Identifying the unknowns by aligning fragmentation trees

Florian Rasche[1], Kerstin Scheubert[1], Franziska Hufsky[1,2], Thomas Zichner[3], Marco Kai[4], Aleš Svatoš[4] and Sebastian Böcker[1]

[1] *Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany*
[2] *Max Planck Institute for Chemical Ecology, Jena, Germany*
[3] *Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*
[4] *Research Group Mass Spectrometry and Proteomics, Max Planck Institute for Chemical Ecology, Jena, Germany*

sebastian.boecker@uni-jena.de

Abstract: Mass spectrometry allows sensitive, automated and high-throughput analysis of small molecules. In principle, tandem mass spectrometry allows us to identify "unknown" small molecules not in any database, but the automated interpretation of such data is in its infancy. Some years ago, fragmentation trees have been introduced for the automated analysis of the fragmentation patterns of small molecules. We have recently presented a method for the automated comparison of such fragmentation patterns, based on aligning the compounds' fragmentation trees. This method enables us to cluster compounds based solely on their fragmentation patterns, and resulting clusterings show a good agreement with known compound classes. We also show that fragmentation pattern similarities are strongly correlated with the chemical similarity of molecules. Finally, we presented a tool for searching a database for compounds with fragmentation pattern similar to an unknown sample compound. Our method allows fully automated computational identification of small molecules that cannot be found in any database.

Extended abstract of Rasche *et al.*, *Anal. Chem.*, 84(7):3417–3426, 2012 [RSH$^+$12].

1 Introduction

Mass spectrometry (MS) is a key analytical technology for detecting and identifying small molecules such as metabolites [CLH$^+$08]. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). Several analytical techniques have been developed, most notably gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS). LC-MS is usually combined with a gentle ionization, that results in minimal fragmentation of the adduct ions formed. Molecules can be further analyzed using tandem MS: Molecules are mass-selected, fragmented, and the mass-to-charge ratios (m/z) of the resulting fragments recorded.

Fragmentation in LC-MS experiments (usually collision-induced dissociation (CID)) is less reproducible than fragmentation by electron ionization for GC-MS. Even the time-consuming manual analysis of such data, as well as searching in spectral libraries, are major problems. Apart from a few pioneering studies, there are few computational methods for the automated analysis of tandem MS data from small molecules.

For decades, MS experts have manually determined fragmentation pathways to explain tandem MS data and determine the molecular structure. In 2008, Böcker and Rasche [BR08] presented an automated and swift method for annotating tandem MS data using a hypothetical *fragmentation tree* (FT). Tree nodes are annotated with the molecular formulas of the fragments and the edges represent (neutral or radical) *losses*. Computing FTs does not require databases of compound structures or of mass spectra. Neither does it require, apart from lists of common and implausible losses, expert knowledge of fragmentation. Expert evaluation suggests that the FTs are of very good quality [RSM$^+$11]. Similar FTs can be identified using visual comparison, which indicates some similarity in the structure of the underlying compounds. Unfortunately, "manual comparison of FTs is also laborious and time-consuming" [RSM$^+$11].

In [RSH$^+$12], we presented an automated method for comparing the FTs of two compounds. This allows us to use FTs in applications such as database searching, where we replace the direct comparison of mass spectra by the comparison of the (annotated and more informative) FTs. Our method is based on local tree alignments, generalizing local sequence alignments. We assume that structural similarity is inherently coded in the CID spectra fragments. FT similarity is defined by its edges, which represent losses and nodes, representing fragments. The local tree alignment contains those parts of the two trees where similar fragmentation cascades occurred.

Aligning FTs when the molecular structure of one compound is known can help elucidate the structure of the unknown compound. In [RSH$^+$12], we presented three workflows based on similarity scores. First, we compute pairwise tree alignments for all compounds and so generate a pairwise similarity matrix. We then cluster the compounds based solely on this similarity measure. We find that the resulting clusters agree well with the structural properties of the compounds. Second, we showed that FT similarities and structural similarities (Tanimoto scores) are strongly correlated. Third, we determine the similarities of a fragmentation tree from an unknown compound with all trees in a database, to search for related compounds. To filter out spurious hits, we presented a statistical evaluation based on decoy database searching. We named this approach *fragmentation tree basic local alignment search tool* or FT-BLAST for short. Finally, as a proof of principle we showed how biological samples from Icelandic poppy (*P. nudicaule*) can be analyzed in this framework.

