ADAP-BIG v0.1 User Manual

Du-lab Team
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte
dulab.binf@gmail.com
http://www.du-lab.org

December 8, 2020

Contents

Chapter 1

Introduction

ADAP-BIG is a free cross-platform software for processing raw untargeted mass spectrometry data, designed to handle a large number of samples on machines with minimal system requirements. Users of ADAP-BIG can choose between two workflows for processing untargeted liquid chromatography (LC-) and gas chromatography coupled to mass spectrometry (GC-MS) data. The graphical user interface provides visualization of the raw data and intermediate results for each step of the data processing.

1.1 Download and Installation

ADAP-BIG is a cross-platform application, that can be used on Windows, Mac OS, and Linux. However, users can download a platform-specific installation package to easily install the application on their workstations. To download the installation package for your platform, please visit the GitHub Release page. Table 1.1 provides the list of files available for download.

Installation of ADAP-BIG on Windows.

1. Download file **Adap-Big App-x.x.x.msi** from the GitHub Release page.
2. Start the installation process by double-clicking on the downloaded file.

File	Description
Adap-Big App-x.x.x.msi	Installation package that installs ADAP-BIG application on Windows.
Adap-Big App-x.x.x.pkg	Installation package that installs ADAP-BIG application on Mac OS.
adap-big-jar-files-x.x.x.zip	Collection of individual jar files for each workflow step (Require **Java 8+**).

Table 1.1: List of the installation files available at the GitHub Release page.

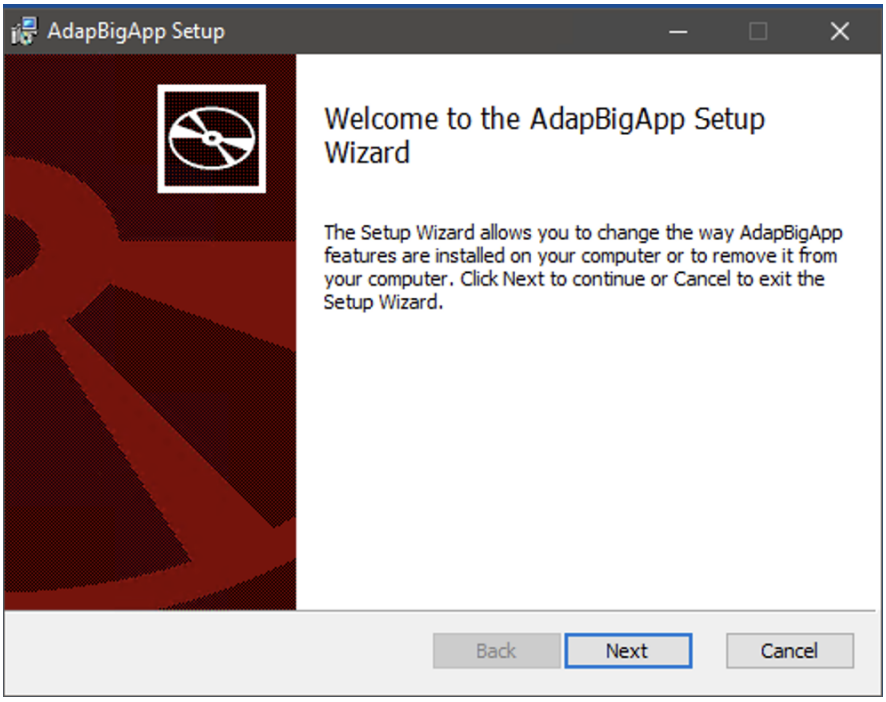

Figure 1.1: Installing ADAP-BIG on Windows

3. Run ADAP-BIG by clicking the Windows Start Button and selecting the ADAP-BIG application.

Installation of ADAP-BIG on Mac OS

1.2 User Interface

User interface of ADAP-BIG is shown on Figure 1.2. The main area displays the processing results of each workflow step, and its content depends on which step is currently selected. For instance, visualization of the *Input* step shows a table of MS scans including the scan number, name, MS level, Polarity, etc., detailed information about a selected scan, spectrum of that scan, total-ion chromatogram, base-ion chromatogram, and *Retention time vs. m/z* plot. Users can select one or several scans by clicking on the scan table to display and compare their raw spectra.

On the top of the ADAP-BIG window, there located the workflow bar, where users can select a workflow step to be displayed in the main area. After users click on any step in the workflow bar, a new tab with the visualization of that step will show up in the main area. However, if a tab for that workflow step has been opened previously, clicking on the corresponding step in the workflow bar will result in switching to the previously opened tab instead of opening a new one. Also, notice that if a workflow step is performed on multiple samples (such as Chromatogram Builder and Peak Detection), then the processing results are shown only for a single automatically-selected sample.

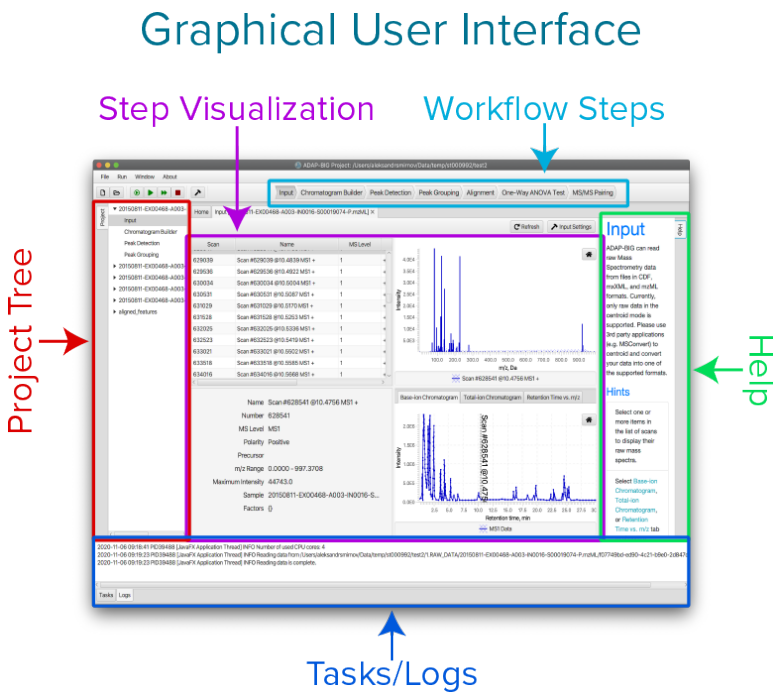

Figure 1.2: ADAP-BIG User Interface

The project tree on the left of the ADAP-BIG let users select specific samples and the associated processing steps to display in the main area. It is more functional than the workflow bar, since users can display processing results for each individual sample. However, it may be harder to navigate the entire project tree than selecting processing steps in the workflow bar. Thus, users can choose their preferred method of navigation between workflow steps, based on their needs and experience.

The bottom of the ADAP-BIG window displays the current progress and the logging information whenever workflow steps are being executed. By clicking on *Tasks* tab, users can see the progress for all workflow steps that are being currently executed or scheduled. If there is no steps being executed, then the tasks area will be empty. As workflow steps are getting executed, they log certain information that is displayed in the *Logs* tab. This information may be useful for keeping track of the user-defined parameters used for each workflow step and reporting any errors that may arise during execution of a workflow step.

Finally, users can see a help information on the right-hand side of the ADAP-BIG window. Here, users will find a short description of the selected workflow step and its visualization, and the suggested actions that users can perform for further evaluation of the processing results.

