Algorithms supporting analysis of mass spectra of metabolites and lipids

Marcin Borowski, Tomasz Weiss, Marek Weber

Institute of Computing Science, Poznań University of Technology

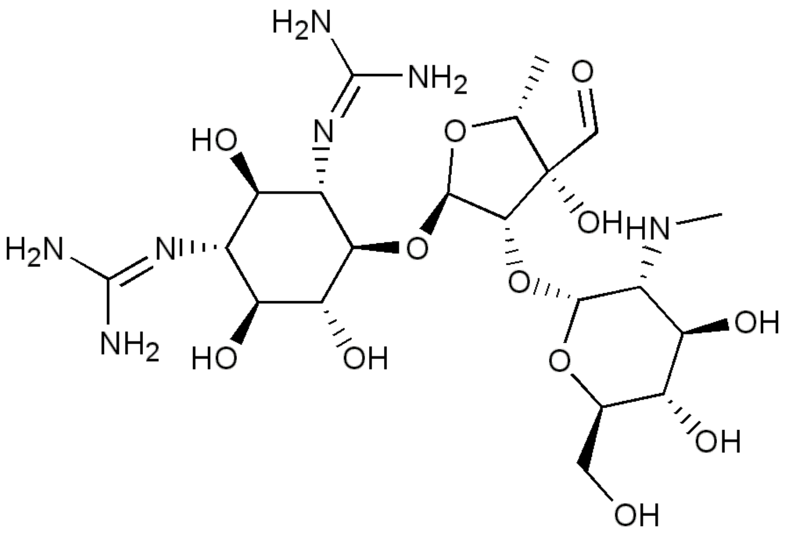
EUROPEAN CENTRE FOR BIOINFORMATICS AND GENOMICS

Outline

1. Background
2. Motivation
3. Method
4. Problem formulation
5. Algorithm
6. Results
7. Summary

Background

- **metabolites and lipids**
  - small chemical molecules that are the product of metabolism in organisms
  - can include a variety of substances such as amino acids, sugars, fatty acids and many other
  - are involved in metabolic processes in living organisms, such as metabolism, energy production, protein synthesis, regulation of cell function and cell-cell interactions



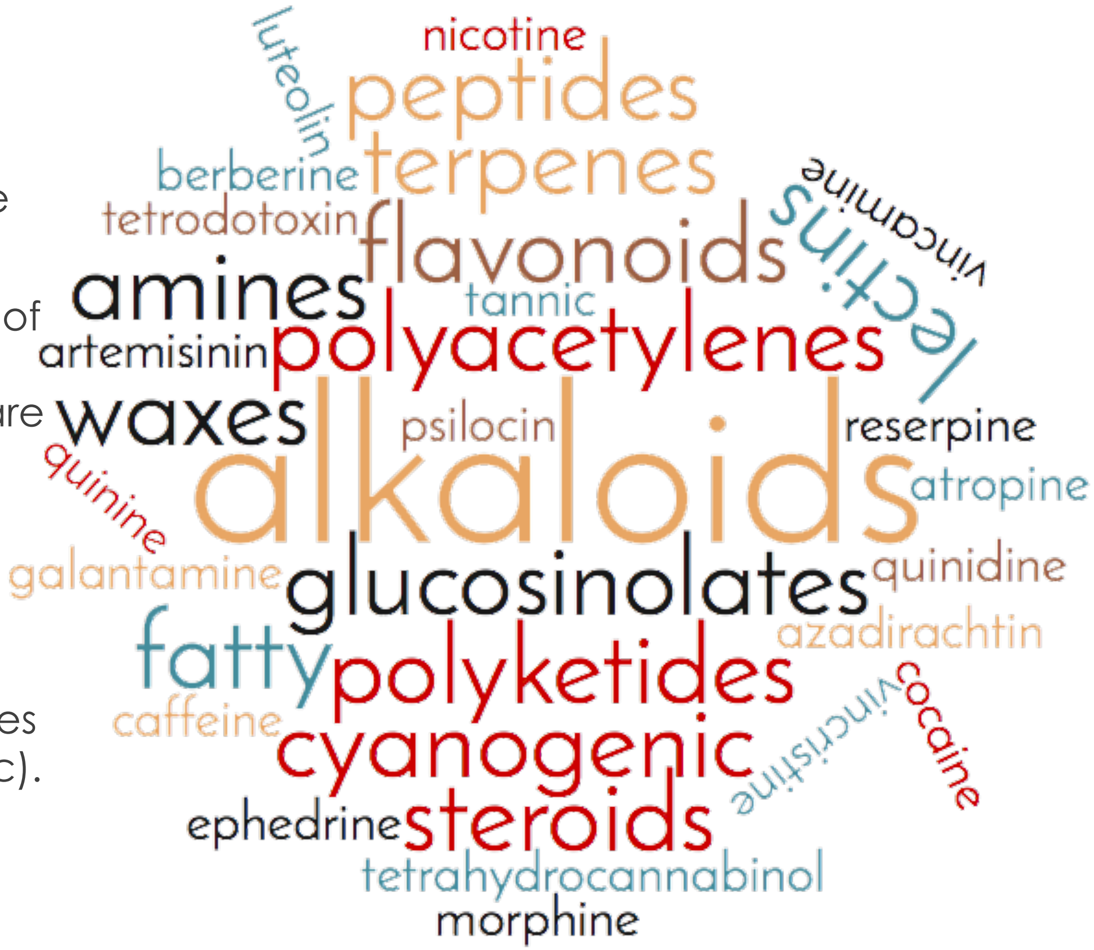
Streptomycin, an important antibiotic drug produced by *Streptomyces* bacteria

