Meeting with Simon Rodgers – 11/03/2015

Introduction

The fellowship project is on how to use mass spectrometry fragmental analysis in an efficient manner to enhance metabolite annotations (and identifications) in complex sample matrices like urine and bacterial extracts.

Show December talk Polyomics as a flavour of what I am doing

Basic idea of the fellowship is:

* To empirically assess the data to find key combinations of fragments and/or losses that represent recurring subunits in the molecular data, i.e., the acylcarnitine combination of a product ion and a fragment (just published a paper on enhanced acylcarnitine annotation)
* Depending on the nature of the biological extract different core substructures will be present in the molecular profile – moreover, once established, all filters can be run on each data set to check if the substructures are present.

What I did so far:

* Tested/optimized different fragmentation settings on the Thermo Elite and Q-Exactive
* Compared spectra at different energies/combined multiple energies/from different machines
* Used molecular networking and MS2analyzer to analyse data:
  + Molecular networking collates very similar spectra in a node and compares all nodes using the cosine scoring
    - + graphical output
    - + unbiased comparison
    - - some isomers are grouped together in one node
    - - need more help/insight in how to manipulate the molecular network using Cytoscape plugins
  + MS2analyzer needs specific neutral losses, mzdifferences, and/or product ions as input and searches ALL the scans to find them and returns an Excel sheet
    - + returns Excel file that contains ‘0’s and ‘1’s (which can be sorted)
    - - needs *known* input (masses and EFs) to search in files
    - - no (direct) information about parent ion (only scan number)
* Started to work on a script following the steps described below

What I would like:

* Pull out key combinations of fragments and losses in an unbiased manner
* Annotate the key combinations with structural subunits using spectral databases and software tools using in silico fragmentation (like MAGMa and MetFrag)
* Use those ‘mass filters’ to ‘enhance metabolite annotation’ in complex sample extracts
  + I will run some data sets in the nearby future of both urine and bacterial extracts, so I can use the developed pipeline to analyse the data

The ideal:

* Perform untargeted fragmentation analysis and upload file into ‘Rscript’ that generates visual output of concurring fragments and losses
* Develop mass filters based on that using spectral database annotations and in silico fragmentation software tools (i.e., annotate mass filters with substructure)
* Annotate
* Report annotations using RMassBank (see below)

What I need:

Assistance with creating an ‘Rbased script’ to facilitate automated search for key combinations:

This will involve some trial and error and optimization of parameters – it will be based on a code written by Fraser Morton (see below under ‘Rscript in development’) which pulls out fragments, but not yet neutral losses or mzdifferences.

Some help with scripting and implementation of existing scripts:

RMassbank: this script helps to convert fragmentation spectra into MassBank (spectral database) – it would be really useful if we can implement that script into the pipeline here so that we can easily publish fragmentation spectra of standards as well as spectra from the urine and bacterial extracts.

Cytoscape: implementation of some plugins and how to run the molecular networking from the command line

Questions regarding the fragmentation data:

* What happens with the ‘scan header ‘000.0000 m/z’ scans? Can we find a way to include those spectra (and thus information!) as well?
* How to implement neutral losses and MZdifferences
* How to effectively pull out the combinations? (start with a heatmap, other ideas?)
* Can we implement existing molecular formula generators to assign masses and fragments with elemental formulas in a meaningful manner?

WIFY (What is in for you)

* I can teach the PhD/MSc student about metabolomics / mass spectrometry and fragmentation
* Author/Co-author of (hopefully) resulting manuscript(s)
* Student (co-)supervision
* Access to data sets
* Application of text mining software/machine learning algorithm (LDA) to mass fragmentation data

WIFM (What is in for me)

* The PhD/MSc student can teach me a bit on how to use ‘R’ (basic stuff – so I can run the scripts and make minor modifications)
* Good discussions on how to automate my fragmentation pipelines
* ‘Rscripts’ that can test my hypothesis that it is possible to find key combinations of fragments/losses in an unbiased (i.e., empirical, data-driven) manner in a data set.

Rscript in development (with/by Fraser Morton) aiming to perform the following steps:

Step 1: read fragmentation mzXML files in ‘R’ and separate MS1 and MS2-n scans

* Define best format and how the data is organized

Step 2: perform peak picking in MS1 [this can then be aligned with the full scan data from samples within a metabolomics batch]

* Output: peak list

Step 3: filter on MS1 level

* Filters that can be used, with values for standard pHILIC runs on the Q-Exactive:
  + Retention time window: 3.0 – 18.0 min
  + Mininum intensity: 3E5 cts (peaks below that intensity are not selected for fragmentation anyway)
  + Mass accuracy MS1: 3 ppm

Step 4\*: clean MS2 spectra

In order to remove noise and low intense fragments:

* Filters that can be used, with values for Q-Exactive:
  + Absolute intensity filter: 5E3 cts
  + Relative intensity filter: ≥5% of the base peak
  + And possibly a maximum number of fragments, i.e. Top20 (intensity wise)….

Step 5\*: assign one MS2 spectrum to each peak picked from MS1

To prevent overrepresentation of some fragmented metabolites in the concurrent analysis *and* select the most intense fragmentation spectrum for search in spectral databases via PiMP. See figure 1 below to get an idea about repeat counts for a fragmentation file in untargeted mode for a QC\_Beer injection in positive mode, also indicating that most mass features are fragmented only once….but(!!) as can be seen in Figure 2, as soon as no charge state can be assigned by Thermo Xcalibur, then the ‘accurate precursor mass’ is set to 0. This is something to take into account if we want to use the precursor value in the scan header.

* I could imagine a search for ms2 scans with the precursor mass based on a retention time window that takes the precursor intensity into account (i.e. > 3E5 in MS1) and the nominal mass values (as the scan title information, i.e., 125.05 is representive for all masses fragmented at 125 m/z) – it then should assign the mass fragmentation spectra with the most intense base peak to the peak picked in MS1.

Step 6: create a large table of precursor ion\_retention time combinations (i.e., precursors) versus mass fragments and neutral losses with their intensities [or presence/absence]

* The mass fragments can be the experimental mass fragments, binned using mass filters:
  + For m/z >70, 3 ppm accuracy can be used, for m/z<70, 6 ppm should be used….
  + In a later stage, we could try to assign elemental formulae to the fragments, which is helpful for elemental formula assignment of the precursor.
* The neutral losses are the mass differences between two mass fragments or the parent ion and the mass fragments. Several combinations are possible, and with 10 fragments present, the list soon becomes very long, so I would restrict this in first instance to the 15 – 20 most intense neutral losses. Using the accurate mass of the precursor from the peak picking in MS1(!), we can define the neutral losses between precursor and fragments.

Step 7: create heat map of mass fragments and neutral losses

* Concurrence analysis: ideally, a heat map of fragments and neutral losses would be constructed, displaying ‘hot areas’ of concurrent fragments/losses that can then be annotated with structural subunits. The value for PiMP is that in future, users could annotate, e.g., histidine related peaks in their data and then see if in their data set anything relevant happens with those metabolites (like metabolic maps, but then based on fragment annotation).

Step 8: implement mass filter search (OR use MS2analyzer by creating a script that produces the text input files for MS2analyzer, and reads in and the resulting Excel file and visualizes it somehow….)

* create script that searches for an annotated combination of mass fragments/neutral losses in fragmentation data

\*these steps can be done in any order, whatever is more convenient for processing….