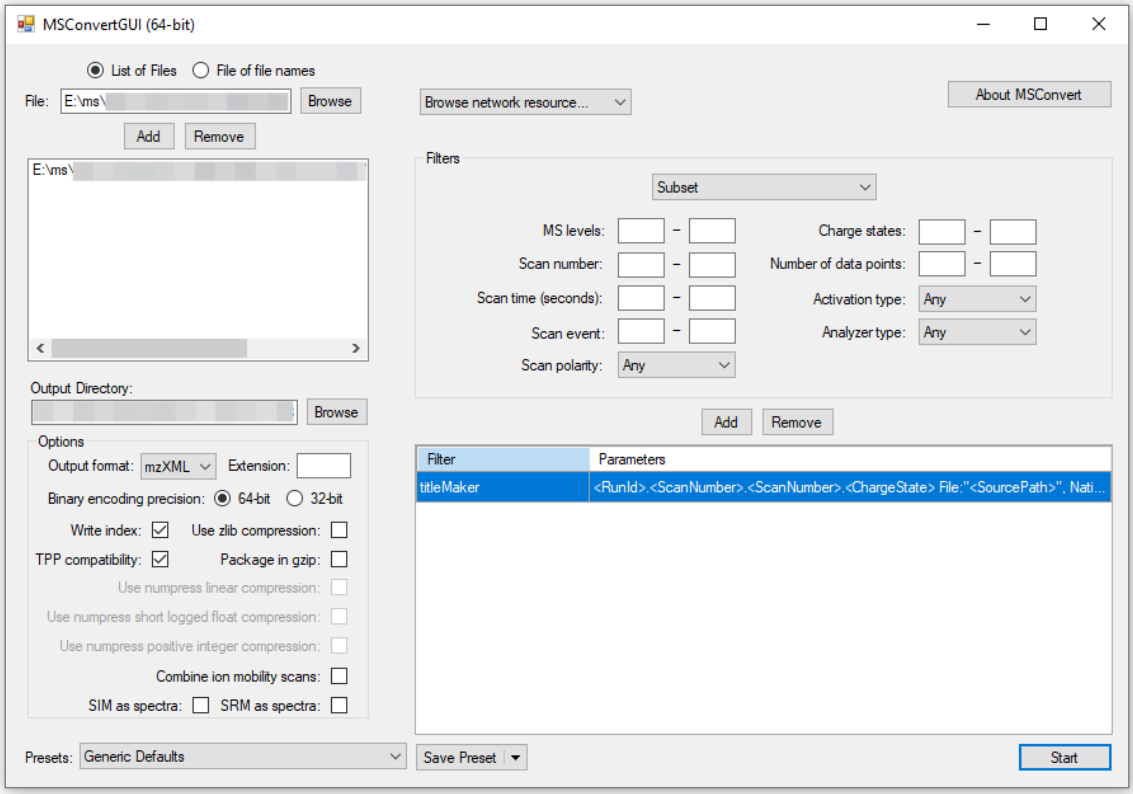
User manual

Data examples could be downloaded from here: https://yadi.sk/d/fswyNmd8taKGmw

1. Convert any mass spectra files to files in mzXML format via MsConvertGUI from the ProteoWizard project, which is available from here http://proteowizard.sourceforge.net/download.html

It requires Microsoft.NET 4.7.2 Framework to be installed.

The program should be used for raw (MS1 level) files to be converted with the parameters that are indicated in the screenshot below:



The parameters could vary due to the version of the MSConverterGUI. The most important parameters are in the Options block.

2. Convert one or multiple mzXML files to .mat format via executable file mzXML2mat.exe (https://yadi.sk/d/sGVGPpEjst9Ujw). Requires 1GB of free space in the system partition for launch. First, select the desired mzXML files in open file dialog, then follow the instructions in the command window and wait until the process is finished. It will create a “data.mat” file in the same directory, where mzXML files were taken from. The file could be renamed in future by the user. The parameters are listed below. Conversion will be processed after setting the parameters. Wait for the "Press any key to exit... " message.

m/z min: **float, optional**

Lowest m/z cutoff: Mz lower than *m/z min* would not be converted and processed. The default value is 200. Acceptable values m/z > 0.

m/z max: **float, optional**

Highest m/z cutoff: Mz higher than *m/z max* would not be converted and processed. The default value is 1000. Acceptable values m/z > *m/z min*.

m/z bin: **float, optional**

The bin width of m/z array: a step of m/z array, which will be used for binning spectra to the vector in space from *m/z min* to *m/z max* with step *m/z bin*. The default value is 0.1. Acceptable values m/z < (*m/z max - m/z min*).

median width: **int, optional**

A number of scans to be averaged with a moving median filter (https://doi.org/10.1038/s41598-018-37560-0). Moving median filter width and step are equal to *median width*. The default value is *None*. If the value is *None* then *median width* is recalculated in a way to binned medianed spectra not to exceed allowed space in RAM. Acceptable values from 0 to the minimum number of scans in measurements.

GB_max: **int, optional**

The allowed maximum of RAM space (in GB) to be allocated for scans stored in measurements. The default value is 2. Acceptable values from 1 to the actual free RAM space.

Notes:

If *n_median* is specified, the number of scans to be taken into account from each measurement will be calculated due to RAM limits. So, only the first scans from each mzXML file will be processed for the RAM limit to be not exceeded. For example, if each scan in any measurement would be 1GB size, n_median=1, and each measurement consists of 10 scans (total 2 measurements), then only the first scan of each measurement will be placed in the output dataset due to the 2GB RAM limit. Whereas skipping n_median (n_median=None) would lead to automated n_median calculation, which means n_median would be set to 10 in this example.

Total number of scans in output (N) affects the time or even possibility of the SSM_calculation to make calculations. Higher N will lead to the size of calculated SSM as N^2. Thus, the user should control the N for not to exceed RAM in the SSM_calculation process.

The python script mzXML2mat.py could be used otherwise for the same purposes in the same way. The script is developed in Python3.5.2 with packages described in Requirements.txt.
The most important function here is read_and_convert_data(filenames, mzmin=200, mzmax=1000, mz_bin=0.1, n_median=None, GB_max=2), which performs calculating m/z array for binned spectra; binned spectra; median filtering of binned spectra.

