Ab Initio Prediction of Molecular Fragments from Tandem Mass Spectrometry Data

Markus Heinonen[a,*] Ari Rantanen[a] Taneli Mielikäinen[a] Esa Pitkänen[a]
Juha Kokkonen[b] Juho Rousu[a]

[a] Department of Computer Science, University of Helsinki, Finland
[b] VTT, Technical Research Centre of Finland

Abstract: Mass spectrometry is one of the key enabling measurement technologies for systems biology, due to its ability to quantify molecules in small concentrations. Tandem mass spectrometers tackle the main shortcoming of mass spectrometry, the fact that molecules with an equal mass-to-charge ratio are not separated. In tandem mass spectrometer molecules can be fragmented and the intensities of these fragments measured as well. However, this creates a need for methods for identifying the generated fragments.

In this paper, we introduce a novel combinatorial approach for predicting the structure of molecular fragments that first enumerates all possible fragment candidates and then ranks them according the cost of cleaving a fragment from a molecule. Unlike many existing methods, our method does not rely on hand-coded fragmentation rule databases. Our method is able to predict the correct fragmentation of small-to-medium sized molecules with high accuracy.

1 Introduction

One of the enabling measurement technologies for the new era of systems biology is mass spectrometry (MS). Mass spectrometer measures the abundances of molecules with different masses in the sample with very high precision [MZSL98]. Mass spectrometry has an integral role in many biological analysis tasks, such as in protein identification [GV00, HZM00, Swe03]. In the study of metabolism mass spectrometry can be used to identify intracellular small molecules by comparing the intensity spectrum of unknown metabolite to a spectra residing in reference library [Fie02, MZSL98, SS94]

More information about an unknown metabolite can be obtained by applying *tandem mass spectrometer* (also known as *MS/MS*) techniques where metabolite molecules are collided with e.g. neutral gas to fragment the molecules and also the abundances of fragments are measured [dH96]. For example, the product ion spectrum produced by tandem MS can be used to improve the accuracy of library-based identification of unknown metabolites [Fie02, JS04] and to deduce structural information about them [KPH$^+$03, SP99,

*Author to whom correspondence should be directed. E-mail: markus.heinonen@cs.Helsinki.FI

vRLDZ$^+$04]. In addition, the elemental composition of a metabolite can be accurately inferred from product ion spectrum [ZGC$^+$05].

Tandem MS has also great potential in the area of ^{13}C metabolic flux analysis [RMR$^+$06, SCNV97, WMPdG01] where the velocities of metabolic reactions are estimated from the isotopomer distributions[1] of the metabolites. The isotopomer distribution of a metabolite can be accurately derived from tandem MS data [CN99, RRKK05]. Before the isotopomer distribution of a metabolite can be computed, the exact structures of molecular fragments produced by tandem MS have to be identified. The identification of fragments produced by tandem MS is also a problem of interest in e.g. structural elucidation [Swe03].

The manual identification of molecular fragments is a very time-consuming process even for an expert [McL80]. In this article we propose a novel method for the identification of molecular fragments produced by tandem MS from a known parent molecule. In the existing commercial tools Mass Frontier [Hig05] and MS Fragmenter [ACD05, Wil02] fragment identification is based on the fragmentation rules stored into a database. However, small changes in the structure of a molecule can result in significant differences in the fragmentation process [McL80]. Rule based systems will err if the fragmentation of a new molecule does not follow the rules found by studying other kinds of molecules. Deduction of fragmentation rules for each molecule and for each different MS technique is also a laborious task.

Our approach for tandem MS fragment identification is not based on a prior knowledge about common fragmentation rules but on the utilization of the combinatorial structure of the problem. Shortly, we first generate candidate fragments whose masses correspond to the observed peaks in a product ion spectrum and rank the candidate fragments according to the cost of cleaving a fragment from a molecule. Our experiments indicate that when molecules are reasonably small and the masses of molecular fragments can be measured with accuracy characteristic to modern high resolution MS devices, tandem MS fragments can be identified with good precision without a priori knowledge about common fragmentation mechanisms.

2 Fragment identification problem

Molecules can be modeled as undirected, connected, weighted and labeled graphs with the vertices being the atoms of the molecules and edges the bonds:

Definition 2.1 (Molecule)**.** A molecule M is an undirected, connected, weighted and labeled graph $\langle V, E, t_V, t_E, w_V, w_E \rangle$, where V is the set of vertices corresponding to the atoms and E is the set of undirected edges corresponding to the bonds between the atoms. The function $t_V : V \to A$ assigns each atom a type (e.g., carbon, hydrogen, etc.) and $t_E : E \to B$ assigns each bond a type (e.g., single, double, triple, aromatic, etc.). Vertices have atomic weights $w_V : V \to \mathbb{R}_+$ and edges have values $w_E : E \to \mathbb{R}_+$ assigning each

[1] By different isotopomers of a metabolite we mean molecules having specific combination of ^{12}C and ^{13}C atoms in different positions of the carbon chain of the metabolite. Isotopomer distribution of the metabolite then gives the relative concentrations of different isotopomers.

edge the strength of the corresponding bond.