Background

They are important for maintaining the body's homeostasis, the internal balance that allows it to function properly.

I.e., in plants the function or importance of these compounds to the organism is usually of an ecological nature as they are mainly used:

- as defenses against predators, parasites and diseases,
- for interspecies competition
- to facilitate the reproductive processes (coloring agents, attractive smells, etc).



Background

- Metabolites
  - have been referred to as "**nature's biological response modifiers**" because of strong experimental evidence of their inherent ability to modify the body's reaction to allergens, viruses, and carcinogens.
  - provided (i.e. flavonoids), show **anti-allergic**, **anti-inflammatory**, **anti-microbial** and **anti-cancer** activity.
  - consumers and food manufacturers have become interested in metabolites for their **medicinal properties**, especially their potential role in the prevention of **cancers and cardiovascular disease**. The beneficial effects of fruit, vegetables, and tea or even red wine have been attributed to some metabolites rather than to known nutrients and vitamins.
- Lipids
  - key components of cell membranes, are the main source of energy in organisms, play an important role in cell membranes and act as signalling lipids
  - One of these, for example, is cholesterol, which is essential for the structure of cell membranes and is a precursor to many hormones such as cortisol, oestrogen and testosterone.

Motivation

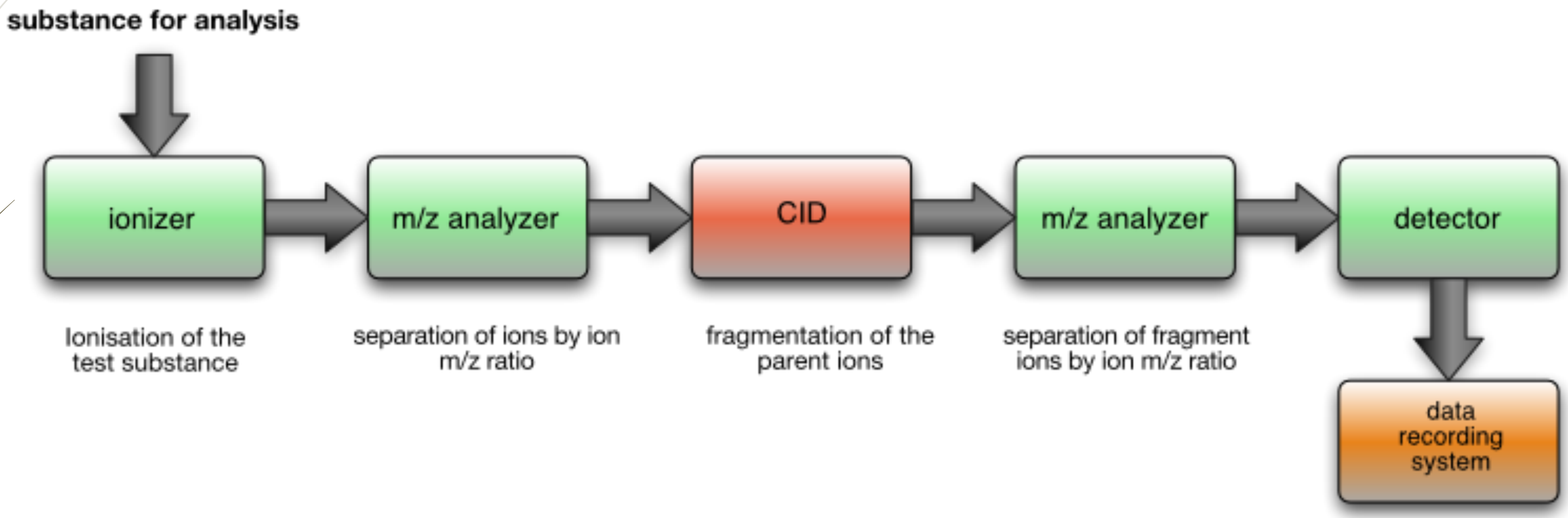
- The primary goal of **metabolomics** and **lipidomics** is the quantitative and qualitative analysis of metabolites and lipids in biological samples, allowing us to understand **metabolic processes** and discern changes in organisms **in response to** various factors, including **diseases**, **diets** or **therapies**.
- Recognition of an metabolites and lipids is a preliminary stage of determination of its **function** in metabolism process.
- The existing methods for a recognition of metabolites and lipids structure **base on** information stored in **databases**.
- In many cases, during high throughput metabolite profiling identification of compounds is difficult or sometimes even impossible, because the identification process requires an **expert analyst knowledge** and **time consuming procedures**.

Methods

- **Mass spectrometry** is a method that allows the precise examination of the chemical composition of biological samples
- Mass spectra of whole extracts give possibilities for detection of ions corresponding to compounds present even in trace amounts.

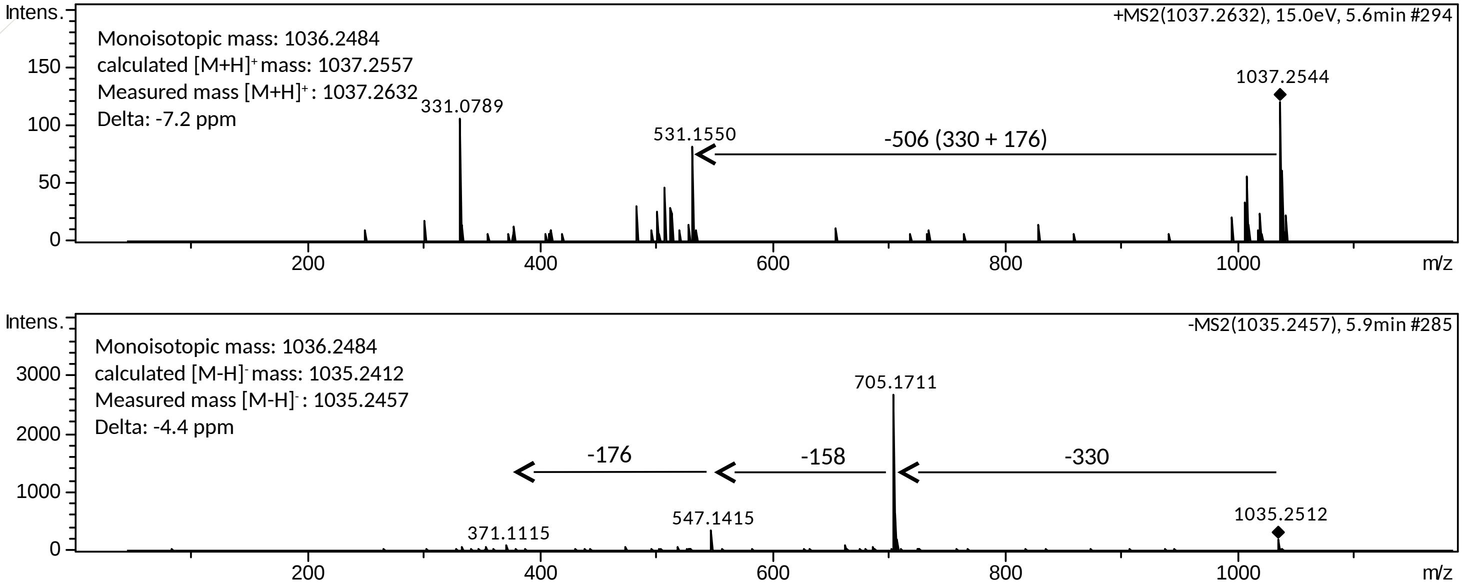
The use of instruments capable for acquisition of MS/MS spectra combined with liquid chromatography or capillary electrophoresis became a user friendly technology for exhaustive identification of secondary metabolites present in extract samples obtained from plant tissue.

