MsMetrix

MsXelerator Tutorial:

High Resolution Peak Matching and Differential Analysis using MsCompare

Introduction:

This document describes how to use the High Resolution Peak Matching algorithm implemented in MsCompare.
As you may well know, the MPeaks Peak Picking module operates on a single sample. Peak Picking can be done at any sensitivity level and will in general be dependent on the user requirements and application. In LC/MS data processing we often find thousands of peaks. The main task of data processing will be to quickly get rid of all non-relevant peaks. This can be a time-consuming task. MPeaks therefore contains many algorithms and filters to remove or check unwanted peaks from the sample (noise, artifacts, baselines, isotopes, broad features, spikes etc.). These filters can be applied at any resolution, ranging from nominal mass values to FT resolution.

MPeaks is your choice to detect and filter peaks in a single sample. The module contains many useful filters that can be applied to accurate result tables at any resolution, e.g.

- Calculate charge states (remove or keep specific charge states)
- Remove all ^{13}C isotopes, keep the monoisotopic peaks only, or keep the largest peak from the isotope cluster
- Calculate accurate peak areas
- Calculate accurate m/z values based on averaging peaks
- Editing of the result table on many different parameters: Intensity, Area, Peak width, m/z range, retention time
- Apply Differential Analysis; find peak in a sample not present in the control
- Cluster and remove peaks that closely co-elute (adducts, fragments, isotopes)
- Check on the presence of frequently observed adducts (Na^+, K^+ etc)
- Apply High Resolution Post Processing filters in one run (check spikes, isotopes, control sample, etc.)
- Identify possible Metabolites or degradation products using High Resolution Metabolite ID Tools
- Check if certain peaks are present by comparison against selected databases

Once the MPeaks result table has been reduced to the most relevant peaks by applying the above filter operations, you can save the result and check whether the peaks that are left are present in all other samples using MsCompare. The result will be an MsCompare table with accurate mass values and precise peak heights based on high resolution extracted mass chromatograms. The samples are checked for all peaks present in the reference sample. This procedure is called **Peak Matching.**

It is advised to run Peak Picking on a sample containing all expected peaks. For Metabonomics and BioMarker Discovery studies you can mix equal aliquots of all samples from the project and use this as the Quality Control Sample. Peak Picking should then be used on this sample.

Peak Matching is different from Peak Picking as it does not require the detection of a peak. It merely determines the peak height or area in high resolution mode across a specified retention time window (allowing for small shifts) in all other samples.

The current Peak Matching algorithm requires that chromatographic shifts between samples are moderate or that shifts after alignment correction (COW or RPW) behave well. For very complex samples, e.g. many closely eluting peaks with the same mass value and strong chromatographic shifts, a different procedure will be needed (not yet implemented).
After the result table from Peak Matching has been calculated in MsCompare, the user can apply all features present in the MsCompare module:

- Find differential peaks between two groups of samples (BioMarkers)

- Apply and view statistical results (t- and p-test, group ratios, uniqueness test etc.)
- Filter results based on one of the statistical tests
- Easy viewing of Mass Chromatograms and Mass Spectra
- Apply Multivariate Analysis techniques like PCA, Clustering and Correlation Analysis to the result table.
- Export result table to other software programs

MsCompare Peak Matching Example:

The Peak Matching example is applied to a Differential Analysis Peptide Mapping study. Also see the document on **Differential Analysis using two Samples.pdf,** which describes the Differential Analysis algorithm in MPeaks.

Sample and Control relate to Trypsin digested Avastin. The initial sample was oxidized using 3.3% tBHP. We want to compare the oxidized sample (Sample) versus the initial sample (Control) to detect peaks being **up-regulated**. Measurements were performed using an Agilent LC1200 system with a Waters Q-TOF Premier. Sample and Control were measured in triplicate.

As we want to find up-regulated or even unique peaks, we will have to perform peak picking on the oxidized sample. If we would like to find down-regulated peaks as well, we should first run Peak Picking followed by a Differential Analysis in MPeaks (Use Algorithm B). When used in this way, the down-regulated peaks will be added to the peak picking table. Another possibility is to split the analysis: 1. run MPeaks on the Sample and save this result table. Then load the Control and run MPeaks again. Using the second approach you will have to run the Peak Matching algorithm twice.

Figure 1 shows an overview of the oxidized sample in the Browser. For a small peak, m/z 619.3, the mass spec and the EIC are plotted. For peak picking it will be important to estimate the width (in scans) of the relevant chromatographic peaks. The scans across the peak are made visible in Figure 1, middle window. The width (near the baseline) is about 30 scans. You can count the dots or use the Width Measurement Tool on the Icon Toolbar. For MPeaks peak picking we will use 11 scans as the minimum peak width.

