

MET-COFEA User Manual

(Beta version - last updated: 03/12/2014)

The Samuel Roberts Noble Foundation, Inc.

MET-COFEA Description

MET-COFEA is a LC-MS Data Processing Platform for METabolite COMpound Feature Extraction and Annotation, which is aiming to extract and annotate the meaningful chromatograph peak based feature that highly associated with metabolite compounds from the inputting LC-MS files. It mainly includes 3 sequential modules (Figure.1): compound feature extraction, compound feature annotation and compound alignment. Compound feature extraction module include 3 sequential sub-modules: EIC extraction and Peak detection and peak filtering while compound feature annotation module include 3 sequential sub-modules: peak grouping, peak annotation and peak refinement. EIC extraction aims to extract the meaningful mass trace slices from the start scan to the end scan. Peak detection aims to detect the local chromatograph peak for each EIC. Peak filtering aims to filter out some ‘bad’ quality peaks. Peak grouping is to cluster the detected peaks with the close retention time and peak shape similarity. Peak annotation is to further cluster the peaks with the close deduced metabolite compound molecular mass, and then annotate the relationship of each peak and the molecular mass. Peak refinement is to use the annotated peak group results and the open HMDB library to further refine the isolated peak that can’t be annotated but with good peak shape quality. The following figure.1 is the flow chart of MET-COFEA data processing.

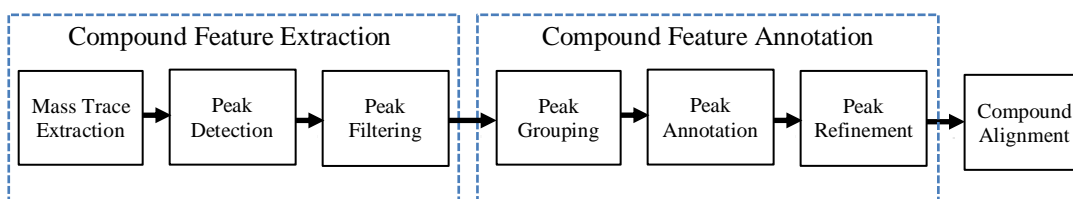


Fig.1 Flow chart of MET-COFEA Data Processing

In the latest version, all of the processing algorithms are coded in C++ and all of visualization parts are coded in C++/CLI. During the data processing, 3 peaklist files named as xxx_chromatograph_peaklist.csv, xxx_grouped_chromatograph_peaklist.csv, and xxx_annotated_grouped_chromatograph_peaklist.csv will be produced for each sample (xxx_ means the sample name). If you choose some samples to align, another aligned peaklist file named as aligned_annotated_grouped_chromatograph_peaklist.csv will also be generated in the end. Additionally, all of the final output peaklist and the intermediate extracted mass traces are stored in database by SQLite, each sample has a corresponding database file named as xxx_annotated_grouped_chromatograph_peaklist.db, all of the files that chosen to align, will have a corresponding database file aligned_annotated_grouped_chromatograph_peaklist.aligndb. The corresponding database file contains the peak shape information. Therefore, it realized the complete separation between data processing and result visualization. For the raw data (.CDF), you can view graphics such as TIC (Total Ion Chromatograph), spectrum data of each scan, 2D display of the raw data, and binning based EIC. For the output results visualization, you can open the database file only to view, the extracted EIC by mass trace method, the detected individual chromatograph peak, peaklist that have been grouped, annotated, and aligned.

The latest version of MET-COFEA support Batch mode and Parallel mode (MPI: Message Passing Interface) to run your multiple samples, depending the core number of your PC. Of course, the data processing time will saved and the required memory will increased, if you run at Parallel mode.

Additionally, considering the compatibility for 32bit and 64 bit CPU in MPI package, we are separated them into two packages. The users should download the corresponding package according to their own CPU hardware.

MET-COFEA Application

The following screenshot is METCOFEA software interface. All the application operation and parameters configuration can be finished by the software. There are 4 main parts (they are displayed as 4 item property page): **Data Process** for raw Data visualization and processing, **Parameter Setup** for processing, **Annotation Result** for individual sample result visualization, and **Alignment Result** for multiple sample visualization after alignment.

Data Process

This property page let user to select the LC-MS data file from d the loaded data file name list(.CDF) and visual the raw data, which include the TIC, spectrum of the specific scan determined by user mouse click position, 2D (spanned by mz-retention time) binarization visualization at the specific cutoff threshold. Additionally, in this property page, user can select the files to run and align from the loaded file name list. After select the parameter file folder (or change the processing parameters according to the property page of Parameter Setting), the user can run the selected files. The following is the normal procedures for this property page:

1. Input the LC-MS data file(s): Click Browse button to select CDF file(s) and click “Load Data” button to display the list of the file name(s) to the table. See Fig.2.

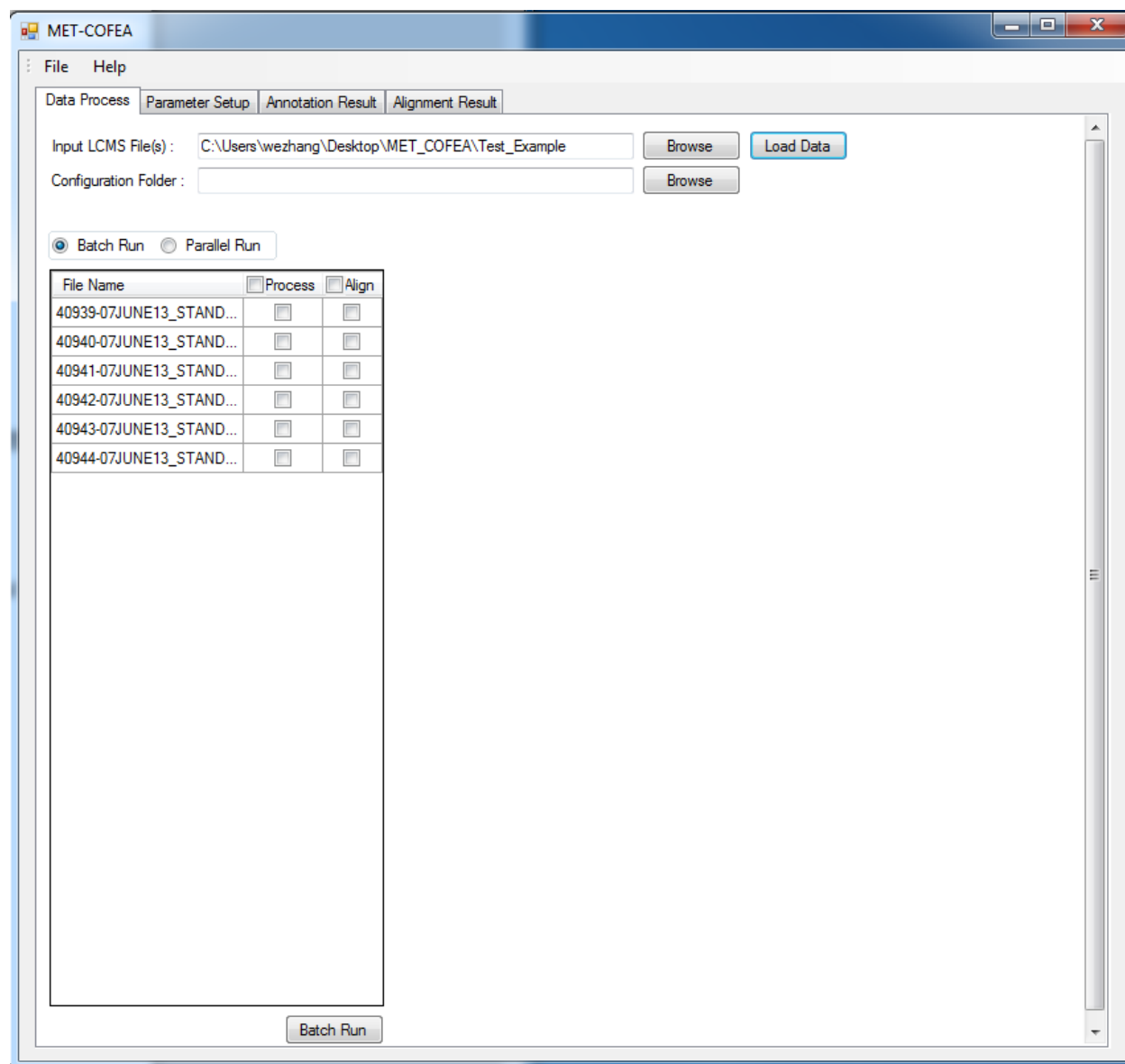


Fig.2 Load Data file name

- Once data file names are loaded, select the file name on the table to visualize the TIC and the spectrum of each scan. TIC can be displayed as scan number mode(see Fig.3) or retention time mode(see Fig.4). Here, the retention time unit is second. Additionally, the raw data also can be displayed as 2D model, if user specific a cutoff threshold(see Fig.5).