The mass of the molecule is the sum of the weights of its atoms, i.e.,

$$w(M) = \sum_{v \in V} w_V(v). \tag{1}$$

We define a *fragment* F of M as a connected subgraph of M.[2]

The output of tandem MS is a spectrum where the locations of peaks correspond to observed weights $W \subset \mathbb{R}_+$ of molecular and fragment ions.[3] On a high level, the fragment identification problem of a molecule M can be formulated as follows:

Problem 2.2. Given a molecule M and a set $W \subset \mathbb{R}_+$ of observed weights of fragments of the molecule, find fragments $F_1, \ldots, F_{|W|}$ of M that most likely correspond to the weights in W.

Formally, a molecule M induces a fragmentation graph G_M containing all fragments of M (see Figure 1 for an example):

Definition 2.3 (Fragmentation graph)**.** A fragmentation graph G_M for a molecule M is a directed acyclic graph $\langle \mathcal{F}, \prec, c \rangle$ where

- $\mathcal{F}$ is the set of nodes corresponding to the fragments of the molecule M, i.e., the subsets of edges in M. That is, $\mathcal{F}$ is the collection of sets $E' \subseteq E$ such that E' forms a connected component in the molecule M;
- $\prec$ is the set of directed edges from each fragment $F \in \mathcal{F}$ to its subfragments $F' \in \mathcal{F}$. Hence, $\prec$ is binary relation over $\mathcal{F}$ such that $F \prec F' \iff F' \subset F$ for all $F, F' \in \mathcal{F}$;
- $c : \prec \rightarrow \mathbb{R}_+$ associates a cost to each edge in the graph giving the cost of producing the fragment F' from the fragment F for each $\langle F, F' \rangle \in \prec$ (i.e., for each $F' \subset F \subseteq E$ where F and F' form connected components).

We use several heuristic cost functions for producing the fragment F' from the fragment F. All functions are based on the assumption that, during the fragmentation process, weak bonds between the atoms of a molecule are more likely to be cleaved than the stronger ones. We approximate the strength of a bond with the standard covalent bond energy.

The simplest cost function for producing F' from F is the sum of energies of all cleaved bonds:

$$c(F, F') = \sum_{C_{F,F'}} w_E(e) \tag{2}$$

[2] Although not all fragments produced by tandem MS are necessary connected subgraphs, the assumption holds quite often. See Section 5 for further discussion.

[3] More precisely, MS separates molecular and fragment ions according to their mass-to-charge (m/z) ratio. However, when analyzing small molecules like metabolites, ions almost always get a single charge. In the following we assume that ions have a single charge.

where $C_{F,F'}$ consists of the bonds that must be cleaved to cut F' from F, i.e., $C_{F,F'} = \left\{e \in F : \left|e \cap \bigcup_{e' \in F'} e'\right| = 1\right\}$.

The total cost of a fragmentation graph G_M is the sum of the costs of its edges:

$$c(G_M) = \sum_{F \prec F'} c(F, F'). \tag{3}$$

With the notion of the fragmentation graph, the task of finding the best fragmentation for a molecule M and the weight set W can be formulated as follows:

Problem 2.4 (Fragment identification). Given a molecule M and a set $W \subset \mathbb{R}_+$ of weights, find a connected subgraph G^*_M of the fragmentation graph G_M such that G^*_M contains at least one fragment for each weight in W and the total cost $c(G^*_M)$ is minimized.

The actual form of the problem relies strongly on the cost function for the fragmentation graphs. We discuss different ways of defining the cost functions and fragmentation models in more detail in Section 3.

3 Models for the fragmentation process

The fragmentation of a molecule in tandem MS is a complex, stochastic and multistep process where ions are decomposed to smaller fragments. In general there exists many competing fragmentation pathways which a single molecule can take. The likelihood of the competing fragmentation pathways depends on many factors, including the amount of internal energy an ion obtains during the fragmentation, the stability of a product ion, steric requirements of fragmentation pathways and charge or radical sites of parent ion [McL80]. The accurate modeling of all these factors is very tedious [RHO00, SHS01] and is not done in practice when fragments are identified in every day laboratory work.

Next we give two alternative models for fragmentation and define the cost $c(G'_M)$ for a connected subgraph $G'_M \subseteq G_M$ of molecule M according to these models.

3.1 Single step fragmentation

Our primary model for fragmentation is based on the consensus that in tandem MS usually weak bonds are cleaved [MFH$^+$99] and that with low collision energies fragments are usually cleaved directly from the parent molecule [dH96]. Thus we can best explain the detected fragment peaks by fragments that can be cleaved from a parent molecule using the smallest amount of energy possible. With the notion of fragmentation graph, *single step fragmentation model* leads to a star-shaped graph, where each fragment originates directly from the original molecular ion in a single reaction. (See Figure 1.)