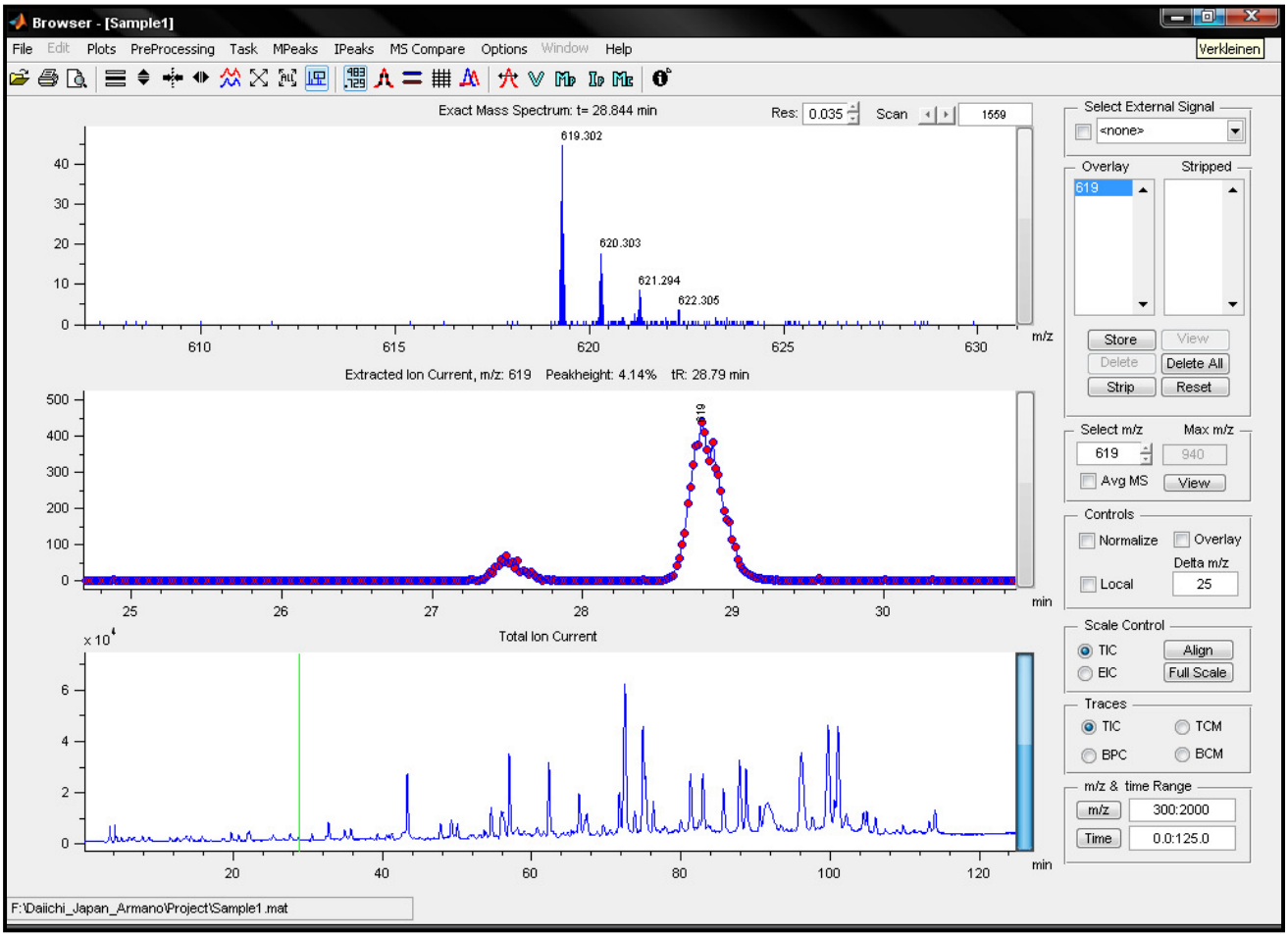


Figure 1: Overview and exploring the Sample in the Browser.

Step 1: Peak Picking using MPeaks:

Start MPeaks from the Browser and load default values. This sample contains many peaks. For the demo we will therefore set the peak threshold to a level of 0.5% (compared to the largest peak present), this is 10 times less sensitive compared to the default value of 0.05%. This means that smaller peaks will be deleted from peak picking. Peak Picking will be done in nominal mode. To convert the peaks automatically to accurate m/z values, mark the **Exact m/z** check box in the Peak Picking Parameters section. Accurate m/z values will be calculated based on a single scan at the peak's retention time. MPeaks has different options to calculate accurate m/z values. Check the Accurate Mass Options or the MPeaks Advanced Tasks from the MsXelerator User Manual.

Press the **Run button** from the buttons below the Table region. A total of 2248 peaks will be detected. At this stage, the table includes all isotopes. You can examine peaks by clicking on the entries in the table. You will be asked if plotting should be done in accurate mass mode (e.g. m/z 619.31 +/- 0.01 Da). The resolution used for plotting EICs can be set from the resolution edit box at the top left of the Figure in the MPeaks window, or from the Options menu. You can easily toggle between EIC and Mass Spec plotting by pressing the Toggle Mass Spectrum icon button on the Toolbar. On the Toolbar you will also find icons to use Auto Zooming and Exact/Nominal EIC plotting.
For Profile Mass Data be sure to use **Auto Zooming** (icon on the right), otherwise the extraction might take some time for data sets having long run times. For this data set a runtime of more than 2 hours was used. See the MsXelerator for an overview of general features for MPeaks and the Browser.

Press the 'h' button to get an overview of all active keyboards keys for zooming, scaling, plotting etc.

As we will only be interested in the larger differential peaks we will first reduce the table by removing all peaks having an absolute intensity less than 20 counts. This seems a fair value for the Q-TOF instrument.
Press Edit Table. An overview will be shown of the basic Peak Characteristics (m/z range, retention time range, Peak Height, % Peak Height, Area %, FWHH etc. To remove the small peaks, enter the value 20 in the PH (Peak Height) **Min Value Box** on the left. **Press enter**. The Number of Peaks shown at the bottom of the Figure will be changed to 535 Peaks. If you want to redo press the Restore MPeaks Table icon from the Icon Bar.

We are not ready yet in the preparations, because we want the comparative analysis to be performed on monoisotopic peaks. For peaks having charge states 1+ to 4+ this seems to be a good approach. The current table can be converted to monoisotopic peaks using the Accurate Mass Utilities. At the same time the charge states will be calculated. From the Menu select: **Accurate Mass > Accurate Mass Conversion.** The Accurate Mass Setting window (Figure 2) will be shown. For this data set select the appropriate settings (Keep Monoisotopic Mass Peak, Profile Mode, Q-TOF Resolution, Calc. Charge State and Recalculate Peak Heights).

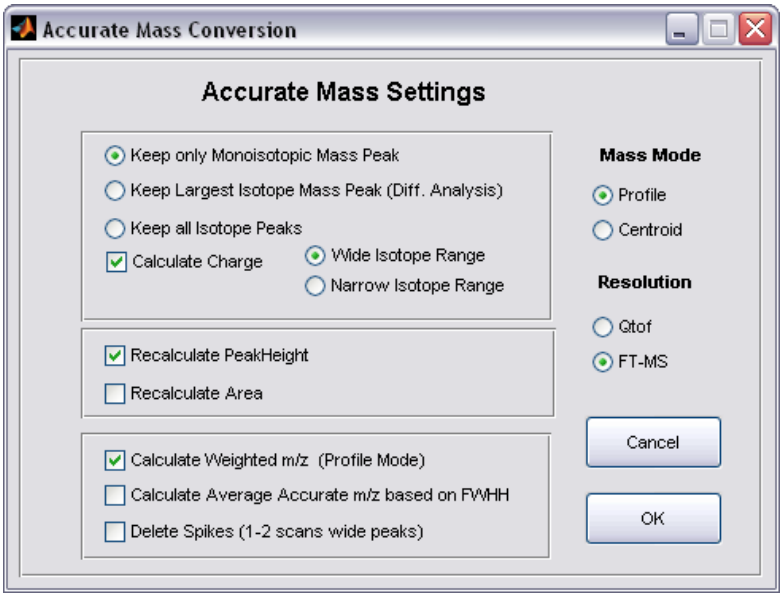


Figure 2: Accurate Mass Settings Window.