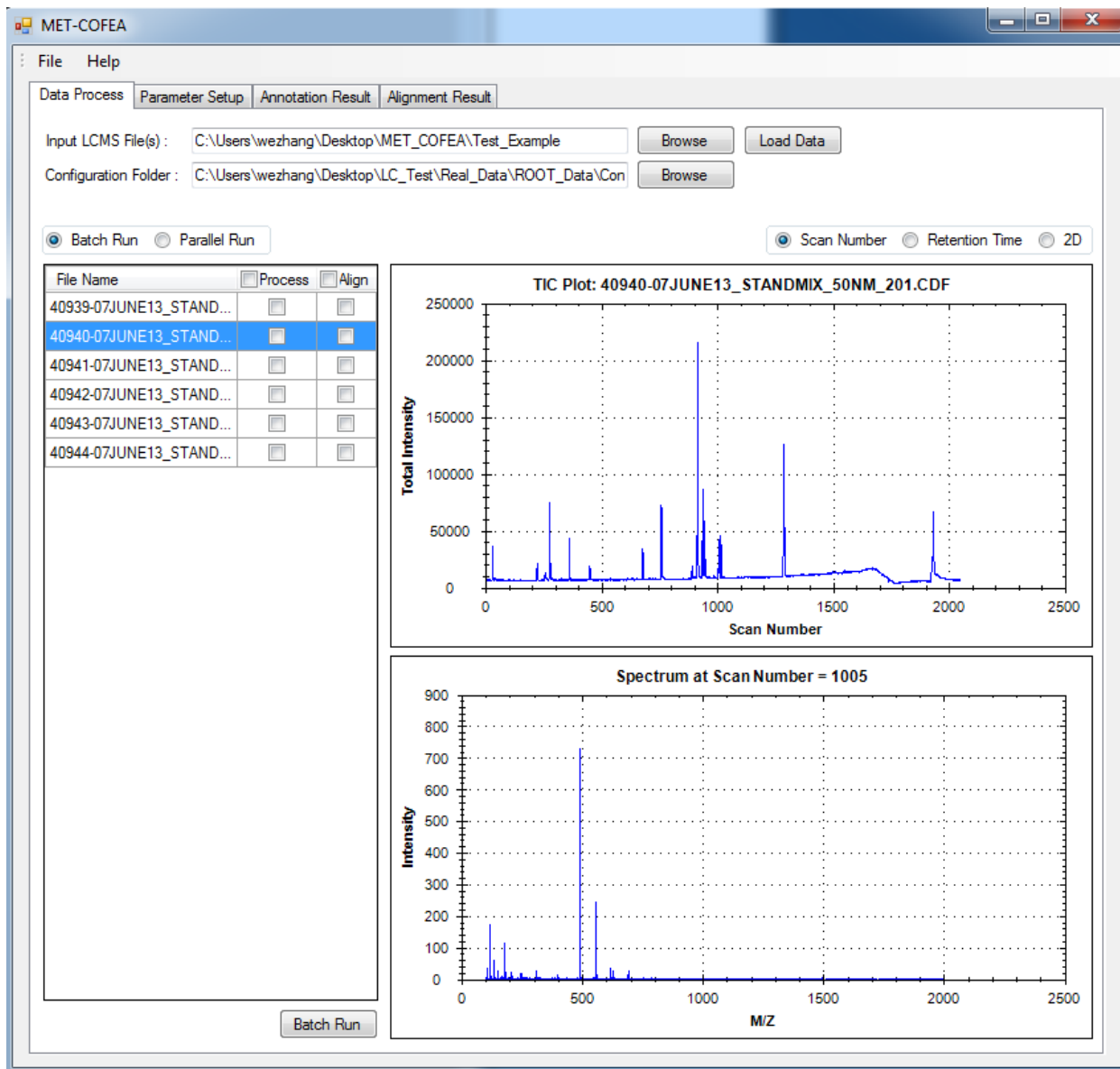


Fig.3 Visualization of Raw LC-MS Data and TIC are plotted at Scan mode

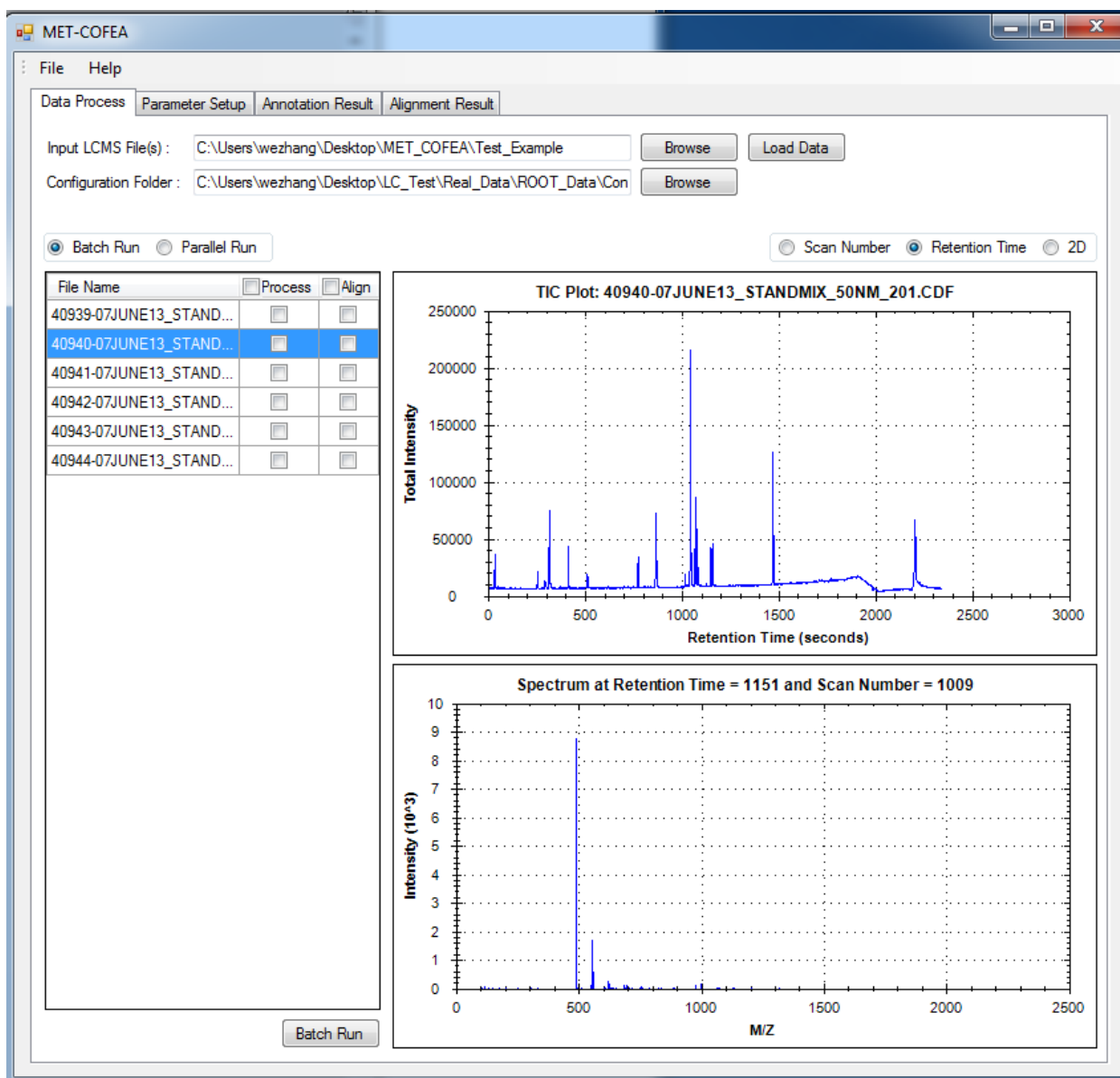


Fig.4 Visualization of Raw LC-MS Data and TIC are plotted at retention time mode

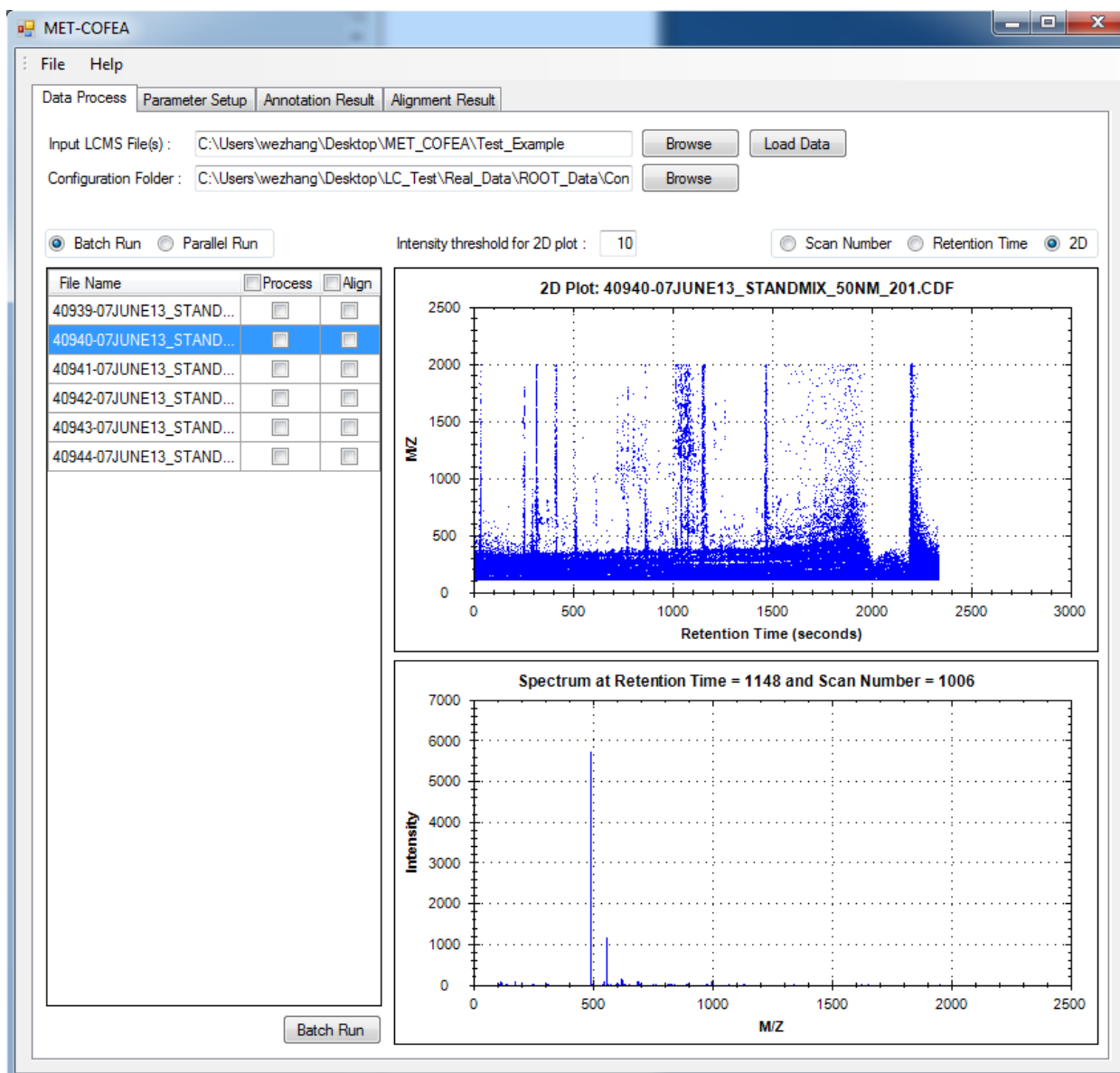


Fig.5. Visualization of Raw LC-MS Data in 2D mode.

