 2.40 GHz with 48 GB RAM. Each instance is started on a single core.

For computing fragmentation tree alignments, we use the sparse DP from [12] and the scoring from [18]. Estimation of Pearson and Spearman correlation coefficients was done using the programming language R.

Running Time Comparison. For the evaluation of running times depending on the number of peaks in the spectrum, we calculate the exact solution (using all four algorithms) for the k' most intense peaks for each molecule. Afterwards, remaining peaks are added heuristically. For each k', we exclude instances with less than k' peaks in the spectrum. For very small instances, the DP algorithms are slightly faster than the ILPs (see Fig. 2 (left)). On moderate large instances (e.g. $k' = 17$), the ILPs clearly outperform the DP algorithms. For $k' > 20$ it is not possible to calculate fragmentation trees

[1] https://sourceforge.net/projects/cdk/
[2] http://openbabel.sourceforge.net/
[3] Gurobi Optimizer 5.6. Houston, Texas: Gurobi Optimization, Inc.

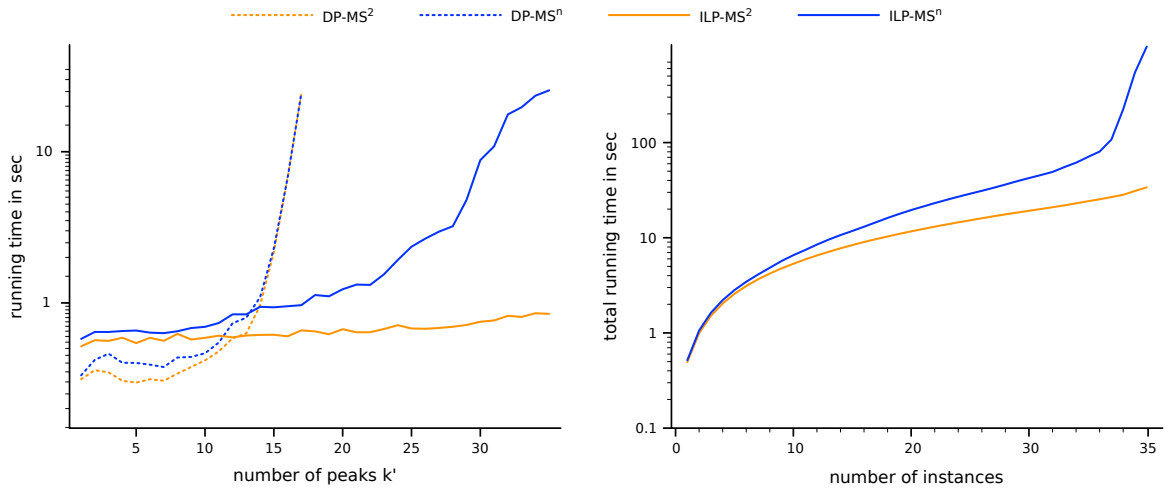

Fig. 2. Running times for calculating fragmentation trees. Times are averaged on 10 repetitive evaluations and given in seconds. Note the logarithmic y-axis. Left: Average running times for calculating one fragmentation tree with exact solution for the k' most intense peaks. The remaining peaks are attached by tree completion heuristic. Right: Total running times for instances of size $k' = 35$. Again, the remaining peaks are attached heuristically. We calculate the total running time of the x instances for which the tree was computed faster than for any of the remaining instances. For each algorithm, instances were sorted separately.

with the DP due to memory and time constraints. On huge instances ($k' > 30$) the ILP-MS^n is slower than the ILP-MS^2.

To get an overview of differences in the running times between hard and easy fragmentation tree computations for tandem MS and multiple MS data, we sort the instances by their running times in increasing order. This is done separately for the ILP-MS^2 and the ILP-MS^n algorithm (see Fig. 2 (right)). We find that solving the Combined Colorful Subtree problem using the ILP-MS^n is still very fast on most instances. Further, we find that for the ILP-MS^n, there is one molecule for which the calculation of the fragmentation tree takes nearly as much time as for the remaining 39 molecules together.

Parameter estimation. In [24] the estimation of parameters was based on the assumption that fragmentation trees change when using the additional scoring of the transitive closure. Here, we want to optimize the scoring of the transitive closure by maximizing the correlation of fragmentation tree alignment scores and the structural similarity scores of the corresponding molecules. For three of the 45 molecules, it was not possible to calculate fragmentation tree alignments due to memory and time constraints. Those compounds were excluded from the analysis.