1.3 Creating a new project

Processing raw data starts with creating a new ADAP-BIG project. Users have multiple options to start a new project: clicking *Create a new project* button on the home page, selecting *New Project...* in the main menu, or clicking the corresponding button in the

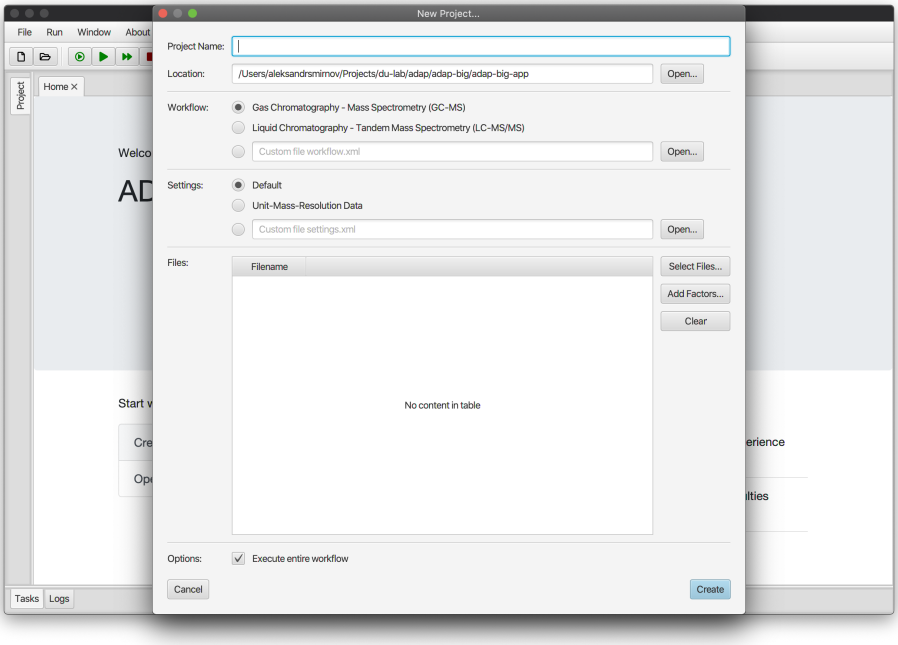

Figure 1.3: Creating a new ADAP-BIG project.

toolbar. After any of these actions, a new window will pop up.

First, users are asked to choose name and location of the new project. Next, users can choose between GC-MS or LC-MS workflow (see Section 1.4). Users can also choose a custom workflow by uploading an XML file that contains the workflow description. See Section 4.2 for the information on how to create a custom workflow. Similarly, users can choose parameters that will be used for the workflow steps. Currently, users can choose between the default parameters, parameters for unit-mass resolution data, and a custom set of parameters. Custom parameters must be listed in an XML format (see Section 4.3 for details). Notice that regardless which parameters are selected when an ADAP-BIG project is created, users can modify all workflow parameters and rerun each workflow step with new parameters later.

Next, users must select raw data files for the new project by clicking on *Select Files...* button. Currently, only files in CDF, mzML, and mzXML formats are supported. Also, the data should be in the centroid mode. After users select raw data files, they will be able to see their filenames in the New Project window. If users intend to perform the ANOVA significance testing, then they also need to upload a CSV file with sample names and the corresponding factors. Such a file can be added by clicking *Add Factors...* button. Figure 1.4 shows an example of a CSV file with sample factors. The first column must contain sample names matching the names of raw data files. Every other column is treated as a factor with the name of a factor in the first row and its values in all other rows. There can be arbitrary number of columns in a CSV file. The added factors will be displayed in the *New Project* window alongside the corresponding samples.

Finally, users can choose whether the workflow steps should be executed immediately after

```
local_sample_id ,Gender ,Metabolic syndrome ,Smoker
S00019041 ,Female ,Yes ,No
S00019063 ,Female ,Yes ,No
S00019078 ,Female ,Yes ,No
S00018999 ,Male ,No ,No
S00019001 ,Male ,No ,No
S00019002 ,Male ,No ,No
```

Figure 1.4: Example of a CSV file with sample factors.

GC-MS Workflow	LC-MS Workflow
1. Input	1. Input
2. Chromatogram Builder	2. Chromatogram Builder
3. Peak Detection	3. Peak Detection
4. Spectral Deconvolution	4. Ion Peak Grouper
5. Alignment	5. Alignment
6. Significance Test	6. Significance Test
	7. MS/MS Pairing

Table 1.2: Data processing workflows available in ADAP-BIG

creating the project. If the *Execute Entire Workflow* checkbox is selected, then all workflow steps will be executed immediately after creating the new project. Otherwise, ADAP-BIG will only import the raw data files, and users can manually perform each workflow step and see its results.

1.4 Data Processing Workflows

Table 1.2 shows the two available workflows for processing raw untargeted mass spectrometry data. The GC-MS workflow consists of the input step, extracted-ion chromatogram (EIC) builder, peak detection, spectral deconvolution, alignment, and significance test. The LC-MS workflow consists of the input step, EIC builder, peak detection, ion peak grouper, alignment, significance test, and MS/MS pairing. Notice that the two workflows share some of their steps (e.g. input, EIC builder, peak detection), while other steps are unique to either GC-MS or LC-MS workflow.

Both workflows start with the *Input* step, which imports raw data files from open data formats. Currently, there are two restrictions on the raw data files that can be used in ADAP-BIG:

- Only CDF, mzML, and mzXML are supported. If you use proprietary or other open data formats, please use third-party software to convert your files into one of the three supported formats.
- The raw data must be in the centroid mode. If your raw data is in the profile mode, please use third-party software to convert the data into the centroid mode.

For both conversion into one of the three supported data formats and writing data in the centroid mode, we recommend using MSConvert from the ProteoWizard package. MSConvert can read from and write to many open and proprietary data formats and has a peak picking algorithm to centroid the data.

The next two steps in the GC-MS and LC-MS workflows are *Chromatogram Builder* and *Peak Detection*. The Chromatogram Builder step constructs extracted-ion chromatograms (EICs) starting from the highest-intensity data point and collecting all data points withing a specified m/z tolerance. Then, Peak Detection is performed on every EIC by applying the Wavelet transform to detect the retention-time ranges corresponding to distinct peaks of an EIC. For more details about these two algorithms, see the paper [1]

Next, *Spectral Deconvolution* and *Ion Peak Grouper* form analytical components by finding the peaks from different EICs, that have similar retention times. Some but not all of these components may correspond to real compounds presented in a sample. The Spectral Deconvolution algorithm produces a pure fragmentation mass spectrum for each component by constructing the elution profile of every component and decomposing every EIC peak into a linear combination of those elution profiles. For more details about the Spectral Deconvolution algorithm, see [2]. The Ion Peak Grouper is a much simpler algorithm that groups EIC peaks of similar shapes and retention times, and forms the pseudo-spectrum out of those peaks.

The *Alignment* and *Significance Test* steps are the same for both GC-MS and LC-MS workflows. The alignment algorithm was first described in [3] and aligns components based on similarities of their retention times and (pure fragmentation or pseudo-) spectra across multiple samples. Then, the ANOVA significance test is performed on each component and for each factor specified during importing the raw data files.

Finally, the *MS/MS Pairing* step is unique to the LC-MS workflow. It finds the MS/MS spectra whose precursor mass is contained in the pseudo-spectrum of a component and assigns those MS/MS spectra to that component. If the raw data does not contain any MS/MS spectra, then this step can be ignored, and its output will just mirror the output of the previous step.

For more details about each workflow step, their parameters and visualization, see Chapter 2 (for GC-MS workflow) and Chapter 3 (for LC-MS workflow).