Unfortunately, even finding one weight-w minimum cost fragment F of a molecule M

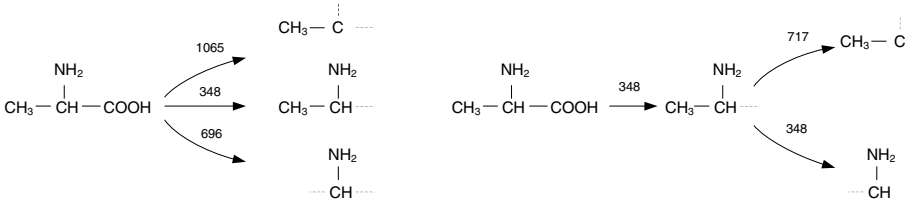

Figure 1: Example of fragmentation graphs of Alanine using single step (left) and multistep fragmentation model (right). Both graphs have four fragment nodes. Fragmentation is indicated by arrows with accompanying weights corresponding to the fragmentation graph edge weights, i.e. sum of cleaved bonds energies. For example, on the left the alanine is fragmented into CH3N (bottom arrow), by two cleavages: the COOH-group with C-C cleavage and the CH3-group with C-C cleavage. Both have energetic value of 348 kJ/mol thus making the total cost of producing CH3N 696 kJ/mol. Dashed arrows indicate cleaved bonds.

for certain weight $w \in \mathbb{R}_+$ is NP-hard. We show that by a polynomial-time computable reduction from the 3-satisfiability problem that is known to be NP-complete [GJ79]:

Problem 3.1 (3-satisfiability). Given a set $U = \{x_1, \ldots, x_n\}$ of boolean variables x_i and a collection $C = \{c_1, \ldots, c_m\}$ of clauses $c_i, |c_i| = 3$, over U, decide whether of not there is a a satisfying truth value assignment for C.

Theorem 3.2. *Given the molecule M, a weight $w \in \mathbb{R}_+$ and a cost c, it is NP-complete to decide whether or not there is a fragment F of M of the weight w with the cost $c(M, F)$ being at most c, where the cost is defined by Equation 2.*

Proof. The problem is clearly in NP, since any subgraph of M is at most as large as M itself and the weight (as defined by Equation 2) of any fragment F of M can be computed in time linear in F.

We reduce the instances $\langle U, C\rangle$ of 3-satisfiability to the instances $\langle M, w, c\rangle$ of finding a fragment of M of weight w and cost at most c.

The set V consists of vertices $c_{i,1}$, $c_{i,2}$, and $c_{i,3}$ for each clause $c_i \in C$, and dummy vertices $d_1, \ldots, d_\delta$. (δ is a constant that shall be determined later.) The vertex $c_{i,k}$ corresponds to setting the truth value of the boolean variable of the kth literal of the clause c_i in such a way that the literal is satisfied.

The weights of the atoms in the molecule are determined as follows. Let $p_1, \ldots, p_{n+1}$ be distinct primes. $w_V(c_{i,k}) = \log p_i$ for each $c_{i,k} \in V$, $w_V(d_j) = \log p_{n+1}$ for each $j = 1, \ldots, n$, and $w = \sum_{i=1}^{n} w_V(c_{i,1})$. Hence, any fragment consisting one vertex for each clause c_i.

Now we need to define the set E of edges and their weights appropriately. There is an edge $\{c_{i,k}, c_{j,l}\}$ in E if and only if kth literal in the clause c_i is not the negation of the lth literal of the clause c_j. Let δ be maximum degree of a vertex in the subgraph induced by the vertices $v_{i,k}$. Each vertex $c_{i,k}$ is connected to so many dummy vertices that the degree of $c_{i,k}$ is δ. The weights of the edges are all one.

The clauses in C are satisfiable if and only if there is a fragment F of weight w with cost $c(M, F) = n(\delta - n + 1)$. To see that, notice that fragment F is a clique if and only if the vertices in F determine a (partial) satisfying truth value assignment for C. □

Fortunately, in practice all fragments of the molecule M can be often generated and computing the cost $c(M, F)$ for a given F is easy. Thus, by generating all fragments (with weights in W) we can solve the problem. This observation leads to a conceptually simple algorithm where for each observed weight $w \in W$ a fragment F of weight w that minimizes $c(M, F)$ is found. The algorithm has three steps for each weight $w_i \in W$:

1. Find a set $\mathcal{F}_i$ of all connected subgraphs of M that have a weight w_i.
2. For each fragment $F \in \mathcal{F}_i$, compute a cost $c(M, F)$ of cleaving F from M.
3. For each $\mathcal{F}_i$, return $F_i \in \mathcal{F}_i$ with the smallest cost among the fragments in $\mathcal{F}_i$.

We find sets $\mathcal{F}_i$ of all fragments of weight w_i by enumerating all fragments, that is, all connected subgraphs induced by M with a depth-first traversal algorithm briefly mentioned in [BV97] and elaborated in [RR00]. The algorithm can easily be modified to give k_i least expensive fragments for each observed weight or all fragments with minimum cost w_i.

