Finding Maximum Colorful Subtrees in practice

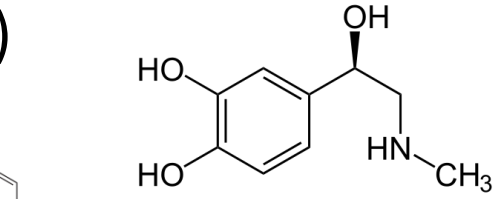
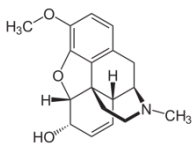
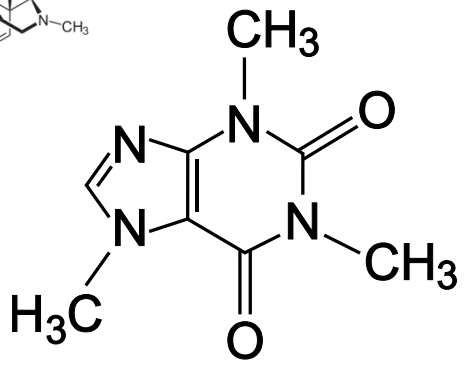
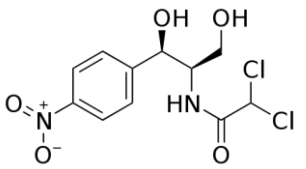
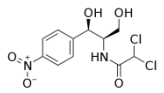
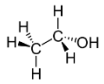
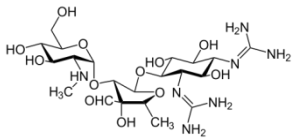
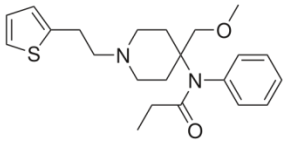
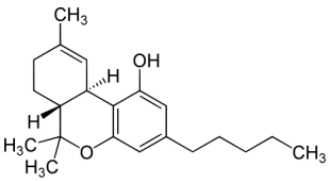
Imran Rauf, Florian Rasche,

François Nicolas, Sebastian Böcker

Friedrich-Schiller-Universität Jena

Metabolites: It's a small small world…

- small molecules (not DNA, proteins, glycans, …)
- typically less than 1000 Dalton
- energy resource, building blocks for cells
- signaling, communication, pathogen defense
- about 4000 metabolites per species
- less than 1% described in literature (**rough** estimate)
- especially diverse in plants and many bacteria

a third of humanity is latently infected with *M. tuberculosis*

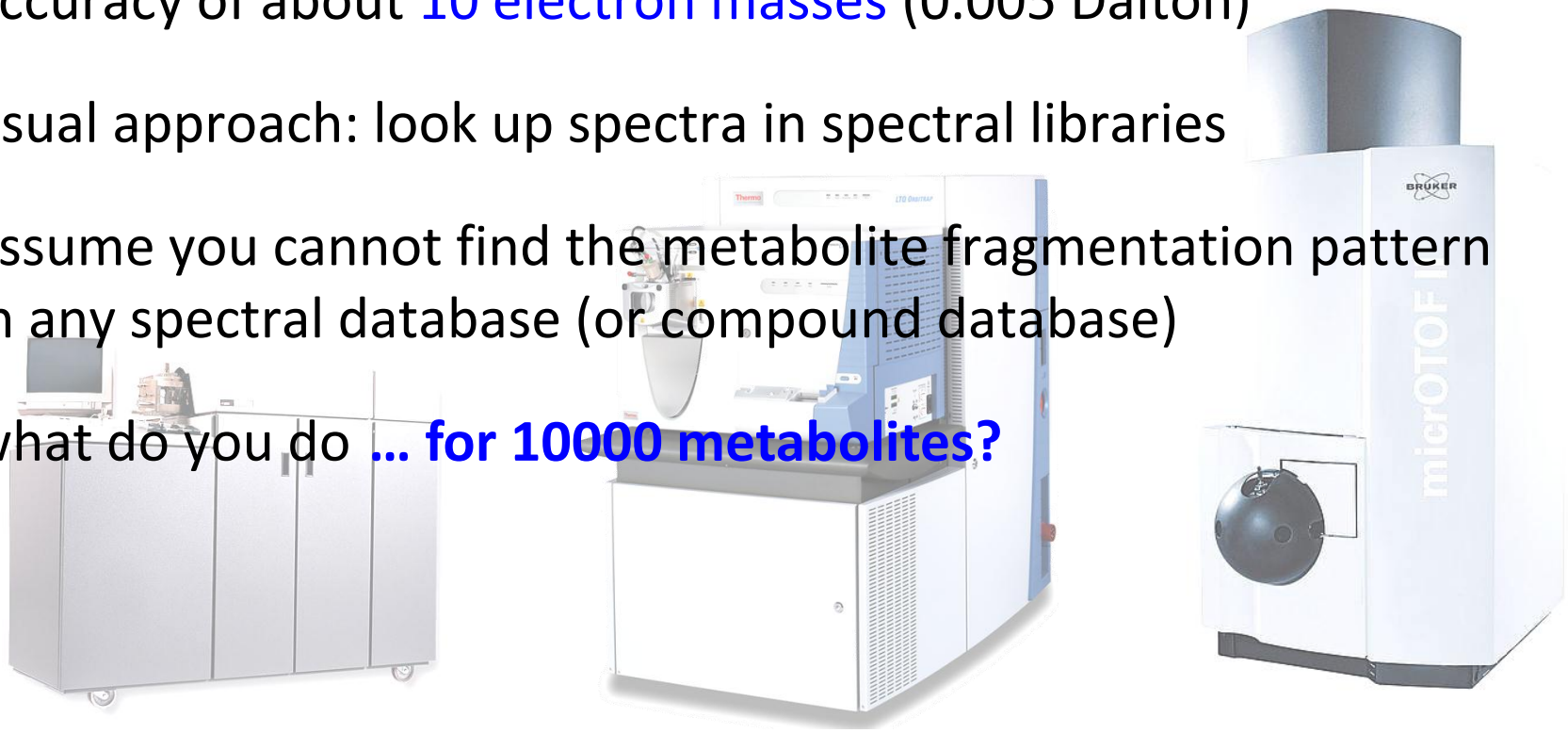
World Health Organization (WHO)

more than 50% of all top-selling drugs, and 80% of all antibiotics are derived from natural products