Users can modify the GC-MS and LC-MS workflows by adding, removing, and rearranging their steps. The custom workflows must be stored in a separate file, so they can be reused in new projects. For details, see Section 4.2.

1.5 Editing workflow parameters

Users can edit parameters for each workflow step either by clicking the *Settings* button in the main ADAP-BIG area or selecting menu *Run/Settings*. In both cases, the Settings window will pop up (see Figure 1.5). Users have several options to edit workflow parameters. First, they can select a workflow step in the list of all processing algorithms, adjust the parameters of that step, and click button *Save*. To process data with the updated workflow parameters, users would need to rerun either the current workflow step or the entire workflow.

Buttons at the top-right of the *Settings* window let users import workflow parameters

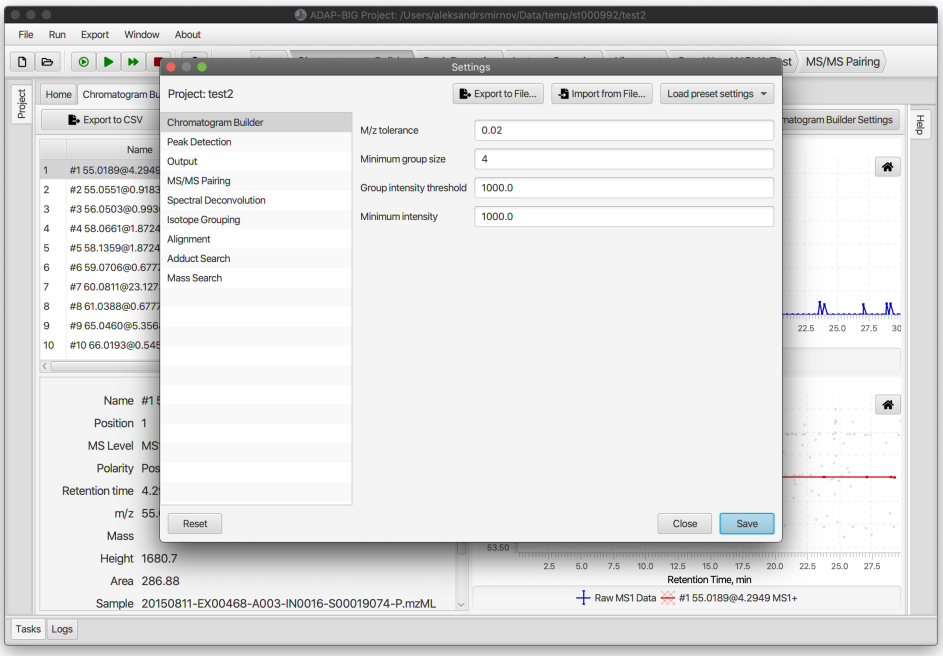

Figure 1.5: Editing parameters of workflow steps.

from an existing XML file, export the current workflow parameters to an XML file, and load either default and unit-mass preset parameters.

Finally, users can restore the parameters specified when the project was created by clicking the *Reset* button. Also, they can discard any parameter changes by clicking the *Close* button.

Chapter 2

GC-MS Workflow

2.1 Creating GC-MS project

For this example of executing the GC-MS workflow, we use raw untargeted data from study ST000897 ("Untargeted metabolomics analysis of ischemia-reperfusion injured hearts ex vivo from sedentary and exercise trained rats") available at the National Metabolomics Data Repository (NMDR). Its data consists of 31 samples split into four factor groups: Exercise Group, Exercise Ischemia/Reperfusion Injury, Sedentary Control, and Sedentary Ischemia/Reperfusion Injury.

Figure 2.1 shows the window for creating a new ADAP-BIG project with "adap-big-app" as the project name, "Gas Chromatography - Mass Spectrometry (GC-MS)" workflow, unit-mass-resolution settings, and sample files with added factor groups. After clicking *Create* button, the data processing starts immediately. The current progress of the data processing is displayed in the *Tasks* area (see Figure 2.2). Also, the current logging information from each workflow step is displayed in the *Logs* area (see Figure 2.3).

It should be noted that users must wait until each data processing step is completed before attempting to look at its results. For instance, if *Peak Detection* step is still in progress and a user clicks on the *Peak Detection* button in the workflow bar, then ADAP-BIG will not be able to display the corresponding results yet. Instead, user will see a message asking to rerun the processing step and wait its completion.

2.2 Input Step

Visualization of the *Input* step is shown on Figure 2.4. The table of scans in the top-left corner displays all scans in a data file. including the following information:

Scan scan number,
Name automatically generated scan name,
MS Level can be either 1 (for MS1 scans) or 2 (for MS2 scans),
Polarity can be either + (for positive mode) or − (for negative mode),
Ret time scan retention time (in minutes),
Precursor m/z scan precursor m/z (if defined).

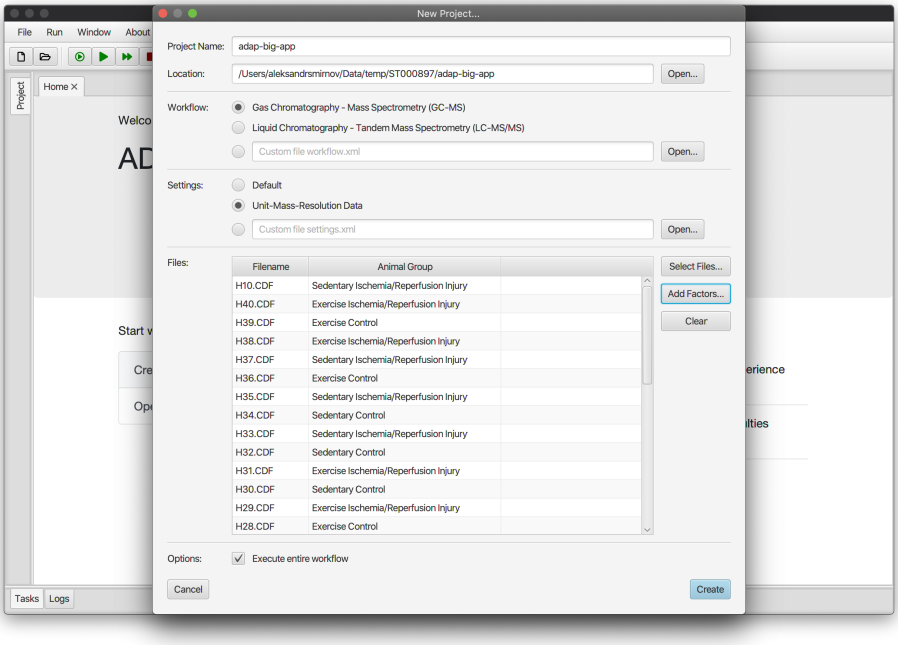

Figure 2.1: Example of GC-MS project.

Users can select one or more scans in this table to see details about each scan (bottom-left corner) and its raw spectrum plot (top-right corner). In addition to the scan number, name, MS-level, polarity, and precursor m/z, the scan details include: range of the m/z value in a scan, its maximum intensity, sample name, and sample factors.

In the bottom-right corner, there located three tabs containing the base-ion chromatogram plot, the total-ion chromatogram plot, and the plot "Retention time vs. m/z". In the base-ion chromatogram plot, intensity at each retention time is equal to the maximum intensity of the scan at that retention time. In the total-ion chromatogram plot, intensity at each retention time is equal to the total intensity of the scan at that retention time. Finally, in the "Retention time vs. m/z" plot, each dot represents pair (ret time, m/z), and its color ranges from white to blue, based on its intensity. When a scan is selected in the scan table, these three plots display a vertical line with the scan name, marking the position of the selected scan. On the other hand, double-clicking anywhere in the boundaries of those plots will result in selecting a scan with the corresponding retention time.