Parameters:

filenames: **list**

A list of mzXML files.

mzmin: **float, optional**

Lowest m/z cutoff: Mz lower than *mzmin* would not be converted and processed. The default value is 200. Acceptable values m/z > 0.

mzmax: **float, optional**

Highest m/z cutoff: Mz higher than *mzmax* would not be converted and processed. The default value is 1000. Acceptable values m/z > mzmin.

mz_bin: **float, optional**

The bin width of the m/z array: a step of the m/z array, which will be used for binning spectra to the vector in space from *mzmin* to *mzmax* with step *mz_bin*. The default value is 0.1. Acceptable values m/z <(*mzmax* - *mzmin*).

n_median: **int, optional**

A number of scans to be averaged with a moving median filter (https://doi.org/10.1038/s41598-018-37560-0). Moving median filter width and step are equal to *n_median*. The default value is *None*. If the value is *None* then *n_median* is recalculated in a way to binned medianed spectra not to exceed allowed space in RAM. Acceptable values from 0 to the minimum number of scans in measurements.

GB_max: **int, optional**

The allowed maximum of RAM space (in GB) to be allocated for scans storing in measurements. The default value is 2GB. Acceptable values from 1 to the actual free RAM space.

Returns:

mz_array, binned_spectra, scans_count: **tuple**

mz_array: m/z bin values.

binned_spectra: ndarray of shape [total_number_of_scans, number_of_bins], which corresponds to binned spectra in *mz_array*.

scan_count: ndarray of shape [number_of_files], each number in the array corresponds to the number of scans in each file.

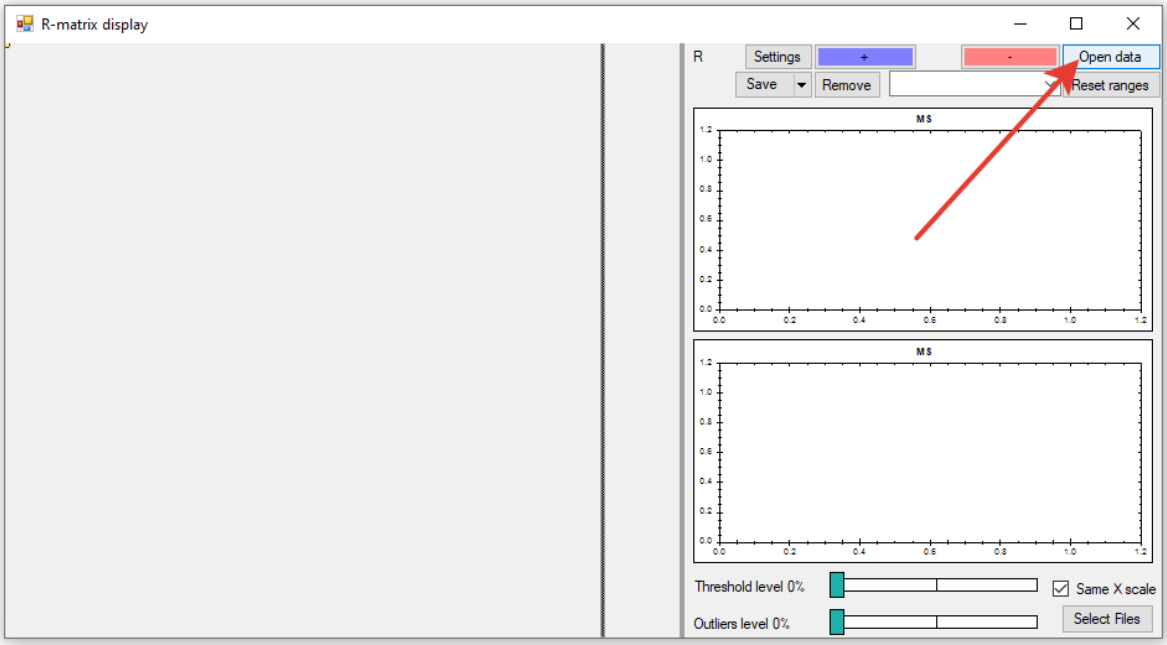
Notes:

If *n_median* is specified, the number of scans to be taken into account from each measurement will be calculated due to RAM limits. So, only the first scans from each mzXML file will be processed for the RAM limit to be not exceeded. For example, if each scan in any measurement would be 1GB size, n_median=1, and each measurement consists of 10 scans (total 2 measurements), then only the first scan of each measurement will be placed in the output dataset due to the 2GB RAM limit. Whereas skipping n_median (n_median=None) would lead to automated n_median calculation, which means n_median would be set to 10 in this example.

Total number of scans in output (N) affects the time or even possibility of the SSM_calculation to make calculations. Higher N will lead to the size of calculated SSM as N^2. Thus, the user should control the N for not to exceed RAM in the SSM_calculation process.

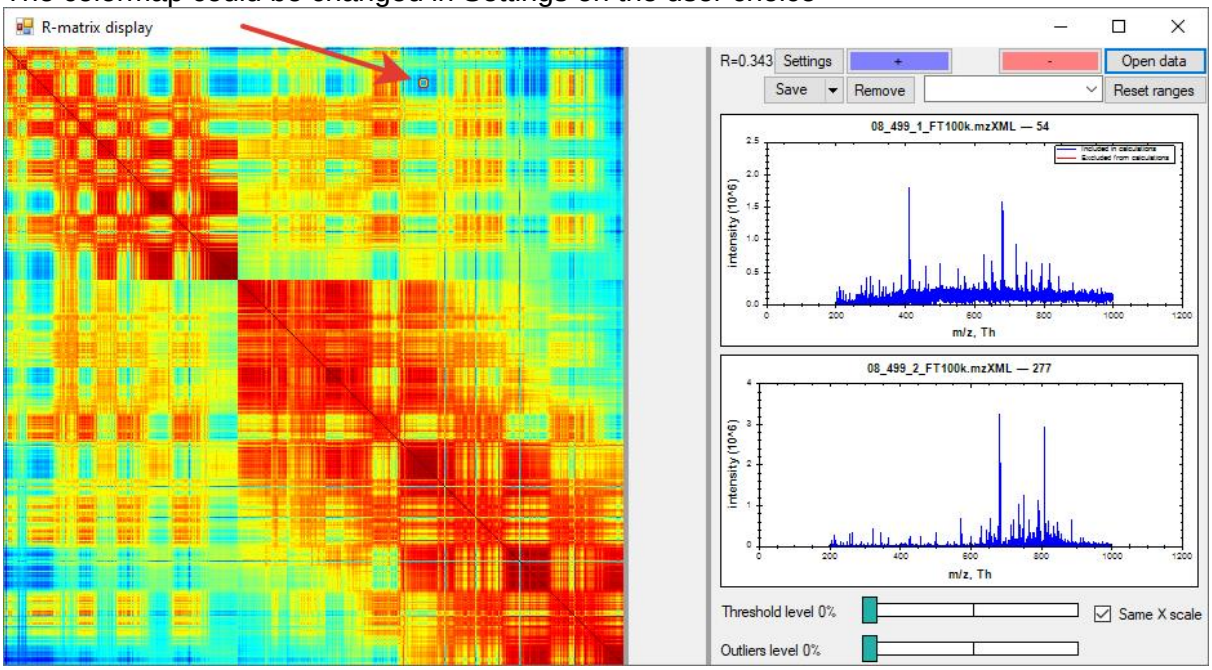
3. Download all files from https://github.com/EvgenyZhvansky/R_matrix/tree/master/mzXMLReader/bin/Release in any directory and open SSM_calculation.exe file.