Newman *et al.*, 2007

Small molecule mass spectrometry

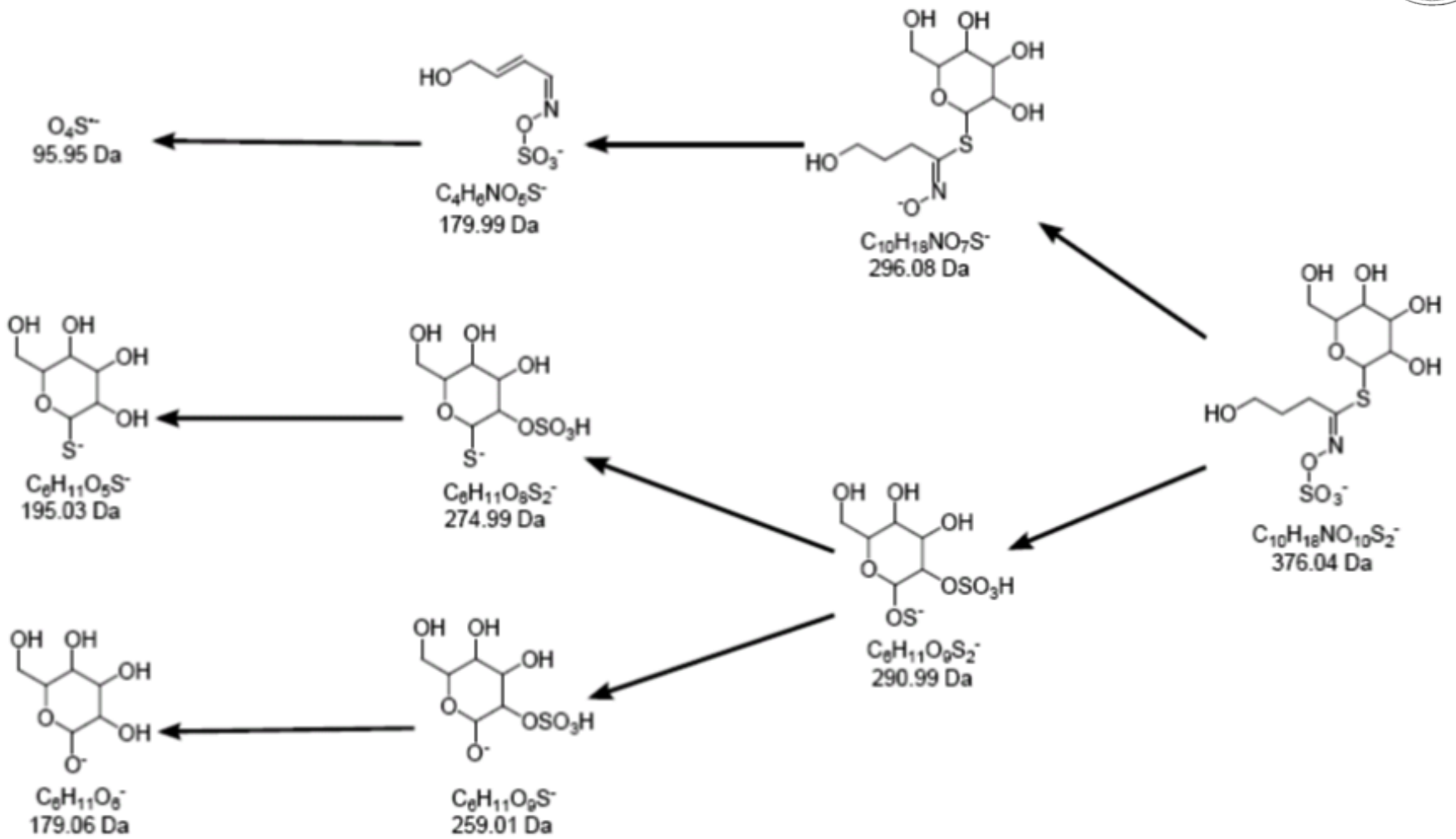
- high throughput setup for data acquisition:
  liquid chromatography (LC) mass spectrometry (MS)
- mass accuracy of 10 ppm: measure metabolite mass with accuracy of about 10 electron masses (0.005 Dalton)
- usual approach: look up spectra in spectral libraries
- assume you cannot find the metabolite fragmentation pattern in any spectral database (or compound database)
- what do you do **… for 10000 metabolites?**



Tandem mass spectrometry of metabolites

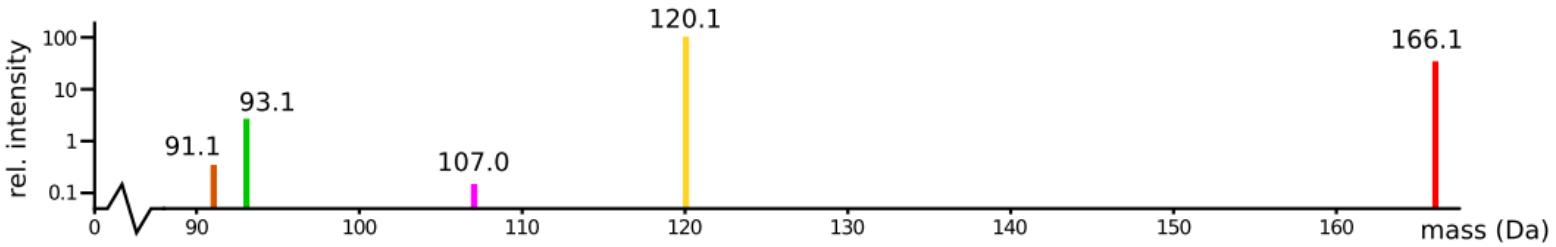
- fragment the metabolite
- identify it by measuring the masses of fragments