In our experiments, the cost $c(M, F)$ was based on five key figures derived from the bonds of M that have to be cleaved to form F from M, that is, bonds that connects elements in F to elements in $M \setminus F$. The key figures are: (1) the number of cleaved bonds, (2) the sum of strengths of cleaved bonds, (3) the strength of strongest cleaved bond, (4) the average strength of cleaved bonds and (5) the difference of strength between strongest intact bond versus the weakest cleaved bond in our candidate fragment. We defined $c(M, F)$ to be an average rank of F according to these key figures among the fragments of same weight.

3.2 Multistep fragmentation

As an alternative to single step fragmentation model, we experimented with a model where we assume that many fragmentation pathways consist of two or more consecutive reactions. Consecutive fragmentation reactions are thought to be common when higher collision energies are applied [dH96]. In this *multistep fragmentation model* we also assume that in intermediate reaction steps of a fragmentation pathway usually not all molecular fragments are further cleaved but some proportion of them is observed as a peak in tandem MS spectrum. These assumptions allow us to construct a model where pathways of consecutive reaction steps that (1) explain observed fragment peaks by intermediates of the pathway and (2) that cleave only weak bonds, are favored. This approach can be thought to mimic the decision process an expert goes through while identifying fragments manually: a proposed fragmentation pathway is more likely correct if peaks matching to intermediate steps of the pathway are present in the spectrum [SHS01].

Multistep fragmentation process can be computationally modeled by allowing fragmentation graphs where fragments are cleaved from other fragments and defining the cost of a

fragmentation subgraph $G'_M = \langle \mathcal{F}', \prec', c \rangle$ to be the sum of the costs of edges in G'_M, i.e.,

$$c(G'_M) = \sum_{e \in \prec'} c(e). \tag{4}$$

We use the sum of all cleaved bonds energies (see Equation 2) as the cost of an edge.

In the multistep fragmentation model the cost of fragment F depends on the other fragments in the fragmentation subgraph while in the single step fragmentation model, where fragments are always cleaved directly from the parent molecule, the cost of F depended only of its own structure. Thus instead of ranking the fragment of observed weight by comparing it to the other fragments of equal weight, we search for the optimal fragmentation subgraph G^*_M that minimizes the cost given in Equation 4.

Proposition 3.1. *The minimum cost connected subgraph G^*_M of the fragmentation graph G_M of a molecule M is a tree with at most $|W|$ leaves, where the cost of G^*_M is defined by Equation 4.*

Proof. Let G^*_M be the minimum cost connected subgraph of G_M. To see that G^*_M is necessarily tree, assume that G^*_M is not a tree.

If G^*_M is not a tree, then there must be a cycle C in G^*_M. However, then also the graph $G^*_M \setminus \{e\}, e \in C$, is connected. As the costs of the edges in G^*_M are strictly positive, the cost of $G^*_M \setminus \{e\}$ strictly smaller than the cost of G^*_M. Thus, if G^*_M is not a tree, then it is not the minimum cost connected subgraph of G_M.

The number of leaves can be at most W, since each leaf corresponds to some weight in W. □

An optimal fragmentation subgraph G^*_M can be found from the fragmentation graph G_M with mixed integer linear programming (MILP) by formulating the problem as a mixed integer linear program. (There exist well-developed techniques for solving MILP reasonably fast in practice [Mar01].)

The MILP formulation of the problem is as follows. We partition the fragments whose weight correspond to observed weights into sets $L_1, \ldots, L_{|W|}$ according to their weights. We denote by $\mathcal{L}$ a collection of sets L_k. Let f_i be a binary variable indicating whether a fragment $F_i \in L_k$ is chosen to be a fragment corresponding an observed weight w_k. We set $f_M = 1$ for the whole molecule. Let binary variable $p_{i,j}$ indicate whether an edge from F_i to F_j in G_M is chosen to G^*_M and $c_{i,j} \in \mathbb{R}$ the cost of $F_i \prec F_j$. The function to be minimized corresponds to the total cost of edges of G_M that are selected to G^*_M (see Equation 4).

We then obtain the following integer linear program:

$$\begin{aligned}
\min & \sum_{F_i \prec F_j} c_{i,j} p_{i,j} & \\
\text{s.t.} & \sum_{f_i \in L_k} f_i = 1 & \forall L_k \in \mathcal{L} \\
& f_j - \sum_{F_i \prec F_j} p_{i,j} = 0 & \forall F_j \in \mathcal{F} \\
& p_{i,j} - f_i \leq 0 & \forall F_i \prec F_j \in G_M
\end{aligned}$$

The first constraint of the above program states that exactly one fragment from each observed weight needs to be selected. The second constraint states that for each selected fragment F_j exactly one parent fragment F_i, from which F_j is cleaved, have to be selected. The third constraint states that if $F_i \prec F_j$ is selected to G_M^*, also F_j have to be in G_M^*. The solution to the above program is a minimal cost set G_M^* of pathways which form a connected tree in the fragmentation graph and cover each weight class of fragments with exactly one fragment. Note that either all f_i's or $p_{i,j}$'s can be relaxed to be real-valued (in the interval $[0, 1]$) in order to speed up the optimization. We relax $p_{i,j}$ as the number of $p_{i,j}$'s is quadratic to the number of f_i's in the worst case.