3. Select the configuration Folder. Usually, 4 files “config_para.csv”, “Adduct_Config_Negative.csv”, “Adduct_Config_Positive.csv”, and “hmdb.csv” will locate in the same folder. This configuration folder recorded all of the necessary configuration parameters. (Parameter for processing can be further configured by software property page of parameter).
Once you configure the necessary parameters, all of the configuration will be stored. If you want to use the same configuration to run the same or other similar data, you only need to specific the configuration folder path.
4. Check process checkbox and (or) align checkbox for desired sample file(s) to run.

5. Click “Batch Run” or “Parallel Run” button to run. See Fig.6 and Fig.7.
Then a file named as “Job_Run_Config.csv” will automatically generated, which recorded the information for this job. See Fig.8.
6. The output files will be created in the “Result” folder where the LC-MS CDF files are loaded.

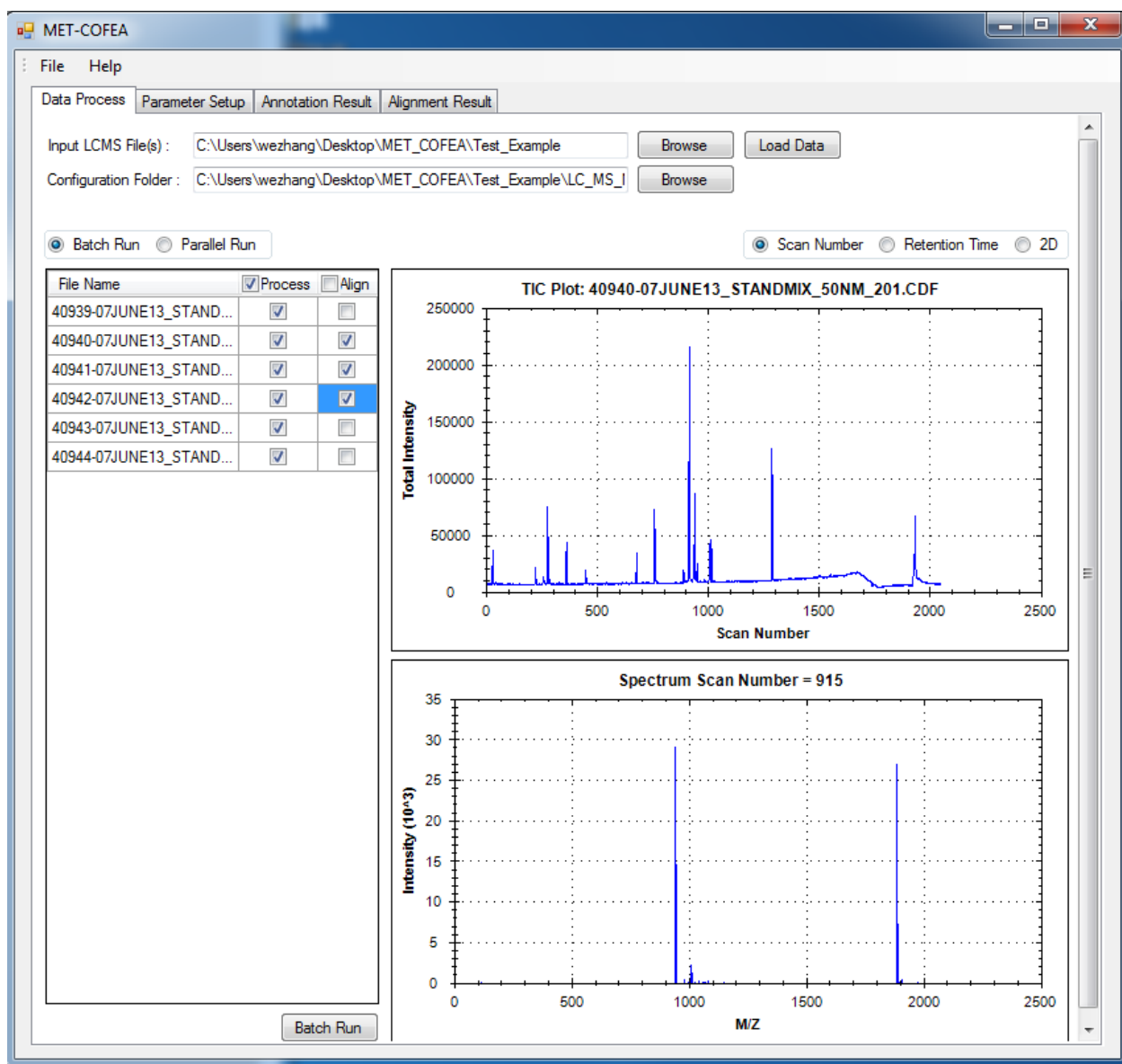


Fig.6 Select files to Batch Run

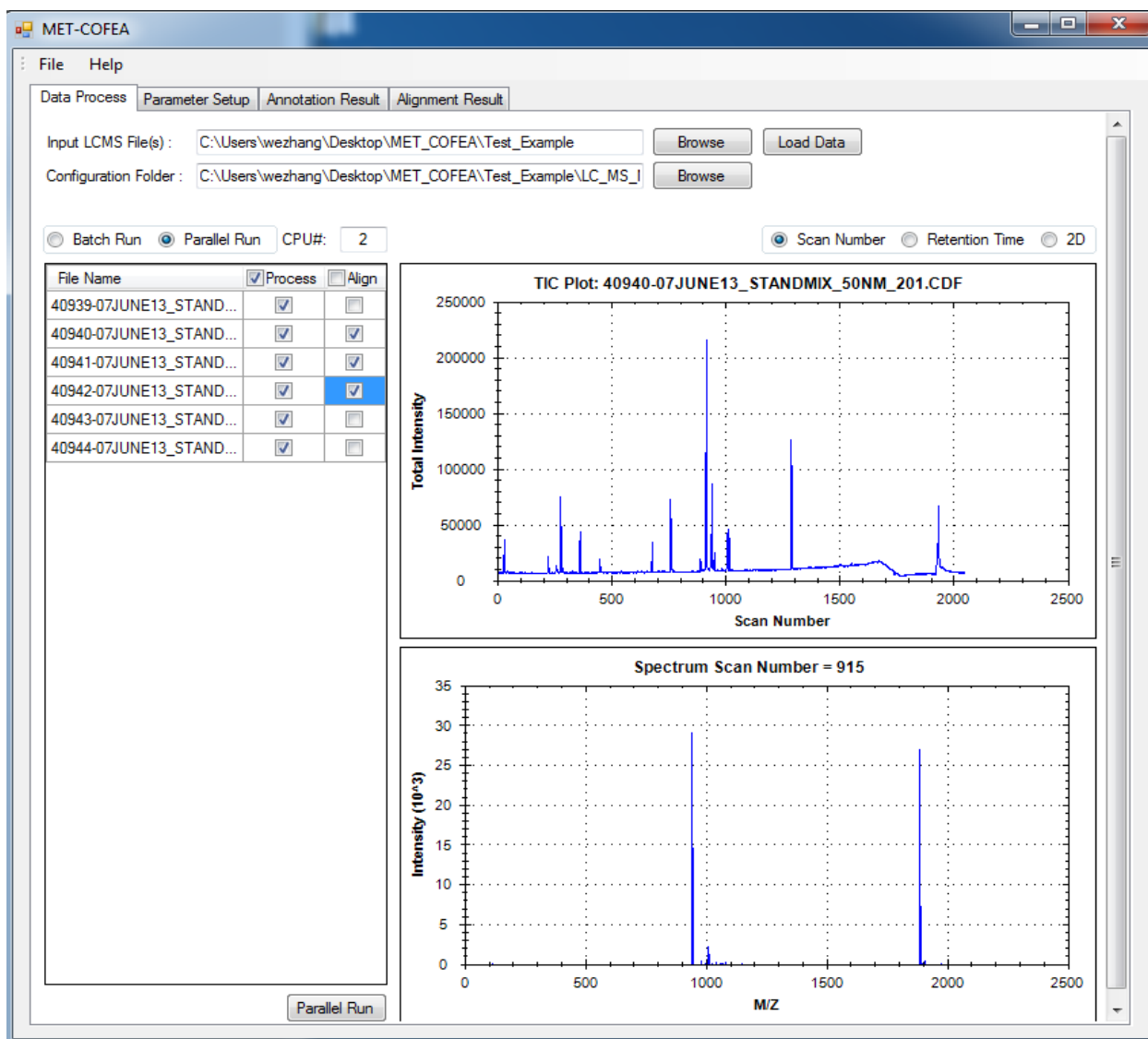


Fig.7 Select files to Parallel Run

	A	B	C	D	E	F	G	H	I	J
1	Job Create Time:	Thu Sep 26 08:59:59 2013								
2	Config Parameter Path	C:\Users\wezhang\Desktop\MET_COFEA\Test_Example\LC_MS_Negative\Config_Files								
3	Data File Path	C:\Users\wezhang\Desktop\MET_COFEA\Test_Example\LC_MS_Negative\Data_File								
4	Output Path	C:\Users\wezhang\Desktop\MET_COFEA\Test_Example\LC_MS_Negative\Data_File\Result								
5	Run Mode	Parallel		M CPU Num	2					
6	Process File Name	Aligned								
7	40939-07JUNE13_STA	0								
8	40940-07JUNE13_STA	1								
9	40941-07JUNE13_STA	1								
10	40942-07JUNE13_STA	1								
11	40943-07JUNE13_STA	0								
12	40944-07JUNE13_STA	0								
13										
14										
15										