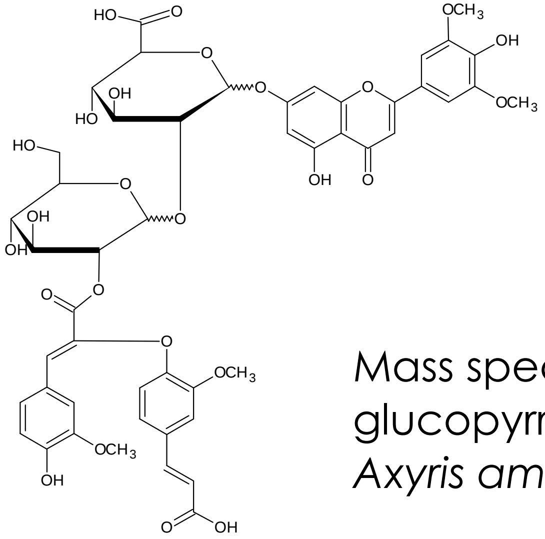
Methods



substance for analysis
ionizer
m/z analyzer
CID
m/z analyzer
detector
Ionisation of the test substance
separation of ions by ion m/z ratio
fragmentation of the parent ions
separation of fragment ions by ion m/z ratio
data recording system

Methods

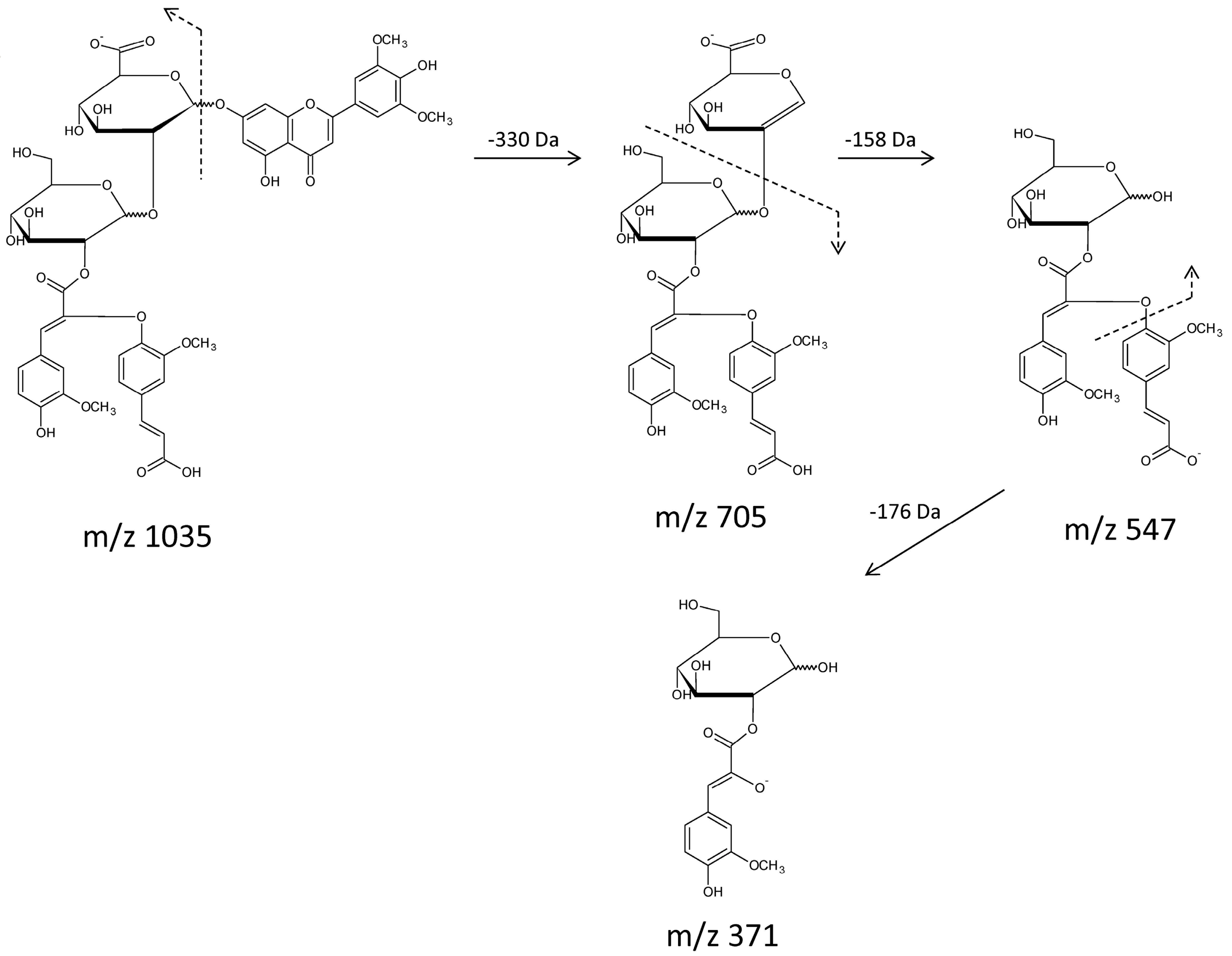




$C_{49}H_{48}O_{25}$

Mass spectra of Tricine 7-O-[2'-O-dehydrodiferuloyl-glucopyrranosyl-(-2)-O-glucuronopyranoside from *Axyris amaranthoides*.

Methods



Problem formulation

Instance: Let *P be a set of pairs* (m_i, w_i)*, where* $m_i \in Q_+$ *and* $w_i \in Z_+$, positive real number *M and positive real number* δ.

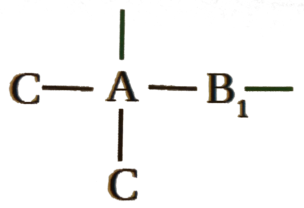
Answer: Set *of vectors* x = $(x_1, x_2, \ldots, x_i, \ldots, x_k)$ with non-negative integer values satisfying following constraints:

$$M - \delta \le \sum_i m_i x_i \le M + \delta, \quad (4.1)$$

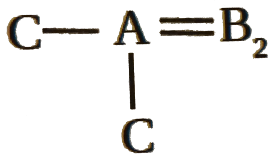
$$\sum_i w_i x_i = 2n, \quad n \in \mathbb{N}, \quad (4.2)$$

$$\sum_i (w_i - 2) x_i + 2 \ge 0, \quad (4.3)$$

$$\sum_i w_i x_i - 2 \max_i \{ w_i | x_i > 0 \} \ge 0 \quad (4.4)$$



invalid solution for valence constraints



valid solution for valence constraints

Problem formulation

Example:

Instance

$P = \{$

$H : (1.007825, 1),$
$C : (12.00000, 4),$
$O : (15.99491, 2),$
$P^{III} : (30.97376, 3),$
$P^{V} : (30.97376, 5),$
$S : (31.97207, 2)$
$\},$

$M = 347.69953$

$\delta = 0.00001$

Answer

$H\ C\ O\ P^{III} P^{V} S$

$\mathbf{x}$: $(0, 0, 0, 3, 1, 7)$

⇩

$P^{III}{}_3 P^{V}{}_1 S_7$

⇩

P_4S_7

- Problem is NP-hard (Money changing problem can be reduced to this problem [BL07]).

