**PeakFilter input file preparation**

Purpose

PeakFilter expects input data to be pre-aligned and framed using SIEVE and stored in a comma separated (csv) format. 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 for processing with PeakFilter.

Sieve file processing

The SIEVE software aligns profile mode mass spectrometry data, which must be input in the .raw format. To begin the process, select ‘small molecule analysis’ and ‘single non-differential analysis’. Upload raw files at the prompt and select a file as the reference.

1. **Frame parameter adjustment**

Retention time start: 01:00

Retention time stop: 57:00

Frame width: 0.70

M/Z minimum: 100

M/Z maximum: 1800

M/Z width: 10

1. **Frame selection**

Max number of frames: 1, 000,000

Peak intensity threshold: 50,000

1. **Identification parameters**

Can modify parameters already set from this screen if necessary.

Tile size: 300

Max retention time shift: 0.7 minutes

Turn PCA analysis off

Min threshold for alignment: 10,000

1. **SIEVE parameters summery**

You can modify parameters already set from this screen if necessary.

Tile size: 300

Max retention time shift: 0.7 minutes

Turn PCA analysis off

Min threshold for alignment: 10,000

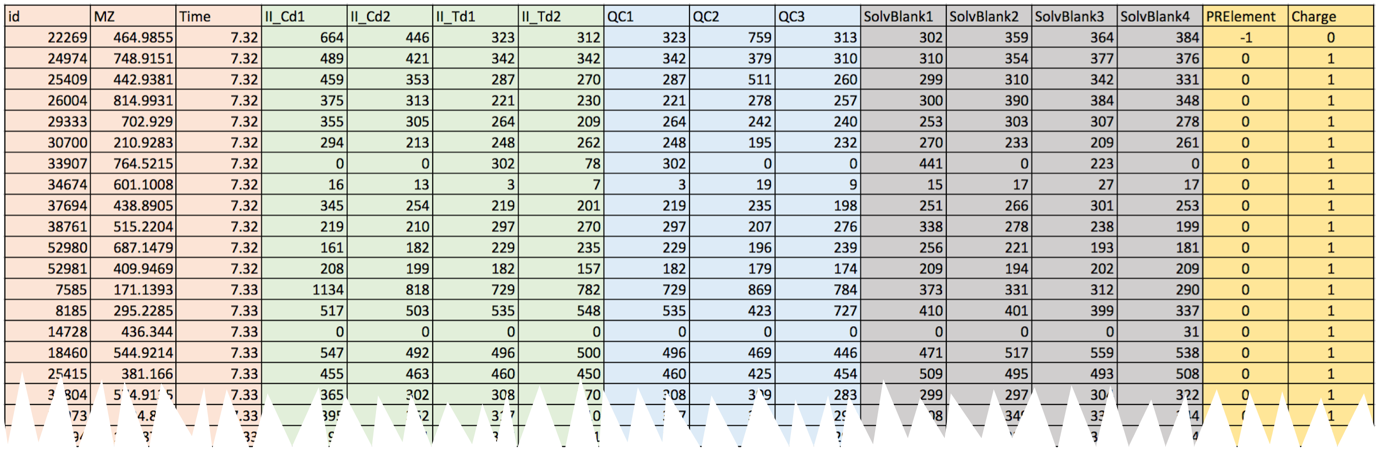
Update to save changes. Run Alignment and then return to the ‘Process’ tab to start the Framing. After this is complete, a table with feature ID, MZ and Time will be presented. Select the appropriate columns. ID, MZ, Time are necessary to continue. Also select the integrated intensity columns by right clicking on any of the features and choose PRelement, PRsize and Charge via the Field Dialog Box, which can be accessed by right clicking on the table headings.

This process may take several days, especially when using a computer with less processing power. Analyses may be split into time periods (see file formatting considerations for details) if a long continuous run time is undesirable.

Exporting data from SIEVE

After processing has completed in SIEVE the data should be arranged and exported as a csv file. There are 6 possible regions of data for each frame within the input file, 3 are optional and 3 are mandatory. The fields **must** appear in the csv in the order below. Field names in quotes must occur exactly as specified - all field names are case sensitive.

1. **Optional additional data:**
   * Any number of columns of additional, uniquely labelled fields that can be used to store additional data about the frame.
2. **Mandatory frame data:**
   * ‘id' – The sieve frame ID
   * 'MZ' – The m/z.
   * 'Time' – The retention time.
3. **Mandatory sample replicate intensites:**
   * Any number of samples replicates are grouped together. Each sample **must** have the same number of replicates.
4. **Optional QC samples**
5. **Optional solvent samples**
6. **Mandatory isotope data**
   * ‘PRElement’ - The number of 13C in the molecule in that frame
   * Charge – The charge on the ion