For estimating the optimal scoring parameters σ_1, σ_2 and σ_3 of the transitive closure, we compute exact fragmentation trees using the $k' = 20$ most intense peaks and attach the remaining peaks by the tree completion heuristic. For scoring the transitive closure of the fragmentation graph, we separately vary $0 \leq \sigma_1 \leq 6$, $-3 \leq \sigma_2 \leq 0$ and $-3 \leq \sigma_3 \leq 0$. We compute fragmentation tree alignments and analyze the resulting PubChem/Tanimmoto as well as MACCS/Tanimoto Pearson correlation coefficients (see Fig. 3). Increasing σ_1 the correlation coefficient increases and converges at approximately $\sigma_1 = 3$. For σ_2 and σ_3 the highest correlation is reached around -0.5. For the further evaluation, we set $\sigma_1 = 3$, $\sigma_2 = -0.5$ and $\sigma_3 = -0.5$. We find, that this result agrees with the original scoring parameters from [24]. Although they were chosen ad hoc, they seem to work very well in practice. We further find, that σ_1 has a larger effect on the correlation than σ_2 and σ_3 (see Fig. 3). This was expected, as the requirement that a fragments is placed below its parent fragment is very strong.

Further, we evaluate the effect of using more peaks for the exact fragmentation tree computation on the correlation. We set $\sigma_1 = 3$, $\sigma_2 = -0.5$ and $\sigma_3 = -0.5$, and vary the number of peaks from $10 \leq k' \leq 35$. We find that the highest PubChem/Tanimoto correlation coefficient $r = 0.5643137$ ($r^2 = 0.31844500$) is achieved for $k' = 21$ (see Fig. 4).

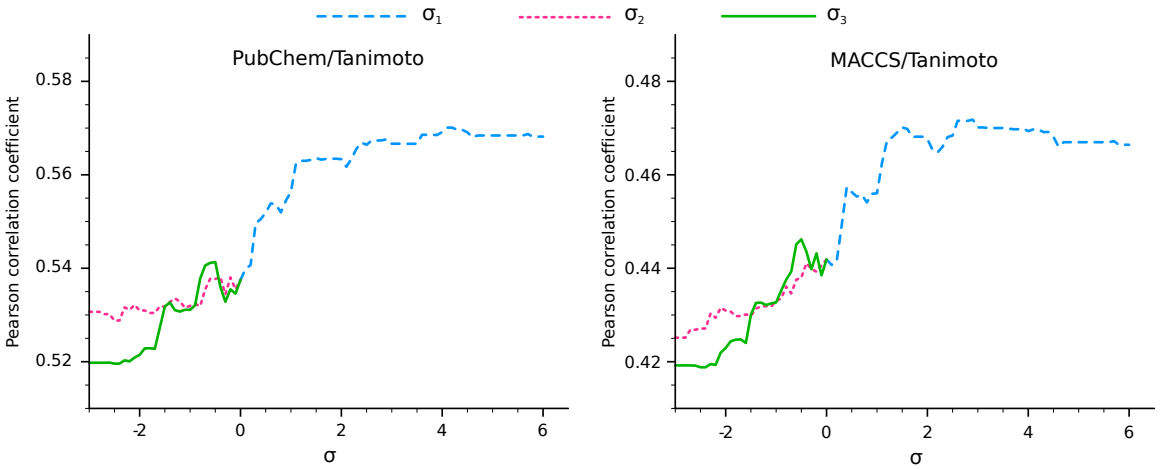

Fig. 3. Pearson correlation coefficients of PubChem/Tanimoto (left) and MACCS/Tanimoto (right) scores with fragmentation tree alignment scores, separately varying the scoring parameters σ_1, σ_2 and σ_3 of the transitive closure for fragmentation tree computation. When varying σ_1, we set $\sigma_2 = 0$ and $\sigma_3 = 0$ and vice versa.

Note that the DP-MS^n is not able to solve problems of size $k' = 21$ with acceptable running time and memory consumption. Thus, only by help of the ILP-MS^n it is possible to compute trees with best quality.

The optimum of k' remains relatively stable in a *leave-one-out* validation experiment: For each compound, we delete the corresponding fragmentation tree from the dataset and repeat the former analysis to determine the best k'. For 30 of the 42 sub-datasets $k' = 21$ achieves the best correlation. For the remaining 11 sub-datasets $k' = 14$, $k' = 20$ or $k' = 25$ are optimal.

Due to the small size of the dataset, it is hard to determine best parameters without overfitting. Hence, these analyzes should not be seen as perfect parameter estimation, but more as a rough estimation until a bigger dataset becomes available.

