Organisation of the suite of RStudio scripts

For running the RStudio suite of programs (**Analysis**, **CheckRep**, **Quantify**, **function**), you MUST have a folder containing four subfolders:

* “*analysis*”, which contains the scripts or programs (analysis, CheckRep, ….)
* “*data*”, which is the default folder where the data files for each script are expected to be found
* “*results*”, with two subfolders names “*figs*” and “*tables*”, where the results files are stored.

**Two input files**, ***TASKTABLE*** and ***TGTABLE***, are expected for running **Analysis**. Running **CheckRep** or **Quantify** requires ***TASKTABLE*** and ***QUANTTABLE***,

**Where to put your MS\_files and DMS\_files? Where to find your output files?**

The file names are listed in ***TASKTABLE***. ***TGTABLE***. In a sake of simplicity, these folders (repositories) are defined in the “functions” script/program as:

* # Define Data and Results repositories
* dataRepo = '../data/'
* figRepo = '../results/figs/'
* tabRepo = '../results/tables/'

Organisation of TASKTABLE and TGTABLE

The ***TASKTABLE*** is a “coma” (,) delimited text file. It can be edited using excel or Rstudio (safer). Note that you can insert “# “ at the beginning of a line, which then will be considered as a comment. ***TASKTABLE*** looks like this:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MS\_file | DMS\_file | t0 | CV0 | dilu | Path |
| C0\_AS\_DV-1800\_1.d.ascii | Fichier\_Dims 20190517-000000.txt | 0.08 | 6 | 0 | Esquire\_MSMS\_Data/2019\_A\_Voir/20190517\_AA/ |

Where:

* C0\_AS\_DV-1800\_1.d.ascii is an ASCII file, extracted using DATAANALYSIS. So far, only the ESQUIRE data files extracted using the “profile” option can be handled.
* Fichier\_Dims 20190517-000000.txt is the corresponding DMS file.
* The t0 and CV0 values are used to convert the ESQUIRE time t values into DMS CV value.
* The dilu value was initially meant to be the dilution factor of the standard metabolites when spiked into a plasma (see CheckRep). When you perform another type of experiment, you can use this dilu value as an index to specify, for example, the flow-rate of the modifier, the day of the experiment, the set of samples, ….
* The Path value allows you to organize your data within the "../data/" folder. Note that the DMS\_files must be in the "../data" directory. In the present case, only the MS\_files are expected to be found in the following directory: "../data/Esquire\_MSMS\_Data/2019\_A\_Voir/20190517\_AA/".

**Important**: the date (here 20190517) from the DMS\_file and part of the name (here “C0\_AS\_DV-1800\_1”) are combined to name the output files as "20190517\_ C0\_AS\_DV-1800\_1.results", for example. You will also find this “tag” in the output figures.

The ***TGTABLE***  is a “semicolon” (;) delimited text file. It can be edited using excel or Rstudio (safer). Note that you can insert “# “ at the beginning of a line, as in the example below. This line will be considered as a comment, which means that in the present case, Glycine will not be analyzed. ***TGTABLE*** looks like this:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | m/z\_EExact | m/z\_exact | CV\_ref |
| # Gly-AA | C2H5NO2H | 76 | -10.7 |
| Ala-AA | 90.054955 | 90.1 | -7.6 |

Where:

* Name is the given name of a metabolite,
* “m/z\_EExact” is actually not used
* m/z\_exact can actually be an approximate m/z value
* CV\_ref is the expected CV value (can be omitted)

Analysis (last version from 2020, July 16)

**Purpose**: for each DMS-MS/MS experiment as given in a series in the ***TASKTABLE*** file, a series of metabolites given in ***TGTABLE*** is analyzed. Ultimately, the aim of the analysis is to integrate the peak (i.e. to derive the area) corresponding to each metabolite.

From the two dimensional data (m/z, CV), the area can be extracted using a 2D\_fit where the fit function is the product of two Gaussian type functions, one in the m/z, the other in the CV dimension. It turns out that we need three types of fit:

* 2D\_fit in the (m/z, CV) space
* 1D\_fit in the (CV) space, assuming that the m/z value is the m/z\_exact given in ***TGTABLE***
* 1D\_fit in the m/z space, assuming that the CV value is the CV\_ref given in the ***TGTABLE***

The choice of fit type is set using the “fit\_dim” variable.

More generally, the important user configuration parameters are listed within the first line of the **Analysis** script as follow:

|  |
| --- |
| # User configuration params ####  taskTable = 'files\_quantification\_2019.csv'  tgTable = 'targets\_paper.csv'  filter\_results = TRUE  fwhm\_mz\_min = 0.1  fwhm\_mz\_max = 0.5  fwhm\_cv\_min = 0.5  fwhm\_cv\_max = 1.5  area\_min = 10  save\_figures = TRUE  plot\_maps = FALSE  fit\_dim = 0 # 2: fit 2D peaks; 1: fit 1D CV line; 0: fit 1D m/z line  fallback = TRUE # Fallback on fit\_dim=1 fit if 2D fit fails  weighted\_fit = FALSE  refine\_CV0 = TRUE  const\_fwhm = ifelse(fit\_dim == 0,NA,0.7)  dmz = 1.0 # Width of mz window around  # exact mz for signal averaging  dCV = 1.2 # Width of CV window around  # reference CV for peak fit |

The general formulae of a Gaussian function is: . Where a is the area, x0 is the position of the peak, and σ is related to the full width at maximum (fwhm) : . Upon the fit process of the data, the area is optimized, as well as peak width (x0 and σ). In order to control the fit, limit ranges are defined (fwhm\_mz\_min, … area\_min). When optimizing CV or m/z position of the Gaussian functions, the window range can be specified using dCV and dmz.