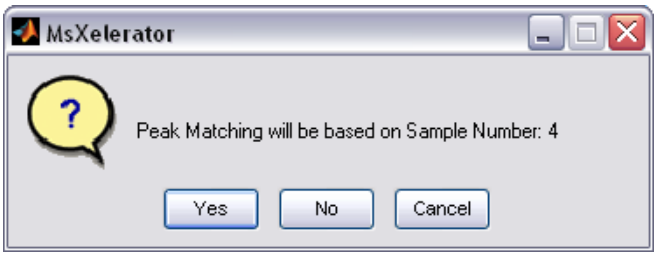
Attention: in Proteomic applications it is sometimes better to keep the largest peak from the Isotope Cluster instead of the mono-isotopic peak. For high masses and higher charged ions (4+ and up) the mono-isotopic peak is often much smaller compared to the largest isotope peak. A comparative analysis might be more precise using the largest signal.

Press OK to start the conversion. When asked to use a broad or narrow range for charge state determination, use Broad. For SILAC based data having difficult isotopic patterns, it is often better to use a narrow range for charge state determination. After completion the table contains 280 entries and will be sorted on m/z value. To apply Peak Matching in MsCompare you will have to save this table (**Menu > File > Save Result Table, Enter a comment), or press the save table button**.

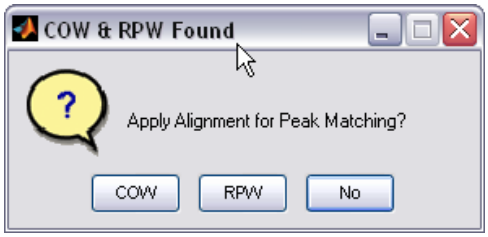
Step 2: Applying Peak Matching using MsCompare:

Start MsCompare from the Browser. Load the Sample List containing the filenames of the 6 samples, **Menu > File > Load Samples from List File**. This will load and visualize the 6 samples from the study. If you have imported multiple samples simultaneously, a default List File is always generated (Importlist.blf). Depending on the type of instrument, the default list file is written to the main directory containing the Data Folders for Bruker and Waters or the Data Files Folder (other vendors), or to the folder of the first sample. You can also create a Sample List manually. Once loaded you can divide your sample into groups e.g. Sample versus Control. Groups are necessary when using the differential statistics of MsCompare. See the manual.

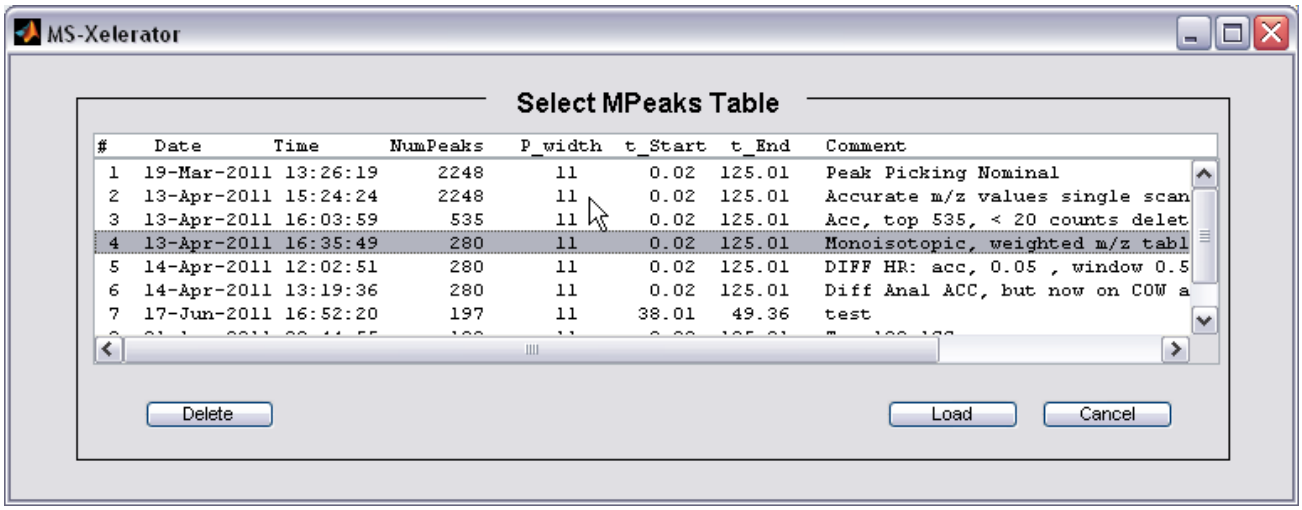
Attention: before starting the Peak Matching algorithm, be sure to select the sample from the sample list for which Peak Picking in MPeaks was performed (Sample number 4, Name: Sample1), see Figure 3. This is also important if you used some form of Alignment Correction.
From the menu select: **Peak Matching > Run Peak Matching**. You will be asked to confirm the selected sample:



Press **Yes.** The program will check if you have applied Correlation Optimized Warping (COW) or Peak Reference Warping (RPW) for Chromatographic Alignment. If corrected time-scales are found for all samples, you will be asked if Peak Matching should be run on aligned data. In general, this will be the case. If both COW and RPW correction has been applied you will have to make a selection. Not strictly necessary, but it is best to also apply the alignment correction when loading the LC/MS data files from disk. Of course the peak picking table and the sample used as a reference for alignment should be the same. The program will check that the active sample was indeed used as reference for alignment. If a message is returned, you can read what sample was actually used for alignment.

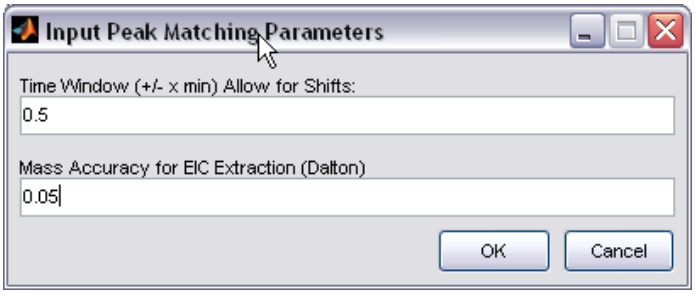


Next, an overview will be presented of the MPeaks result Tables belonging to the active sample. Select the fourth table from the list (280 mono-isotopic peaks having intensity values larger than 20) and press **Load**.