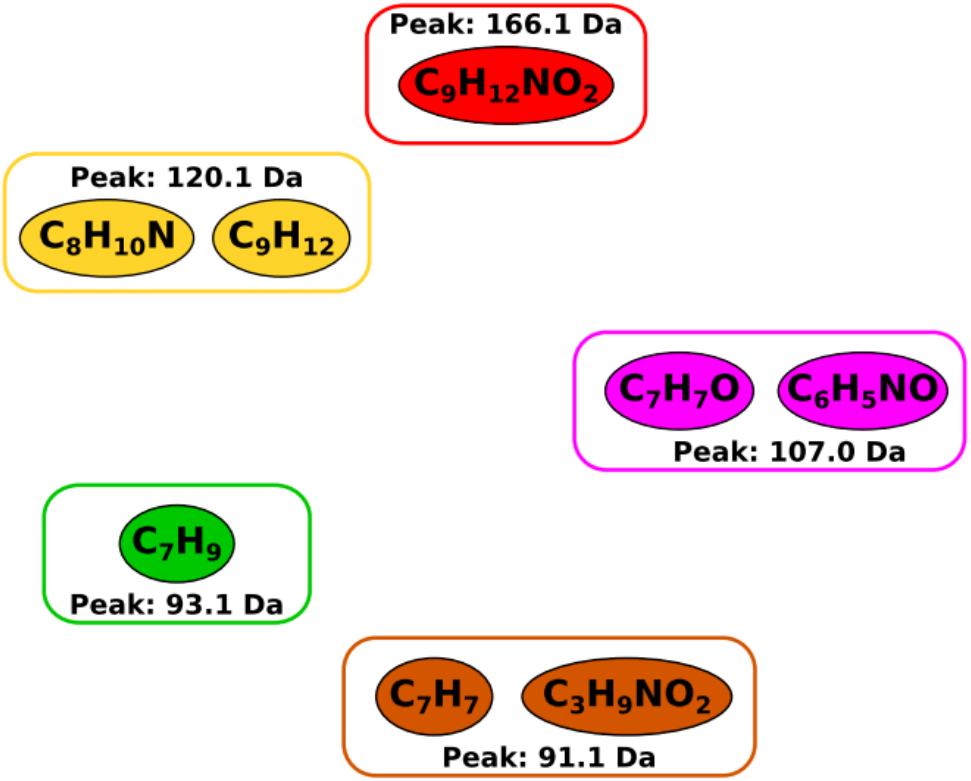
Fragmentation trees



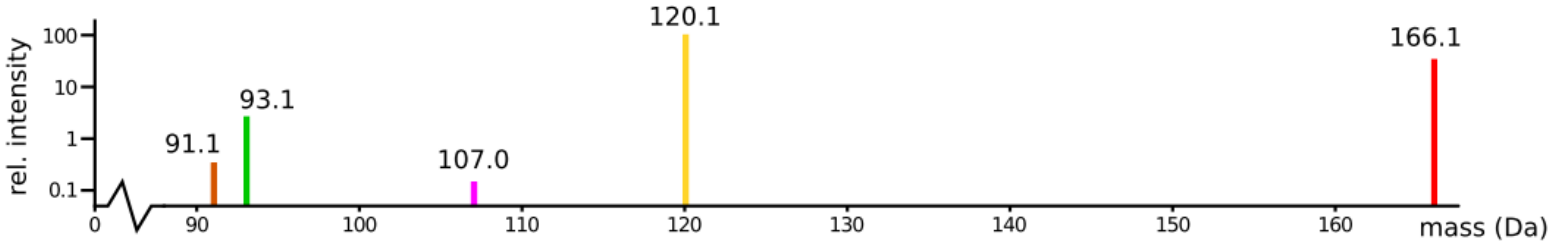
You need: MS data & ~~molecular structure~~ & ~~MS expert~~

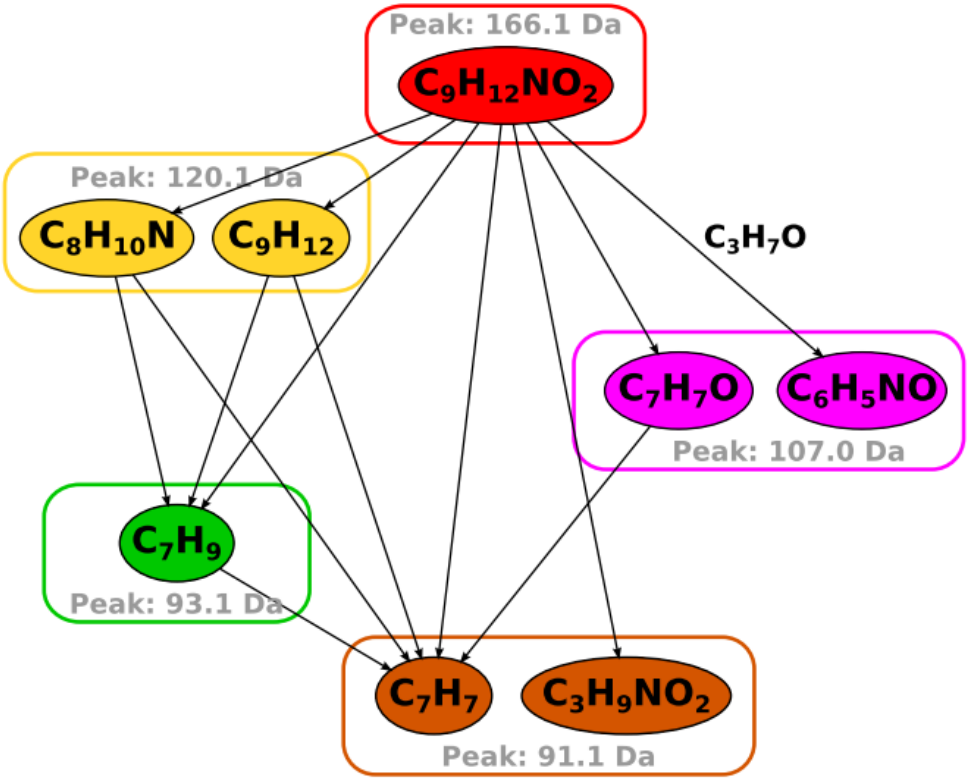
Fragmentation graph





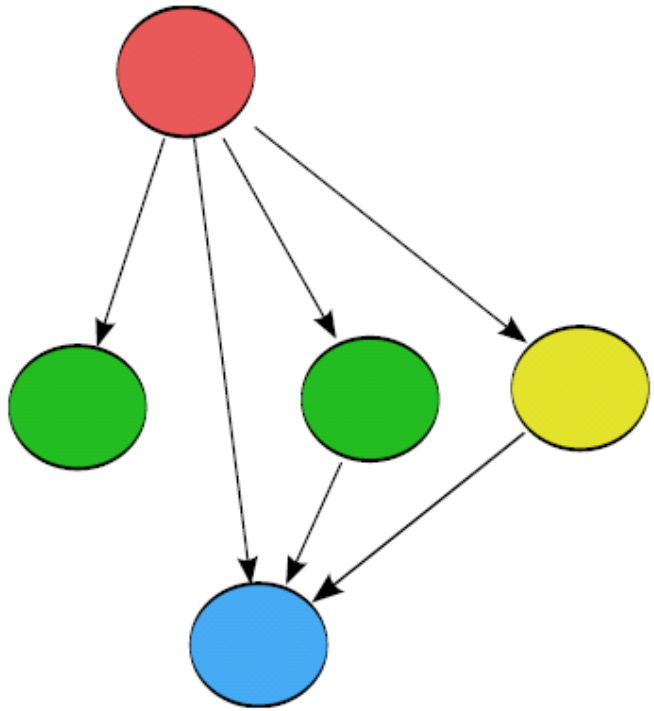
Fragmentation graph





Formal problem definition

- every peak is created from only one fragment (no peak double counting)
- every fragment originates from exactly one other fragment or the parent ion
- edges are weighted with log likelihood that they are real