The fit dimension and type is specified using the fit\_dim variable:

|  |  |
| --- | --- |
| Fit\_dim = 2 | A two\_dimensional fit is performed |
| Fit\_dim = 1 | A 1D fit in the CV dimension is performed. If fallback=true, and if the Fit\_dim=2 fails (does not converge), a 1D\_fit (Fit\_dim=1) is performed |
| Fit\_dim = 0 | Initially named “fast”, this is also a 1D fit in the m/z dimension at fixed CV\_ref. |

The save\_figures and plot\_maps logical variables can be used to save diskspace and time, respectively.

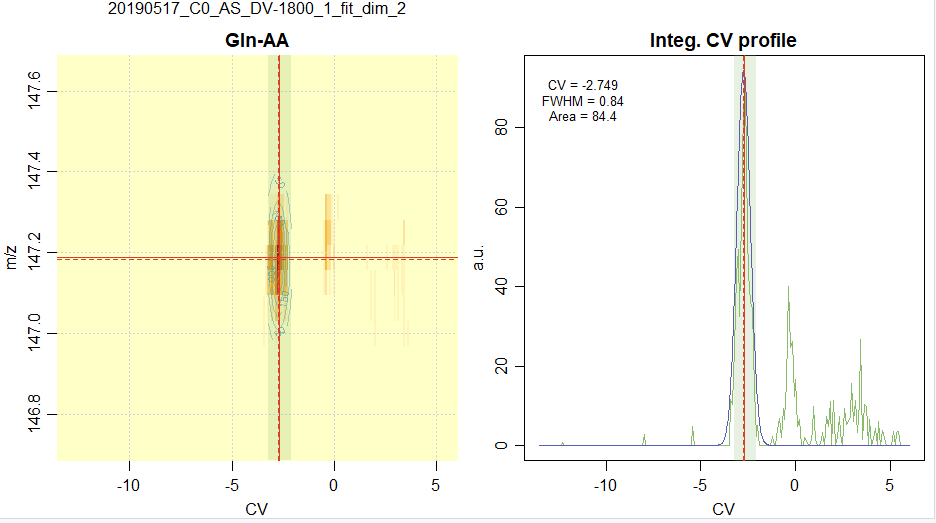
Analysis: screen and files outputs

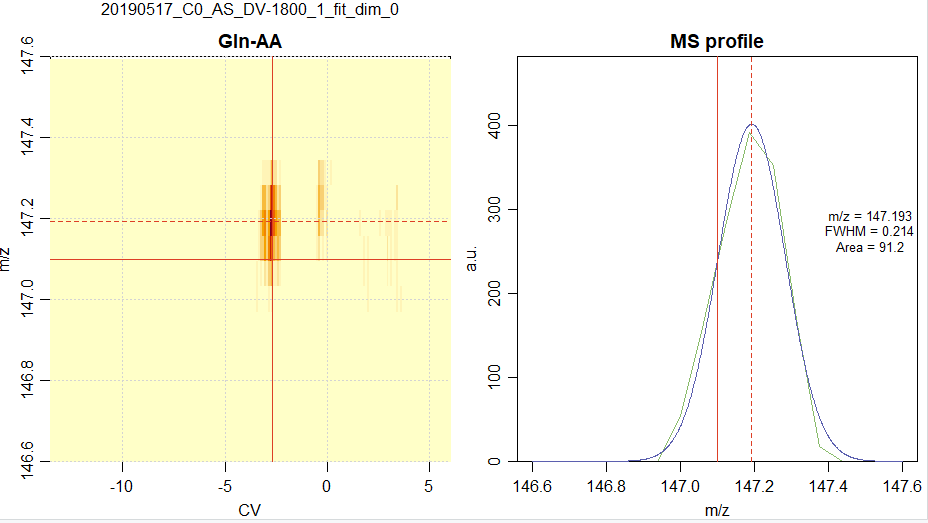
The output files can be found in the following repositories:

* figRepo = '../results/figs/'
* tabRepo = '../results/tables/'

**Figures**

For each experiment and target, you get a figure (on the screen and as a file) which looks like:

For a 2D fit: 

For a 1D fit (m/z), i.e. “fast”: 

**Tables**

For each experiments associated with (MS\_file, DMS\_file), three '.csv' files are generated. If your data are (MS\_file= C0\_AS\_DV-1800\_1.d.ascii, DMS\_file= Fichier\_Dims 20190517-000000.txt), and if fit\_dim=2, you get:



The results file, you will get the following columns. The first 4 columns are copies of the **TGTABLE** data:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | m/z\_EExact | m/z\_exact | CV\_ref |

The next 8 correspond to the position, width and corresponding dispersion values of the optimized Gaussian in the m/z and CV dimensions.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| m/z | u\_m/z | CV | u\_CV | FWHM\_m/z | u\_FWHM\_m/z | FWHM\_CV | u\_FWHM\_CV |

The next two columns are the results for the optimized Area values, and corresponding dispersion values.

|  |  |
| --- | --- |
| Area | u\_Area |

Finally, you will find the “fit\_dim” value, the “dilu” index, and the “tag” wich is a concatenation of date+MS\_filename+dim\_fit.