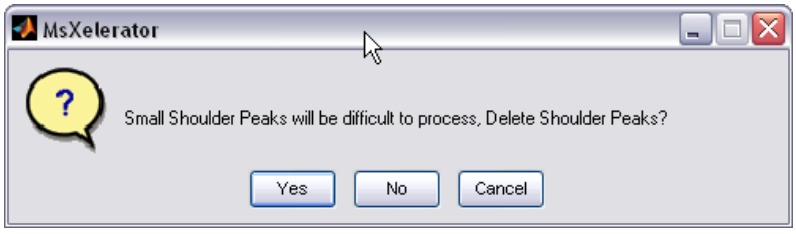


After selection of the table you will be asked to specify a time window (in minutes). The algorithm will determine the peak heights of all peaks (defined by the reference sample) in all data sets using a specified retention time window and resolution. To estimate the Time Window, you should explore the sample set in MsCompare before starting the algorithm. Based on a number of selected peaks it was observed that retention time shifts were not much larger than +/- 0.25 minutes. Therefore a window of 0.5 minutes was selected. If Alignment correction has been applied, you might want to use a smaller value.

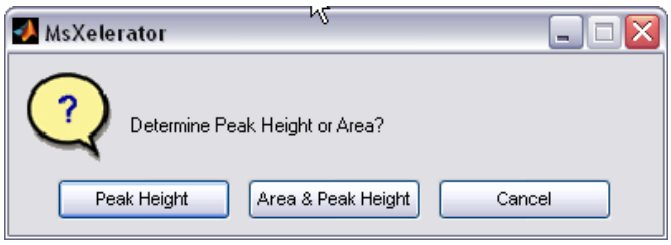
For the Q-TOF instrument we generally apply a resolution of 0.05 Dalton. This means that Accurate EICs will be extracted using the specified m/z value of the reference sample e.g. m/z 434.1254 +/- 0.05 Dalton. For FT instruments like the Orbitrap you can enter a value of 0.005. Specifying a too high value might result in the extraction of multiple peaks in the same EIC which actually have a different mass. Specifying a too low value, might result in badly extracted peaks (missing parts of the peak). You can experiment with the resolution in the Browser or in MPeaks. Target will be to find a value that gives high specificity and good peak shapes.



Press OK to accept the values.



Next you will get a question if small shoulder peaks should be deleted from the table. These are the small peaks having the same mass and closely elute next to another larger peak. These are the ones most difficult to operate on. When using a relatively large time window, it is better to delete the small shoulder peaks. If the separation of closely eluting peaks having the same mass, is larger compared to the applied time window, select No.



Next, you will get a question to determine Peak Heights or Peak Areas and Height. When selecting Area and Height, two MsCompare result tables will be written to disk. Using only Peak Height is of course faster.

The algorithm will start and you will be informed about the progress on every sample. When finished a message will be displayed telling the user that the result table was automatically saved to disk. You will be able to retrieve it from disk whenever needed. The result table has a fixed name and the end of the name a marker is added which kind of intensities were calculated, accurate peak heights or accurate areas.

Finally the Table will be displayed in MsCompare, as shown in Figure 3. The order of the peaks will be equal to the MPeaks table that was used. Figure 3 shows that 6 samples have been loaded, 3 controls and 3 samples. The Sample List contains the filenames, class names and full path of each sample. The **Class Colors check box** has been activated. This will plot EICs of both groups in two different colors, which will make a visual recognition of differential peaks easier.

The table contains the following columns: peak number, retention time, m/z value, followed by the intensities for all six samples (order equal to the Sample List). Clicking on an entry of the table will plot the EICs in overlay to the bottom window. Plots can be made in overlay, stacked or stacked offset mode (see the manual). The selected sample from the table will be plotted bold in the lower window. This sample will also be marked in the Sample List Window.

As an example the peak having an m/z value of 851.4419 at 49.23 minutes was selected from the table. As a default, the nominal mass chromatograms for all samples will be plotted in the lower window. It can be seen from the table and the overlay plot that the peaks in the samples are highly up-regulated compared to the controls. From the table it can also be seen that most of the peaks from the sample and control group are very similar in intensity. If you want to plot the accurate EICs of m/z 851.4419 peak directly use: **Menu > Plots > Exact Mass Chromatogram**. You will be asked to specify the resolution and time range for plotting. You can also directly press the **Plot Accurate Mass Chromatograms icon** on the Toolbar (see icon on the right). Please use the same resolution as used for Peak Matching.