Fig.8 Content of Job Run File

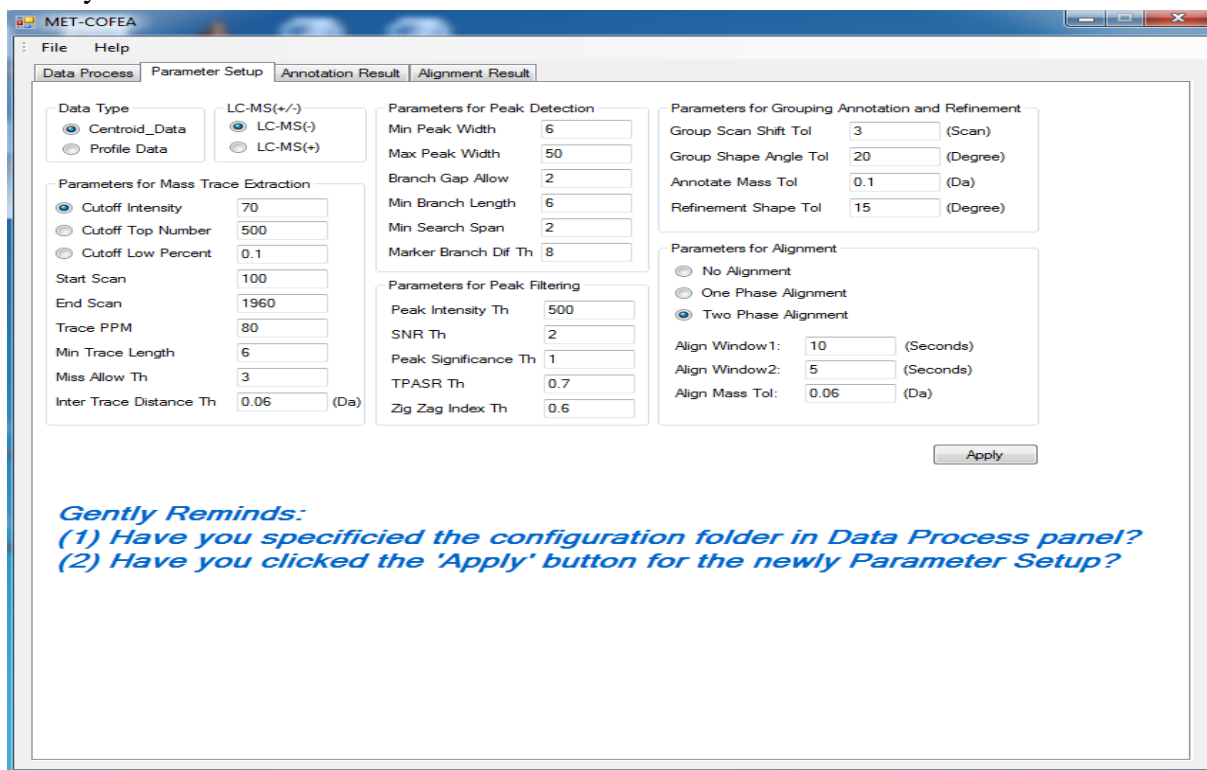
Parameter Setup

In this property page, the user can configure the processing parameter (See Fig.9). The different parameter setting will generate very different results. If the user wants to acquire the optimal results by setting the optimal parameters, please read the parameter explanation section and parameter optimization section first.

After parameter configuring, then Click “Apply” button, it will save modified parameters to file named “config_para.csv” in the configuration Folder you select (see Data Process property page). Only the user click the button of ‘Apply’, the newly configured parameters can be loaded into parameter setup panel. In the Data Process property page, if the user specific the configuration Folder, all of configuration including the parameter file in that folder will be loaded and applied to Data process. If the user wants to use the new configured parameter, only need to click the “Apply” and then go back to “Data Process” page, and click Batch Run or Parallel Run.

The parameters configuration includes 4 parts: Data type configuration, Parameter configuration for Mass traces (EIC) Extraction, Parameter configuration for CWT based Peak detection, Parameter configuration for peak quality filtering, Parameter configuration for peak grouping, annotation and refinement, Parameter configuration for Compound alignment. Regarding Data type configuration, you need to select the inputting data as profile data or centroid data, LC-MS (+) or LC-MS (-). For profile data, the centroid processing module will be called at first. These data configuration should be correct, otherwise, can’t get the correct meaningful run result.

The same Job (Job_Run_Config.csv) and the same processing parameter file (config_para.csv) usually can ensure the same result for the same data file set.



Gently Reminds:

- (1) Have you specified the configuration folder in Data Process panel?
- (2) Have you clicked the 'Apply' button for the newly Parameter Setup?

Fig.9 Parameter Configuring

In the folder of file config_para.csv, another three files Adduct_Config_Negative.csv, Adduct_Config_Positive.csv and hmdb.csv should also be existed during the whole data processing period. The first two files are used for peak annotation according to the metabolite type of the data. The user can configure them according to their experiment object, but in the current version, the user only can configure them at edit mode. File hmdb.csv is a library that contains some metabolite peak fragment, which will be searched in peak refinement module. We think user will less touch library file hmdb.csv in their real application, so, we suggest user not to touch it. Once the metabolite fragment/ adduct type are fixed, file Adduct_Config_Negative.csv, Adduct_Config_Positive.csv should be kept untouched.

Annotation Result

In this property page, the user can view the chromatograph peak shape for each detected peak, peaklist for a group with the same Group_ID, peaklist for a compound with the same Compound_ID. We separated peaklist into different Group_ID, if the peaks' retention time (corresponds to peak apex) and peak shape meets some criteria. We separate the peaklist further into different Compound_ID, if their observed m/z value has some relationship to the same compound's molecular mass. The peaklist with the same Group_ID may contain several sub-peaklist with different Compound_ID. So, in this property, MET-COFEA visualization can clearly provide the relationship of an individual peak with the related peaklist with the same Group_ID, the relationship of an individual peak with the related peaklist with the same Compound_ID. The user only need to click the radio check button at "Peak Plot", "Group Plot", "Compound Plot", the detailed visualization for the individual peak, peaklist with the same Group_ID, peaklist with the same "Compound_ID" will be plotted.

Additionally, MET-COFEA visualization also can clearly provide the relationship of the individual peak and the related whole extracted EIC by mass trace method or binning method (you need to specific the raw CDF file).

According to the peaklist with the same Group_ID, or Compound_ID, the corresponding pseudo-spectrum can be constructed and plotted. So, the user can export the constructed pseudo-spectrum. Compared with the raw mixed spectrum, the pseudo-spectrum constructed with the same Group_ID become more pure by separation according to the peak shape, however, the pseudo-spectrum constructed with the same Compound_ID become further pure by further separation according to the annotation based on the each peak's observed m/z value and the common molecular mass.

The following is the normal procedures for this property page:

1. Select the result file (xxx_annotated_grouped_chromatograph_peaklist.db) from result directory. Usually, after the data processing, the peaklist file will be generated in the same folder of the original LC-MS data files located.
2. Click a cell in the table to visualize the individual peak (Check Radio Peak Plot), associated peaks with the same Group_ID (Check Radio Group Plot), associated peaks with the same Compound_ID (Check Radio Compound Plot).
3. Export the constructed pseudo-spectrum into a txt file for future library building.

4. Select the raw CDF file and manually configure the m/z tolerance (ppm or Da) to view the binning based EIC.

If the user wants to view the meaningful peak or peaklist, you can click the head of the column with the Group_ID, Compound_ID, or Apex_MZ, Compound_MolecularMass, and then the whole table displayed in the left part will be sorted as ascending or descending order. The user can view the associated peaks in the lower right part at the normalized (focus on peak shape similarity) or standard, Scan Number or Retention time mode, see Fig.10, Fig.11 and Fig.13.

For Group Plot and Compound Plot, a pseudo-spectrum will be generated according to the m/z values and apex intensity values of the associated peaks. Compared with the original spectrum of specific scan, the constructed pseudo spectrum with the same Group_ID will be more pure, because only the part with the near chromatograph shape is kept, however, the constructed pseudo spectrum with the same Compound_ID will be even more pure, because only the part that have some relationship between the observed m/z and molecular mass are kept. The constructed pseudo spectrum can be considered as a multi feature of a metabolite compound or finger pattern, and can be used to identify the same metabolite in different samples; even we don't know their chemical structure. So, the constructed pseudo-spectrum can be used to build a library. The constructed pseudo-spectrum using peaklist with same Compound_ID and Group_ID are showed in Fig.12 and Fig.14.

From the extracted EIC, the user can know the shape information of the front and back of the specific peak. If the raw CDF file data is selected, the binning EIC and TIC also can be displayed. Compared the difference between the extracted EIC and binning EIC (specific m/z and tolerance), the user can optimize parameter for mass trace extraction. See Fig.15.

