ShinyLipidSearchMatrixParser

# What it does

1. Reads positive matrix
2. Reads negative matrix
3. Merges matrices
4. Calculates aggregate values per feature
5. Filters
6. Exports

# Output

1. Merged sample list
2. Merged feature list with aggregate values for Grade, Occupancy …, Mz, Rt
3. Feature/Sample matrices for all the numeric/ordinal data (Area, Height, S/N, but also Grade etc)

# Aggregate values

1. Grade = best (lowest) ‘Grade’
2. Occupancy = best (highest) ‘Occupancy’
3. Rt, ObsMz = median
4. Number of points = median

For a given feature, ‘Grade’ is the best grade of all samples, and in practice the best grade of the MS^2 runs.

We need these single (scalar) values for filtering.

# Filter

1. Grade, separately for Pos and Neg mode
2. Occupancy, number of points

Grade is very different in Pos and Neg mode. E.g. PCs, Pos mode median is “C”, Neg mode all better than “C”.

The filtering is based on the aggregate values calculated in ‘Features’.

In addition, adducts that conflict with the mode (Pos/Neg) are excluded always. For example, if M+H appears in the Neg mode set, that feature will always be filtered out.

# Issues

1. If the sample sets differ, as they do in the Maarouf data set (pools are named differently in pos and neg), NAs are produced. For example, if samp6-1 was run in Pos mode but not in Neg (called samp5-1 perhaps?), then Area, Height, etc for all Neg features will be NA for this column.
2. Some columns may be missing in ‘Features’ in either pos or neg mode. Will be filled by NAs. Probably trivial, e.g. ‘FA4’, when there are no cardiolipins.
3. There is no topRT that could be used in a linear model. Need to fit against aggregate value, e.g. median(Rt).
4. When are the control samples (MS^2) to be removed?

# What’s next

The output of this shiny is intended to be used by a choice of additional tools that still need to be developed.

1. Linear model filtering
2. Feature aggregation
3. Big-data heatmap viewer
4. Statistics and plots, essentially supershiny

The idea is to read from and write to a generic format, which consists of an excel sheet with a Samples table, a Features table, and then numeric data in matrix form.

The new strategy is to break down the workflow. In the end the user needs to decide at what stage she wants to e.g. merge pos and neg features of the same lipid molecule, or remove features that do not agree well with the linear model.

Technical issues

* Remove controls
* Filename without \_pos or \_neg more reliable than sample\_id
* Where are the group matrices like GroupTopPos?
* Maybe look at Cem data (many tissues …)