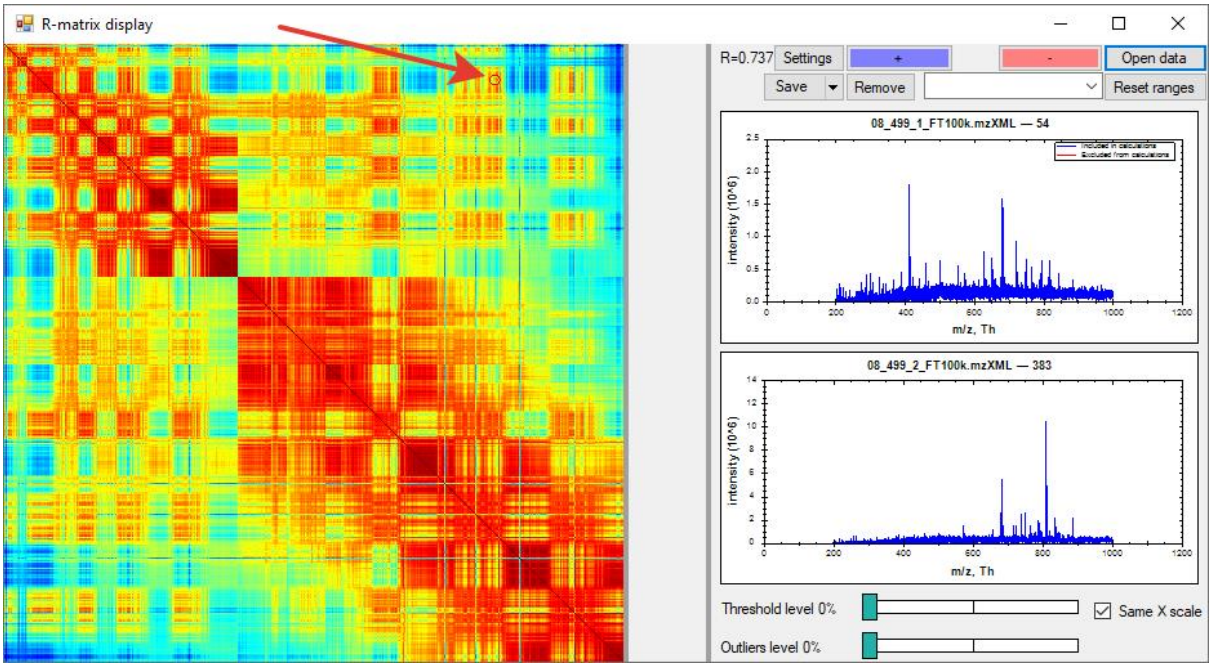
4. Create a cosine similarity spectra matrix with the SSM_calculation software by selecting .mat file that was created in a previous step.



5. One could select any pixel on the cosine measure matrix to view the corresponding spectra of that pixel. Also one could move a selected pixel with arrow keys.

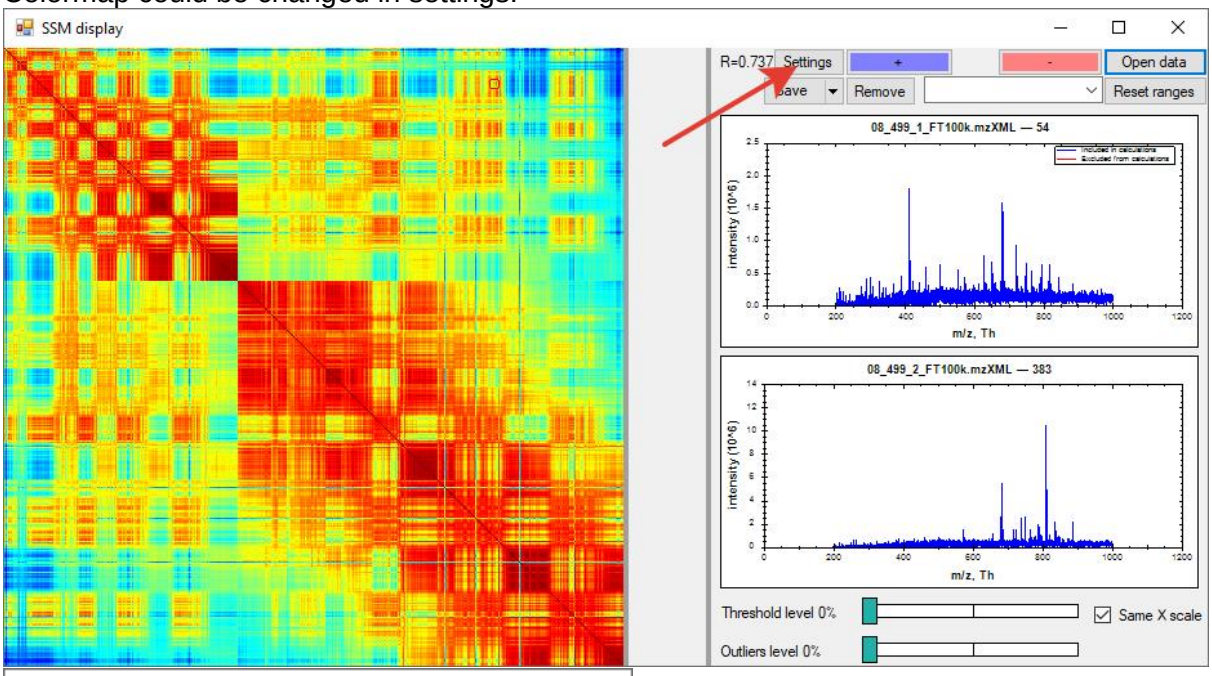
Titles of figures contain the names of files and the number of a spectrum from this file (after moving median filtration if it was applied). The cosine similarity measure is also reflected above the spectra (in the top-left corner of the right panel with spectra as R=...). The colormap could be changed in Settings on the user choice