6: Mandatory isotope data

4: Optional QC samples

2: Mandatory frame data

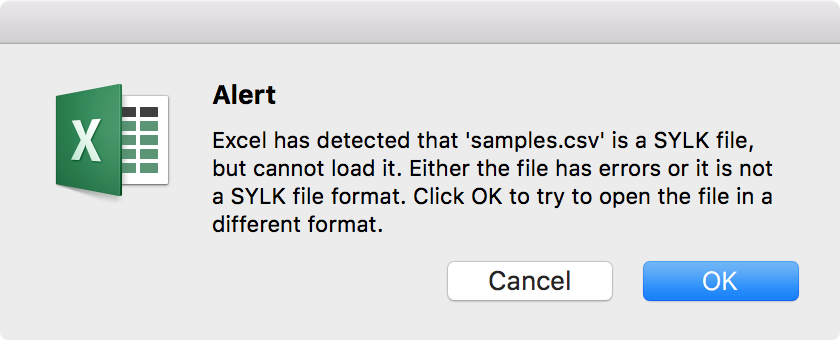
3: Mandatory sample replicates

5: Optional solvent samples

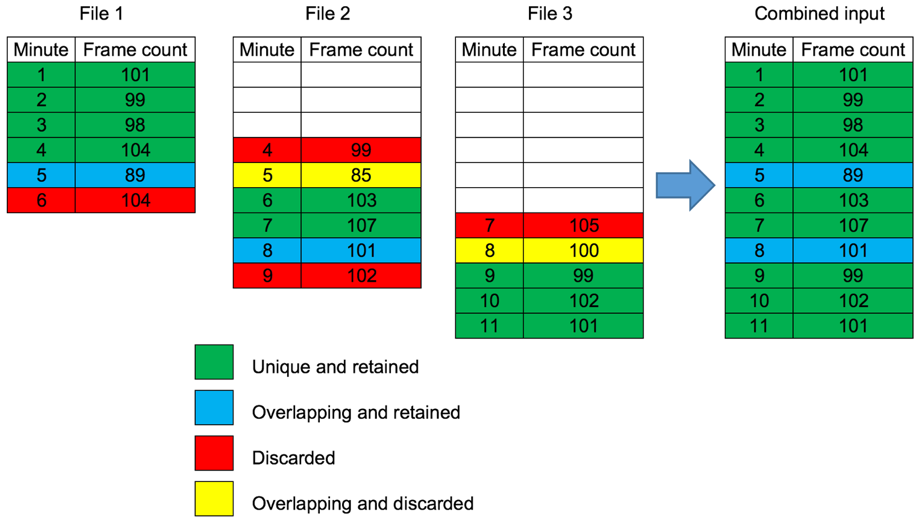
1: Optional additional data

File formatting considerations

1. When exporting the csv file from SIEVE the frame ID will have the heading ‘ID’ in upper case. This should be changed in a text editing package to ‘id’ especially if there is no optional additional data (see input file preparation below). Attempting to open the csv file in excel if cell A1 contains ‘ID’ will result in the following error.



1. Column wise there should be no gaps between sections or between columns within sections.
2. All retention times should be greater than 0.
3. 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 as 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. (see trivial example below)



1. 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...
2. Where samples have single replicates ensure that if their name does not end in a number.
3. Field names should only consist of alphanumeric characters, ‘-‘ and ‘\_’.
4. If the cvs file is manipulated in microsoft excel prior to PeakFilter ensure no additional ghost columns are saved , this usually occurs by deleting the contents of a column but not deleting the column and can be seen in a text editor as trailing commas after the final data column.

Isotopic and charge filtering

A feature of SIEVE is its ability to identify isotopic distributions of the same ion and infer it’s charge. Currently for processing with PeakFilter we use only the monoisotopic masses and ions with a single charge. Higher masses and charges are filtered out by following the procedure below. Typically a small number of frames may be allocated a PRElement of -1 and a charge of 0, currently we retain these peculiar data.

Pre-process the csv file in Microsoft excel as follow:

1. Ensure the ‘id’ header is in lower case.
2. Open the file in excel.
3. Add filters to the ‘PRElement’ and ‘Charge’ columns.
4. Filter the ‘Charge’ column where value <2.
5. Filter the ‘PRElement’ column where value <1.
6. Delete the remaining rows.
7. Delete the ‘PRElement’ and ‘Charge’ columns.
8. Save the resulting dataset as a csv file.

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. Where there are no optional additional data columns the value is 3 and should be changed to 3 + n where n is the number of additional data columns.
* 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. The example above has 2 samples.
* numberOfTechReps – The number of technical replicates of each sample. The above example has 2 replicates per sample.
* numberOfQCReps – The number of optional QC replicates. The example has 3 QC replicates.
* numberOfSolventReps - The number of optional extracted solvent replicates. The example has 4 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 dependant on this parameter.
* columnType – This is an indication of the column type, it is used purely for labelling in the final output summary file so if a column is used that can separate both polar and non-polar analytes then either column type may be chosen.