2 Methods

We shortly recall the most important principle of our FT alignment method introduced in [RSH$^+$12], see there for all details. For the automated comparison of FTs we followed the paradigm of pairwise *local alignments*. We defined a simple similarity measure on the edges (losses) and nodes (fragments) of the two FTs. We generalized this similarity measure to trees of identical topology and summed the similarity of tree edges. We also allowed for the insertion and deletion of edges. We searched for *subtrees* in the two FTs that maximized our similarity measure.

Similarity of subtrees was defined as the sum of similarities of edges which, in turn, was chosen to reward identical losses and penalize distinct losses and insertions or deletions. Edge similarities were modified based on the number of non-hydrogen atoms contained. Similarity between fragments (nodes) was also rewarded or penalized. We modified a known recurrence for the problem in three ways. First, we also considered edge similarities. Second, we computed local alignments for maximum subtree similarity by adding a "zero-case" to the recurrence, corresponding to the leaves of the subtree. Third, we scored *join nodes* where two losses were combined into one, corresponding to the non-appearance of intermediate fragmentation steps. Alignment scores will clearly be large for large trees and small for small trees, so we normalized similarities by perfect match scores. To do this we computed for each FT the alignment score against itself, then used the minimum of the two scores, taken to the power of 0.5. We refrained from using the similarity matrix directly. Instead, for each compound we viewed its similarity matrix column as a fingerprint (or feature vector), as is done with gene expression data. See Fig. 1 for an example.

3 Discussion

To achieve the full potential of small molecule MS analysis and to overcome limitations of spectral libraries, we need methods for the computational analysis of fragmentation spectra from unknown compounds. Rule-based approaches for analyzing compound fragmentation spectra may suffer from the tremendous number of rules, both known and unknown. In addition, completely unknown compounds may not necessarily follow the

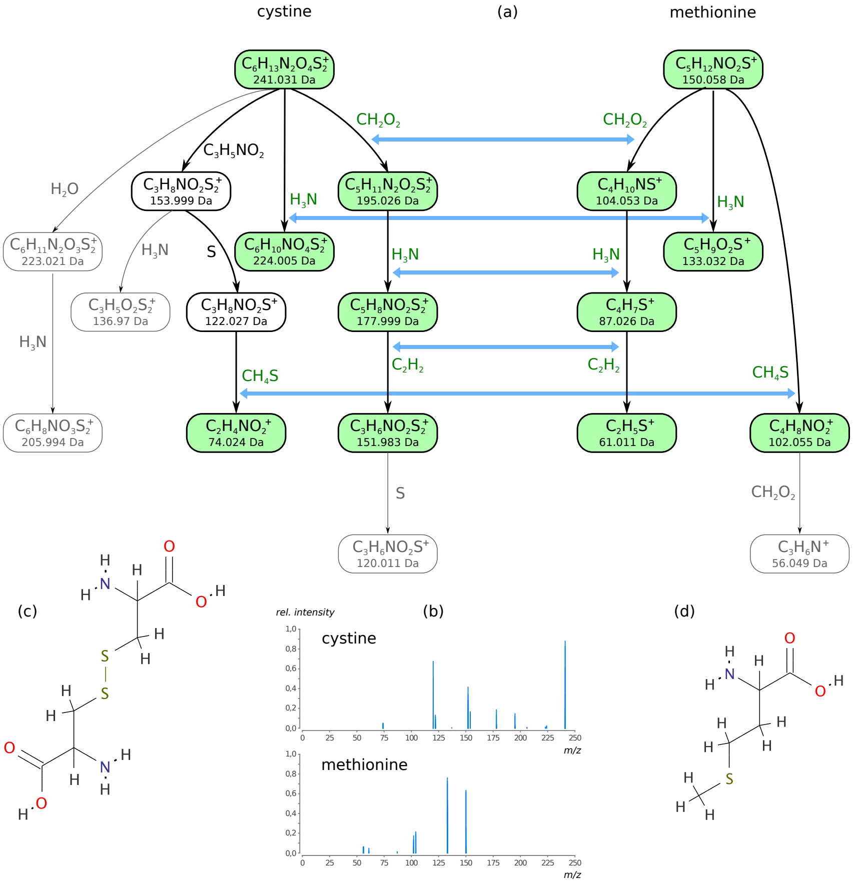
cystine
(a)
methionine
$C_6H_{13}N_2O_4S_2^+$ 241.031 Da
$C_5H_{12}NO_2S^+$ 150.058 Da
H_2O
$C_3H_5NO_2$
CH_2O_2
CH_2O_2
$C_3H_8NO_2S_2^+$ 153.999 Da
H_3N
$C_5H_{11}N_2O_2S_2^+$ 195.026 Da
$C_4H_{10}NS^+$ 104.053 Da
H_3N
$C_6H_{11}N_2O_3S_2^+$ 223.021 Da
H_3N
S
$C_6H_{10}NO_4S_2^+$ 224.005 Da
H_3N
H_3N
$C_5H_9O_2S^+$ 133.032 Da
$C_3H_5O_2S_2^+$ 136.97 Da
$C_3H_8NO_2S^+$ 122.027 Da
$C_5H_8NO_2S_2^+$ 177.999 Da
$C_4H_7S^+$ 87.026 Da
H_3N
CH_4S
C_2H_2
C_2H_2
CH_4S
$C_6H_8NO_3S_2^+$ 205.994 Da
$C_2H_4NO_2^+$ 74.024 Da
$C_3H_6NO_2S_2^+$ 151.983 Da
$C_2H_5S^+$ 61.011 Da
$C_4H_8NO_2^+$ 102.055 Da
S
CH_2O_2
$C_3H_6NO_2S^+$ 120.011 Da
$C_3H_6N^+$ 56.049 Da
(c)
(b)
(d)
rel. intensity
cystine
methionine
1,0
0,8
0,6
0,4
0,2
0,0
0 25 50 75 100 125 150 175 200 225 250
m/z