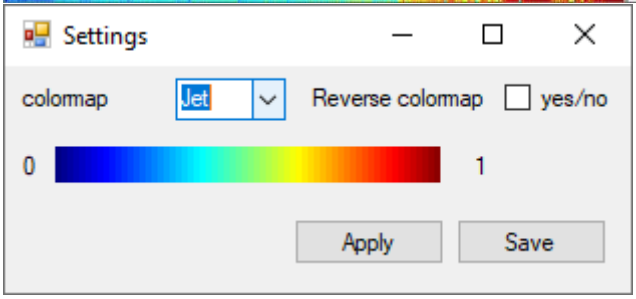


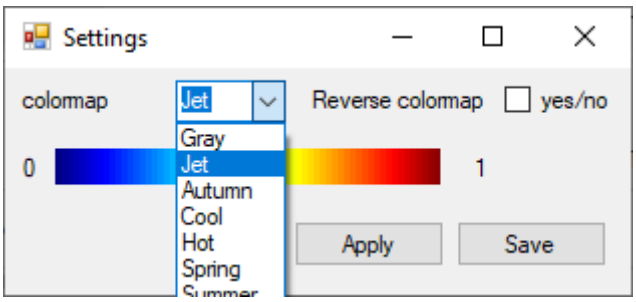


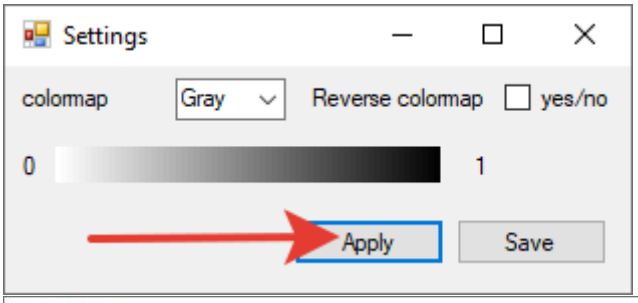
The upper graph pane reflects the spectrum in the column, and the lower graph pane reflects the spectrum in the row of SSM.

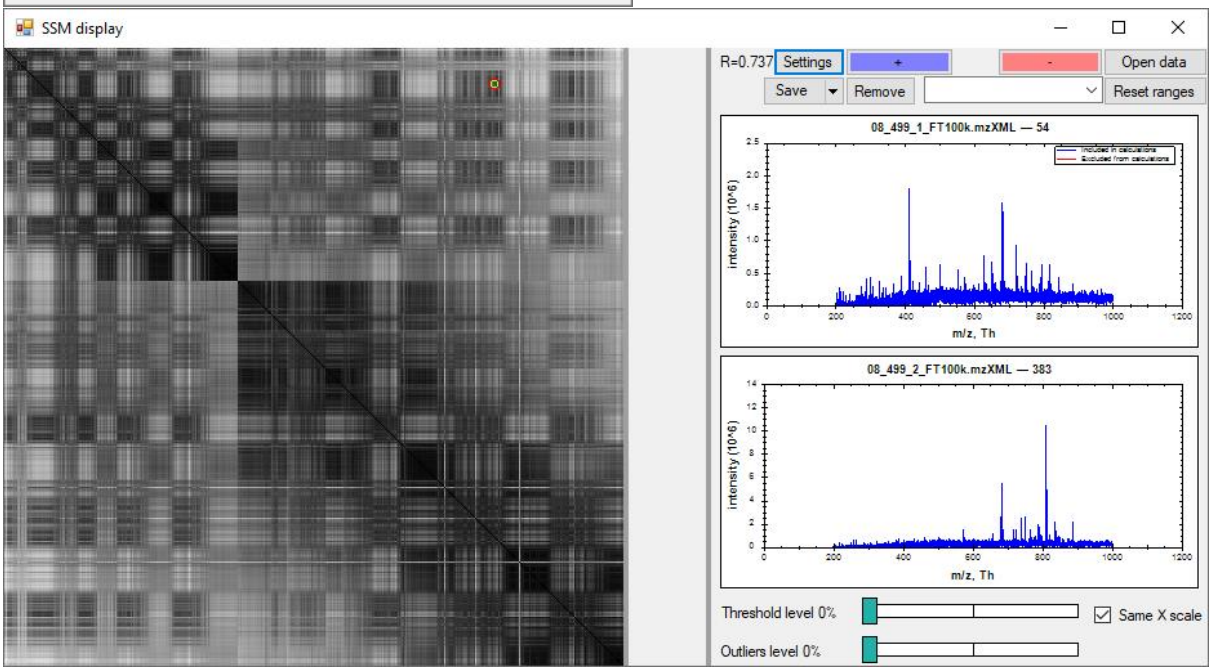
6. Colormap could be changed in settings.