The *Input* step doesn't have any user-defined parameters. So, clicking on the *Input Settings* button will results in opening the Settings window with no workflow step selected.

2.3 Chromatogram Builder Step

This step builds extracted-ion chromatograms (EICs) for the masses that are present in the raw data continuously over a certain duration of time. Users can adjust parameters of the Chromatogram Builder either by clicking button *Chromatogram Builder Settings* or by selecting menu *Run/Settings...*.

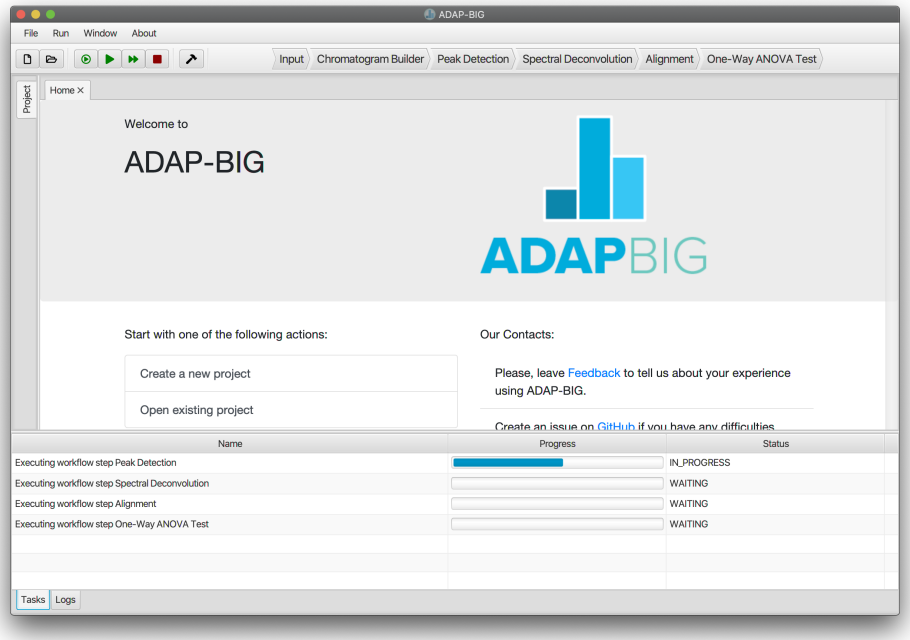

Figure 2.2: Data processing progress

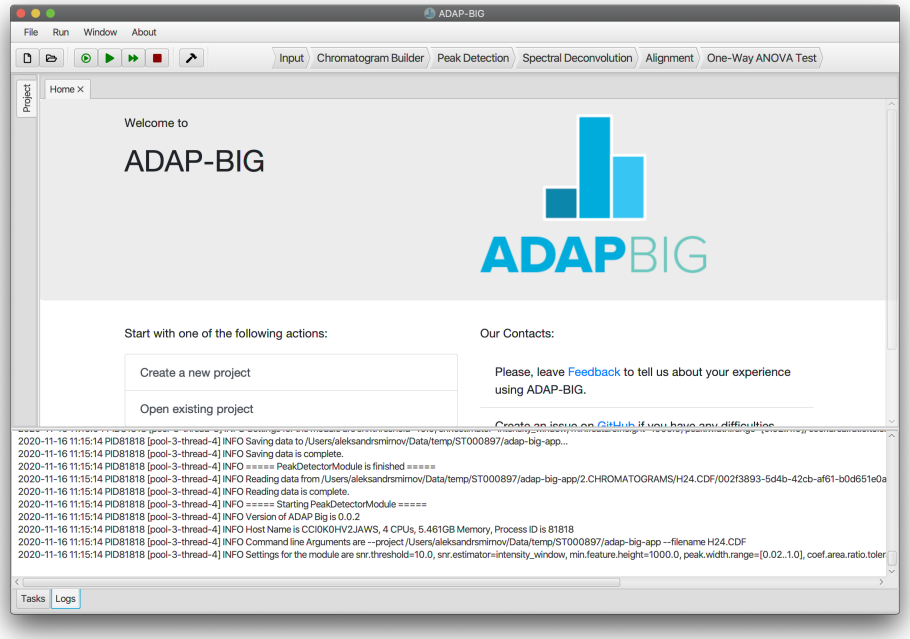

Figure 2.3: Data processing logging information

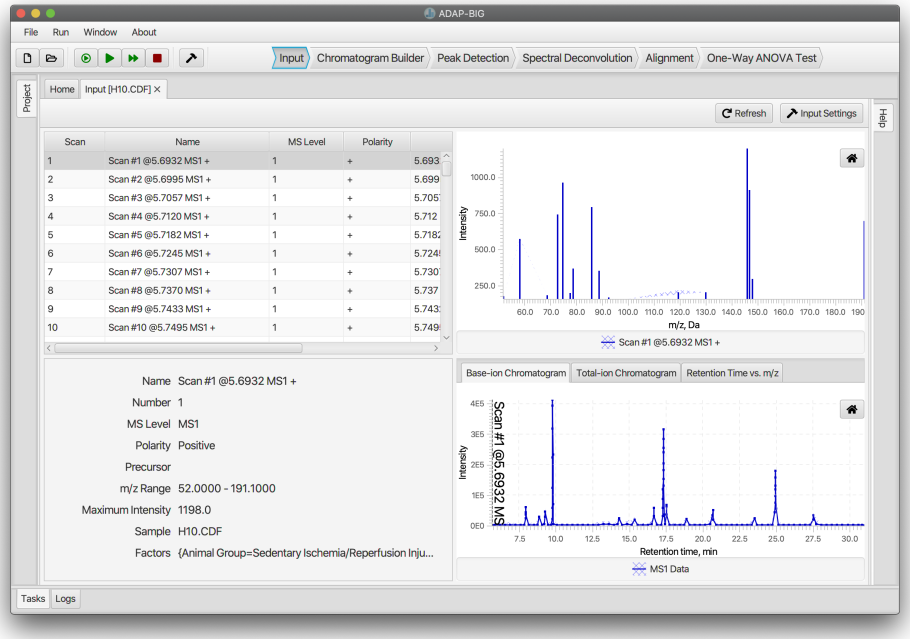

Figure 2.4: Input Step visualization.

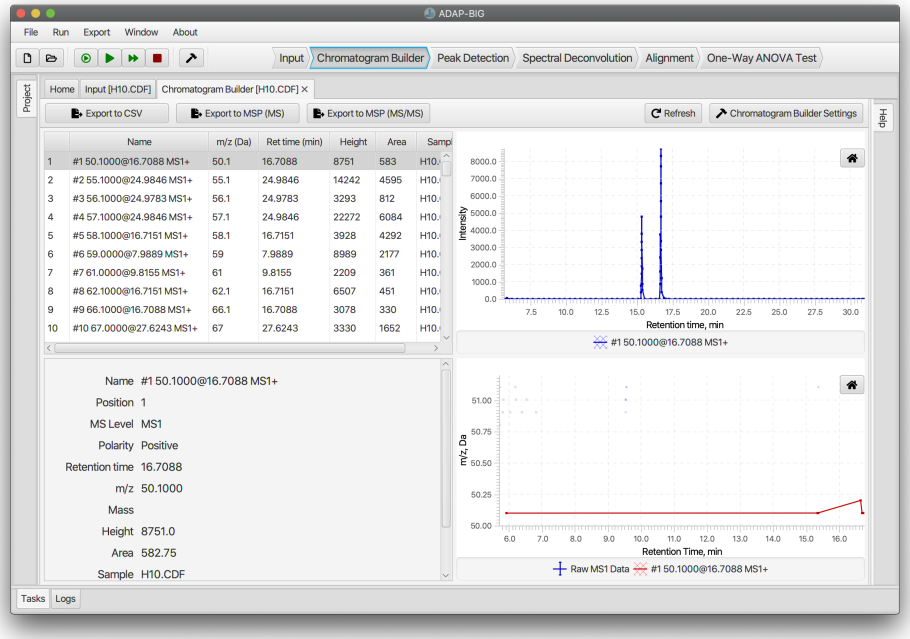

Figure 2.5: Chromatography Builder Step visualization.

