Similarity-based learning

One of the most challenging tasks in non-targeted (LC or GC)/HRMS is unraveling the structure of the detected compounds. The main player here often is MS/MS spectra.

Task 1: matching an unknown MS/MS spectrum with a library

You are given a library of MS/MS spectra. It is a small part of the MassBank library (massbank.eu), so we can visually observe what is going on with the data and verify if the data treatment makes chemical sense. You are also given an MS/MS spectrum of an unknown compound. It is expected that this compound is present in the library.

PS! While solving the task, do not forget the mass accuracy!

Write functions to:

(1) to read in all files in the library directory and combine these into one tibble;

(2) obtain the similarity of two spectra;

(3) to calculate the similarity of the unknown spectra and the spectra of all compounds in the library.

Use these functions to identify the unknown compound with the given MS/MS spectrum.

Task 2: structural similarity from MS/MS spectra

The true unknown-unknowns are not in the libraries and can not be identified based on the spectral matching. However, here similarity-based learning is useful. Common features in the MS/MS spectra can be used to find the similarity of the compound and assign compounds that are similar to the unknown-unknowns.

(1) Based on your chemical knowledge, decide which MS/MS features (not specific masses, but mathematical manipulations of the MS/MS spectra) would make sense in finding similar compounds.
(2) Write functions to calculate these features for the library.
(3) Estimate which of the compounds are linked with which other compounds through these features.
(4) Visualize through a graph (ask for help from the supervisor).
(5) Now add the spectrum of unknown 2 and unknown 3 to the library and repeat the feature calculation and visualization. What do you think about which functional groups do these compounds contain?

Anneli