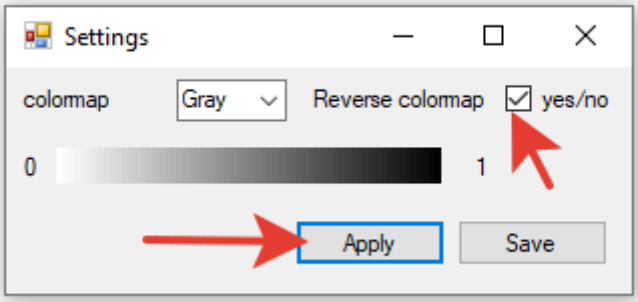


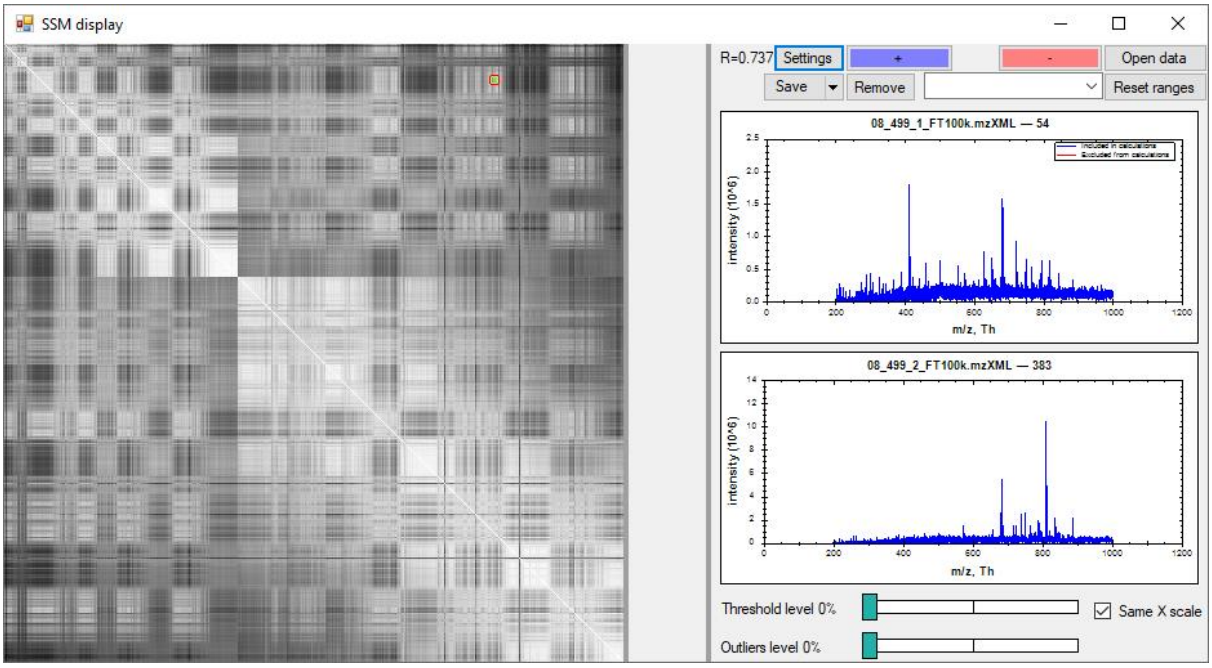



Settings
colormap
Jet
Reverse colormap
yes/no
Gray
Jet
Autumn
Cool
Hot
Spring
0
1
Apply
Save