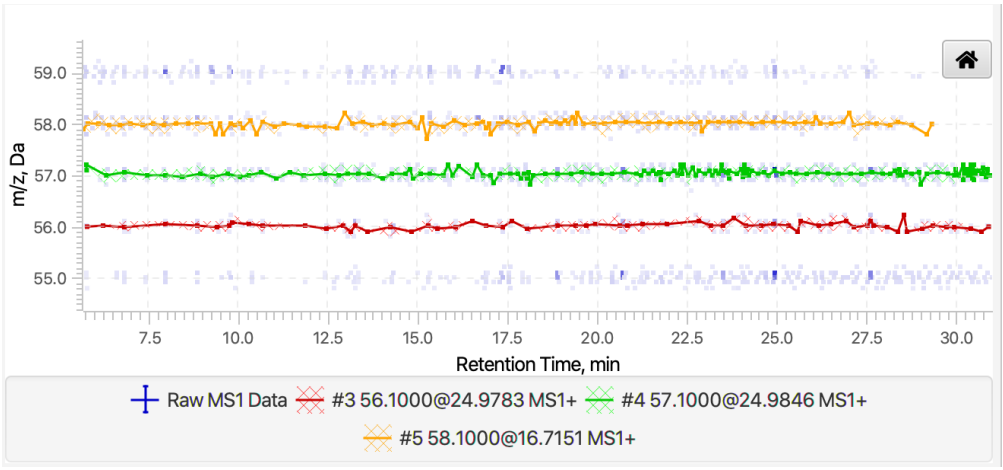

Figure 2.6: Three chromatograms in the *Retention time vs. m/z* plane.

M/z tolerance Minimum m/z difference of data points in consecutive scans in order to be connected to the same chromatogram. Twice the *m/z tolerance* set by the user is the maximum width of a mass trace.

Minimum group size In the entire chromatogram, there must be at least this number of sequential scans having points above the *Group intensity threshold* set by the user. The optimal value depends on the chromatography system setup. The best way to set this parameter is by studying the raw data and determining what is the typical time span of chromatographic peaks.

Group intensity threhsold See above.

Min hieghest intensity There must be at least one point in the chromatogram that has an intensity greater than or equal to this value.

The visualization of the *Chromatogram Builder* step consists of four parts. The table in the top-left corner shows all built chromatograms with their IDs, names, m/z values, retention times (i.e. the retention time of the highest point in a chromatogram), heights, areas, and sample names. A more detailed information is displayed in the bottom-left corner for the currently selected chromatogram.

The figure at the top-right corner, shows the currently selected chromatogram(s), and the figure at the bottom-right corner, shows the shape of the selected chromatogram(s) in the *Retention time vs. m/z* plane. The latter figure also displays the near-by raw data points whose color depends on their intensities. This figure can be useful to access the quality of chromatogram. For instance, users can select sort the chromatogram table by the m/z values and select two or more close chromatogram to see how those chromatograms are built (see Figure 2.6). If there is a significant number of data points in between chromatograms, then some chromatograms were missed and users may want to adjust the parameters to build more chromatograms. If there are several chromatograms built over the same cluster of data points, then users may want to adjust the parameters to build fewer chromatograms.

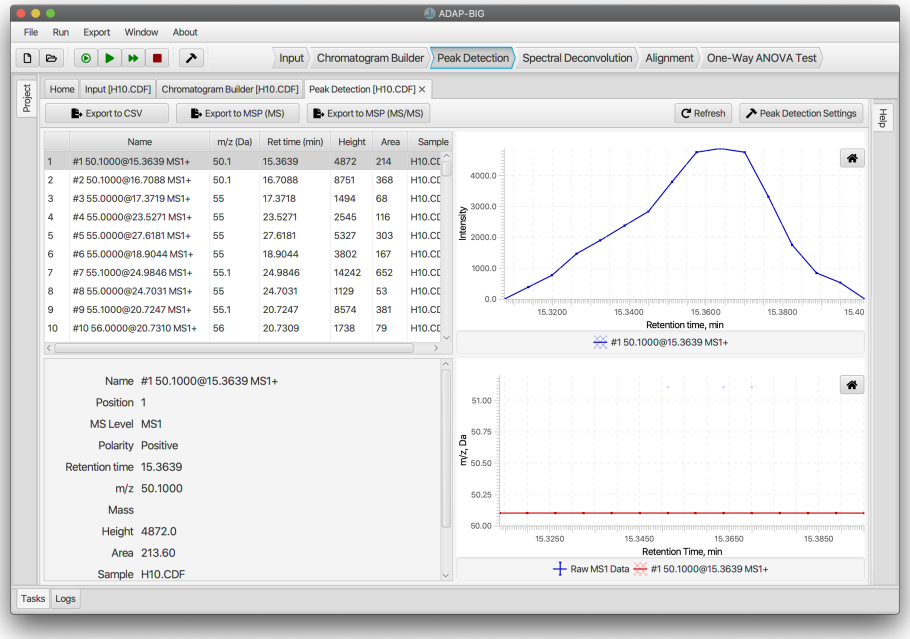

Figure 2.7: Peak Detection Step visualization.

2.4 Peak Detection Step

Each EIC that has been constructed spans the entire duration of the chromatography. *Peak Deteciton* step detects individual peaks in those EICs. Users can adjust parameters of the Peak Detection either by clicking button *Peak Detection Settings* or by selecting menu *Run/Settings....*

S/N threshold the minimum signal-to-noise ratio a peak must have to be considered real. Values grater then or equal to 7 will work well and will only detect a very small number of false positive peaks.

S/N estimator users can choose of two estimators of the signal-to-noise ratio: `intensity_window` uses the peak height as the signal level and the standard deviation of intensities around the peak as the noise level; `wavelet_coefficient` uses the continuous wavelet transform coefficients to estimate the signal and noise level. Analogous approach is implemented in R-package *wmtsa*.

Minimum height the smallest intensity a peak can have and be considered real.

Peak duration range minimum and maximum widths (in minutes) of a peak to be considered real.

Coefficient/Area threshold this coefficient is found by taking the inner product of the wavelet at the best scale and the peak, and then dividing by the area under the peak. Values around 100 work well for most data.