Figure 1: Optimal FT alignment for cystine (10 losses) and methionine (6 losses). (b) Fragmentation mass spectra used for computing FTs. Molecular structures of cystine (c) and methionine (d).

known rules of fragmentation. Unfortunately, real fragmentation patterns are extremely complicated, and new "rules" are constantly being introduced. This makes manual compound classification and structure elucidation cumbersome. In contrast, the approach presented here is fully automated and "rule-free", both when computing and aligning FTs. It only requires sufficiently information-rich fragmentation spectra.

Clustering results in [RSH$^+$12] show the potential of the method to differentiate compound classes. In many cases, large compound classes formed almost perfectly separated clusters; smaller compound classes were distributed among several clusters, but clusters contained few outliers. Hierarchical clustering was applied as a proof-of-concept and to demonstrate clustering results. Better results can possibly be achieved by other clustering methods and supervised Machine Learning. Nevertheless, our results indicate how to deduce the compound class of an unknown when a reasonable number of knowns are clustered simultaneously.

We found strong correlation between FT similarity and chemical similarity. FT similarity must not be understood as a *prediction* of chemical similarity in the sense of Machine Learning methods. However, FT similarity, expert knowledge, and other sources of information can be combined to permit the accurate prediction of chemical similarity.

Our method for searching spectral libraries (FT-BLAST) achieves a "larger profit" than classical spectral comparison methods, as it searches for similar, not identical, compounds. We achieved excellent search results for most compounds: Even when FT-BLAST returned only a single hit it was often meaningful. Cases where no hits or spurious hits were returned could often be attributed to small FTs, low quality measurements, or the absence of similar compounds from the database. FT-BLAST individually selects the size of the output for each query compound. For this purpose, we proposed a method for generating a decoy database of FTs that can be searched simultaneously [RSH$^+$12]. Database searching by spectral comparison has been in use for decades; but even today, no sensible methods for generating decoy databases for spectral comparisons have been developed.

By applying FT-BLAST and clustering to an unknown sample from poppy, we confirmed eight manual identifications and suggested compound classes for some other unknowns, as they were unquestionably members of a well-defined cluster. We also identified the biosynthetic precursor of several alkaloids, which come from mixed biosynthetic pathways.

FT alignments open a way to a fast classification/identification of metabolites, limiting work spent on ubiquitously occurring “uninteresting” molecules. Areas of application include natural product discovery, dereplication, or even inferring biosynthetic pathways and metabolic networks.

References

[BR08] Sebastian Böcker and Florian 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).

[CLH$^+$08] Qiu Cui, Ian A Lewis, Adrian D Hegeman, Mark E Anderson, Jing Li, Christopher F Schulte, William M Westler, Hamid R Eghbalnia, Michael R Sussman, and John L Markley. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat Biotechnol*, 26(2):162–164, 2008.

[RSH$^+$12] Florian Rasche, Kerstin Scheubert, Franziska Hufsky, Thomas Zichner, Marco Kai, Aleš Svatoš, and Sebastian Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal Chem*, 84(7):3417–3426, 2012.

[RSM$^+$11] Florian Rasche, Aleš Svatoš, Ravi Kumar Maddula, Christoph Böttcher, and Sebastian Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*, 83:1243–1251, 2011.