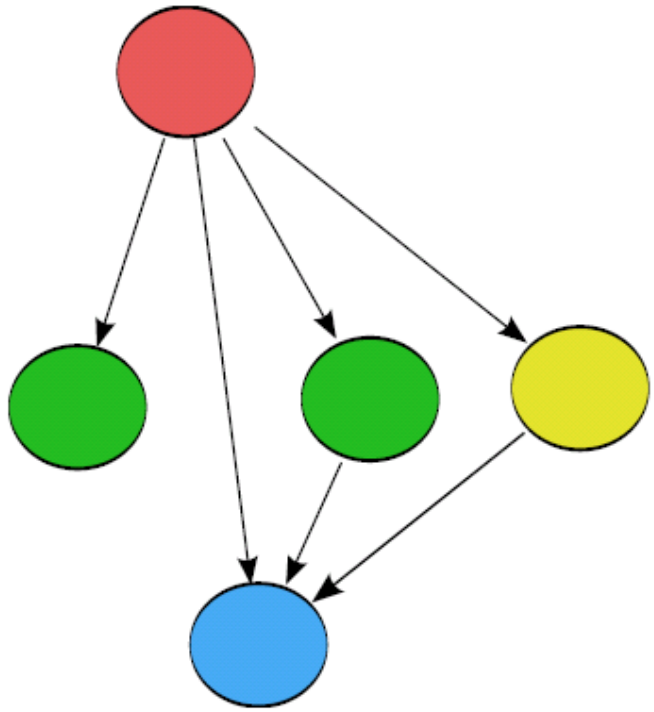


Maximum colorful subtree problem.

Given a colored directed acyclic graph *G*, find the subtree of *G* with maximal weight that uses every color of *G* at most once.

Intractability of the problem

- every peak is created from only one fragment (no peak double counting)
- every fragment originates from exactly one other fragment or the parent ion
- edges are weighted with log likelihood that they are real



Theorem.

There is no $O(|V|^{1-\varepsilon})$ approximation algorithm for the Maximum Subtree problem for any $\varepsilon > 0$, unless P = NP.

Two heuristic algorithms

Greedy heuristic

- search for the edge with the highest score (can be negative)
- in the end, remove leaves with negative edge weight

Insertion heuristic

- search for the vertex that has the largest *gain*
- the *gain* takes into account both the weight of incoming edge, and “re-wiring” of existing edges to increase the score

Dynamic programming revisited

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

[Böcker and Rasche, ECCB 2008]

- works only with relatively few (up to 15) peaks
- use exact method for 10-15 most intense peaks (DP10, DP15)
- greedily insert remaining peaks using insertion heuristic

Integer Linear Programing

$$\max \sum w(u,v) \cdot e_{uv} \tag{1}$$

$$\text{s.t.} \quad \sum_u e_{uv} \leq 1, \qquad \text{for all } v \in V \setminus \{r\}, \tag{2}$$

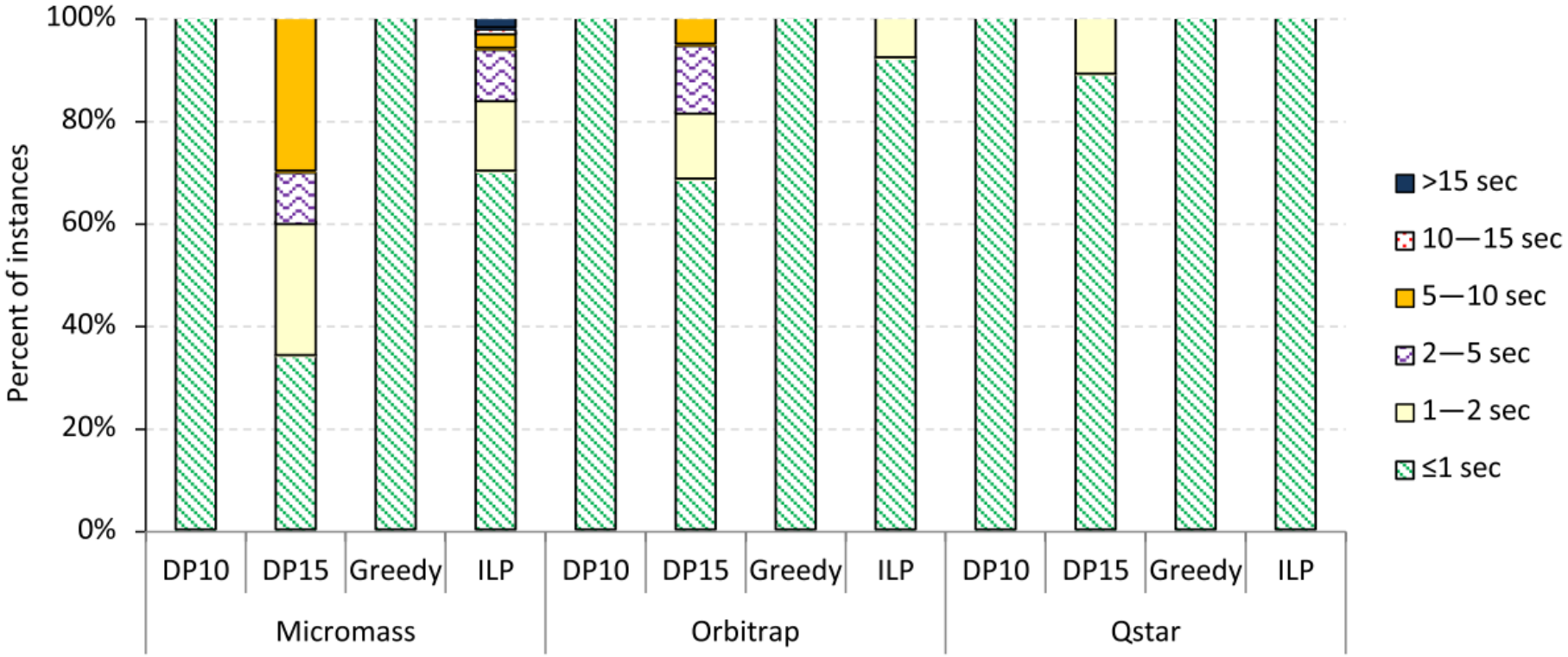
$$e_{vw} - \sum_u e_{uv} \leq 0, \qquad \text{for all } vw \in E \text{ and } v \neq r, \tag{3}$$