RT wavelet range minimum and maximum widths (in minutes) of the wavelets used for detecting peaks. The *Peak Detection* algorithm is highly sensitive to the upper limit of the *RT wavelet range*. Also, the `wavelet_coefficient` S/N estimator is sensitive

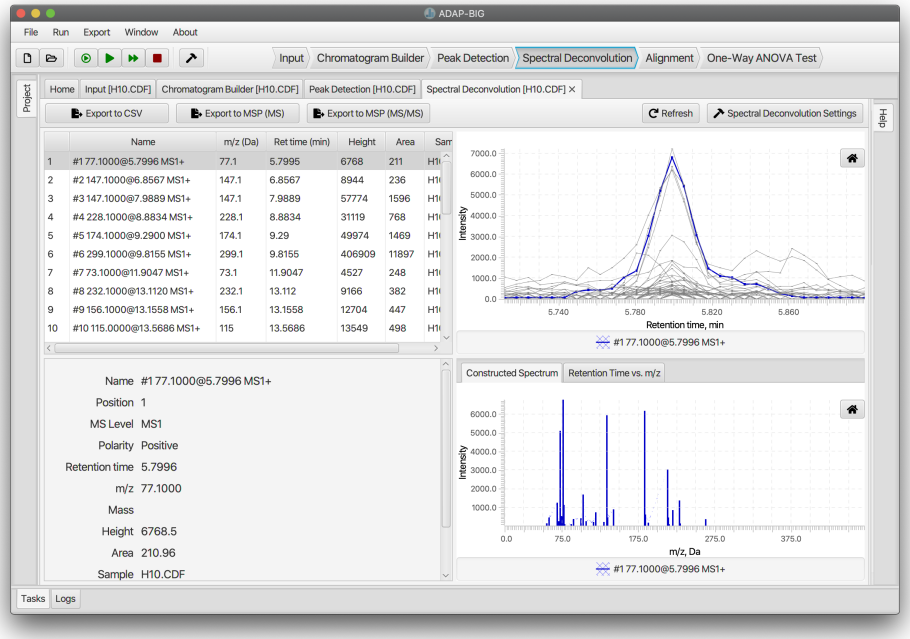

Figure 2.8: Spectral Deconvolution Step visualization.

to the lower limit of the *RT wavelet range*.

The visualization of the *Peak Detection* step is similar to the visualization of the *Chromatogram builder* step. See the previous section for details.

2.5 Spectral Deconvolution Step

Spectral Deconvolution finds the analytes that are present in a sample and constructs their pure fragmentation mass spectra. Location of analytes is performed in two steps. First, all peaks are assigned to deconvolution windows based on their retention times. Then, in each window, clusters of similar peaks are determined using the similarities of the peak shapes. Then, a model peak is constructed for each cluster, and all peaks are decomposed into a linear combination of the constructed model peaks. Users can adjust parameters of the Spectral Deconvolution either by clicking button *Spectral Deconvolution Settings* or by selecting menu *Run/Settings...*.

Maximum window width is the maximum length (in minutes) of clusters after the first clustering step. This window width can be chosen based on the width of detected peaks. Typically, value 0.2 works well in most cases.

Retention time tolerance is the smallest time-gap between any two analytes. The value of this parameter should be a fraction of the average peak width. In our tests, we use 0.04 minutes.

Minimum cluster size the smallest number of peaks in a single analyte. This parameter depends on a dataset and the number of peaks detected by the previous workflow steps.

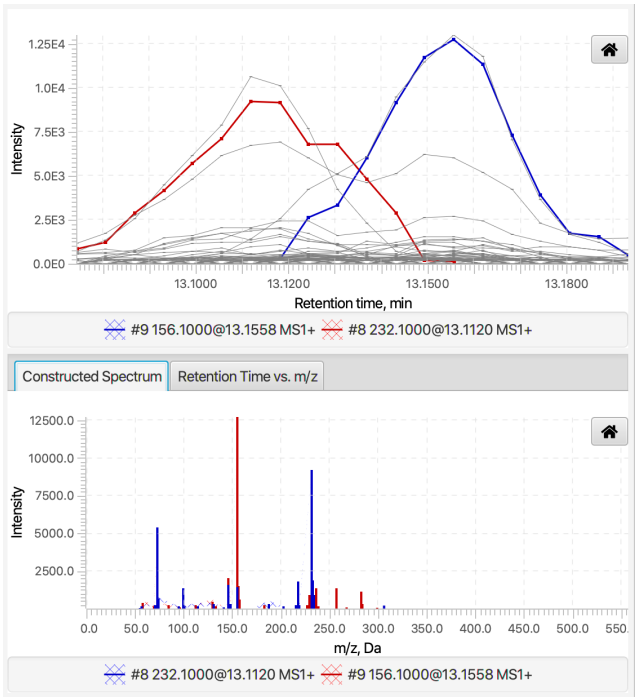

Figure 2.9: Decomposition of peaks of two coeluting components.

Typically, its value would range from 1 (if only a few peaks are detected) to 10 or more (if the number of detected peaks is large).

Adjust apex retention times For a unit-mass-resolution data, where coeluting analytes may be present, and a peak typically consists of a hundred or more points, this parameter should be off. For high-mass-resolution data, where coeluting compounds are rare and a peak consists of a few points, this parameter can be turned on.

The visualization of the *Spectral Deconvolution* step consists of four parts. In the upper-left and bottom-left corners are located the table of components and a panel with detailed information about the selected component, respectively. The figures on the right plot the elution profiles and pure fragmentation mass spectra of one or more components (colored curves) and the constituent peaks (grey). This figures help evaluate how peaks are decomposed into linear combinations of the model peaks. To do the later, users can sort the table of components by the retention time and select two or more near-by components. The figures will show the elution profiles of coeluting components and their constructed pure fragmentation spectra (Figure 2.9).

2.6 Alignment Step

The *Alignment* step uses uses similarity between constructed mass spectra to find similar components across multiple samples. For this reason, the alignment is performed **after** the spectral deconvolution step. Users can adjust parameters of the Alignment either by clicking button *Alignment Settings* or by selecting menu *Run/Settings...*.

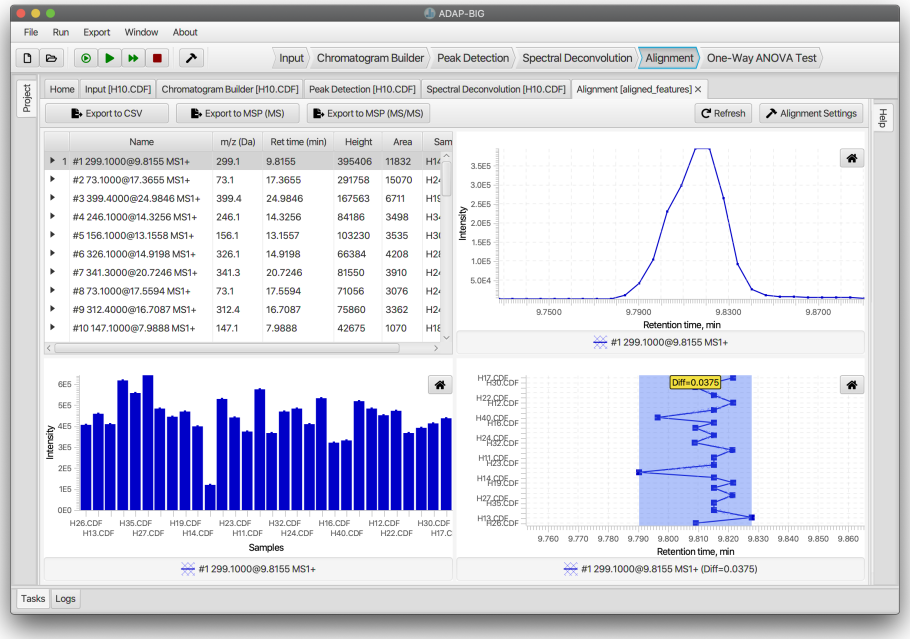

Figure 2.10: Alignment Step visualization

Minimum sample frequency takes values between 0 and 1 and equals to the minimum fraction of samples containing aligned components. I.e. if similar components are observed in the fraction of samples less than the *minimum sample frequency*, then those components are not aligned. If the sample factors are provided than *minimum sample frequency* is the minimum fraction of samples on one factor group.