In practice the mixed integer linear programs tend to be very large. A major optimization for the model is to notice the specialty of hydrogen atoms in the fragments. As hydrogens connect to at most one other element, their removal from the model do not split a molecule or fragment to two fragments. Thus hydrogens do not need to be included when all fragments are enumerated. By using hydrogen-suppressed fragments, the amount of fragments drops drastically.

To cover the loss of hydrogen specificity in fragments, we add variables and constraints to integer linear program requiring that the correct number of hydrogens is cleaved from each selected fragment and that the cleaved hydrogen of parent fragment in G_M^* stays cleaved in its daughter fragments. Also, the objective function is modified such that the costs of hydrogen cleavages are correctly accounted for.

Let $h_{n,j}$ be a binary variable indicating whether a hydrogen n directly connected to fragment F_j is cleaved. Let $\mathcal{H}$ be the set of all hydrogens in M and $|H_i|$ the (precomputed) number of hydrogens connected to F_i that should be cleaved in order to obtain F_i.

We add to MILP a constraint to ensure that the correct amount of hydrogens will be chosen for the fragment:

$$\sum_{n \in \mathcal{H}} h_{n,i} - |H_i| \, f_i = 0 \qquad \forall F_i \in \mathcal{F}.$$

We also add a constraint ensuring that a hydrogen cleaved in F_i is cleaved in all F_j's that have selected to be its children in the solution:

$$p_{i,j} + h_{n,i} - h_{n,j} \leq 1 \qquad \forall F_i \prec F_j, \forall n \in \mathcal{H}.$$

Finally, the cost of the solution is modified to take the costs of cleaved hydrogens into account:

$$\min \sum_{F_i \prec F_j} c_{i,j} p_{i,j} + \sum_{h \in \mathcal{H}} \sum_{F_i \prec F_j} c_h (h_{n,j} - h_{n,i}).$$

Again, the variables $p_{i,j}$ can be relaxed to be in $[0, 1]$.

4 Experiments

We tested our method of identifying tandem MS fragments with 20 amino acids and 7 sugar phosphates. Molecular masses ranged from 75 Da to 340 Da, 160 Da being the average. In particular, the most massive molecule Fructose-1,6-bisphosphate had 34 atoms and 34 bonds. Out of the 27 molecules, 8 were cyclic. The number of connected subgraphs of the molecules varied from hundreds to millions, depending on the cyclicity and size of the molecules. The run times of the above algorithms for candidate fragment enumeration and ranking varied accordingly from seconds to days.

Compounds were fragmented with the collision-induced dissociation (CID) method by using a Micromass Quattro II triple quadrupole MS equipped with an electrospray ionization interface. The spectra of compound were measured in a positive ionization mode. The collision gas for CID fragmentation was argon and collision energies varied between $10 - 50$ eV. The number of peaks in the product ion spectra of the molecules varied from one to 15, average being 7.1 peaks/molecule. Domain experts first manually identified the fragmentation pathways for each of the 27 molecules and the weights of the manually identified fragments were calculated with high precision for comparison of the effect of measurement accuracy to fragment identification. We then predicted the fragments with both of our models and compared the results against the manually identified fragments. A predicted fragment was deemed correct if its chemical formula and carbon backbone matched the manually identified one as this level of accuracy is sufficient for applications such as ^{13}C metabolic flux analysis. We used the off-the-shelf MILP solver `lp_solve` [BEN05] to solve the MILPs introduced in Section 3.2.

Our methods for identifying fragments agreed well with the domain experts when atom weights of peaks were assumed to be measurable at 0.01 Da (mass) accuracy. This is a realistic assumption in the current high resolution mass spectrometers and in our dataset. In high accuracy there were 6.5 fragments for each peak in fragment spectra, on average ($\sigma = 9.8$). If the fragments corresponding to observed peaks were selected randomly from the sets of fragments with the lowest cost suggested by the single step fragmentation method (Section 3.1), the fragmentations of the metabolites would be 88.7% correct, on average. If the best fragment among the fragments with the lowest cost was selected for each peak, metabolites would get 90.8% of correct fragments, on the average. On average, there were 1.4 fragments with the equal lowest cost per peak ($\sigma = 0.9$).

With the multistep fragmentation method (Section 3.2) fragmentation subgraphs with the lowest cost consisted of 82.8% correct fragments, on the average. The fragmentation subgraph in best agreement with manual identification among the subgraphs whose cost

was among the top-3 costs consisted 93.8% of correct fragments, on average. (There were 17.0 subgraphs in top-3 cost classes.)