[BL07] A Fast and Simple Algorithm for the Money Changing Problem, S. Bocker, Z. Liptak, Algorithmica, August 2007, Volume 48, Issue 4, pp 413-432

Algorithm

- The proposed algorithm is based on steps:
  - configuration of the algorithm

    MIN_INTENSITY_PERCENTAGE

    FRAGMENTATION_GAP_BETWEEN_PEAKS

    PRECURSOR_MASS_DELTA

    SEARCH_COMPOUND_MASS_DELTA

    KEGG_COMPOUNDS_LOCATION

    KEGG_LIPIDS_LOCATION

    COMBINATION_MASS_DELTA

    EXCLUDED_FUNCTIONAL_GROUPS

    EXCLUDED_ELEMENTS

    DATA_LOCATION

Algorithm

- loading data for defined possible compounds from **PubChem** database (***cid, cmpdname, mw, mf, canonicalsmiles, exactmass, monoisotopicmass)***
- loading of mass spectra data.
- analysis of the mass spectrum in terms of mass-to-charge ratio (m/z) and signal intensity to determine the decay of the compound.
- generation of possible decay combinations and match compounds to them.
- selection of generated combinations with respect to compound masses, molecular structure and combination methods (**when no relationships are found for any distance, combinations containing this distance are discarded from further analysis**)
- **calculation of basic information such as summary/structural formula, compound mass and SMILES notation for the resulting combinations**
- selection of combinations against the calculated mass.
- generation an analysis report

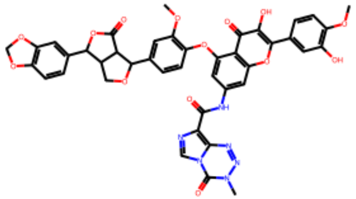
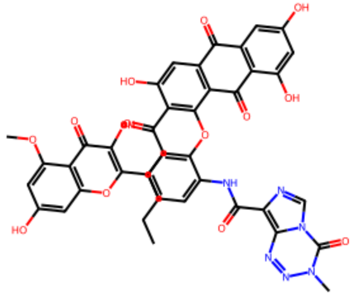
Algorithm
Experimental data

Spectra of 5 compounds from the Auxiris plant were used for testing

- Isorhamnetin_3-O-3-O-coumaroyl-glucuronopyranosyl-1-2-O-glucopyranosyl-1-2-O-glucuronopyranoside
- Isorhamnetin_3-O-2-O-feruloyl-glucuronopyranosyl-1-2-O-glucuronopyranoside
- Tricin_7-O-2-O-dehydrodiferuloyl-glucuronopyranosyl-1-2-O-glucuronopyranoside
- Apigenin_7-O-_2-O-feruloyl-glucopyranosyl-1-3-O-glucuronopyranosyl-1-2-O-glucuronopyranoside
- Chrysoeriol_7-O-_3-O-feruloyl-glucuronopyranosyl-1-2-O-glucopyranosyl-1-2-O-glucuronopyranoside

Results

Utworzone możliwe związki:

Wzór sumaryczny:	Masa: *[m/z]*	Delta: (względem szukanej masy) *[m/z]*	Znalezione związki:	SMILES:	Wzór strukturalny
C42H32N6O14	844.1888	0.0002	Zakres: 1 -> 2 Masa zakresu: 352.0918 -- Związek: C20H18O7 Masa związku: 370.105 PubChem: link Zakres: 2 -> 3 Masa zakresu: 176.0301 -- Związek: C6H6N6O2 Masa związku: 194.055 PubChem: link Zakres: 3 -> 0 Masa zakresu: 316.0633 -- Związek: C16H12O7 Masa związku: 316.058 PubChem: link	COc1ccc(-c2oc3cc(NC(=O)c4ncn5c(=O)n(C)nnc45)cc(Oc4ccc(C5OCC6C(c7ccc8c(c7)OCO8)OC(=O)C56)cc4OC)c3c(=O)c2O)cc1O	
C42H32N6O14	844.1888	0.0002	Zakres: 1 -> 2 Masa zakresu: 352.0918 -- Związek: C20H18O7 Masa związku: 370.105 PubChem: link Zakres: 2 -> 3 Masa zakresu: 176.0301 -- Związek: C6H6N6O2 Masa związku: 194.055 PubChem: link Zakres: 3 -> 0 Masa zakresu: 316.0633 -- Związek: C16H12O7 Masa związku: 316.058 PubChem: link	CCCCCC(=O)c1c(O)cc2c(c1Oc1cc(-c3oc4cc(O)cc(OC)c4c(=O)c3O)ccc1NC(=O)c1ncn3c(=O)n(C)nnc13)C(=O)c1c(O)cc(O)cc1C2=O	

Algorithm - results

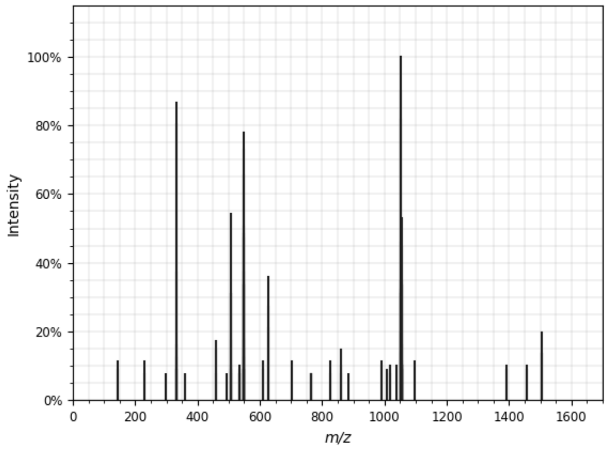
- in addition, the algorithm generates a log file in which we can read, among other things, the number of combinations for each spectrum range found and the total number of combinations resulting

Name	MIP = 30 [%]	MIP = 20 [%]
Spectra 1	423	177195
Spectra 2	1660	1723
Spectra 3	30	30
Spectra 4	872	262302
Spectra 5	0	147

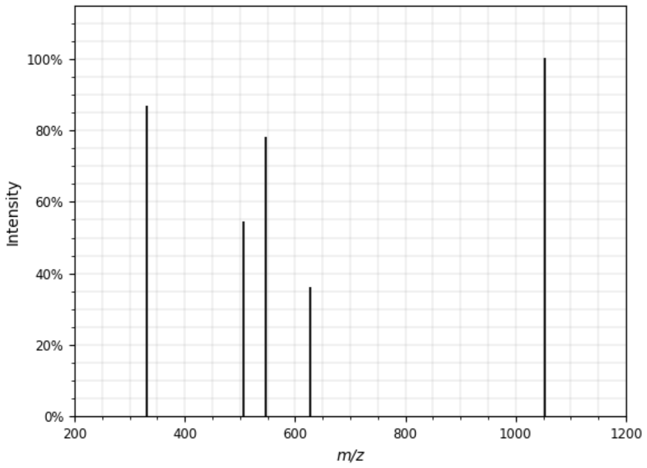
MIP - MIN_INTENSITY_PERCENTAGE

Algorithm - results

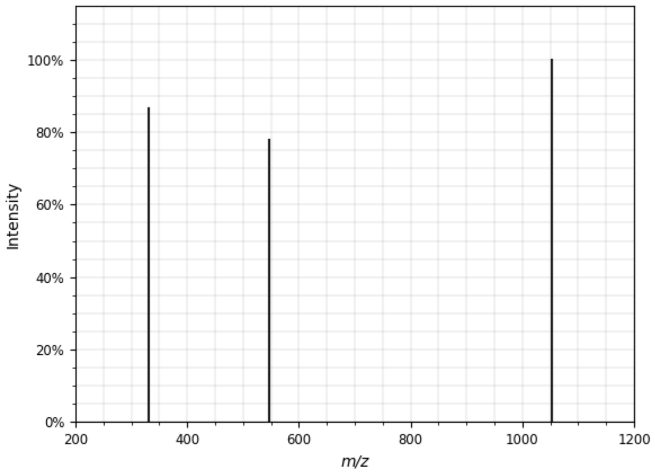
Oryginalne widmo:



Po odfiltrowaniu pików:
(MIN_INTENSITY_PERCENTAGE = 30)

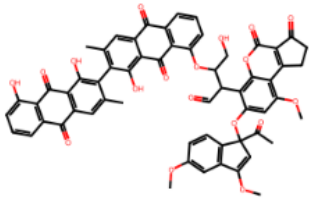
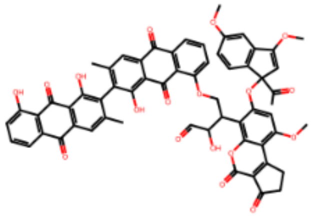


Po odfiltrowaniu pików:
(MIN_INTENSITY_PERCENTAGE = 60)



Algorithm - results

Utworzone możliwe związki:

Wzór sumaryczny:	**Masa:** *[m/z]*	**Delta: (względem szukanej masy)** *[m/z]*	**Znalezione związki:**	**SMILES:**	**Wzór strukturalny**
$C_{60}H_{44}O_{18}$	1052.2448	0.0152	Zakres: 1 -> 2 Masa zakresu: 506.0984 -- Związek: $C_{30}H_{18}O_8$ Masa związku: 506.1 PubChem: link Zakres: 2 -> 3 Masa zakresu: 216.0544 -- Związek: $C_{13}H_{14}O_4$ Masa związku: 234.089 PubChem: link Zakres: 3 -> 0 Masa zakresu: 330.0858 -- Związek: $C_{17}H_{16}O_8$ Masa związku: 348.085 PubChem: link	COC1=CC(Oc2cc(OC)c3c4c(c(=O)oc3c2C(C=O)C(CO)Oc2cccc3c2C(=O)c2c(cc(C)c(-c5c(C)cc6c(c5O)C(=O)c5c(O)cccc5C6=O)c2O)C3=O)C(=O)CC4)(C(C)=O)c2ccc(OC)cc21	
$C_{60}H_{44}O_{18}$	1052.2448	0.0152	Zakres: 1 -> 2 Masa zakresu: 506.0984 -- Związek: $C_{30}H_{18}O_8$ Masa związku: 506.1 PubChem: link Zakres: 2 -> 3 Masa zakresu: 216.0544 -- Związek: $C_{13}H_{14}O_4$ Masa związku: 234.089 PubChem: link Zakres: 3 -> 0 Masa zakresu: 330.0858 -- Związek: $C_{17}H_{16}O_8$ Masa związku: 348.085 PubChem: link	COC1=CC(Oc2cc(OC)c3c4c(c(=O)oc3c2C(COc2cccc3c2C(=O)c2c(cc(C)c(-c5c(C)cc6c(c5O)C(=O)c5c(O)cccc5C6=O)c2O)C3=O)C(O)C=O)C(=O)CC4)(C(C)=O)c2ccc(OC)cc21	

Summary

Conclusions

- The computational experiment whose results have been presented has been performed **on real biological data**.
- The results of the computational experiment clearly show that this approach **can be used only for small molecules**, which in fact is suitable to simple applications identifying the elemental composition of metabolites/lipids
- additionally, the researcher can obtain information on how the **chemical compound decomposed**

Future work

- Since˙ the number of potential aggregates for a given mass is significant, it would be feasible to widen the problem definition with **additional constraints**, i.e. taking into account information about **the structure of the fragmentation**.
- Conduct a computational experiment on a broader scale - with **other chemical compounds** in this group of compounds

Thank you…