Retention time tolerance is the maximum time-gap between similar components from different samples.

Matching score threshold takes values between 0 and 1. Similarity score between components in different samples is determined as follows:

$$Score = w \cdot S_{time} + (1 - w) \cdot S_{spec},$$

where S_{time} is the relative retention time difference between two components and S_{spec} the spectral similarity between two components. The score threshold defines the minimum similarity score between two components to be aligned.

Matching score weight takes values between 0 and 1. This parameter is the coefficient w in the similarity score. If $w = 0$, then only the spectral similarity is used for calculating the similarity of two components. If $w = 1$, then only the retention time difference is used for calculating the similarity of two components. If w is between 0 and 1, then a weighted combination of the spectral similarity and the retention time difference is used.

m/z tolerance is used for matching masses in two mass spectra. This parameter may affect the spectral similarity score.

The top-left corner of the visualization of the *Alignment* step contains a table of aligned

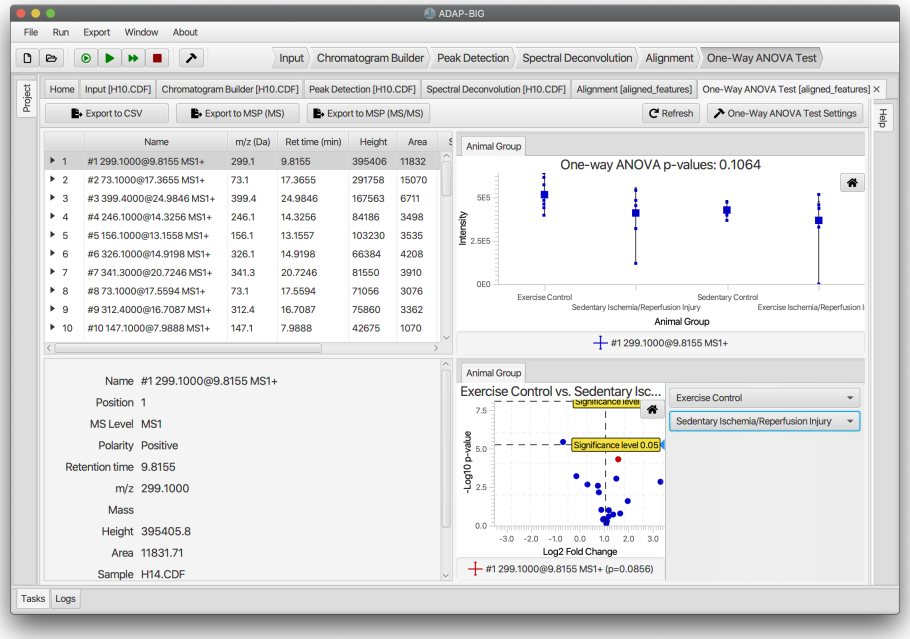

Figure 2.11: One-way ANOVA Test visualization.

components. In addition to regular table features, users can expand each component to see aligned components from different samples. The figure in the top-right corner displays the elution profiles of aligned components. Users can select all aligned components from different samples to look at their elution profiles together. The figure in the bottom-right corner displays the retention time of the aligned components in the *Retention time vs. Sample* plane. This figure also displays the maximum difference between retention times of the aligned components. Finally, the bottom-right figure displays intensities of aligned component from each sample.

2.7 One-way ANOVA Test

The *One-way ANOVA* test doesn't have any user-defined parameters. So, clicking on the *One-way ANOVA Settings* button will result in opening the Settings window with no workflow step selected.

In the top-left there located the table of aligned components. It is similar to the table of aligned components in the *Alignment* step visualization, but has additional columns. For each provided sample factor, this table has an addition column named after that factor, and its values are the p-values of the ANOVA tests computer for each aligned component. If there are more than one factor provided, the table will contain a column for each factor. If there are no factors provided, this table will be identical to the table in the *Alignment* step.

The top-right corner contains tabs with the figures displaying intensity values of the selected aligned components for each factor group. The intensity of an aligned component from a single sample is marked as a small square. The average intensity of each group is

Workflow Step	CSV	MSP		Content
		MS	MS/MS	
Input	–	–	–	None
Chromatogram Builder	✓	–	–	Extracted-ion chromatograms (one file per sample)
Peak Detection	✓	–	–	EIC peaks (one file per sample)
Spectra Deconvolution	✓	✓	–	Components and their spectra (one file per sample)
Alignment	✓	✓	–	Aligned components and their spectra (single file)
One-way ANOVA Test	✓	✓	–	Aligned components with p-values and their spectra (single file)

Table 2.1: Content exported by buttons *Export to CSV*, *Export to MSP (MS)*, and *Export to MSP (MS/MS)* for each workflow step.

marked as a large square, and the span of intensities within each group is displayed as a vertical line segment. If there are more than one sample factor provide, users should click on the tab with a desired factor name to see the corresponding figure.

The bottom-right corner contains tabs with the figures displaying the volcano plots for each factor. The selected components are displayed with colored circles and all other components are displayed with blue circles. To help with reading the volcano plot, there are also displayed two horizontal lines for p-values 0.05 and 0.01, and one vertical line marking the zero logarithmic fold change. Since the volcano plot can be plotted only for two factor groups at a time, user can select those factor groups on the right-hand side of the plot. Finally, if there are more than one sample factor provided, users should click on the tab with a desired factor to see the corresponding figure.

2.8 Export feature tables and mass spectra

Users have three options to export the processing results. They can export a feature table with retention times, m/z values, intensities, and areas for each feature. Next, they can export MS1 spectra into an MSP file. Finally, they can export MS/MS spectra into an MSP file. This can be done either by clicking buttons *Export to CSV*, *Export to MSP (MS)*, *Export to MSP (MS/MS)*, or by selecting the corresponding items in the *Export* menu.

Notice that the content of the exported files depends on the workflow step currently selected in ADAP-BIG. I.e., if the current workflow step is *Chromatogram Builder*, then ADAP-BIG will export a table of extracted-ion chromatograms into a separate CSV file for each sample. Also, it will export nothing into MSP files because the chromatograms contains no spectra. Similarly, if the current workflow step is *Peak Detection*, then ADAP-BIG will only export a table of detected EIC peaks. If the current workflow step is *Spectra Deconvolution*, then ADAP-BIG will export a table of components into a CSV file and their constructed spectra a MSP file. Table 2.1 described the exported content for all workflow steps. Notice that *Export to MSP (MS/MS)* is not used in the GC-MS workflow.

Chapter 3

LC-MS Workflow

As an example of processing LC-MS raw untargeted data, we use study ST001122 ("Identification of urine metabolites in patients with interstitial cystitis using untargeted metabolomics (part II)") from the National Metabolomics Data Repository. The data consists of 43 samples with no factor information.

Figure 3.1 shows the window for creating a new ADAP-BIG project with "adap-big-app" as the project name, "Liquid Chromatography - Mass Spectrometry (LC-MS)" workflow, default settings, and added sample files. The current progress of the data processing is displayed in the *Tasks* area (see Figure 3.2). Also, the current logging information from each workflow step is displayed in the *Logs* area (see Figure 3.3).

It should be noted that users must wait until each data processing step is completed before attempting to look at its results. For instance, if *Peak Detection* step is still in progress and a user clicks on the *Peak Detection* button in the workflow bar, then ADAP-BIG will not be able to display the corresponding results yet. Instead, user will see a message asking to rerun the processing step and wait its completion.