In comparison, randomly constructed fragmentation subgraph of fragments whose weight match with observed peaks would have 36.8% ($\sigma = 36.3$) of correct fragments, on average.

If we assume that the mass spectrometer can separate compounds only at integer accuracy, the number of fragments with the same weight is considerably larger, namely 19.3 versus 6.5 fragments/peak on the average. This makes combinatorial identification of fragments much harder. With integer accuracy and single step model the fragmentations of the metabolites would be 66.4% correct on average, if the fragments corresponding peaks were selected randomly from the sets of fragments with lowest cost. Again, there were for each observed peak 1.4 fragments that had the lowest cost, on average. With multistep model the fragmentation subgraphs with the lowest cost yield an average accuracy of 55.9% and with the best subgraph among the subgraphs with top three lowest cost an average accuracy of 70.7%. (There were 25.7 subgraphs in the three lower cost classes on average.) Randomly constructed fragmentation subgraph of fragments that have an observed weight, has an average accuracy of 12.3% ($\sigma = 9.9$).

Figure 2 and Table 1 summarize the results of the experiments. Table 1 shows the prediction accuracies of fragmentation subgraphs with the lowest costs. In Figure 2, prediction accuracies of fragmentation subgraphs that had the cost among k lowest costs are shown. For example, with high mass accuracy and the single step model and examining the best fragmentation subgraphs with the cost in $k = 3$ lowest cost classes for each peak, 94.6% of predicted fragments match the manually identified ones. The reported accuracies are averages over 27 metabolites.

As a conclusion, most of the molecules can be resolved without difficulties and near 90% prediction rates are achieved, when high resolution MS is available. With our dataset the single step fragmentation model gives more accurate prediction than the multistep model.

Table 1: Single step and multistep model accuracies with integer and high mass (0.01 Da) accuracy. The best, the worst and the average accuracies of the fragmentation subgraphs that had the lowest cost according to single step or multistep models are shown. Reported accuracies are averages over 27 metabolites.

Scheme	Best	Average	Worst	σ_B	σ_A	σ_W
Single step, integer	68.2%	66.4%	64.5%	19.2%	21.2%	23.9%
Single step, high	90.8%	88.7%	86.3%	11.3%	12.0%	14.3%
Multistep, integer	62.0%	55.9%	51.1%	22.4%	23.5%	26.1%
Multistep, high	87.0%	82.8%	78.0%	20.4%	21.6%	24.8%

