Towards de novo identification of metabolites by analyzing tandem mass spectra

Sebastian Böcker, Florian Rasche

Chair for Bioinformatics, Friedrich-Schiller-University Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
{boecker, florian.rasche}@minet.uni-jena.de

Abstract. Mass spectrometry is among the most widely used technologies in proteomics and metabolomics. 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. In this work, we analyze 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. This is achieved by calculating the possible formulas of the fragment peaks and then reconstructing the most likely fragmentation tree from this information. We present tests on real mass spectra showing that our algorithms solve the reconstruction 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.

Keywords. Tandem mass spectrometry, metabolomics, de novo interpretation

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 [1]. 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. In contrast, most of the metabolites not directly involved in the aforementioned functions remain unknown. These *secondary metabolites* are especially abundant in plants.

In this work, we evaluate a method for the automated de novo identification of metabolites from quadrupole time-of-flight tandem mass spectra recently presented in [2]. Mass accuracy of these instruments is approximately 20 ppm. The

Dagstuhl Seminar Proceedings 08101
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2008/1783

metabolite is fragmented using collision-induced dissociation (CID) [3], 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, so our method *does not* use isotopic pattern to identify the molecular formula. Such information can be easily integrated into the method and will further increase its identification accuracy.

We have developed a model for the fragmentation process resulting in a graph theoretical problem called Maximum Colorful Subtree problem [2]. Unfortunately, we can show that this problem is NP-hard. Despite this negative result, we developed several exact and heuristic algorithms for its solution. One of these exact algorithms is fixed-parameter tractable (FPT) [4]. The FPT algorithm and the heuristics show good performance in practice both with respect to identification accuracy and running times, as our 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 minutes to process all mass spectra.

2 Empirical Results

We compare four exact algorithms and two heuristics presented in [2]. First, we tried a *branch and bound* approach, which can solve the problem exactly, but is very slow. The *brute force* algorithm splits the problem into many MST problems, what works well for small instances. The dynamic programming approach, which is *fixed parameter tractable*, worked best for large instances. But it generates too much overhead for small instances, therefore we tested a *combined algorithm* using brute-force for small molecules and dynamic programming for the larger ones. Two *heuristics* are also evaluated. They differ in the order in which they select fragments into the calculated fragmentation tree. The heuristics slightly improve running times.

We implemented all six 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 Arabidopsis 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 [1]. The test set was analyzed with the following options: Masses were decomposed using a relative

Table 1. The identification rates of the exact algorithm, the greedy heuristic, and the top-down heuristic.

Mass range	# comp.	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%

Table 2. The total running times of the algorithms.

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

mass error of 20 ppm over the standard CHNOPS-alphabet containing the six elements most abundant in living organisms.

Identification results can be found in Table 1. 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.

Fig. 1 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. 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. 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.

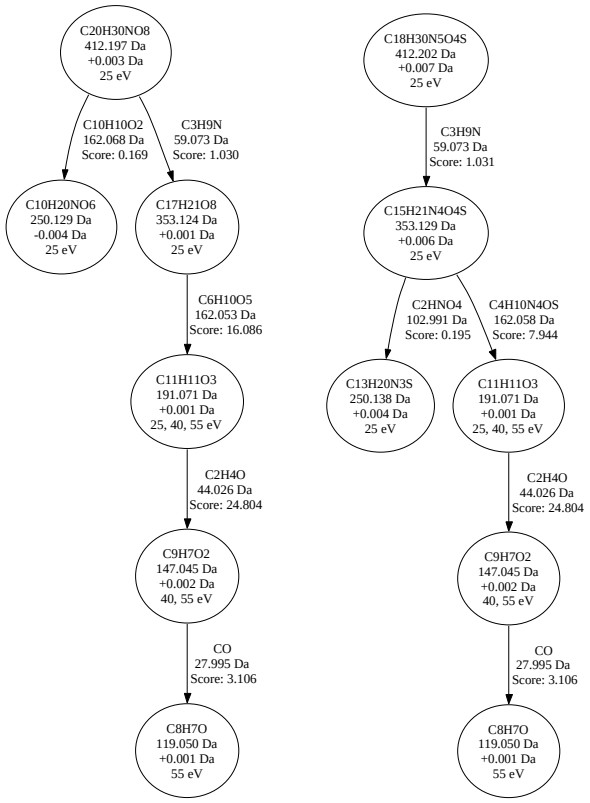

Fig. 1. Two fragmentation trees calculated from the spectra of hexosyloxycinnamoyl choline. Left: Fragmentation tree of the correct molecular formula ranked at first position. Right: Fragmentation tree of an incorrect molecular formula ranked at seventh position.

Running times of the different approaches can be found in Table 2. The speed of both heuristics and the DP+BF exact algorithm is sufficient to analyze data on the fly. It takes around 3 seconds 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.

3 Conclusion

We have presented an evaluation of an approach for the automated de novo identification of metabolites using tandem mass spectra [2]. We analyze an exact FPT-algorithm as well as two heuristics to solve the problem. Experiments on

real mass spectra show that all algorithms achieve very good identification results in application.

Acknowledgments

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.

References

1. von Roepenack-Lahaye, E., Degenkolb, T., Zerjeski, M., Franz, M., Roth, U., Wessjohann, L., Schmidt, J., Scheel, D., Clemens, S.: Profiling of arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. Plant Physiol **134** (2004) 548–559
2. Böcker, S., Rasche, F.: Towards de novo identification of metabolites by analyzing tandem mass spectra. In: Proc. of European Conference on Computational Biology (ECCB 2008). (2008) To be presented.
3. Wells, J.M., McLuckey, S.A.: Collision-induced dissociation (CID) of peptides and proteins. Methods Enzymol **402** (2005) 148–185
4. Niedermeier, R.: Invitation to Fixed-Parameter Algorithms. Oxford University Press (2006)