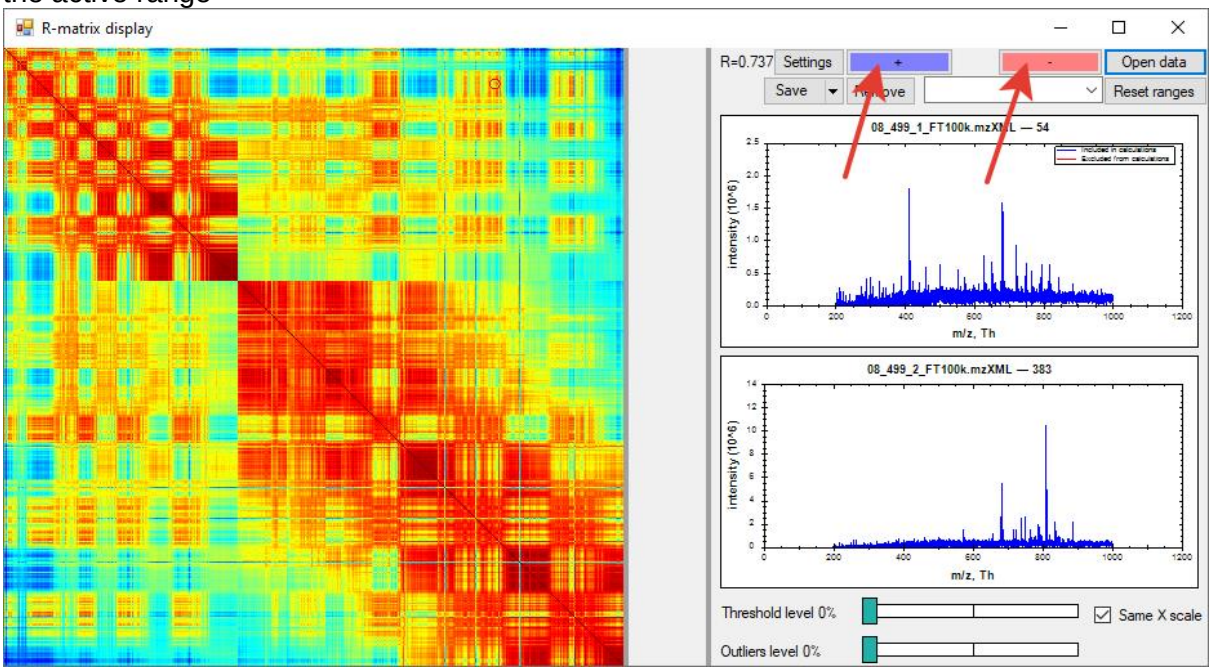

Settings
colormap
Gray
Reverse colormap
yes/no
0
1
Apply
Save

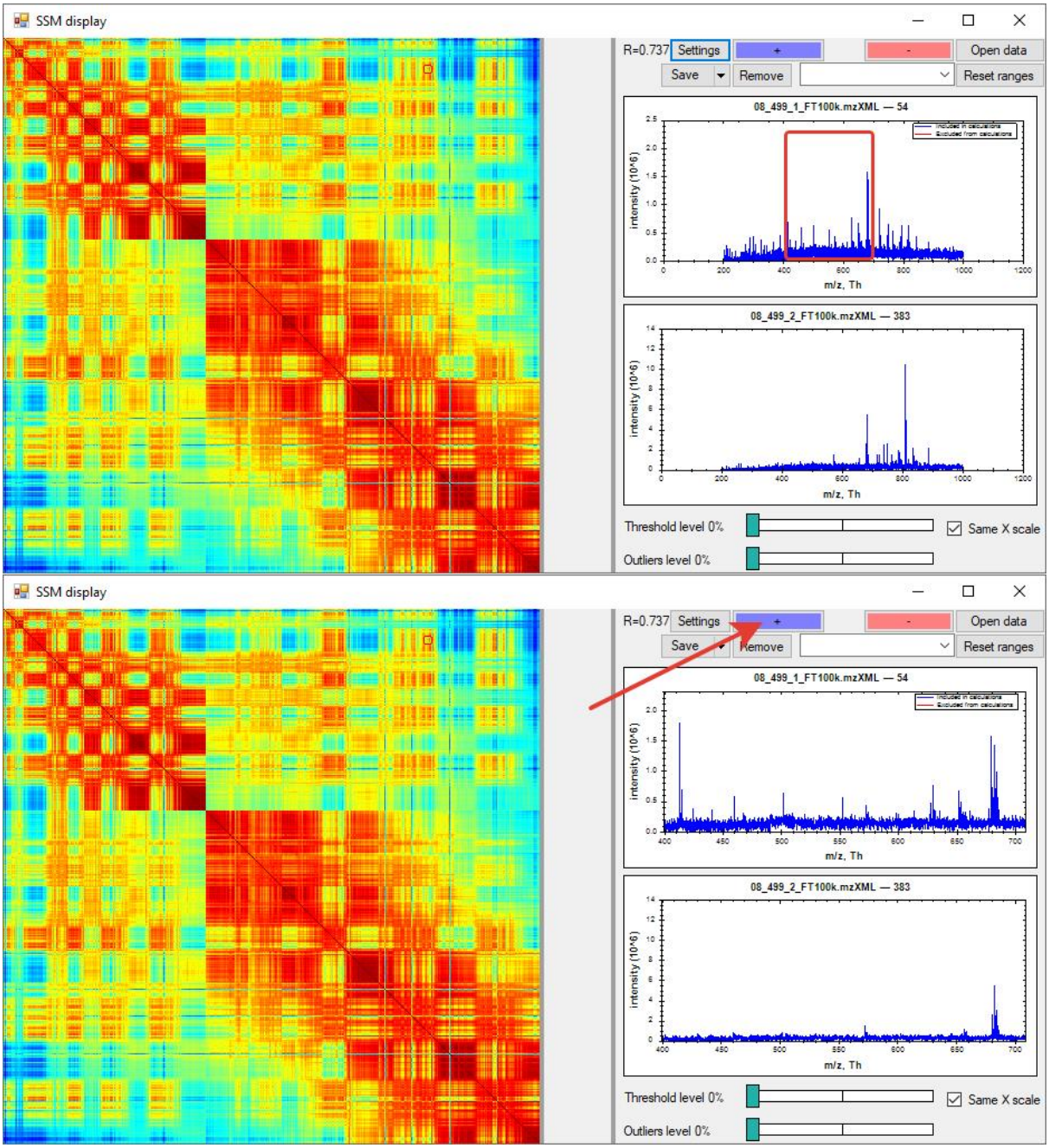

SSM display
R=0.737
Settings
Open data
Save
Remove
Reset ranges
08_499_1_FT100k.mzXML — 54
intensity (10^6)
m/z, Th
08_499_2_FT100k.mzXML — 383
intensity (10^6)
m/z, Th
Threshold level 0%
Same X scale
Outliers level 0%


Settings
colormap
Gray
Reverse colormap
yes/no
0
1
Apply
Save

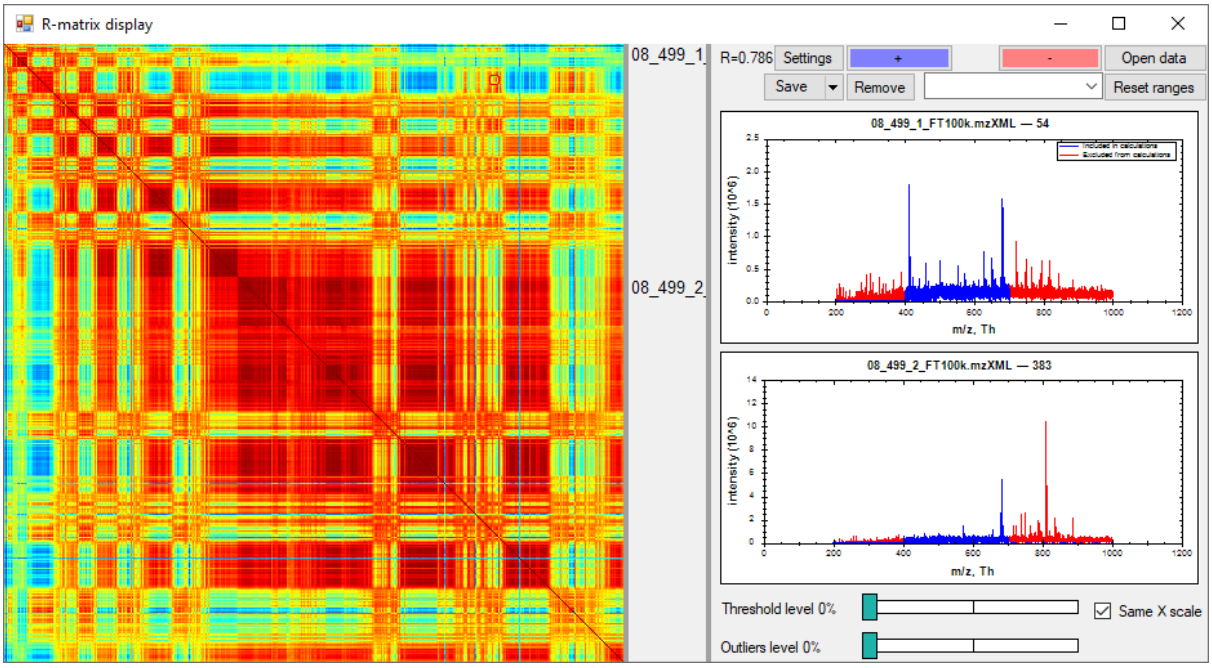


7. One could zoom in graph pane for one spectrum or for both spectra by checking the checkbox “Same X scale”. To unzoom spectra, one could use double-click. While zooming the area one could press “+” or “-” button to include or exclude the current m/z range from the cosine measure calculation. If the “+” button is clicked before excluding anything, it excludes everything except the active range