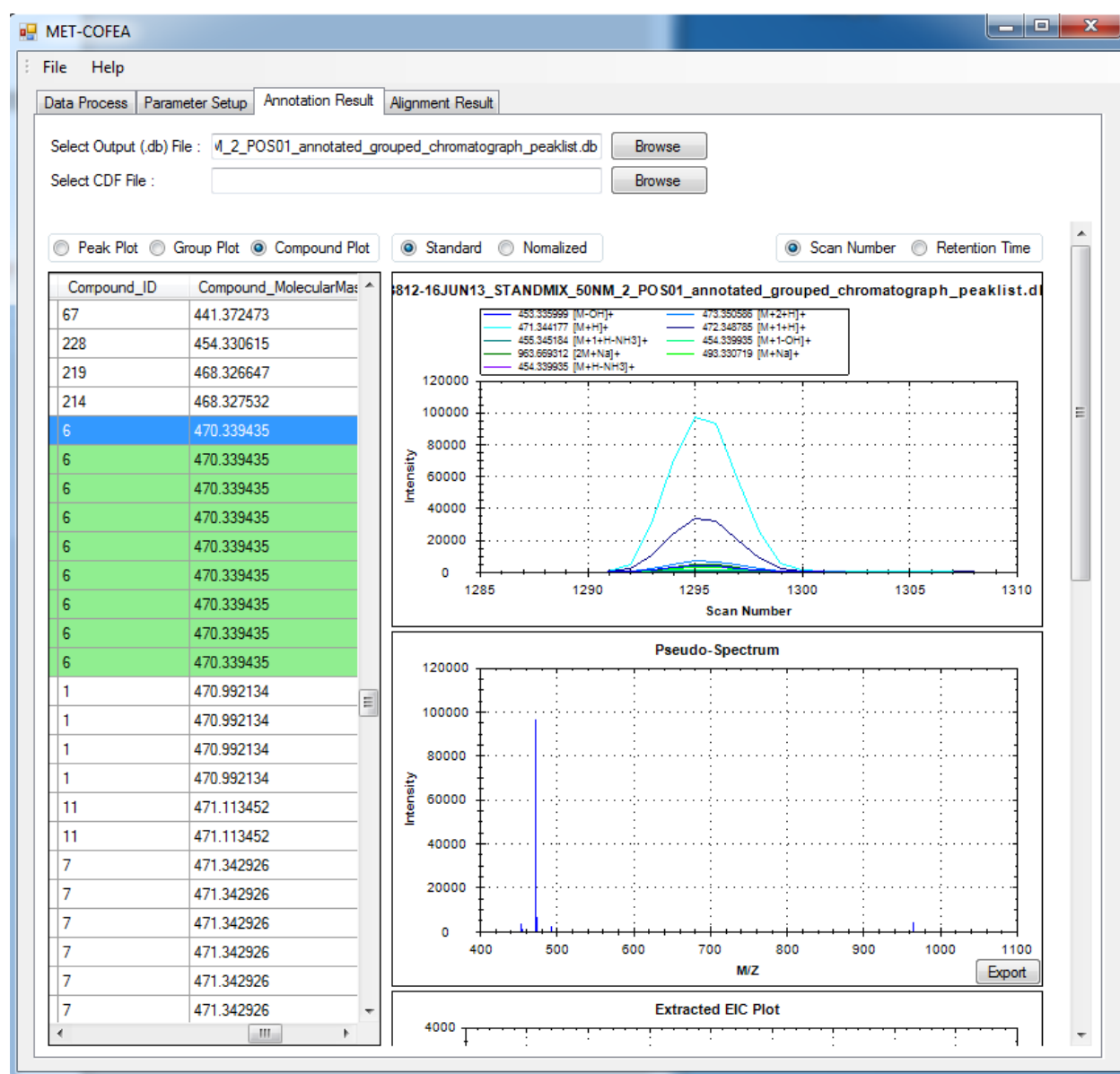


Fig.10 Visualization of associated un-normalized peaks with the same Compound_ID and its Pseudo-Spectrum

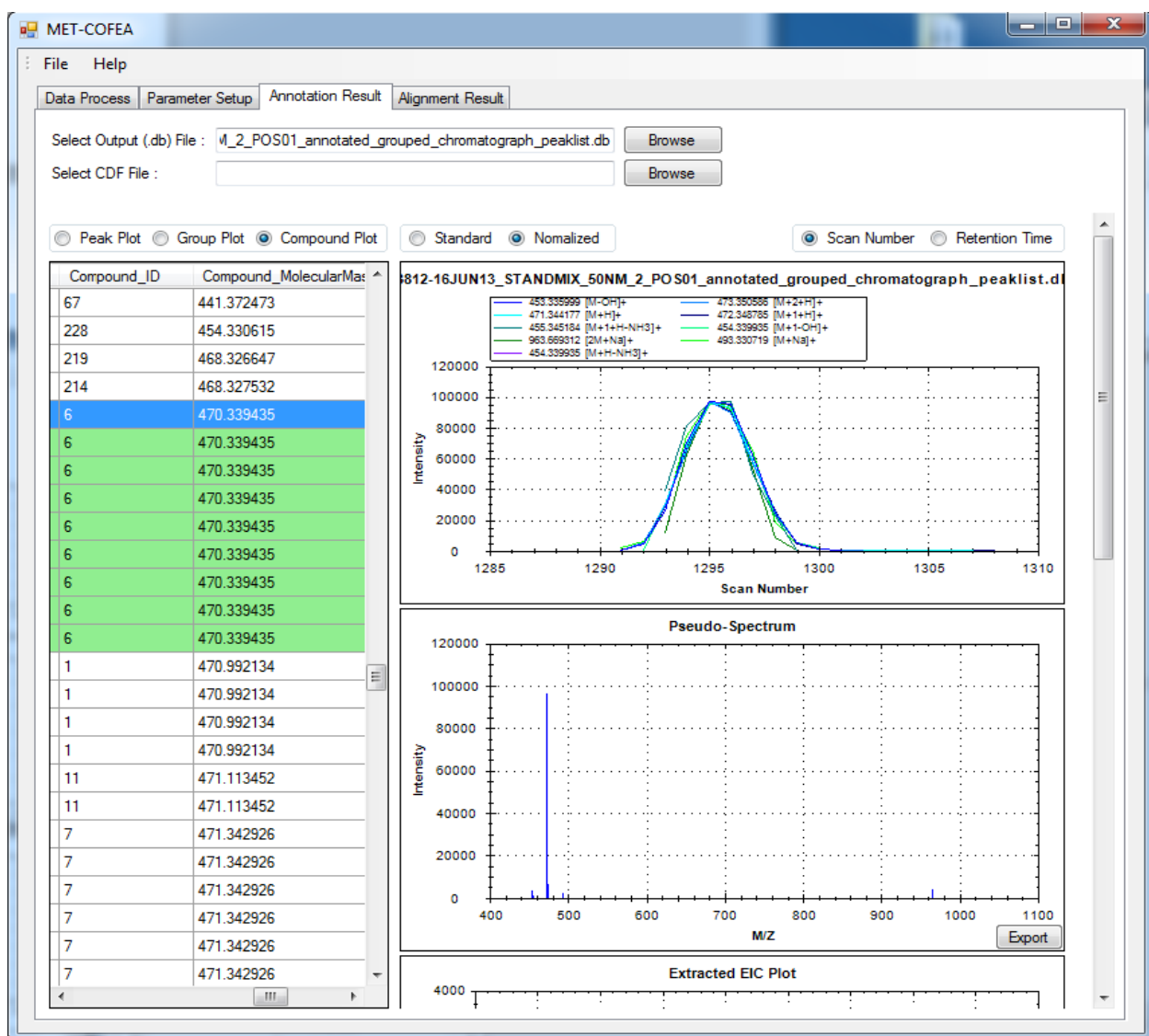


Fig.11 Visualization of associated Normalized peaks with the same Compound_ID and its Pseudo-Spectrum

pseudospectrum_compound_id_6.txt - Notepad

M/Z	Intensity	Annotation
453.335999	3588.564453	[M-OH] ⁺
473.350586	6803.632813	[M+2+H] ⁺
471.344177	96814.8125	[M+H] ⁺
472.348785	33629.6875	[M+1+H] ⁺
455.345184	249.274048	[M+1+H-NH ₃] ⁺
454.339935	1237.451172	[M+1-OH] ⁺
963.669312	4373.191406	[2M+Na] ⁺
493.330719	2600.849609	[M+Na] ⁺
454.339935	1237.451172	[M+H-NH ₃] ⁺

