**PeakFilter input file preparation**

# Purpose

PeakFilter expects input data to be pre-aligned and framed using SIEVE™ or XCMS and stored in a comma-separated values (CSV) file. There is currently little error handling within PeakFilter so it is imperative that the input files are in the exact format expected to avoid crashes or erroneous processing. This guide details how to prepare and format data ready to be processed with PeakFilter.

# SIEVE™ file pre-processing

The SIEVE™ software aligns profile mode mass spectrometry data, which must be input in **raw** format. The following steps will guide you through the pre-processing phase:

1. Open Sieve 2.2, create a **New Experiment** and click on **Next**.
2. In the next window, select:

***Domain:*** Small molecules.

***Experiment Type:*** Non-differential Single Class Analysis.

1. Upload the desired **raw** files. Afterwards, choose one of them as reference.
2. Frame parameter adjustment:

Retention Time Start (min): 01.00

Retention Time Stop (min): 57.00

M/Z Min: 100

M/Z Max: 1800

Frame Time Width (min): 0.70

M/Z Width (ppm): 10

1. Frame selection:

Maximum Number of Frames: 1000000

Peak Intensity Threshold: 50000

1. Click on **Next** on Scan filter selection.
2. In Identification Parameters, select **Defer Identification**.
3. Click on **Finish**.
4. SIEVE Parameters:

*Can modify parameters already set from this screen if necessary.*

0. Global Settings > PCA Process: DISABLE

1. Alignment Parameters > AlignmentMinIntensity: 10000

1. Alignment Parameters > MaxRTShift: 0.7

1. Alignment Parameters > Tile Size: 300

1. Click on **Update** and **Run as Workflow**. It will run the alignment and the framing automatically.

This process may take several days, especially when using a computer with less processing power. Analyses can be split into time periods if a long runtime is undesirable (see [*File formatting considerations*](#_File_formatting_considerations) section for details).