Comparison between trees from MS^2, Pseudo-MS^2 and MS^n data. To evaluate the benefit of scoring the additional information from MS^n data, we compare the correlation coefficients of using only the MS^2 spectra, using Pseudo-MS^2 data, and using MS^n data. As mentioned above, Pseudo-MS^2 data means mapping all peaks into a single spectrum and ignoring the additional information gained from the succession of fragments in MS^n, that is not scoring the transitive closure. For fragmentation tree computation from MS^2 and Pseudo-MS^2 data we use the ILP-MS^2, for MS^n data we use the ILP-MS^n. For a fair evaluation, we again vary the number of peaks from $10 \leq k' \leq 35$ to choose the k' with the highest correlation coefficient. The highest Pearson correlation coefficient with PubChem/Tanimoto fingerprints for MS^2 data is $r = 0.3860874$ ($r^2 = 0.1490635$) with $k' = 21$ and for Pseudo-MS^2 data $r = 0.5477199$ ($r^2 = 0.2999970$) with $k' = 25$ (see Fig. 4).

Further, we compare the Pearson correlation coefficients between the three datasets MS^2, Pseudo-MS^2 and MS^n (see Table 1). We find that the benefit of MS^n data is huge in comparison to using only MS^2 data, which is expected since the MS^2 spectra contain too few peaks. The question that is more intriguing is whether scoring the transitive closure improves correlation results. Comparing Pseudo-MS^2 with MS^n data, we get an increase in the coefficient of determination r^2 by up to 6.7 % for PubChem fingerprints and 6.3 % for MACCS fingerprints. The results for Spearman correlation coefficients look similar. When restricting the evaluation to large trees (at least three edges, five edges, seven edges), we cannot observe an increase in correlation.

When fragmentation trees are used in database search the relevant accuracy measure is not Pearson correlation, but identification accuracy. The dataset used in this paper is small and there is only one measurement per compound. Thus we cannot evaluate the identification accuracy. Instead we analyze the Tanimoto scores $T(h)$ of the first h hits with h ranging from one to the number of compounds (see Fig. 5). We exclude the identical compound from the hitlist and then average over the hitlists of all compounds in the dataset. We compare the results from MS^2, Pseudo-MS^2 and MS^n data with pseudo hitlists containing randomly ordered compounds (minimum value, RANDOM) and compounds

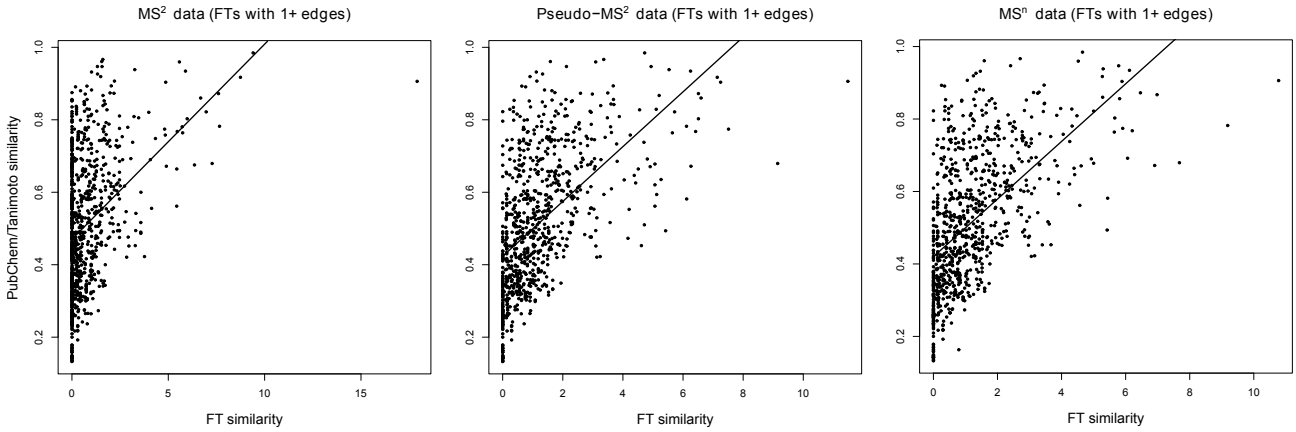

Fig. 4. Correlation and regression line for the complete datasets. Fragmentation tree similarity (x-axis) plotted against structural similarity measured by PubChem/Tanimoto score (y-axis). (a) Fragmentation trees for MS^2 data ($k' = 21$). Pearson correlation is $r = 0.386$. Spearman correlation is $\rho = 0.364$ (b) Fragmentation trees for Pseudo-MS^2 data ($k' = 25$). Pearson correlation is $r = 0.548$. Spearman correlation is $\rho = 0.615$ (c) Fragmentation trees for MS^n data ($k' = 21$). Pearson correlation is $r = 0.564$. Spearman correlation is $\rho = 0.624$.