Fig.12 Content of Pseudo-Spectrum exported by peaklist with Compound_ID=6

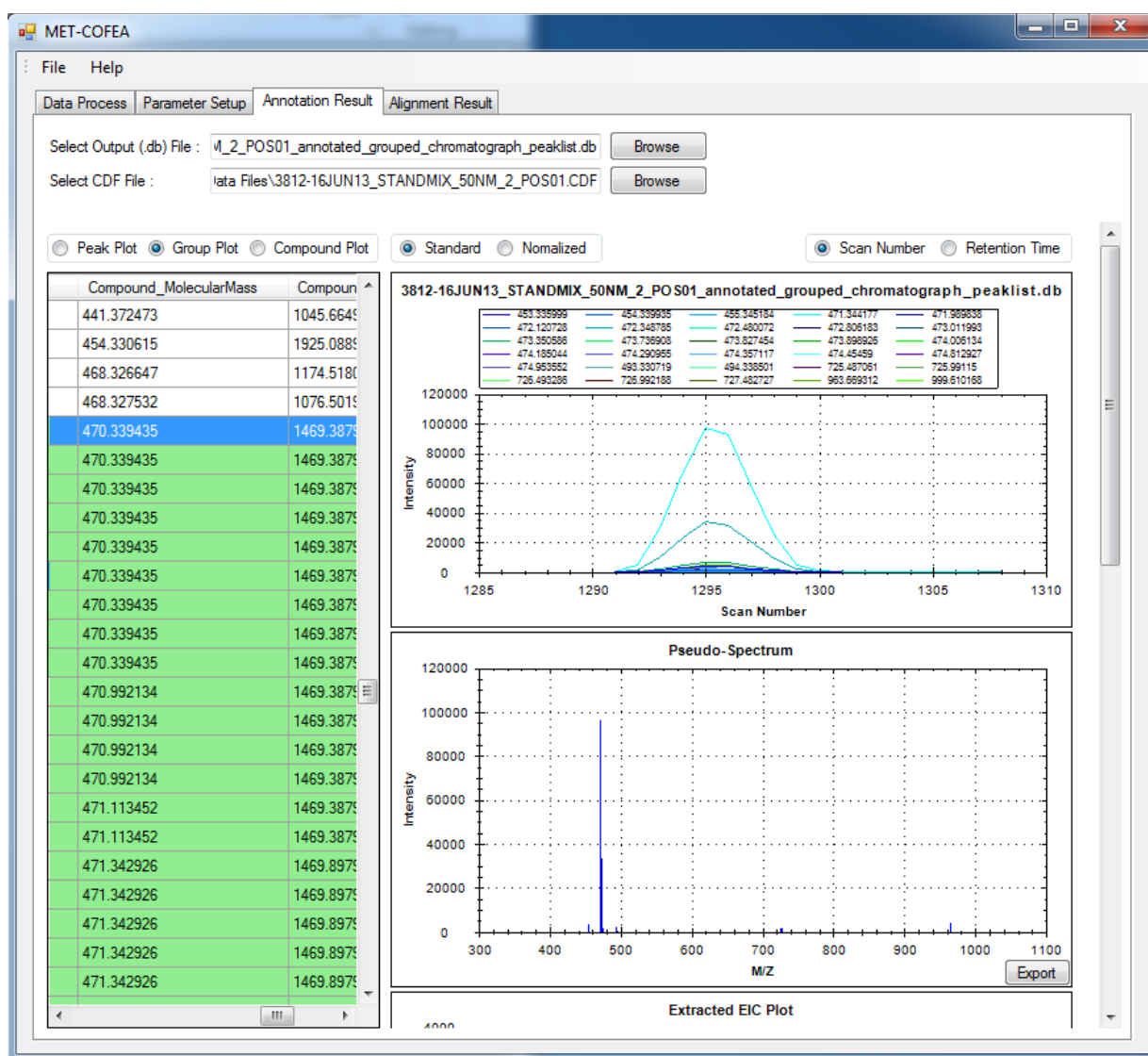


Fig.13 Visualization of associated un-normalized peaks with the same Group_ID and its Pseudo-Spectrum

pseudospectrum_group_id_1.txt - Notepad

M/Z	Intensity
453.335999	3588.564453
454.339935	1237.451172
455.345184	249.274048
471.344177	96814.8125
471.989838	389.853271
472.120728	685.449219
472.348785	33629.6875
472.480072	716.913086
472.806183	919.86084
473.011993	834.037109
473.350586	6803.632813
473.736908	851.044922

Fig.14 Content of Pseudo-Spectrum exported by peaklist with Group_ID=1



Fig.15 Visualization of the extracted EIC for the specific peak and the binning EIC from raw CDF file

Alignment Result

In this property page, the user can visualize the peaklist with the same Align_ID across different samples. From the visualization (see Fig.16), all of the peaks with the same Align_ID across different samples are plotted and the peaks from the same Sample are plotted with one specific color (Right top panel). Additionally, the peak associated with the mouse click can also be plotted individually, or with the associated peaklist with the same Compound_ID, Group_ID (Right bottom panel). So, in this property page, the user can clearly know the relationship of the same compound associated peaks across different samples, and the relationship between the individual peak and the peaklist with the same Compound_ID, Group_ID.

The following is the normal procedures for this property page:

1. Select the result file name “aligned_annotated_grouped_chromatograph_peaklist.aligndb”.
(This database file will be generated only if you choose some files for alignment.(See the Property Page of Data Process)
2. Switch the retention time mode between RT_Original and RT_Corrected to check the alignment results.
3. Click a cell in the table to visualize the peaklist that have been aligned into the same Align_ID across different sample files.

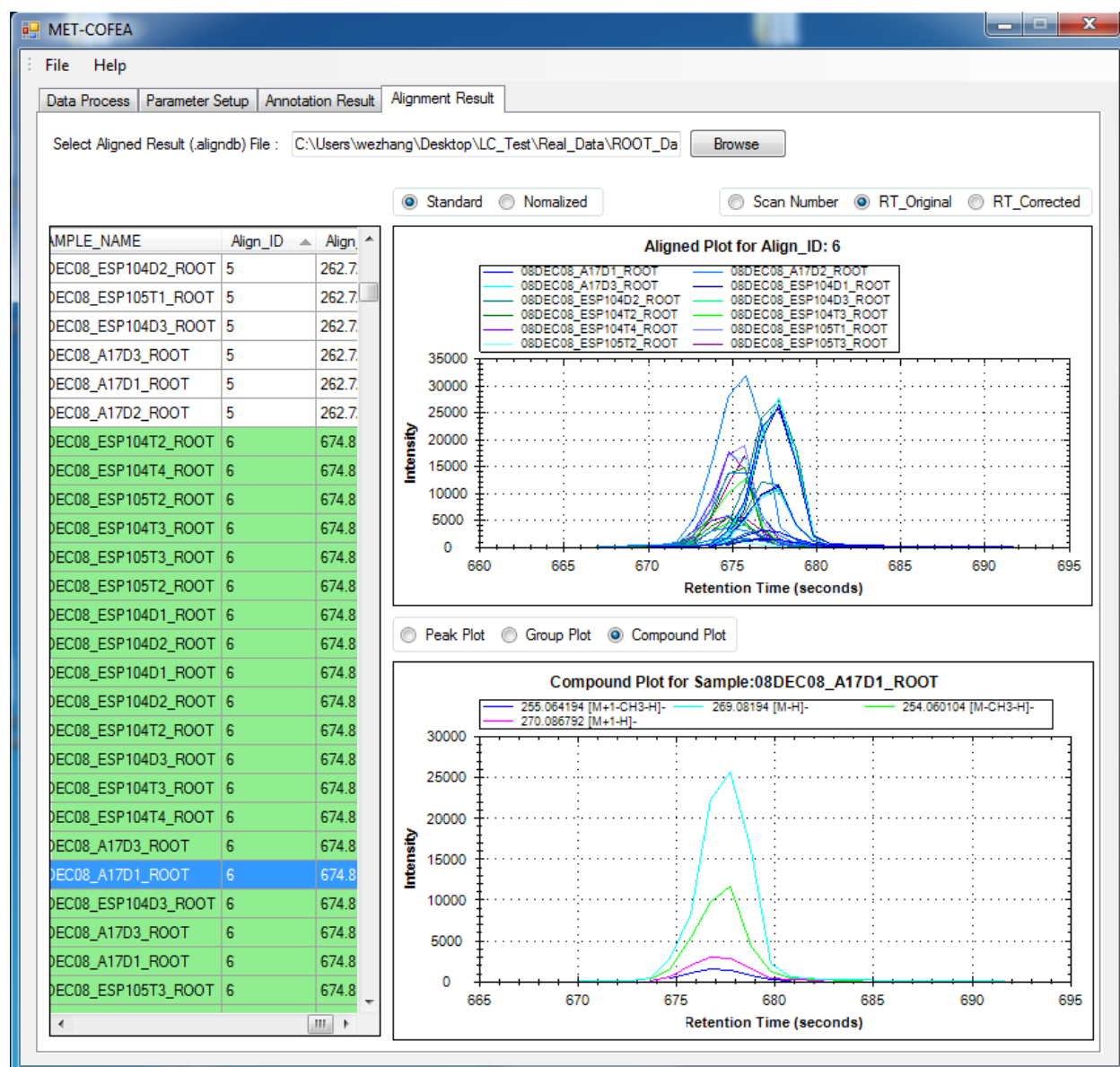


Fig.16 Visualization of peaklist with the same Align_ID across different samples displayed at RT_Original mode

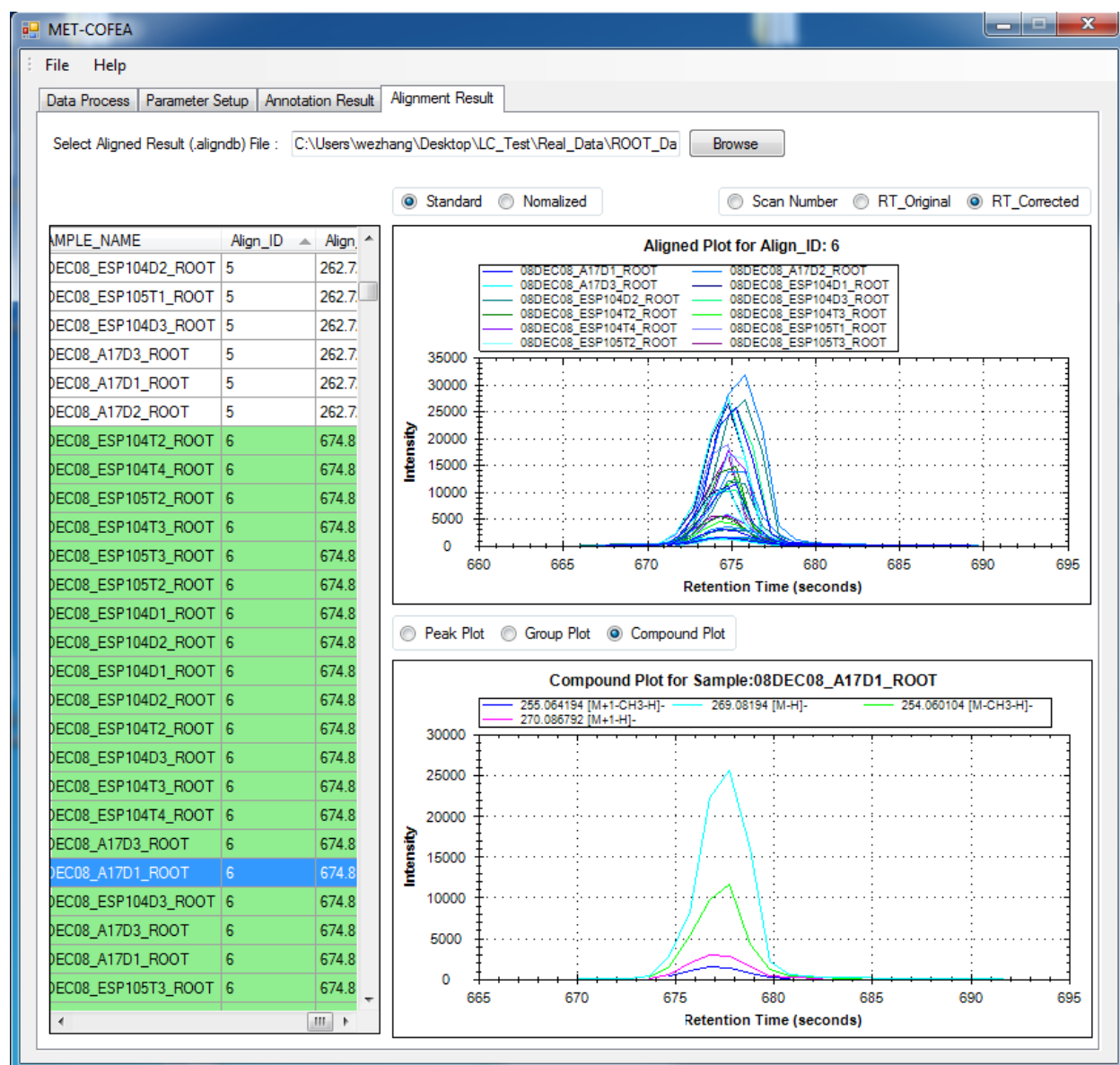


Fig.17 Visualization of peaklist with the same Align_ID across different samples and displayed at RT_Corrected mode