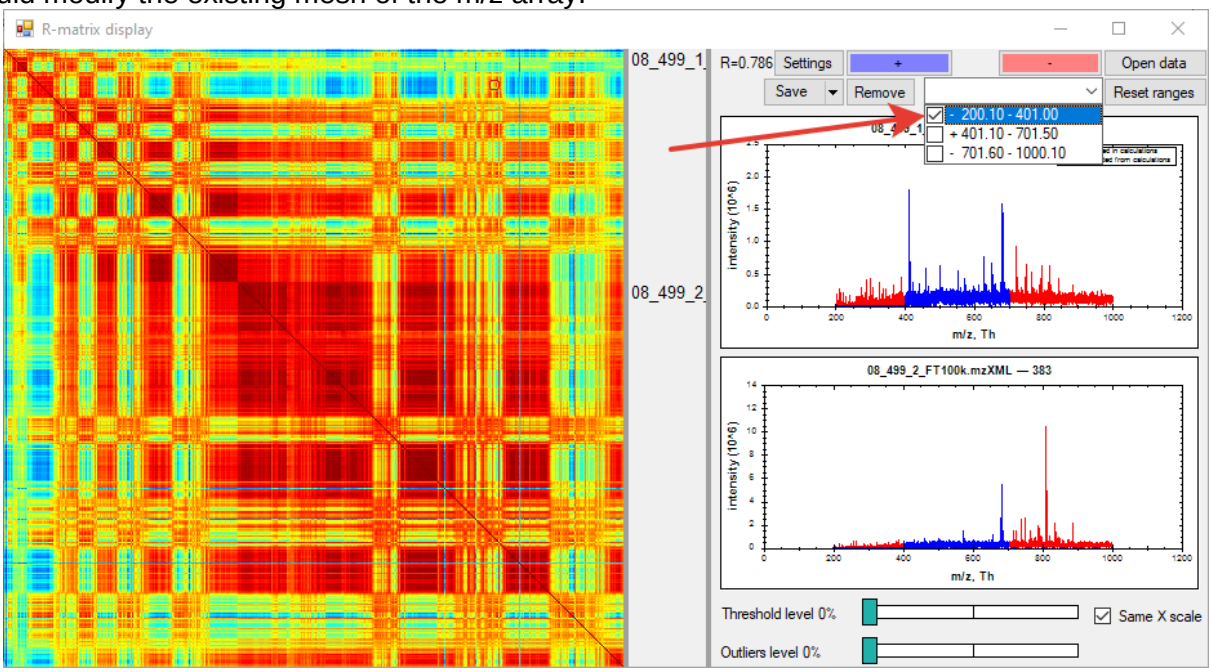




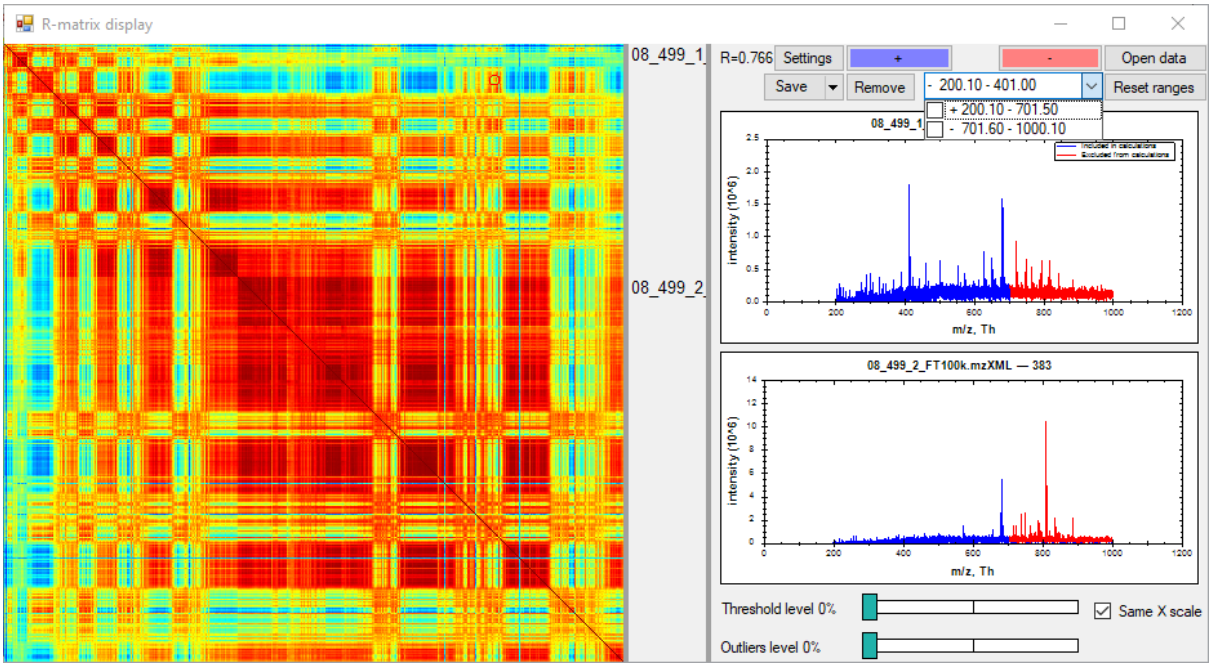
Unzoom after button “+” click (by double-clicking on a graph pane):



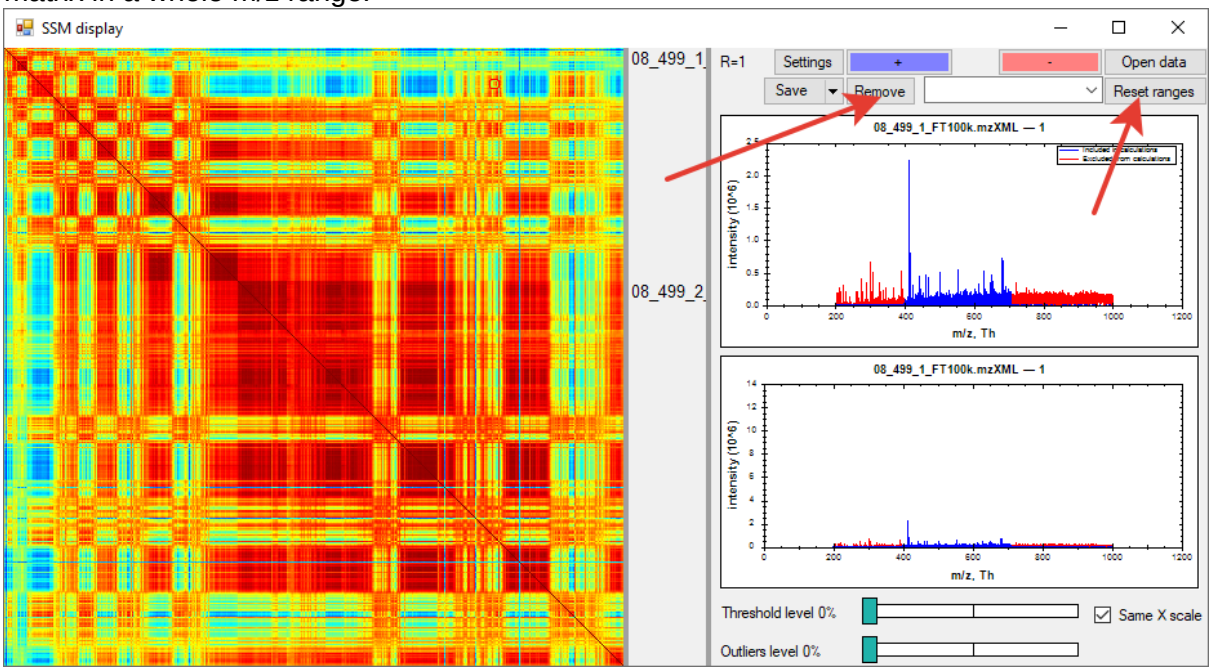
8. By removing the checked ranges of included and excluded m/z ranges from the list-box, one could modify the existing mesh of the m/z array.