$$\sum_{v \in V(c)} e_{uv} \leq 1, \qquad \text{for all } c \in C, \tag{4}$$

$$e_{uv} \in \{0,1\}. \tag{5}$$

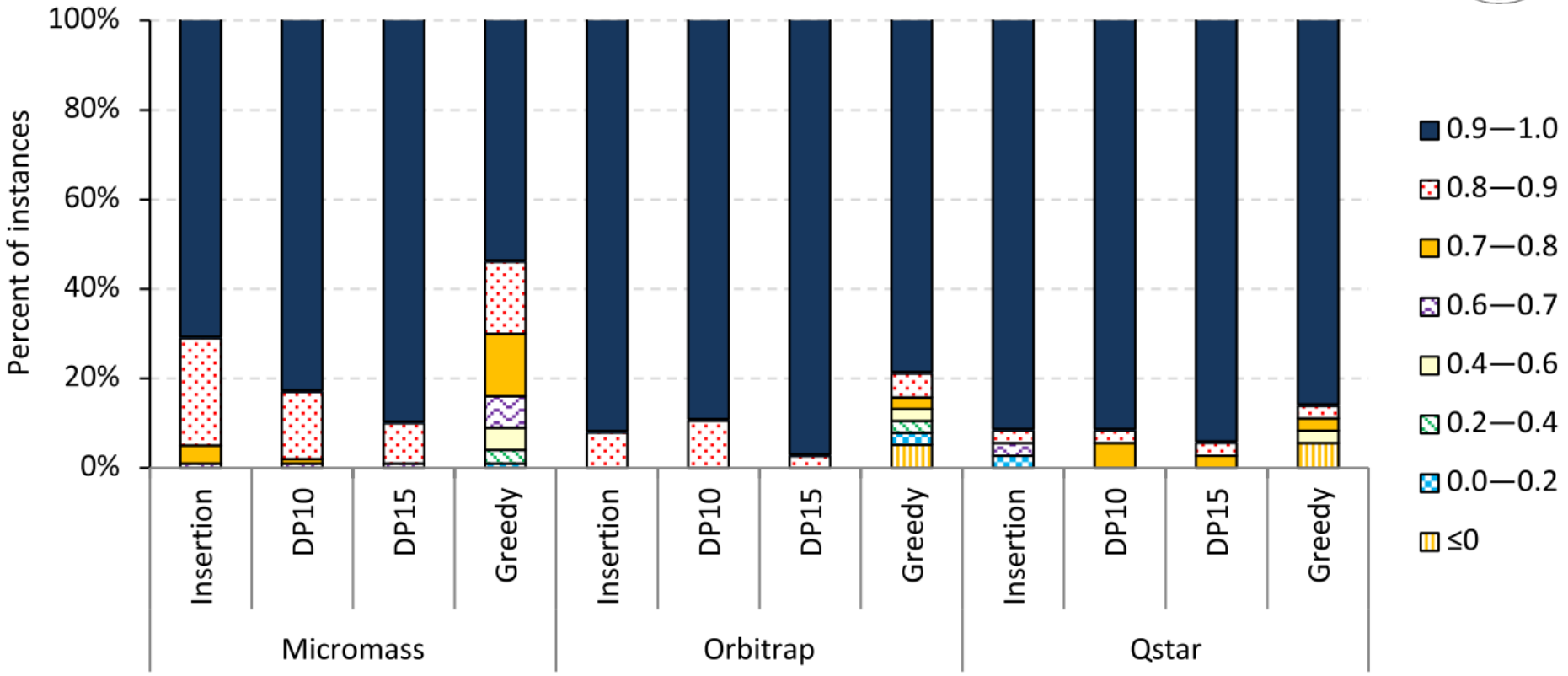
- similar to Ljubić *et al*., 2005 (Prize-collecting Steiner trees)
- differences: color constraints, input graph is DAG