Parameter Explanation

1. Cutoff Intensity: Intensity cutoff threshold for each scan. Only the data point with the intensity is larger than the threshold is used to do mass trace extraction, by which the low intensity noisy data point can be filtered
2. Cutoff Top Number: For each scan, only the Top Number intensity data points are considered to do mass trace extraction, by which the low intensity noisy data point can be filtered.
3. Cutoff Low Percent: For each scan, only the percentage of low intensity data points is filtered.
4. Start Scan: Specific the start scan for mass trace extraction.
5. End Scan: Specific the end scans for mass trace extraction.
6. Trace PPM: Specific the PPM threshold of m/z value variation for a valid mass trace.
7. Min Trace Length: Specific the minimum mass trace length.
8. Miss Allow Th: Specific the maximum allowed miss data point number during the mass trace extraction.
9. Inter Trace Distance Th: Specific the minimum distance between two neighboring mass trace

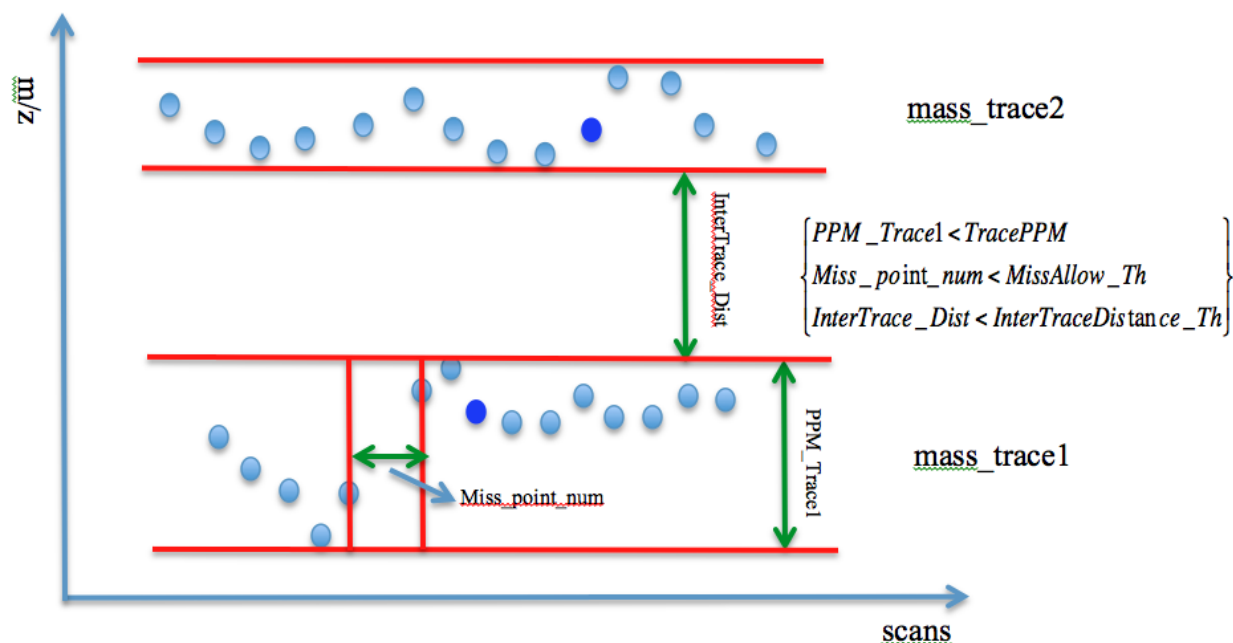


Fig.18 Illustration for some Parameter of Mass Trace

10. Min Peak Width: The minimum width of a valid peak, the very thin peaks with peak width smaller than MinPeakWidth will be filtered.
11. Max Peak Width: The maximum width of a valid peak, the very fat peaks with peak width larger than MaxPeakWidth will be filtered.

12. Peak intensity Th: The minimum intensity value of a valid peak, the very low intensity peaks with intensity smaller than Peakintensity_Th will be filtered.
 13. SNR Th: The minimum S (Signal)/N (Noise) value of a valid peak, the peaks with SNR smaller than SNR Th will be filtered. SNR is defined in the wavelet domain by the ratio of the CWT coefficient at marker point to 95% quintile of the absolute CWT coefficient in scale 1.
 14. Peak Significance Th: The minimum Peak Significant level of a valid peak, the peaks with the Peak Significant level smaller than Peak Significance Th will be filtered. Peak significant level is defined by the ratio between the mean intensity value of data points near the peak apex and the mean intensity value of data points near the two boundaries.
 15. TPASR Th: The maximum Triangle Peak Area Similarity Ratio (TPASR) of a valid peak, the peaks with TPASR larger than TPASR Th will be filtered.
- Triangle Peak Area Similarity Ratio (TPASR) is defined as the following formula,

$$\begin{cases} TPA = 0.5 * Peak_Width * Intensity(Peak_Apex) \\ RPA = \sum_{i=Left_Boundary}^{Right_Boundary} Intensity(i) \\ TPASR = \frac{|TPA - RPA|}{TPA} \end{cases}$$

Here, TPA is the Triangle peak area and RPA is the real peak area. TPASR provides an index for the closeness of the detected peak and triangle peak in area. The TPASR value is more close to 0, the better of the peak quality.

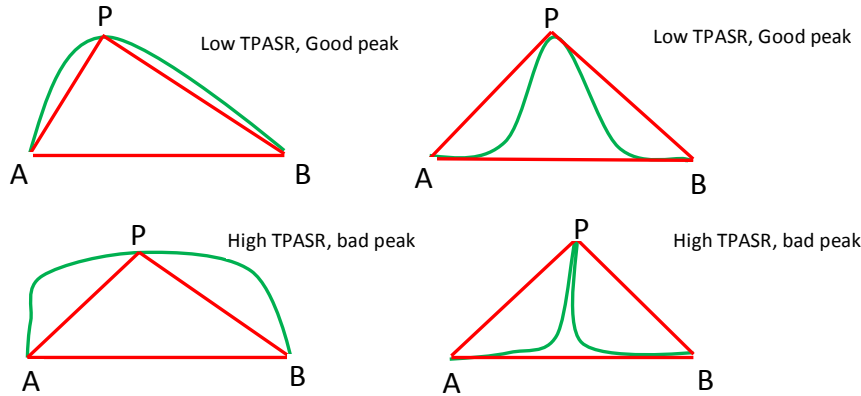


Fig. 19 Illustration of Parameter of TPASR for fat peak and thin peak

16. Zig Zag Index Th: The maximum Zig Zag Index of a valid peak, the peaks with Zig Zag Index larger than Zig Zag Index Th will be filtered. Zig Zag Index is adopted to evaluate the degree of zig-zag of a chromatograph peak. Zig Zag Index can be calculated by the following procedure.

Suppose the intensity array of a chromatograph peak is represented as

$$I_1, I_2, \dots, I_{n-1}, I_n, I_{n+1}, \dots, I_N.$$

- 1) Calculate the effective peak intensity by subtract the baseline at the peak apex.
 $EPI = \text{Max}(I_1, I_2, \dots, I_{n-1}, I_n, I_{n+1}, \dots, I_N) - \text{Baseline}(\text{Apex}).$
- 2) Calculate the first order derivative of the peak and acquire the increment for each data point pair.
 $d_n = I_n - I_{n-1}, d_{n+1} = I_{n+1} - I_n \quad n = 2, 3 \dots N.$
- 3) Calculate the variance of each two neighbor increment pair as:
 $v(d_n, d_{n+1}) = (d_n - d_{n\text{mean}})^2 + (d_{n+1} - d_{n\text{mean}})^2$ and $d_{n\text{mean}} = \frac{(d_n + d_{n+1})}{2.0}$
 After some simple deducing, the variance can be represented as:
 $v(d_n, d_{n+1}) = 0.5 * (2I_n - I_{n-1} - I_{n+1})^2$
 Here $(2I_n - I_{n-1} - I_{n+1})^2$ indicate the local zig zag degree of data point I_{n-1}, I_n, I_{n+1} .
- 4) Sum all of the local zig zag, we get
 $\text{Sum_zig_zag} = \sum_{n=2}^{n=N-1} (2I_n - I_{n-1} - I_{n+1})^2$
- 5) Calculate the average and normalized Sum_zig_zag then get Zig Zag Index as following:

$$\text{Zig_Zag_Index} = \frac{\sum_{n=2}^{n=N-1} (2I_n - I_{n-1} - I_{n+1})^2}{N * EPI^2}$$

Based on the real data's testing, the proposed Zig_Zag_Index can evaluate the zig zag degree of a chromatograph peak shape, and the lower the Zig_Zag_Index, the better the peak quality.