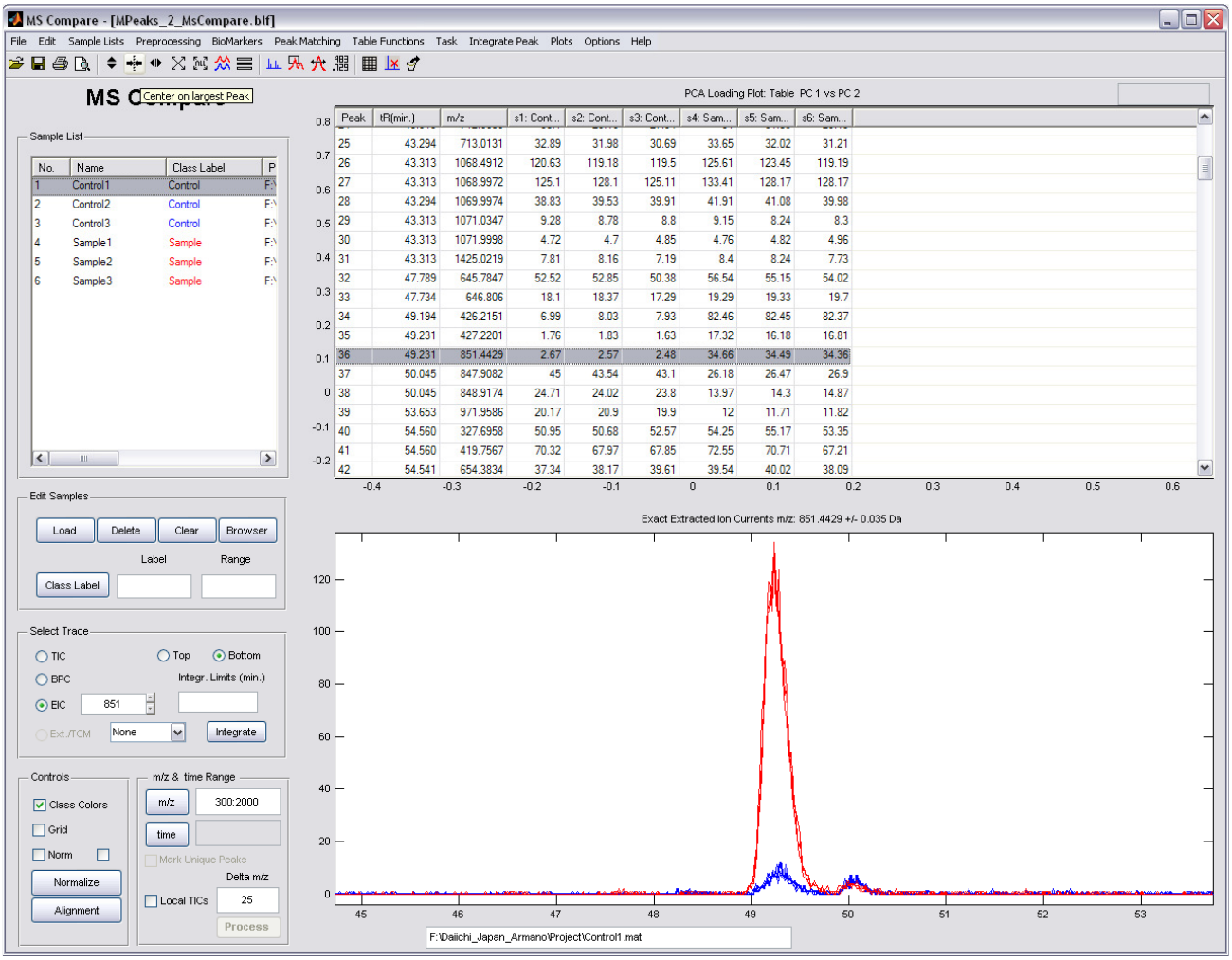


Figure 3: MsCompare Overview after Peak Matching. A differential peak was selected from the result table. EICs are plotted in accurate mode (m/z +/- 0.035 Da).

Finding Differential Peaks using MsCompare:

Now that relevant peaks have been determined in accurate mass mode, we can use the differential analysis tools from MsCompare to find up- or down-regulated peaks using a number of tools/statistics.

- Table Profile Plot: plot all intensities from all samples in one plot. Visually scan differences using color coded profile traces. Click on any of the peaks numbers in the Profile plot and extract all EIC traces for all samples.
- Ratio Line Plot, Ratio/Intensity Scatter Plot. Define how to create ratio's first. (e.g. group1/group2 or group 2/group1) . Calculate all ratio's and create a ratio line-plot. Click on any of the peak numbers to extract the EIC traces. Also available are t- and p- statistics and the uniqueness plot. See Menu > BioMarkers > Table: Statistical Plots.
- Apply Statistical Filtering to reduce to table to up-regulated peaks only, e.g. keep only peaks having ratio's of 3 or higher.
- Apply PCA and Cluster analysis on the result table. Loading plots are interactive. Click on any of the peak number and the EICs or the concentration profiles will be plotted. More detail is given below.

Table Profile Plot: based on the calculated table you can plot all intensities for all samples and all peaks in one plot. This together with group color coding and interactivity with EIC plotting makes this plot powerful for analyzing your results. Figure 4 shows the Table Profile Plot (**Menu > Table Functions > Table Profile Plot**).
Along the x-axis all detected peaks are plotted (peak number as it relates to the table entry). The y-axis shows the intensity of the peaks. In this case, 6 lines are plotted of which the color relates to the two classes. The red lines refer to the

samples, the blue line to the controls. Up- or down-regulated peaks are easily detected by looking at separation between the blue and red lines.

At the position of the arrow one can see that all intensities of the samples from the Sample Group are above the Control Group. Clicking at this position extracts the mass chromatograms of all samples to the top window. From the profile plot we can also see that the majority of the peaks have similar intensities. Also notice that some peaks are present in which the intensities of the Controls are significantly larger compared to the Sample Group.