Running time evaluation



- Three experimental datasets
- Exact ILP usually requires less than one second

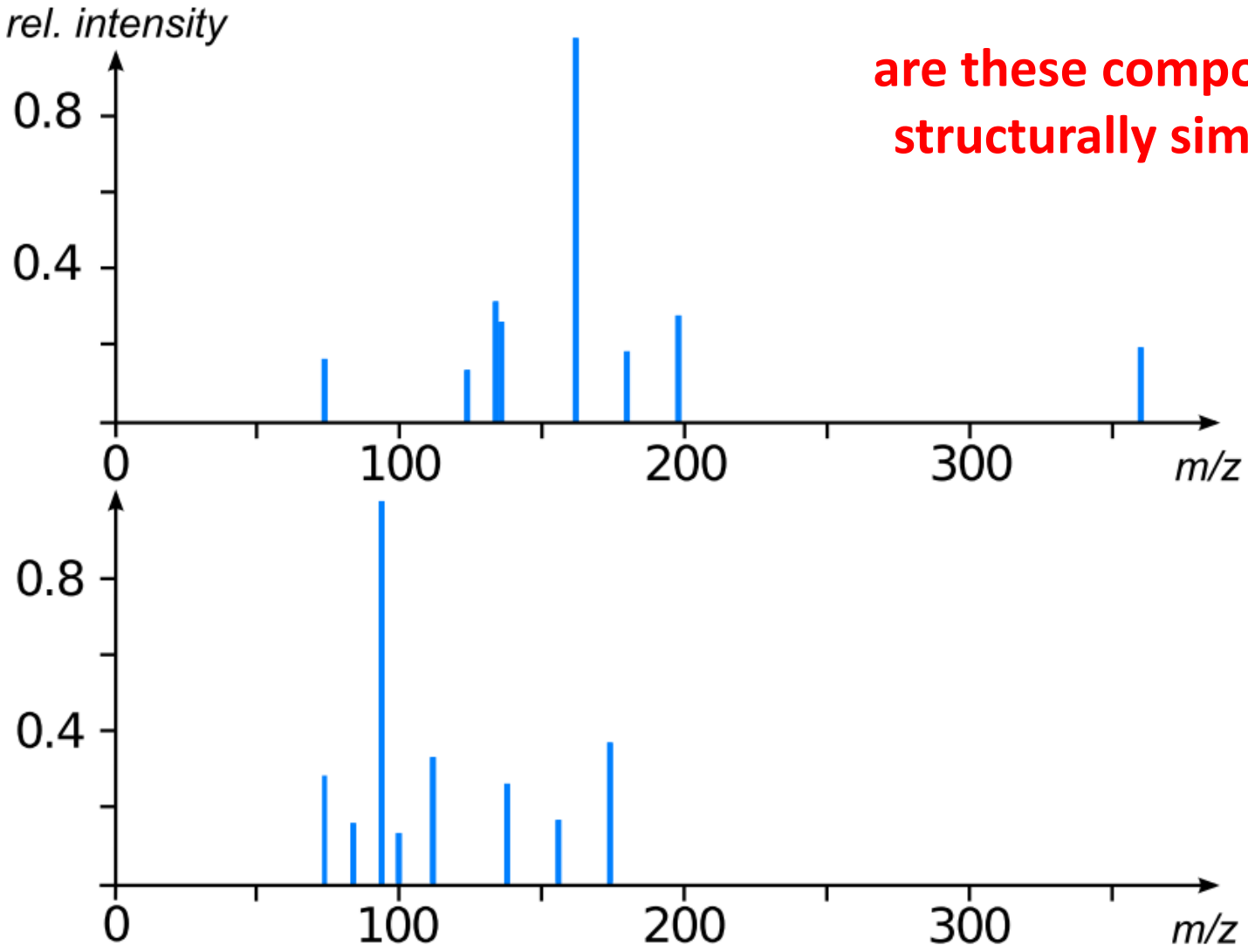
Heuristic quality evaluation



- We do not judge structure of trees, as the “true” tree structure is usually somewhat uncertain
- To this end, we only compare the reached scores

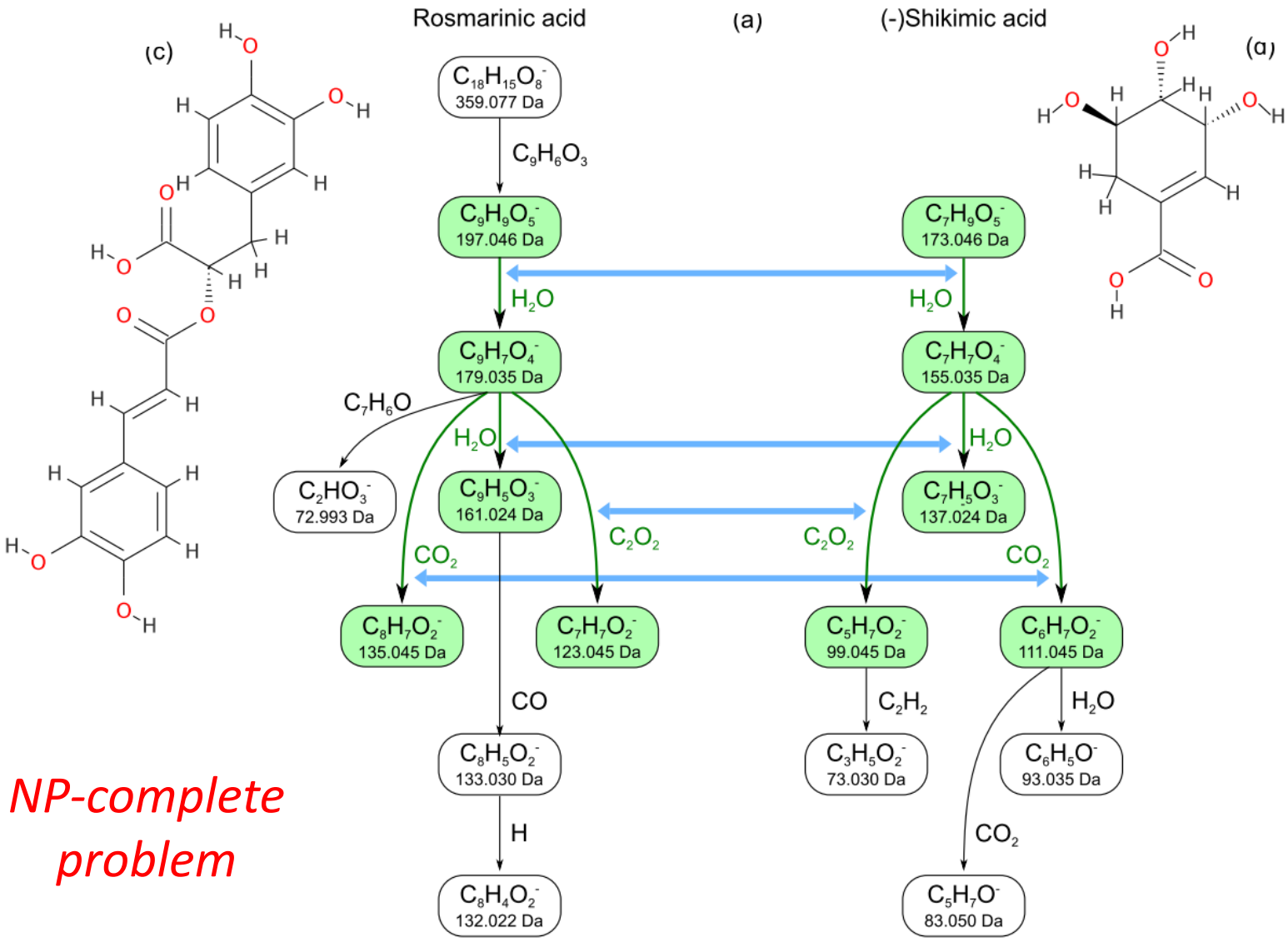
Aligning fragmentation trees

Tandem MS from two compounds



are these compounds structurally similar?

Aligning fragmentation trees



NP-complete problem

Sparse Dynamic Programming

$$S(u,v) = \max_{A \subseteq C(u), B \subseteq C(v)} S'_{u,v}[A,B]$$

$$S'_{u,v}[A,B] = \max\Big\{ match'_{u,v}[A,B], deleteL'_{u,v}[A,B], deleteR'_{u,v}[A,B], joinL'_{u,v}[A,B], joinR'_{u,v}[A,B] \Big\}$$

$$S'_{u,v}[A,*] := \max_{B' \subseteq C(v)} \Big\{ S'_{u,v}[A,B'] \Big\},$$

$$S'_{u,v}[*,B] := \max_{A' \subseteq C(u)} \Big\{ S'_{u,v}[A',B] \Big\},$$

$$match'_{u,v}[A,B] := \max_{a \in A, b \in B} \Big\{ S(a,b) + S'_{u,v}\big[A - \{a\}, B - \{b\}\big] + \sigma(ua, vb) \Big\}$$

$$deleteL'_{u,v}[A,B] := \max_{a \in A, B' \subseteq B} \Big\{ S'_{a,v}[*,B'] + S'_{u,v}[A - \{a\}, B - B'] + \sigma(ua, -) \Big\}$$

$$deleteR'_{u,v}[A,B] := \max_{A' \subseteq A, b \in B} \Big\{ S'_{u,b}[A',*] + S'_{u,v}[A - A', B - \{b\}] + \sigma(-, vb) \Big\}$$