Table 1. Pearson correlation r and coefficient of determination r^2 (in brackets) of structural similarity (PubChem/Tanimoto and MACCS/Tanimoto) with fragmentation tree similarity, for all three datasets and different minimum tree sizes (at least one edge, three edges, five edges, seven edges). We report the number of alignments (molecule pairs) N for each set. The subsets with different minimum tree sizes are determined by the tree sizes of the MS^n trees (that is, the MS^2 and Pseudo-MS^2 subsets contain the same molecules).

		only molecules with at least			
fingerprint	dataset	1 edge	3 edges	5 edges	7 edges
PubChem	MS^2	0.386 (0.149)	0.386 (0.149)	0.374 (0.140)	0.384 (0.147)
	Pseudo-MS^2	0.548 (0.300)	0.549 (0.301)	0.530 (0.281)	0.549 (0.301)
	MS^n	0.564 (0.318)	0.567 (0.321)	0.547 (0.299)	0.565 (0.319)
MACCS	MS^2	0.379 (0.143)	0.371 (0.138)	0.371 (0.138)	0.373 (0.139)
	Pseudo-MS^2	0.453 (0.206)	0.445 (0.198)	0.438 (0.192)	0.439 (0.193)
	MS^n	0.466 (0.217)	0.456 (0.210)	0.449 (0.202)	0.449 (0.201)
	no. molecule pairs N	861	820	630	561

arranged in descending order in accordance with the Tanimoto scores (upper limit, BEST). There is a significant increase of average Tanimoto scores from MS^2 data to MS^n data, and a slight increase from Pseudo-MS^2 data to MS^n data especially for the first $h = 5$ compounds.

6 Conclusion

In this work, we have presented an Integer Linear Program for the Combined Colorful Subtree problem, that outperforms the Dynamic Programming algorithm that has been presented before [24]. Solving this problem is relevant for calculating fragmentation trees from multistage mass spectrometry data.

Quality of fragmentation trees is measured by correlation of tree alignment scores with structural similarity scores of the corresponding compounds. Experiments on a dataset with 45 compounds revealed that trees computed with transitivity scores $\sigma_1 = 3$, $\sigma_2 = -0.5$ and $\sigma_3 = -0.5$ achieve the best quality. The highest correlation of $r = 0.564$ was achieved when computing exact fragmentation trees for the $k' = 21$ most intense peaks and attaching the remaining peaks heuristically. Using the additional information provided by multiple MS data, the coefficient of determination r^2 increases by up to 6.7 % compared to trees computed without transitivity scores. Thus, we could show for the first time that additional information from MS^n data can improve the quality of fragmentation trees.

For the computation of those trees with highest quality ($k' = 21$), our ILP needs 1.3 s on average. In contrast, the original DP is not able to solve those instances with acceptable running time and memory consumption. The ILP for MS^n is, however, slower than the ILP for MS^2 that has been presented

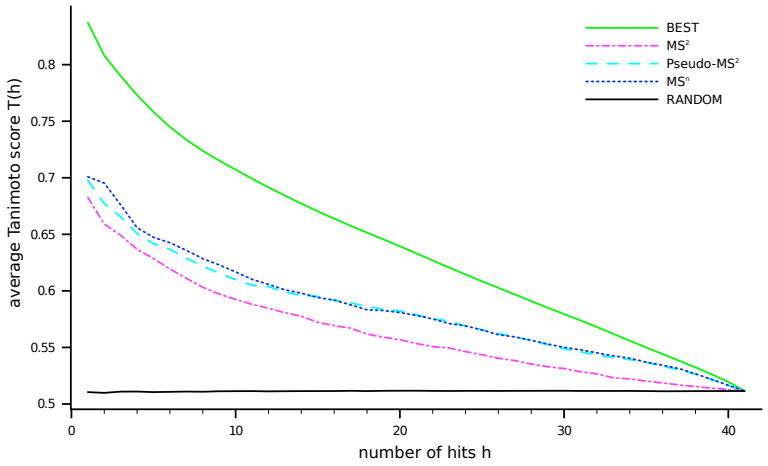

Fig. 5. Average Tanimoto scores $T(h)$ between query structures and the first h structures from hitlists obtained by FT alignments (MS^2, Pseudo-MS^2, MS^n data), pseudo hitlists containing the structures with maximum Tanimoto score to query structure (BEST) and randomly selected pseudo hitlists (RANDOM).

before [20]. This is due to the number of constraints which increases by an order of magnitude from MS^2 to MS^n. White *et al.* [30] suggested rules to speed up computations for the ILP on MS^2 data. These rules may also improve the running time of our algorithm.