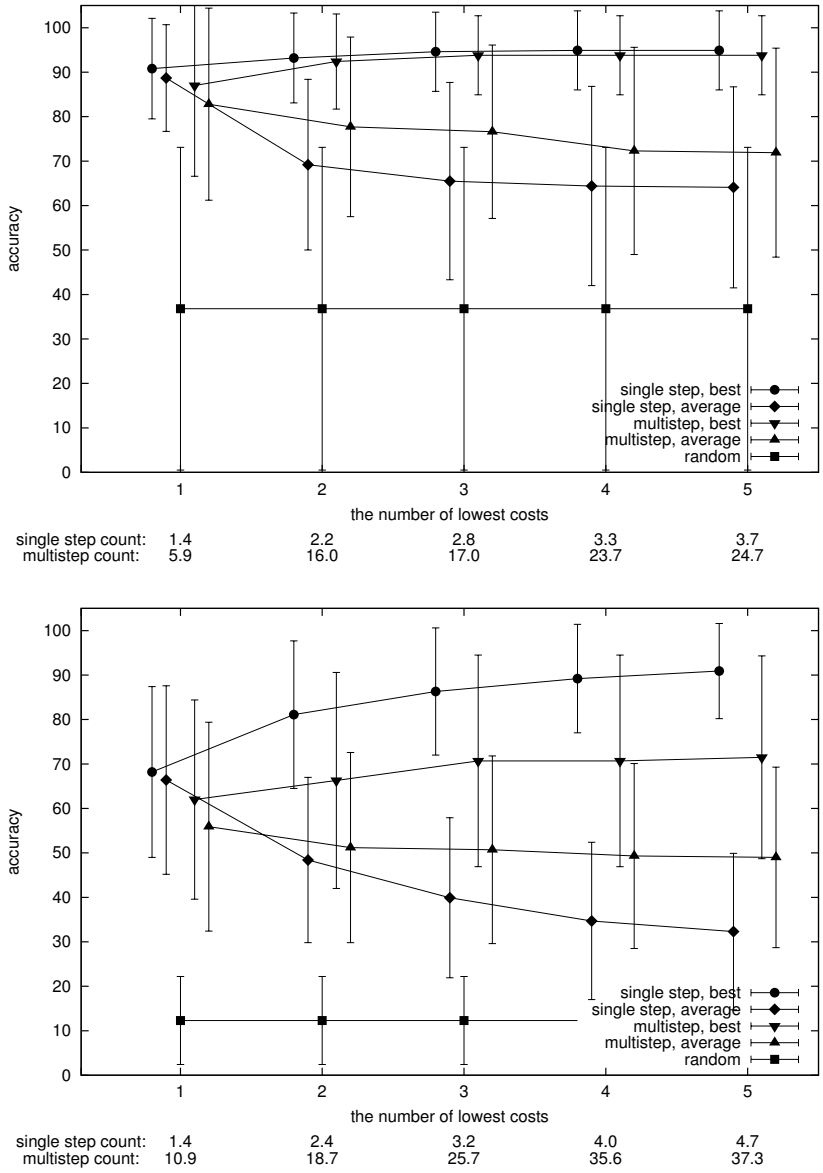

Figure 2: Figures depict the accuracy of single step and multistep fragmentation models when fragmentation subgraphs whose cost was among k lowest costs (x-axis) were taken into account. On top, fragment weights are assumed to be measurable with 0.01 Da accuracy, on bottom with integer accuracy. *Single step count* and *multistep count* below x-axes show the cumulative number of fragments per peak (single step) and the cumulative number of fragmentation subgraphs (multistep) with the cost among k lowest costs. The reported accuracies are averages over 27 metabolites. The lines connecting the points are only to improve the readability of the figures.

5 Discussion

The fragmentation of a molecule in mass spectrometer is a complex process which is not fully understood. We have shown that a combinatorial approach gives good results when the molecules analyzed are sufficiently small and the resolution of the mass spectrometer is characteristic to modern mass spectrometers. The combinatorial method given above automatically generates good hypotheses of the fragmentation patterns, thus aiding an experimentalist to evaluate all relevant possibilities of the fragmentation. Furthermore, our approach does not make assumptions on the MS technique used and is thus potentially applicable to a wide variety of problems.

The number of connected subgraphs of a molecule graph easily explodes when the size of the graph grows, even if hydrogen atoms are disregarded. Thus, the applicability of the combinatorial method is limited to small or medium–sized molecules. The number of connected subgraphs depends heavily on the cyclicity of the graph. As a rule of thumb, the method requires that the size of the molecule does not exceed 50 atoms, excluding hydrogens. Thus the method is suitable for many metabolites, but unsuitable for proteins. Additionally, as a result of element rearrangements, that is, by formation of new bonds during the fragmentation [MZSL98], not all fragments are necessarily connected subgraphs of the parent molecule. Fortunately, the most common example of such bond formation is hydrogen rearrangement. Again, hydrogen rearrangements can be handled as special cases as hydrogen atoms can only be transferred from one position to another, not creating cycles. For more complex rearrangements involving cyclizations, our software implementation of the above methods provides the user a tool to manually add bonds that are formed during the fragmentation to the molecule. Comparing our method against the commercial rule based systems proved problematic. To the authors knowledge, no public data on the performance or accuracy of existing tools is available.

Taking advantage of fragment intensities provides an interesting direction for further development of our combinatorial fragment identification method. In addition, we are investigating the possibility of combining the combinatorial approach with stochastic modeling to improve the accuracy of identification. Also combining the local ranking heuristics in a more advanced way than computing the average rankings is a promising direction [FISS03, FKM+04]. The software implementing the methods described in this paper is available from the authors and from a web site `http://www.cs.helsinki.fi/group/sysfys/software/fragid/`.

Acknowledgments. This work was supported by grant 203668 from the Academy of Finland (SYSBIO program) and by European Commission IST programme FET arm, contract no. FP6-508861 (APrIL II).

References

[ACD05] ACD/Labs. ACD/MS Fragmenter. `http://www.acdlabs.com`, 2005.

[BEN05] Michel Berkelaar, Kjell Eikland, and Peter Notebaert. lp_solve: Open source (Mixed-Integer) Linear Programming system., 2005. Multi-platform, pure ANSI C / POSIX source code, Lex/Yacc based parsing. Version 5.5.0.0 dated 17 may 2005. GNU LGPL (Lesser General Public Licence). http://groups.yahoo.com/group/lp_solve/.

[BV97] Richard G. A. Bone and Hugo O. Villar. Exhaustive Enumeration of Molecular Substructures. *Journal of Computational Chemistry*, 18(1):86–107, 1997.

[CN99] Bjarke Christensen and Jens Nielsen. Isotopomer analysis using GC-MS. *Metabolic Engineering*, 1:E6–E16, 1999.

[dH96] Edmond de Hoffmann. Tandem Mass Spectrometry: a Primer. *Journal of Mass Spectrometry*, 31:129–137, 1996.

[Fie02] Oliver Fiehn. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, 48:155–171, 2002.

[FISS03] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[FKM+04] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. Comparing and Aggregating Rankings with Ties. In Alin Deutsch, editor, *Proceedings of the Twenty-third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 14-16, 2004, Paris, France*, pages 47–58, 2004.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, 1979.

[GV00] Kris Gevaert and Joël Vandekerckhove. Protein identification methods in proteomics. *Electrophoresis*, 21:1145–1154, 2000.