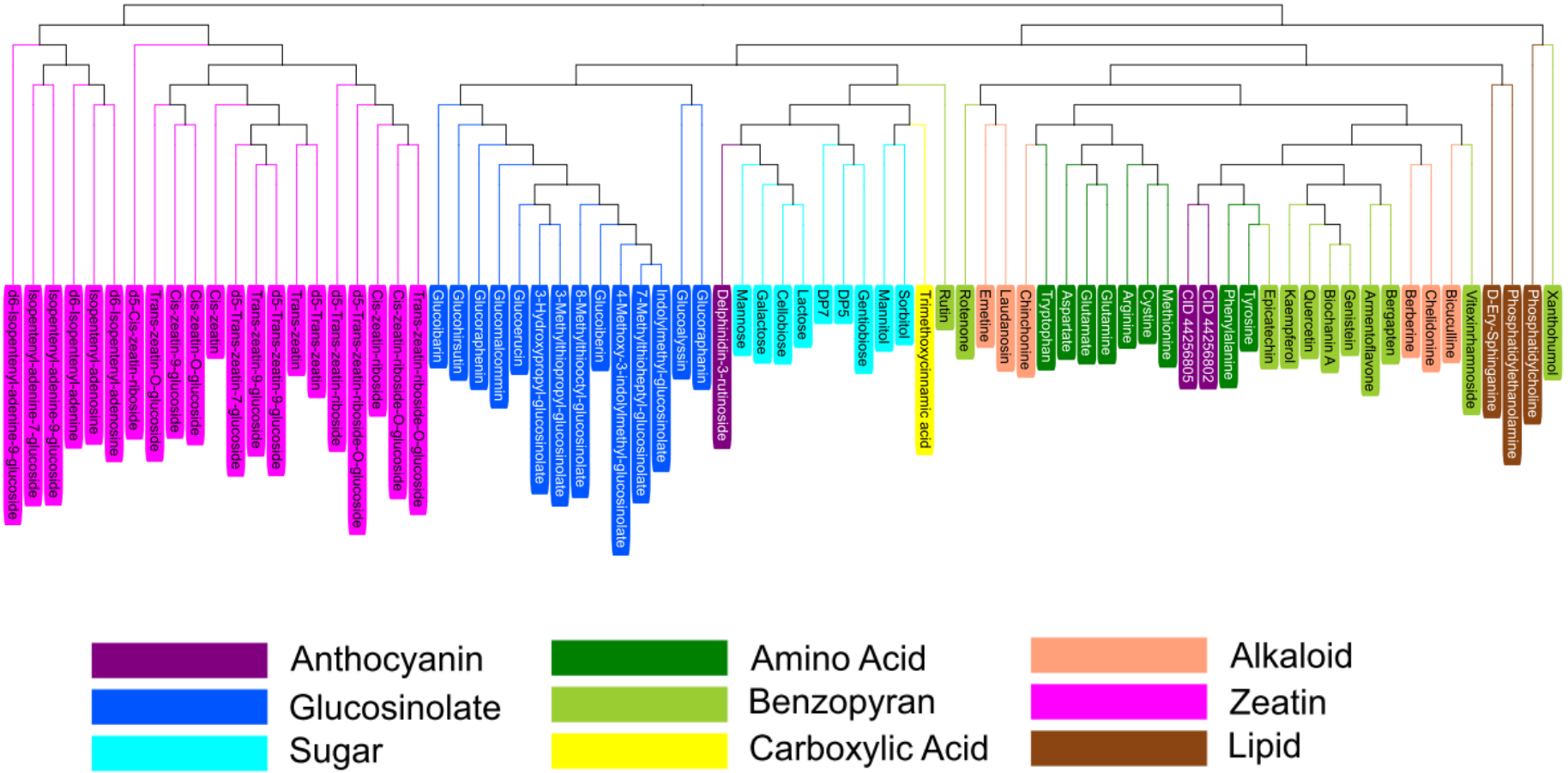
$$prejoinL'_{u,v}[A,B] := \max_{a \in A, b \in B} \Big\{ S(a,b) + prejoinL'_{u,v}\big[A - \{a\}, B - \{b\}\big] + \sigma(p(u)a, vb) + \sigma_{\text{join}} \Big\}$$

$$prejoinL'_{u,v}[*,B] := \max_{A' \subseteq A} \Big\{ prejoinL'_{u,v}[A',B] \Big\}$$

$$joinL'_{u,v}[A,B] := \max_{a \in A, B' \subseteq B} \Big\{ prejoinL'_{a,v}[*,B'] + S'_{u,v}[A - \{a\}, B - B'] \Big\}$$

[Hufsky *et al.*, ISMB 2012, to be presented]

Clustering „unknown“ compounds: Orbitrap data



[Rasche *et al.*, *Anal. Chem.* 2012]

Conclusion

Conclusion and outlook

- Fragmentation trees enable *de novo* interpretation of small molecule mass spectra
- Together with alignments: analysis of unknown compounds
- Database searching: FT-BLAST
- Compound class classifier: Is this an alkaloid?
- Reconstruct metabolite networks
- Analyze drug degradation products
- Dereplication, searching for novel drugs

Poster 51 on aligning fragmentation trees

Identification of Unknown Metabolites by Aligning Fragmentation Trees

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

Max Planck Institute for Chemical Ecology

[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, Heidelberg, Germany

seit 1558

Introduction

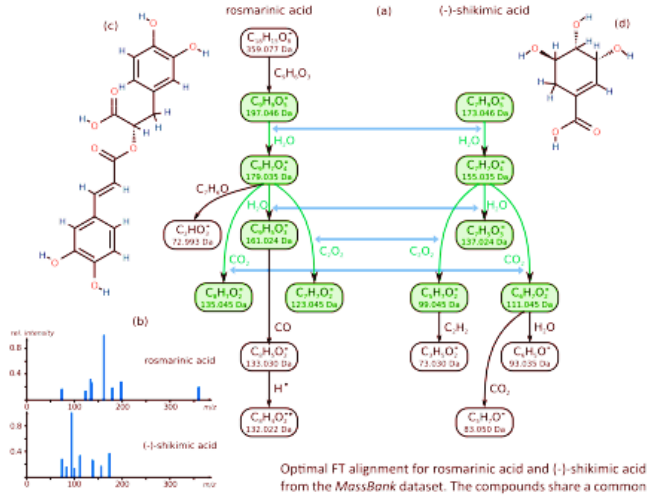
Small molecules play an essential role in metabolism and serve as drugs in pharmacology. Tandem mass spectrometry allows sensitive and high-throughput analysis of small molecules. Unfortunately, the automated interpretation of the data is still in its infancy. In 2008, Böcker and Rasche [1] introduced fragmentation trees (FT) for the automated analysis of fragmentation pattern, and in 2011 Rasche et al. [2] showed that the FTs are of excellent quality.