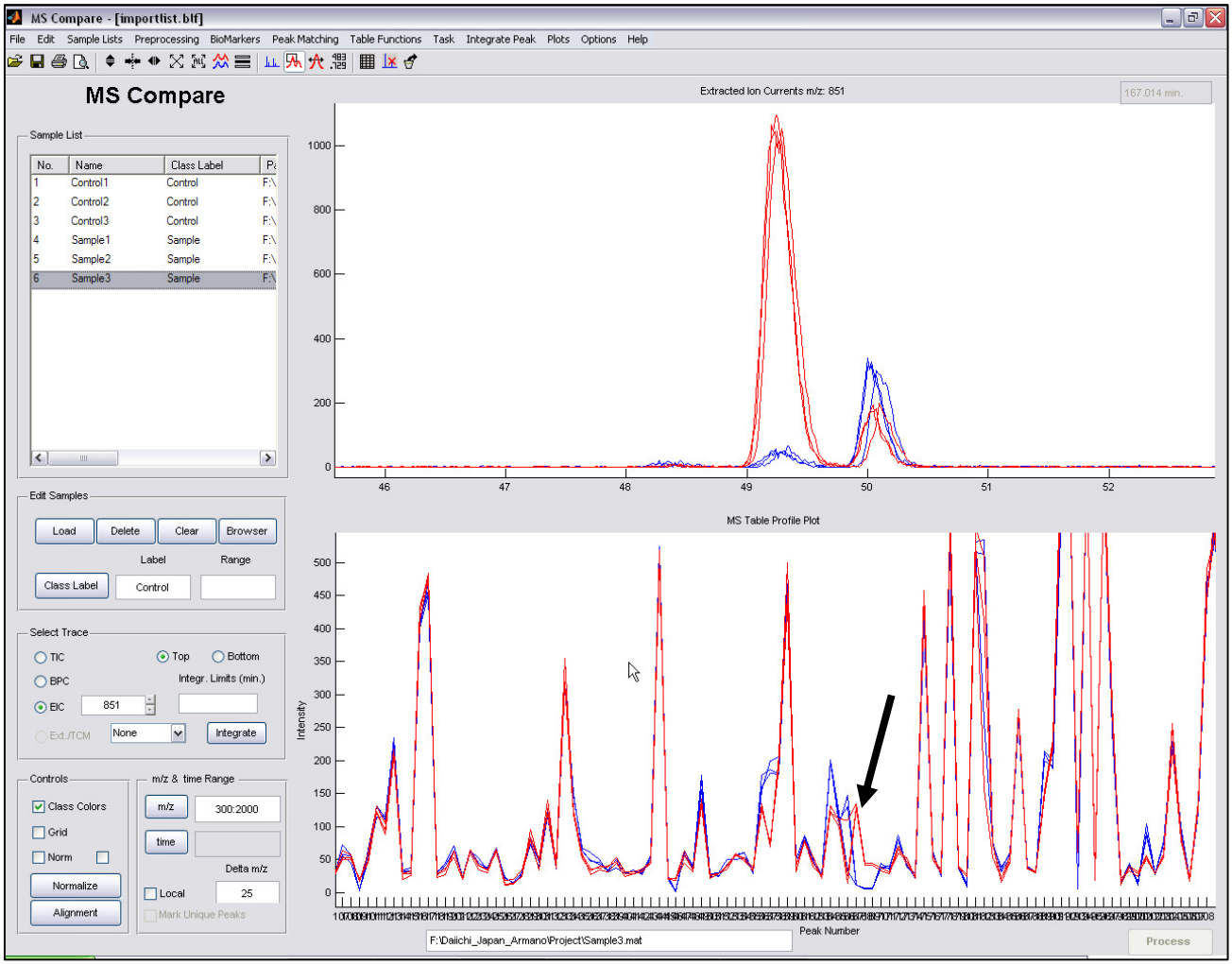


Figure 4: MsCompare – Table Profile Plot showing result for all samples and all peaks. The Profile line plot was zoomed in to better detect the differences between both groups.

Ratio Line Plot: an easier and faster way to find differential peaks is to make use of the ratio line plot, shown in Figure 6. However, before creating this plot you will have to specify how the ratios have to be calculated (Sample/Control or Control/Sample). First you will have to specify the Class Selectivity Rules. Select from the menu: **BioMarkers > Set Selectivity Rules.** This will bring up the window shown in Figure 5. Displayed are the two Classes, their respective Class Names and the number of samples in each class. Since our goal is to find peaks in the Sample group that are larger compared to the Control group, we will select the second option, **Peaks Class B > Peaks Class A,** (Class B is the Sample Group, Class A the Control Group).
Now you generate the ratio line plot by selecting: **Menu > BioMarkers > Table: Statistical Plots > Ratio value line plot**. The calculated ratios from the group averages are shown in Figure 6. Clicking on one of the peaks with a high ratio will automatically extract the EICs to the top window for a visual exploration. **To explore down-regulated peaks**, reverse the group assignment in Figure 5 and select – **Peaks in Class A > Peaks in Class B**.

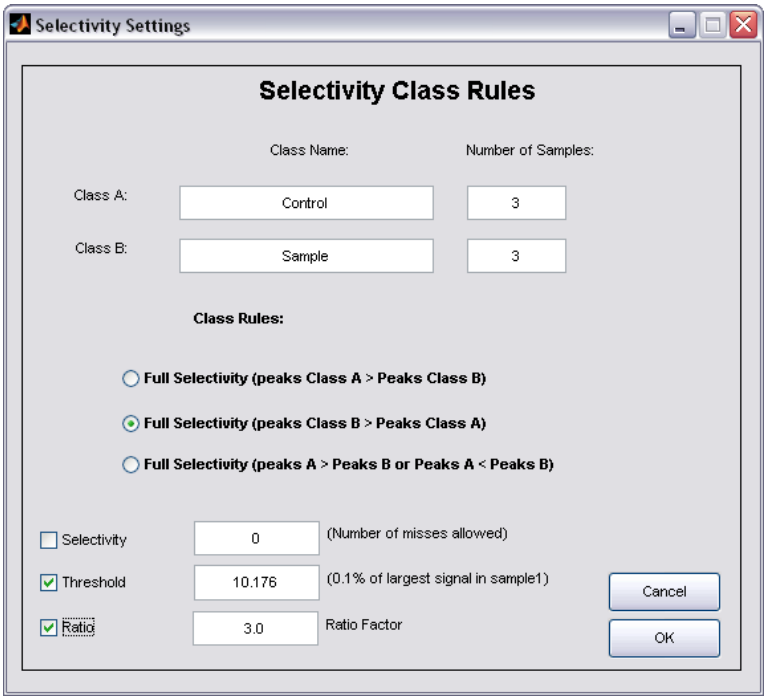


Figure 5: Specify Selectivity Class Rules to calculate correct group ratios

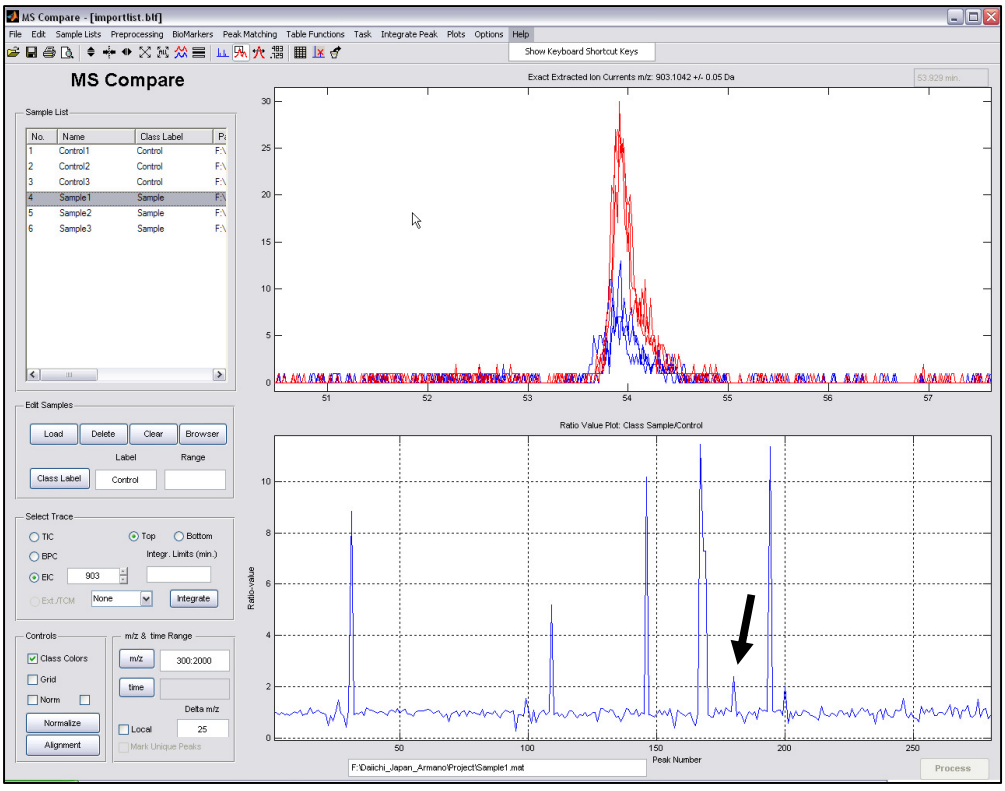


Figure 6: Ratio Line Plot. Shown are the ratios calculated from the intensity group averages for all peaks (Sample/Control). In the top window the EICs are plotted in high resolution mode (+/- 0.05 Da.) for a low intensity up-regulated peak (ratio = 2.4).