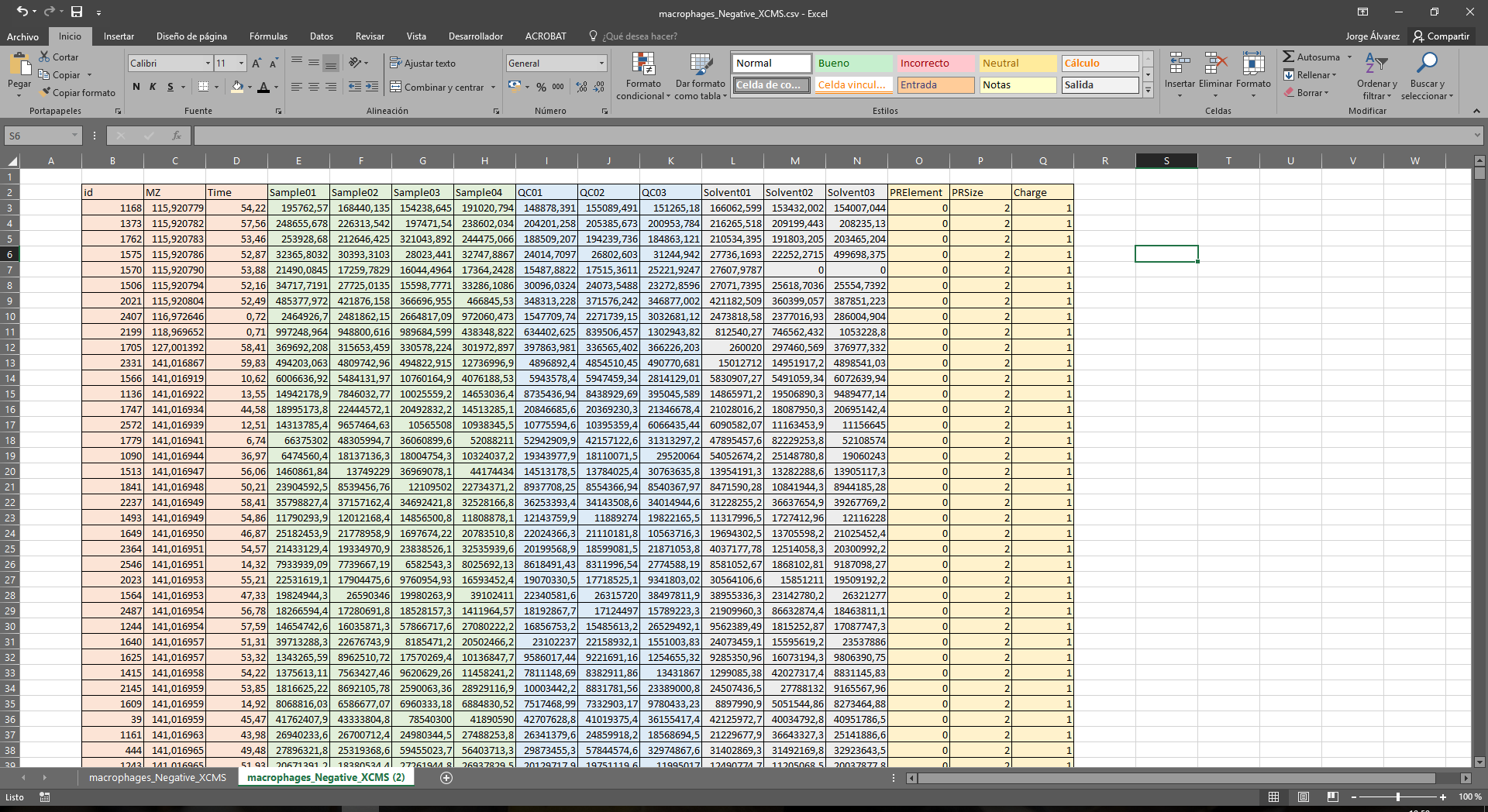
1. Once the workflow has finished a table with feature *ID*, *MZ* and *Time* will be presented. Choose the desired columns via the **Field Dialog Box** (right clicking on the table headings) together with *ID*, *MZ*, *Time*, *PRElement*, *PRSize*, *Charge*, the integrated intensity columns, which are necessary to continue.
2. Finally, **Export top level table view** to Excel.

A feature of SIEVE™ is its ability to identify isotopic distributions of the same ion and infer its charge. Currently for processing with PeakFilter we use only the monoisotopic masses and ions with a single charge. We need to filter out higher masses and charges with the following steps:

1. Open the exported Excel file.
2. To avoid future problems, ensure that the *id* header is in **lower case**.
3. Apply the following filters to *PRElement*, *PRSize* and *Charge* columns:
   1. PRElement **<** 1
   2. PRSize **>** 1
   3. Charge **<** 2
4. The rows shown now are the ones you need **to keep**.
5. Delete *PRElement*, *PRSize* and *Charge* columns.
6. **Save the file as** comma-separated values (**CSV**) file.

Typically, a small number of frames may be allocated a *PRElement* of -1 and a *Charge* of 0. With the previous procedure, we retain these peculiar data.

At the end, the CSV file should look like the example below, where there are 2 mandatory and 3 optional regions that **must follow the order displayed**.