We present a method that allows fully automated computational classification and identification of small molecules that cannot be found in any database [3]. This method is based on the assumption that similar molecules fragment in a similar way.

Method

For each compound, we computed a hypothetical FT [1,2]. For the automated comparison of FTs we follow the paradigm of pairwise local alignments. We defined a simple similarity measure on the edges (losses) and nodes (fragments) of the two FTs. We also allow for the insertion and deletion of edges, and for combining two losses into one, corresponding to the non-appearance of intermediate fragmentation steps.

Aligning unordered trees is computationally hard [4]. To compute alignments of unordered trees, we used an exact algorithm based on dynamic programming that guarantees the optimal solution is found.



Optimal FT alignment for rosmarinic acid and (-)-shikimic acid from the *MassBank* dataset. The compounds share a common biosynthetic polyketide origin.

Data

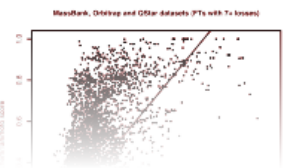
We tested our method on three datasets: 97 compounds measured on a Thermo Scientific *Orbitrap* XL instrument (< 5ppm), 370 compounds from *MassBank* measured on a Waters Q-Tof Premier spectrometer (≈ 50 ppm), and 44 compounds measured on an API *QSTAR* QTOF spectrometer by Applied Biosystems (20 ppm). The masses of all compounds ranges from 75 Da to 1258 Da.

1 - Clustering

For each dataset, we performed all-

2 - Correlation

We found strong correlation between FT similarity and chemical similarity (PubChem/Tanimoto [5]). We did not compare any compound against itself and restricted ourselves to compounds with 3+ losses. Pearson correlation was r=+0.60 for the *MassBank* dataset, r=+0.62 for *QSTAR*, and r=+0.67 for



Thank you for the data

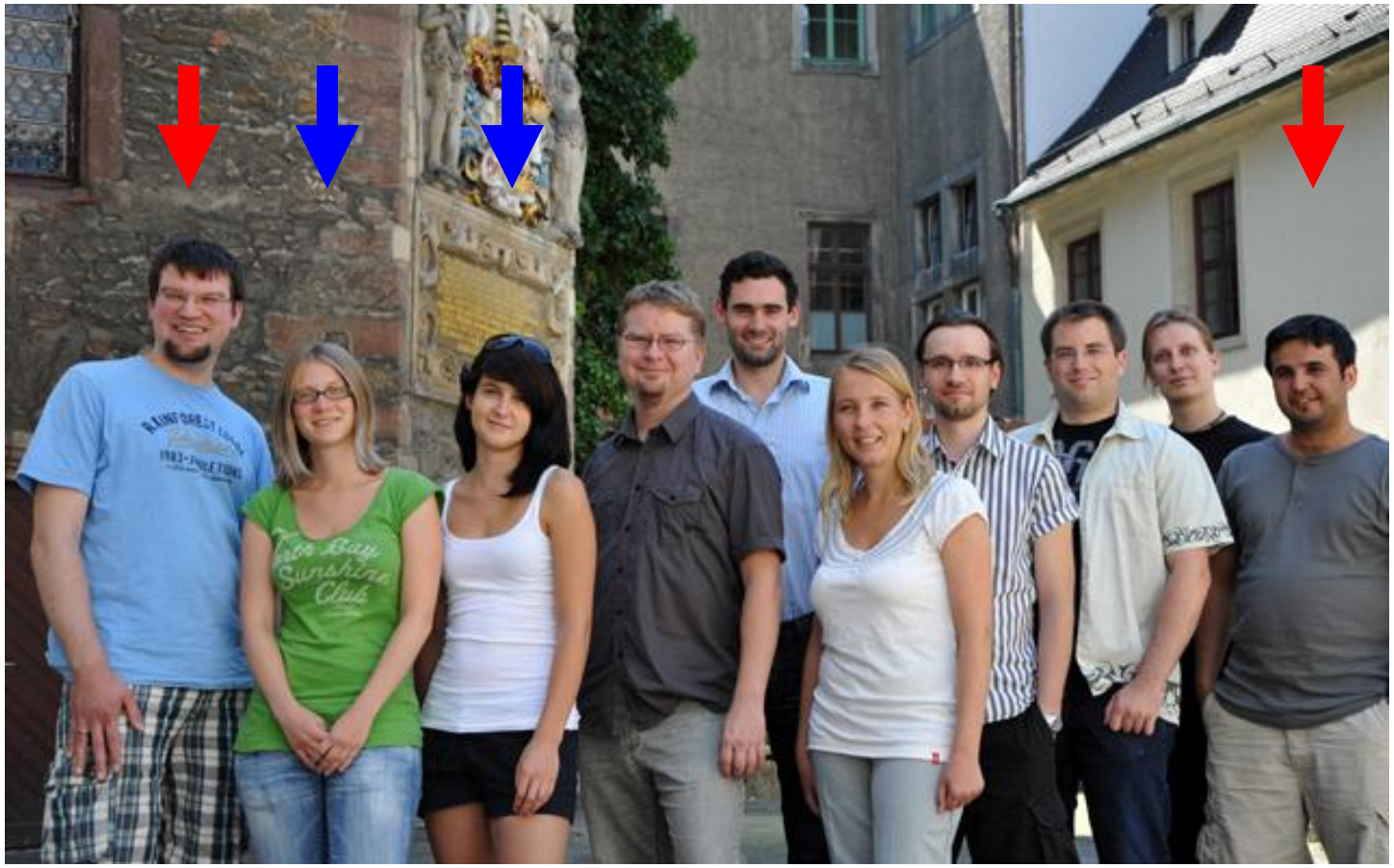
Our MS experts and collaborators

- Aleš Svatoš (MPI Chemical Ecology, Jena)
- Christoph Böttcher (Leibniz Inst. of Plant Biochemistry, Halle)
- Mass Bank consortium, in particular M. Arita (RIKEN Plant Science Center, Yokohama, Japan)

Thank you for the funding

Deutsche Forschungsgemeinschaft

Chair for Bioinformatics, FSU Jena



Thank you for your attention! It's a wrap! Enjoy your meal!