Intensity – Ratio Scatter Plot: another useful plot is the so-called Intensity – Ratio Scatter Plot, shown in Figure 7. The scatter plot shows both the ratio displayed along the x-axis and the intensity along the y-axis. The marked peaks are probably the ones most interesting as these are high up-regulated and also quite large. A ratio limit control line of 3 and 0.33 is added to the figure. Clicking a dot will identify the peak in the plot and also jump to this position in the result table. Using the statistical filters from the Biomarkers Menu, you can directly filter on both intensity and ratio.

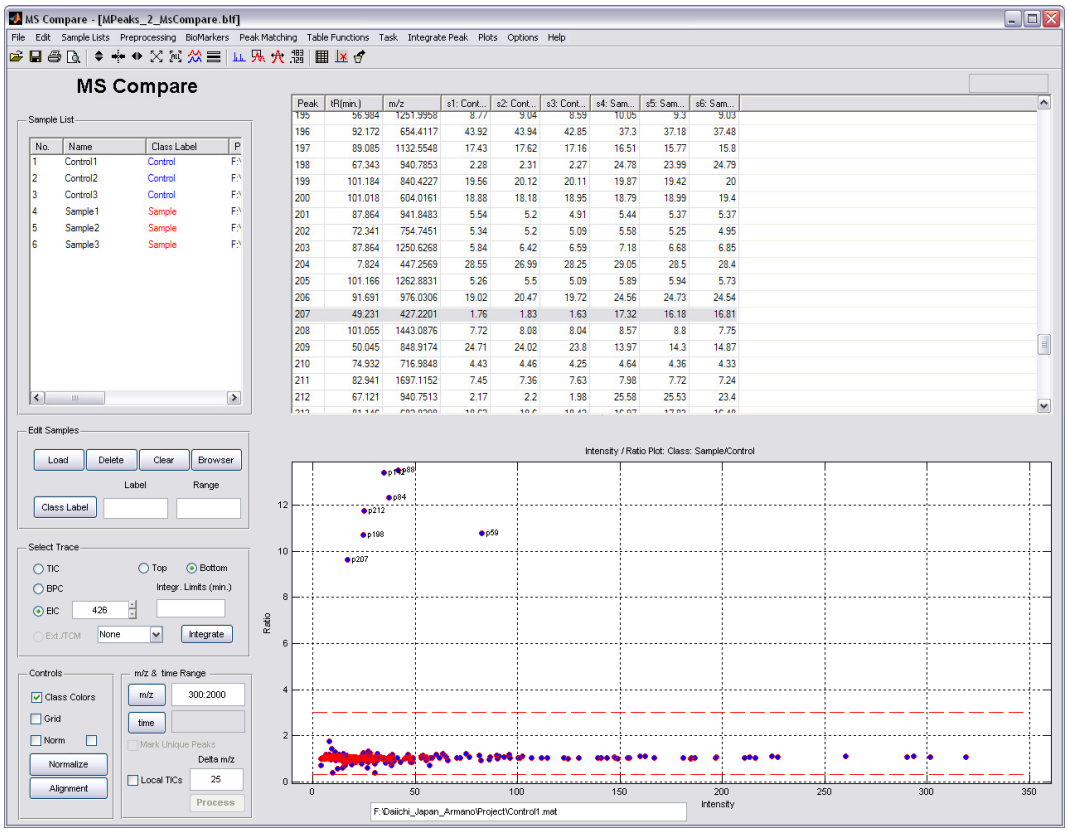


Figure 7: Intensity - Ratio Scatter plot. Shown are both the ratios and absolute intensities for all peaks. High values along the y-axis have large ratios; along the x-axis the intensity values are plotted.

MsCompare has other useful plots e.g. the uniqueness plot. It will directly show both up- and down-regulated peaks. High or values account for very unique peaks, e.g. almost absent in the control group.

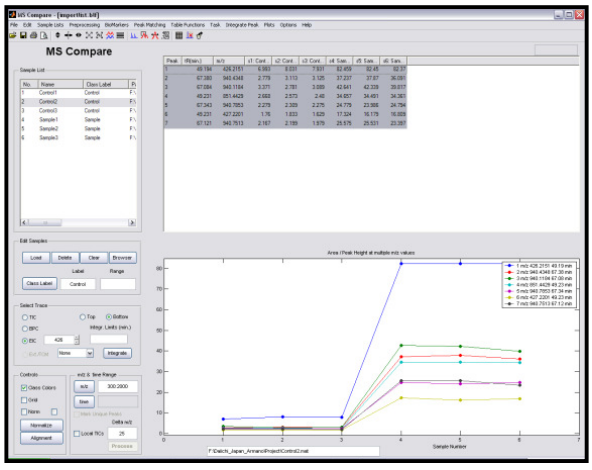


Table Profile Plot: Select Table > Peak Profile Plot to plot the samples against the intensity values for a number of selected peaks. Easily observe trends or confirm that the peaks distinguish both groups. From the same menu a few other plots can be created: Peak Correlation Plot – see which peaks are highly correlated, Peak 2D Scatter Plot – plot two peaks in a scatter plot and observer their correlation. Peak tR index plot – this is a useful plot to see which peak cluster or co-elute at the same retention time. The table will be sorted on retention time. (Figure 8, inset).

Statistical Filtering: based on statistical calculations the user can directly remove all non relevant peaks. From the menu select: **BioMarkers > Table: Statistical Filters > Ratio Value Filter**. Enter a value of 3.0. A message will be displayed that 271 peaks will be removed from the table. Press **Yes** to continue. The final table is shown in Figure 8.

One of the plots often used is the so-called Peak Retention Time Index Plot, show as an inset in Figure 8. Plotted are the peaks as a function of retention time. From this plot one can directly determine which peaks are co-eluting. In Figure 8 two groups of two peaks are marked. These peaks might be adducts or fragments which need special attention during interpretation. Especially when many peaks remain after differential analysis, this plot helps to find the major interesting peaks. Final results can be exported to Excel or text files for reporting or alternative analysis.

Finding Down-Regulated Peaks: at the end of page 7 it was mentioned that you can find down-regulated peaks by reversing the Class Rules. This is only partly true. You will not be able to find unique peaks in this case. This would mean that we are looking for peaks absent in the Sample group, which are not present because the peak picking was based on a sample from the Sample Group. The reversed procedure only works for peaks that are detected.