Parameter 17-20 is about peak branch pattern detection in CWT domain. In MET-COFEA, 1D mass trace (EIC) is firstly transformed into 2D CWT coefficients. Then local maximum detection is utilized for each scale. Several continuous meaningful local maximum points across the 2D scan-scale space are defined as a meaningful peak pattern branch; in general, a meaningful branch should be composed of the local maximum points across several continuous scales, and corresponds to one valid peak of the original EIC. A meaningful peak branch pattern should be larger than a specific length, and its searching span should be limited to a specific value, and all of its Branch pattern searching gap should be smaller than a specific value.

17. Min Branch Length: The minimum value for a meaningful branch. The branch with its final length is smaller than Min Branch Length will not be considered a valid peak branch pattern.
18. Min Search Span: The minimum search span for search another local maximum point in its neighboring scale.
19. Branch Gap Allow: The maximum gap across several continuous neighboring scales. Only the branches with its maximum gap is smaller than Branch Gap Allow are considered as a meaningful branch, and finally a meaningful profile peak.
20. Marker Branch Dif Th: For all data points of a branch, the point with its coefficient value larger than its neighboring scales is defined as a marker points. Usually, for the good quality peak shape, there is only one marker point for a branch. But at the case of peak overlapping or low peak shape quality, there maybe exists several Marker points for one branch. If the distances of two marker points across scans are larger than Marker Branch

Dif Th, the branch should be split into two meaningful branches, and finally two peaks should be identified.

A: The original EIC signal. B: CWT coefficients at different scales. C: local maximum detection for each scale and 3 meaningful branches can be recognized, for the branch 2, there exist two marker points and also the distance of the two marker points are larger than Marker Branch Dif Th. So, there are 4 meaningful peaks are detected. D: The peak's parameters such as peak's apex, left/right boundary, etc. are retrieved according to the detected marker points.

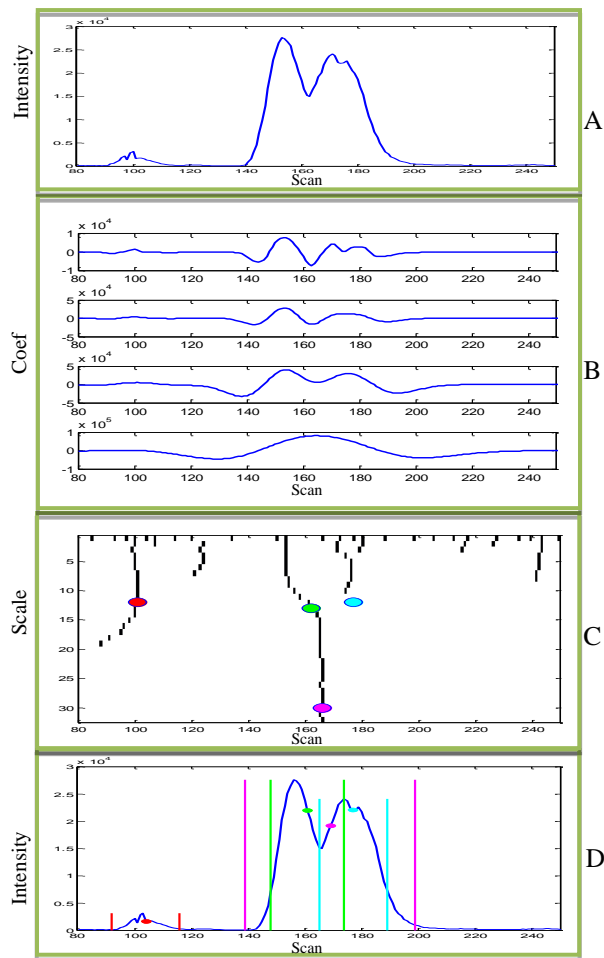


Fig. 20 CWT-based peak detection

21. Group Scan Shift Tol: The peaks with its peak apexes fall in the ranges of Group Scan Shift Tol can be considered a peak group.
22. Group Shape Angle Tol: The peaks with their peak shape similarities (defined by dot product and then \cos^{-1}) are smaller than Group Shape Angle Tol can be considered as a meaningful peak group. Here, HCA (Hierarchical Cluster Analysis) are adopted.

23. Annotate Mass Tol: Only the molecular mass form the peak can be considered as a meaningful compound group. The relationship of the observed mass signal S (m/z) and its possible molecular mass are.

$$M_{i,j} = \frac{S_i * Para_CS_j - Para_Mass_Shift_j}{Para_N_j}$$

24. Refinement Shape Tol: For the isolate peaks that can't be divided into any group by the peak annotation process, if their average peak similarity (defined by dot product and then \cos^{-1}) between the isolate peak and one compound group is smaller than Refinement Shape Tol, then the isolate peak is added into that group, and annotated as a unknown fragment peak.
25. None/One/Two Phase alignment: If you need alignment to align the compound associated peaklist across samples, please choose one or two phase alignment strategy.
26. Align Window1: The align window for the first phase alignment across retention time.
27. Align Window2: The align window for the second phase alignment across retention time. The second alignment means a more accurate align. So, $AlignWindow2 < AlignWindow1$.

Parameter Setting and Optimization Procedure

Given one LC-MS Data sample, parameter configuration will affect the processing speed and performance greatly. Usually, the loosen parameter can get more peak information but with the high computation burden and longer computation time. The user needs to balance it in real application. If there are not too much dataset, the user always wants to acquire the optimal analysis results. Here, I provide a normal procedure for parameter setting and optimization.

1. Open a representative CDF file (refer to Data Process property page). Determine the Start Scan and End Scan from the TIC curve.
2. Move mouse to a representative scan, from the corresponding spectrum, determine the intensity cutoff threshold. Bear in mind, only the higher intensity point of a scan will be used for EIC extraction and peak detection.
3. Still from the opened representative spectrum, zoom out, you can easily know the data is centroid data or profile data. In profile data, the spectrum displayed as many spectrum peaks while centroid data displayed as sticks.
4. From the mass spectrometry instrument facility personnel, you should know the data type, whether it is LC-MS (+) or LC-MS (-), the polarization type is usually determined by the structure type of metabolite compounds you are interested. In human metabolomics, usually adopt LC-MS (+) while in plant, usually adopt LC-MS (-).
5. Parameter configuring for EIC extraction.

Bear in mind, MET-COFEA adopt a more advanced method to extract EIC, mass trace (not binning) based method, in principle, it is same to object tracing problem in

video tracing. So, a meaning mass trace (EIC) should be a continuous point trace across several continuous scans and the m/z value varies (shift) in the specific tolerance.

5-a, from the mass spectrometry instrument facility personnel, you should know the mass spectrometry accuracy/tolerance, and then you can configure Trace PPM.

5-b, also, you can know the minimum meaningful chromatograph peak width from the mass spectrometry instrument facility personnel, and then you can configure the Min Trace Length.

5-c, Allow some point missing during EIC tracing, you should configure Miss Allow Th (default value=3.)

5-d, Allow the tolerance between two neighboring mass trace, you should configure Inter Trace Distance Th.

You can refer to fig.18 to know the real physical meaning at first and then configure the parameters for mass trace extraction.

6. Parameter configuring for Peak detection

Bear in mind, MET-COFEA adopt a more advanced method to detect meaningful profile peak. The peak detection is implemented in 2D CWT domain spanned by retention time-scan by detect the peak branch pattern. One meaningful peak corresponds to a meaningful branch pattern. A meaningful branch pattern should be across several scales and the continuity is not too bad.

6-a, Min Branch Length: the minimum branch length in 2D CWT domain for a peak. Default value =6;

6-b, Branch Gap Allow: allow the branch have a limited gap across neighboring scales. Default value =2-3

6-c, Min Search Span: specific the branch searching range in the neighboring scales. Default value =2-3.

6-d, Marker Branch Dif Th: for sharing peak issues, the branch pattern is a little challenge; you can consider it a very wide peak, or two sharing peaks. So, one peak branch pattern in the high scale will split into two at some scale. In this case, there will two marker points. When the distance between the two marker points is larger than Marker Branch Dif Th, MET-COFEA will consider it as two sharing peaks.

You can refer to fig.20 to know the real meaning of this method and then configure the parameters for peak detection.

7. Parameter configuring for peak quality filtering

Bear in mind, MET-COFEA is aiming to output the high-quality chromatograph peak by some criteria.

7-a: Min Peak Width (default =6), Max Peak Width (default =50), Peak intensity Th (default =50), can be known from mass spectrometry instrument facility personnel, or, you can tentative configure this parameter as default values and run MET-COFEA, if found some peaks missing, try to configure a loosen parameters to find more peaks.

7-b: SNR Th (default =2.0), Peak Significance Th (default =1.0), TPASRTh (default =0.7), Zig Zag Index (default =0.6)

You can tentative configure this parameter as default values and run MET-COFEA, if found some peaks missing, try to configure a loosen parameters to find more peaks.

8. Parameter configuring for peak grouping, annotation, refinement and alignment

8-a: Group Scan Shift_Tol (default=3), Group Shape Angle Tol (default=20), Annotate Mass Tol (default=0.1), Refinement Shape Tol (default=10).

8-b: One/Two Phase alignment, Align Window1, Align Window2. Usually, use One Phase alignment, if you find some compounds are misalign, you can use two phase alignment strategy to refine the alignment results, keep in mind, the AlignWindo2 for Two Phase alignment should be narrower than AlignWindow1 for One Phase alignment. Regarding the configured value for Align Window1/ Align Window2, you should know these parameters from the mass spectrometry instrument facility personnel, because they are related to the real dynamic range for retention time shift.

Additionally, Start Scan, End Scan, and Cutoff Intensity are mainly factor to affect the computation burden.

Potential problems in application

MET-COFEA development team tried to find any potential bugs or problems in application, if you find any problems, please contact us by pzhao@noble.org or wezhang@noble.org. We are very appreciated for your feedback and suggestion.

During our test, the potential problems that have been found include:

1. MET-COFEA may fail in parallel processing mode if the specific file folder name and file name have any space. This problem is generated by the MPI command line parsing module.
Solutions: remove any space in the folder name and file name.
2. In some Windows 7 machine, MPI based parallel processing may fail,
Solutions: Check Windows OS and update into the latest version, ensure it support MPI, .NET.