1. Mandatory frame data

2. Mandatory sample replicates

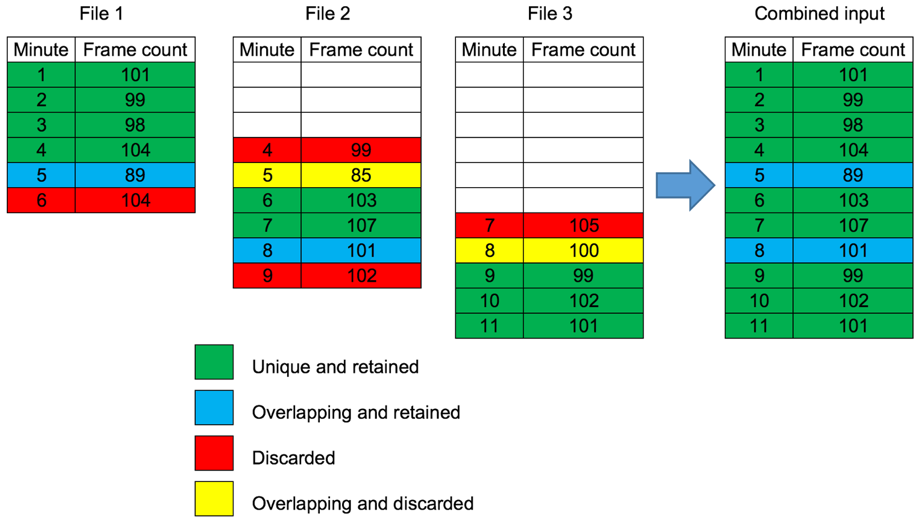
3. Optional QC samples

4. Optional solvent samples

5. Optional additional data

1. Mandatory frame data (in this exact order, **case sensitive**):
   1. *id*: SIEVE™ frame ID
   2. *MZ*: m/z values
   3. *Time*: retention times
2. Mandatory sample replicate intensities: the **sample replicates** must be **grouped** together and each sample must have the **same number of replicates**.
3. Optional QC samples
4. Optional solvent samples
5. Optional additional data: any number of columns of additional and uniquely labelled fields that store additional data about the frame.

### File formatting considerations

1. All **retention times** should be **greater than zero**.
2. PeakFilter supports the use of multiple files split by time ranges to represent a single run. However, note that except for the first retention time minute of the first file and the last retention time minute of the last file, all first and last minutes are trimmed from the data since they are unreliable. The file import procedure supports overlap (after trimming). Where retention times overlap (in minute chunks), the frames retained are those from the file with the most frames for that minute. We show a trivial example below:

* Ensure that samples with multiple replicates are given names in the format name1, name2, etc. such that name is unique for each sample. Replicates should be suffixed 1, 2, 3, 4...
* Where samples have single replicates, ensure that if their name does not end in a number.
* Field names should only consist of alphanumeric characters, “-” and “\_”.
* Ensure there are no additional ghost columns: this usually occurs by deleting the contents of a column but not deleting the column itself, and can be seen in a text editor as trailing commas after the final data column.

# XCMS file pre-processing

The following steps will guide you through the pre-processing phase with XCMS:

1. Generate **mzXML** files from your **raw** files with a software like *ProteoWizard*.

There is a useful guide provided by XMCS about this topic [here](https://xcmsonline.scripps.edu/docs/fileformats.html).

1. The mzXML files must then be located in two subfolders: one must contain the biological samples (and quality control samples if any), and the other must contain the blank (solvent) samples.
2. Run the R script we have developed to pre-process your files with XCMS (included in the same folder as this document). The script will ask for the folder were the two previous subfolders are located, as well as the name to give to the output CSV file.

Note that the functions included in the script have **several parameters that should be tailored** to your experimental conditions. Depending on the number of files, their size and the computer where it is being run, the script might take several hours to finish.

1. Ensure the **retention time** is in **seconds** before running PeakFilter.

# Input file parameters

Although PeakFilter prompts for details about the input file(s) format at runtime it is convenient to alter the parameters file to reflect the input file if a number of files are to be processed.

* **firstRepOffset**: this parameter tells PeakFilter where to find the 1st sample intensity replicate. Its minimum value is 3 (SIEVE™ files) and it should be changed to 3 + ***n***, where ***n*** is the number of additional data columns, in XCMS files.
* **numberOfSamples**: the number of samples in the dataset where a sample is a group of technical replicates. Each sample must have the same number of replicates.
* **numberOfTechReps**: the number of technical replicates of each sample.
* **numberOfQCReps**: the number of optional QC replicates.
* **numberOfSolventReps**: the number of optional extracted solvent replicates.
* **filePolarityMode**: this is the mode that the samples were processed on the mass spectrometer. Not only is this used to label columns in the final output summary file ready for amalgamator but the way the file is processed in regards to adducts and stack is dependent on this parameter.
* **columnType**: this is an indication of the column type and it is used purely for labelling the final output summary file.