To be able to find unique peaks in the Control Group there are two approaches:

1. After running Peak Picking in MPeaks, perform a Differential Analysis using Algoirhtm B. This algorithm can perform a dual analysis, check up-regulated peaks from the table and next do a reversed analysis using the control sample to detect peaks not present in the current MPeaks table. These peaks from the control will be added to the result table. So the result table will have peaks from the sample and control.
2. The second approach is to use the sample from the Control Group for peak picking in MPeaks. Then proceed as usual, and search for peaks unique or highly up-regulated in the Control Group in MsCompare after Peak Matching has been run.

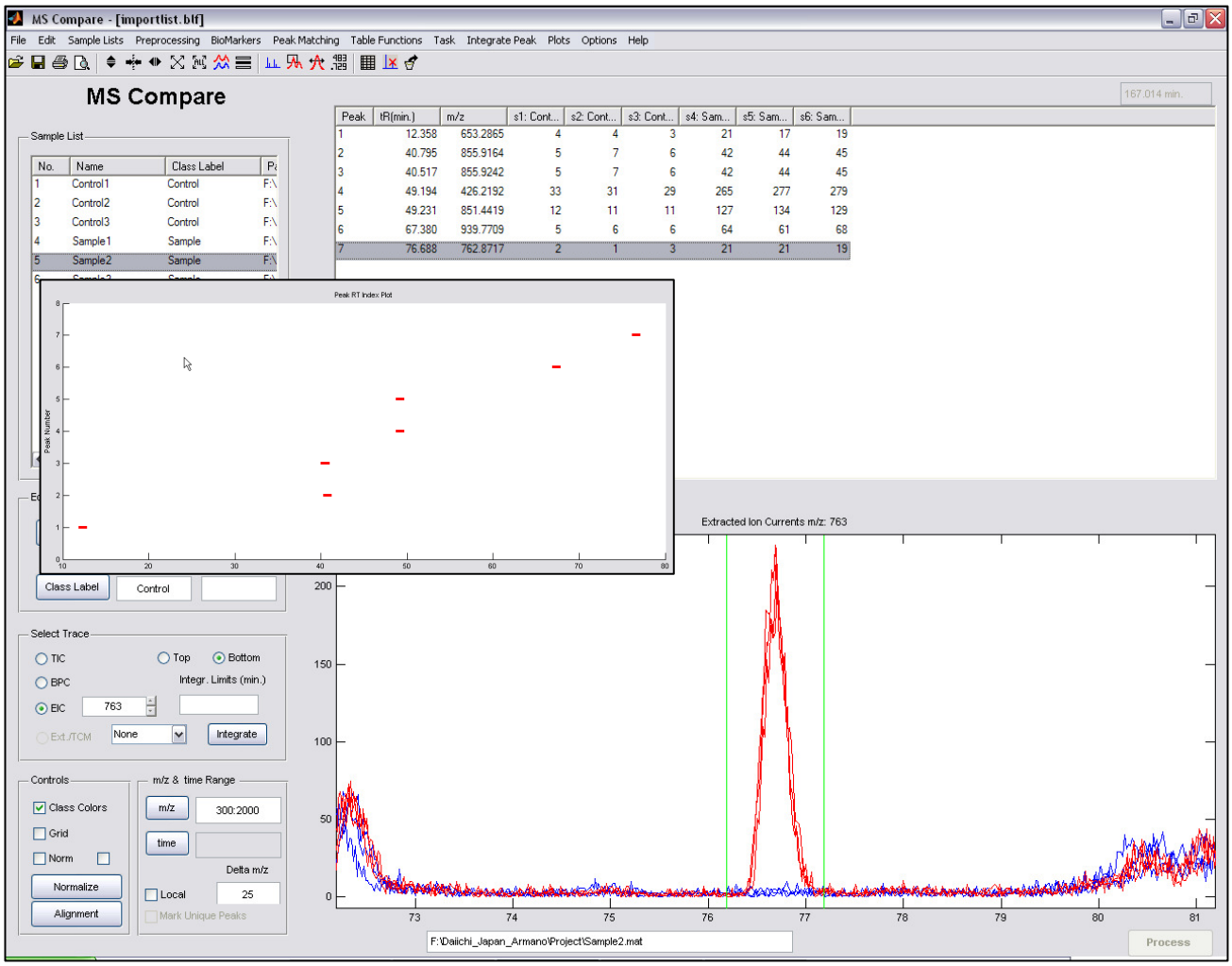


Figure 8: Final table showing 7 up-regulated peaks from the Sample Group as a result from statistical filtering using a Ratio criterion. The inset shows the Retention Time Index plot to visualize co-eluting peaks.

Multivariate Analysis: once the result table has been calculated, the user may want to explore the results using PCA or Clustering. These algorithms are available from the Table section. Examples are giving below for some studies involving more samples. Loading Plots are interactive; clicking on a peak number extracts the EICs or the Intensity Profiles.

Using multivariate analysis tools the user can observe groupings or outliers graphically more easily.

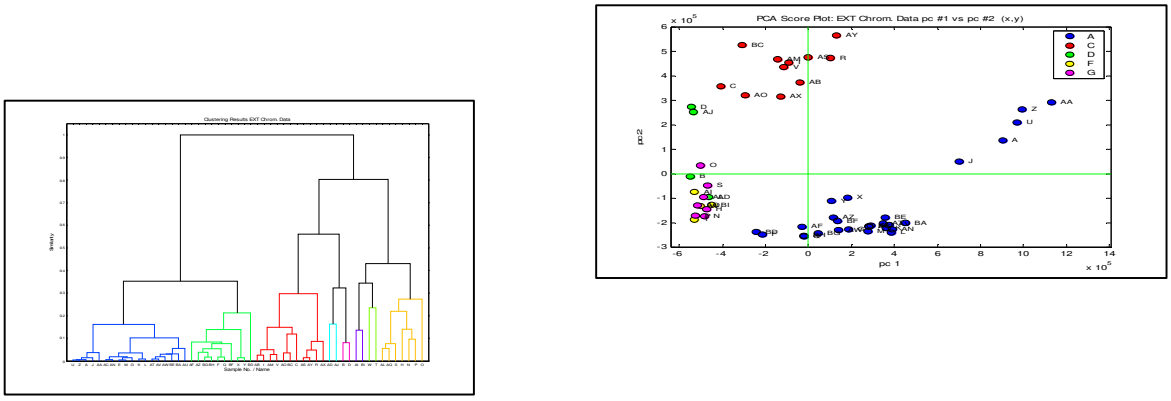


Figure 9: Clustering and PCA Scores Plot