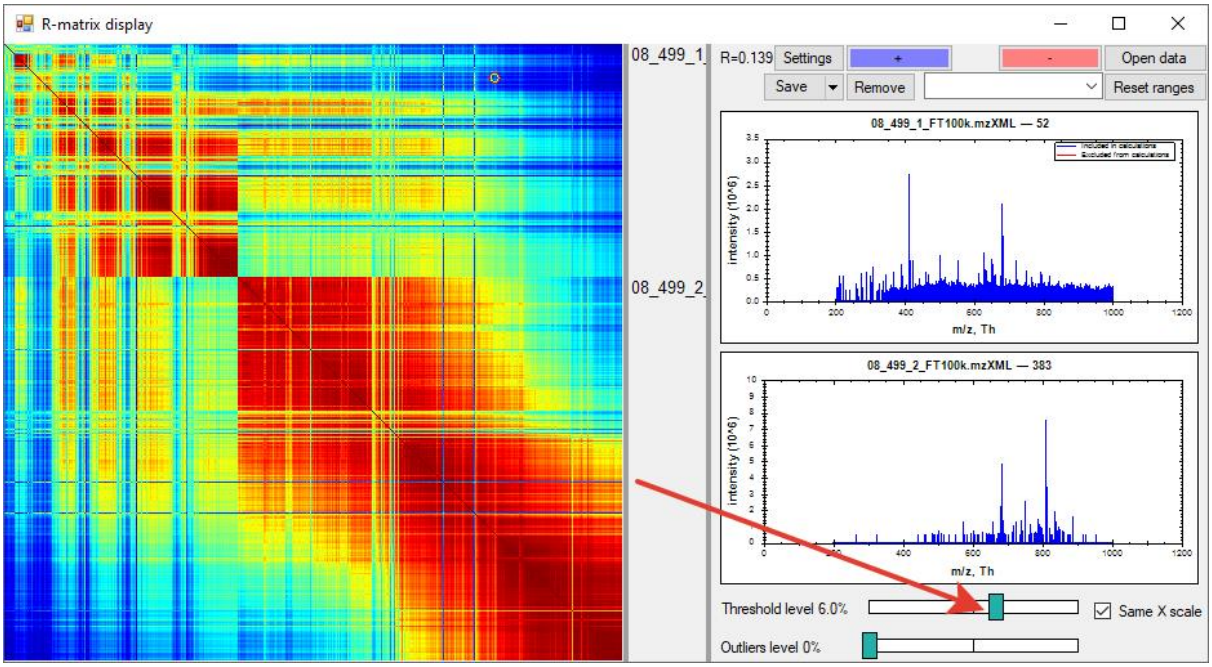
Press “Remove” to delete checked range (so the range is altered from “+” to “-” and vice versa):



Press the “Reset ranges” button to clear all ranges and to calculate the cosine measure matrix in a whole m/z range.

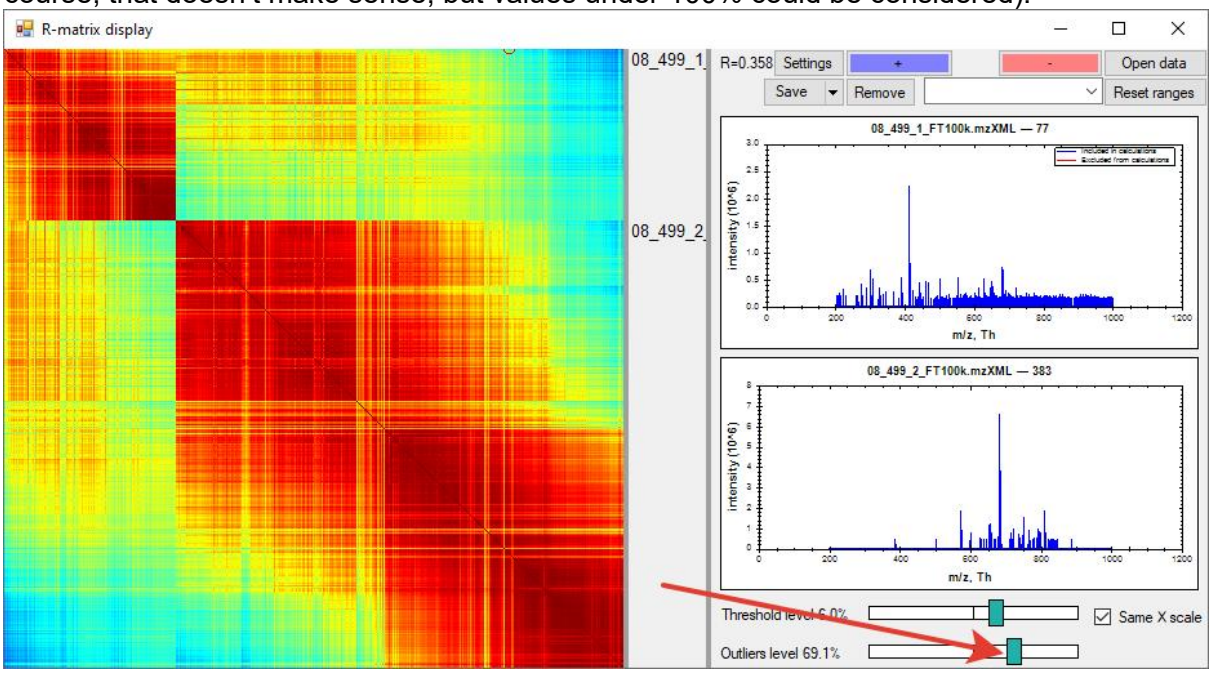


9. Threshold level.
Slider could take the values between 0% and 10%. 0% percent means that all intensities are taken into account for SSM calculation. 10% means that in each scan, intensity values lower than 10% of their maximum in the scan will be set to zero.



10. Outliers level.

Outliers slider allows excluding those pixels (columns and rows), whose sum of R values is less than value of the outliers level. Outliers level varies from 0%, which means all of the scans will be considered, to 100%, which means that only one scan will be left (of course, that doesn't make sense, but values under 100% could be considered).



The range of outlier levels is 0 - 100%. 0 corresponds to the minimum of the sum of R values in columns (or rows as they are equal), and 100 corresponds to the maximum.

11. "Save" button allows saving the cosine measure matrix view as a .png, .jpg, .bmp or .gif file or SSM view with two spectra or stream the cosine measure matrix and other data in .txt file. Also one could call the context menu by clicking the right mouse button on the graph panes with spectra.

There are four options to save the results: 1) saving figures of both spectra similarity matrix and two spectra corresponding to selected pixel in automated way by "Save images"; 2) saving figures of the spectra similarity matrix with choosing name and extension of the file by "Save image as…"; 3) save two spectra in .emf vector format through the context menu of the graph pane; 4) saving data in a text file in automated mode by "Save data". Data is

tab separated in the last option and includes coordinates of clicked pixel, filenames of the corresponding measurements and scans, mass spectra of selected pixels (both included and excluded from calculations separately), m/z array and SSM.