[Hig05] HighChem. HighChem Mass Frontier 4.0. `http://www.highchem.com`, 2005.

[HZM00] David Horn, Roman Zubarev, and Fred McLafferty. Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proceeding of the National Academy of Sciences*, 97(19):10313–10317, 2000.

[JS04] Jonathan Josephs and Mark Sanders. Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers. *Rapid Communications in Mass Spectrometry*, 18:743–759, 2004.

[KPH+03] Katerina Klagkou, Frank Pullen, Mark Harrison, Andy Organ, Alistair Firth, and John Langley. Approaches towards the automated interpretation and prediction of electrospray tandem mass spectra of non-peptidic combinatorial compounds. *Rapid Communications in Mass Spectrometry*, 17:1163–1168, 2003.

[Mar01] Alexander Martin. General Mixed Integer Programming: Computational Issues for Branch-and-Cut Algorithms. In Michael Jünger and Denis Naddef, editors, *Computational Combinatorial Optimization: Optimal and Provably Near-Optimal Solutions*, volume 2241 of *Lecture Notes in Computer Science*, pages 1–25. Springer, 2001.

[McL80] Fred McLafferty. *Interpretation of Mass Spectra*. University Science Books, 3rd edition, 1980.

[MFH+99] Fred McLafferty, Einar Fridriksson, David Horn, Mark Lewis, and Roman Zubarev. Biomolecule Mass Spectrometry. *Science*, 284(5418):1289–1290, 1999.

[MZSL98] Fred McLafferty, Mei-Yi Zhang, Douglas Stauffer, and Stanton Loh. Comparison of Algorithms and Databases for Matching Unknown Mass Spectra. *American Society for Mass Spectrometry*, 9:92–95, 1998.

[RHO00] Françoise Rogalewicz, Yannik Hoppilard, and Gilles Ohanessian. Fragmentation mechanisms of α-amino acids protonated under electrospray ionization: a collision activation and ab initio theoretical study. *International Journal of Mass Spectrometry*, 195/196:565–590, 2000.

[RMR+06] Ari Rantanen, Taneli Mielikäinen, Juho Rousu, Hannu Maaheimo, and Esko Ukkonen. Planning optimal measurements of isotopomer distributions for estimation of metabolic fluxes. *Bioinformatics*, 22(10):1198–1206, 2006.

[RR00] Gerta Rücker and Christoph Rücker. Automatic Enumeration of All Connected Subgraphs. *MATCH Communications in Mathematical and Computer Chemistry*, 41:145–149, 2000.

[RRKK05] Juho Rousu, Ari Rantanen, Raimo Ketola, and Juha Kokkonen. Isotopomer distribution computation from tandem mass spectrometric data with overlapping fragment spectra. *Spectroscopy*, 19:53–67, 2005.

[SCNV97] Karsten Schmidt, Morten Carlsen, Jens Nielsen, and John Villadsen. Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and Bioengineering*, 55:831–840, 1997.

[SHS01] Tamer Shoeib, Alan Hopkinson, and Michael Siu. Collision-Induced Dissociation of the AG^+ –Proline Complex: Fragmentation Pathways and Reaction Mechanisms – A Synergy between Experiment and Theory. *The Journal of Physical Chemistry B*, 105:12399–12409, 2001.

[SP99] Bernhard Seebass and Ernö Pretsch. Automated Compatibility Tests of the Molecular Formulas or Structures of Organic Compounds with Their Mass Spectra. *Journal of Chemical Information and Computer Sciences*, 39:713–717, 1999.

[SS94] S.E. Stein and D. Scott. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *Journal of American Society of Mass Spectrometry*, 5:859–866, 1994.

[Swe03] Daniel Sweeney. Small molecules as Mathematical Partitions. *Analytical Chemistry*, 75:5362–5373, 2003.

[vRLDZ+04] Edda von Roepenack-Lahaye, Thomas Degenkolb, Michael Zerjeski, Mathias Franz, Udo Roth, Ludger Wessjohann, Jürgen Schmidt, Dierk Scheel, and Stephan Clemens. Profiling of Arabidopsis Secondary Metabolites by Capillary Liquid Chromatography Coupled to Electrospray Ionization Quadrupole Time-of-Flight Mass Spectrometry. *Plant Physiology*, 134:548–557, 2004.

[Wil02] Antony Williams. Applications of Computer Software for the Interpretation and Management of Mass Spectrometry Data in Pharmaceutical Science. *Current Topics in Medicinal Chemistry*, 2:99–107, 2002.

[WMPdG01] Wolfgang Wiechert, Michael Möllney, Sören Petersen, and Albert de Graaf. A Universal Framework for ^{13}C Metabolic Flux Analysis. *Metabolic Engineering*, 3:265–283, 2001.

[ZGC+05] Jingfen Zhang, Wen Gao, Jinjin Cai, Simin He, Rong Zeng, and Runsheng Chen. Predicting Molecular Formulas of Fragment Ions with Isotope Patterns in Tandem Mass Spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:217–230, 2005.