Acknowledgements

We thank Aleš Svatoš and Ravi Kumar Maddula from the Max Planck Institute for Chemical Ecology in Jena, Germany for supplying us with the test data. We thank Kai Dührkop for helpful discussions on the ILP. F. Hufsky was funded by Deutsche Forschungsgemeinschaft, project "IDUN".

References

1. F. Allen, M. Wilson, A. Pon, R. Greiner, and D. Wishart. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res*, 2014.
2. S. Böcker, S. Briesemeister, and G. W. Klau. On optimal comparability editing with applications to molecular diagnostics. *BMC Bioinformatics*, 10(Suppl 1):S61, 2009. Proc. of *Asia-Pacific Bioinformatics Conference* (APBC 2009).
3. S. Böcker and Zs. Lipták. Efficient mass decomposition. In *Proc. of ACM Symposium on Applied Computing (ACM SAC 2005)*, pages 151–157. ACM press, New York, 2005.
4. S. Böcker and Zs. Lipták. A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, 48(4):413–432, 2007.
5. 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).
6. R. Dondi, G. Fertin, and S. Vialette. Complexity issues in vertex-colored graph pattern matching. *J Discrete Algorithms*, 9(1):82–99, 2011.
7. S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1(3):195–207, 1972.
8. M. R. Fellows, J. Gramm, and R. Niedermeier. On the parameterized intractability of motif search problems. *Combinatorica*, 26(2):141–167, 2006.
9. M. Gerlich and S. Neumann. MetFusion: integration of compound identification strategies. *J Mass Spectrom*, 48(3):291–298, 2013.
10. M. Heinonen, H. Shen, N. Zamboni, and J. Rousu. Metabolite identification and molecular fingerprint prediction via machine learning. *Bioinformatics*, 28(18):2333–2341, 2012. Proc. of *European Conference on Computational Biology* (ECCB 2012).
11. D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman, and D. F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal Chem*, 80(14):5574–5582, 2008.
12. 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).
13. A. R. Leach and V. J. Gillet. *An Introduction to Chemoinformatics*. Springer, Berlin, Dordrecht, The Netherlands, 2005.
14. J. W.-H. Li and J. C. Vederas. Drug discovery and natural products: End of an era or an endless frontier? *Science*, 325(5937):161–165, 2009.

15. I. Ljubić, R. Weiskircher, U. Pferschy, G. W. Klau, P. Mutzel, and M. Fischetti. Solving the prize-collecting Steiner tree problem to optimality. In *Proc. of Algorithm Engineering and Experiments (ALENEX 2005)*, pages 68–76. SIAM, 2005.
16. H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *J Mass Spectrom*, 44(4):485–493, 2009.
17. G. J. Patti, O. Yanes, and G. Siuzdak. Metabolomics: The apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, 2012.
18. 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.
19. F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher, and S. Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*, 83(4):1243–1251, 2011.
20. 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.
21. D. J. Rogers and T. T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.
22. M. Rojas-Chertó, P. T. Kasper, E. L. Willighagen, R. J. Vreeken, T. Hankemeier, and T. H. Reijmers. Elemental composition determination based on MS^n. *Bioinformatics*, 27:2376–2383, 2011.
23. K. Scheubert, F. Hufsky, F. Rasche, and S. Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. In *Proc. of Research in Computational Molecular Biology (RECOMB 2011)*, volume 6577 of *Lect Notes Comput Sci*, pages 377–391. Springer, Berlin, 2011.
24. 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.
25. M. T. Sheldon, R. Mistrik, and T. R. Croley. Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *J Am Soc Mass Spectrom*, 20(3):370–376, 2009.
26. H. Shen, K. Dührkop, S. Böcker, and J. Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 2014. Accepted, Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2014).
27. F. Sikora. *Aspects algorithmiques de la comparaison d'éléments biologiques*. PhD thesis, Université Paris-Est, 2011.
28. C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen. Recent developments of the Chemistry Development Kit (CDK) - an open-source Java library for chemo- and bioinformatics. *Curr Pharm Des*, 12(17):2111–2120, 2006.
29. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 37(Web Server issue):W623–W633, 2009.
30. W. T. J. White, S. Beyer, K. Dührkop, M. Chimani, and S. Böcker. Speedy colorful subtrees. Submitted to European Conference on Computational Biology (ECCB 2014), 2014.
31. 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.