3.1 Input Step

Visualization of the *Input* step is identical to the input step of the GC-MS workflow. See Section 2.2 for details.

3.2 Chromatogram Builder Step

Visualization of the *Chromatogram Builder* step is identical to the chromatogram builder step of the GC-MS workflow. See Section 2.3 for details.

3.3 Peak Detection Step

Visualization of the *Peak Detection* step is identical to the peak detection step of the GC-MS workflow. See Section 2.4 for details.

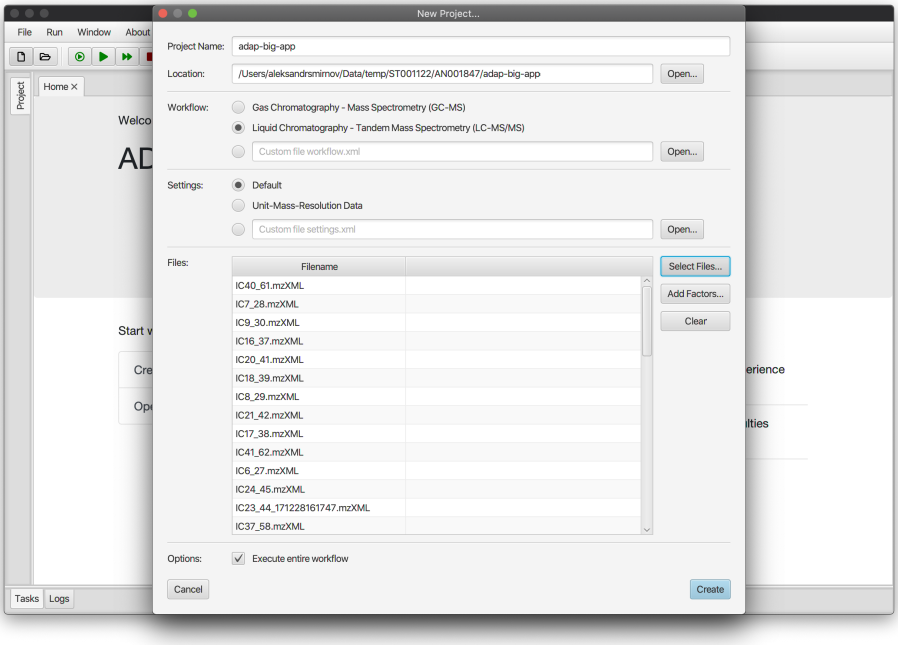

Figure 3.1: Example of LC-MS project.

3.4 Ion Peak Grouper Step

The *Ion Peak Grouper* step finds the detected peaks with close retention times and similar shapes. Then, it assigns all similar peaks to a separate component. Finally, the peaks withing each component are used to construct the pseudo-spectrum for that component. Users can adjust parameters of the Ion Peak Grouper either by clicking button *Ion Peak Grouper Settings* or by selecting menu *Run/Settings...*.

Peak similarity threshold ranges from 0 to 1. Peak similarity is calculated as the similarity of peak shapes with values close to 1 indicating high similarity and values close to 0 indicating low similarity of the peak shapes.

Retention time tolerance is the largest difference (in minutes) of two peak to be considered similar.

Minimum number of peaks can be 1 or more. The minimum number of similar peak that can form a component.

The visualization of the *Ion Peak Grouper* step consists of four parts (Figure 3.4). In the upper-left and bottom-left corners are located the table of components and a panel with detailed information about the selected component, respectively. The figures in the right-hand side plot the elution profiles and pseudo-spectra of one or more components (colored curves) and the constituent peaks (grey).

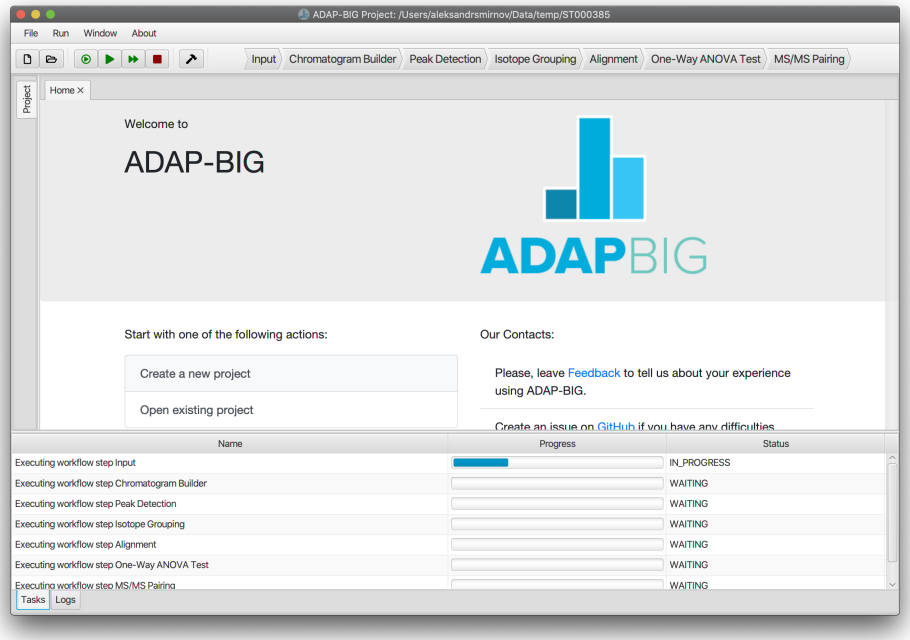

Figure 3.2: Data processing progress

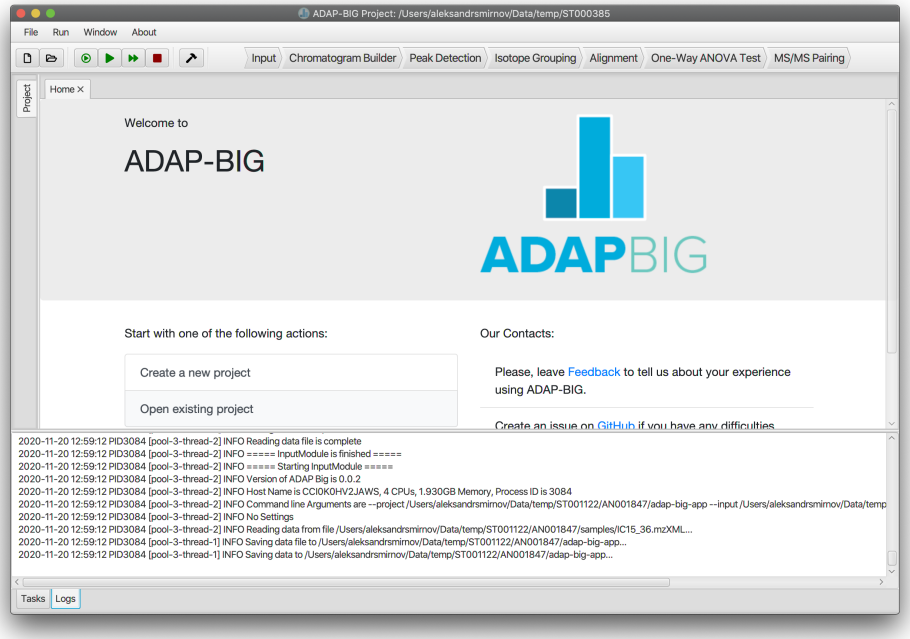

Figure 3.3: Data processing logging information

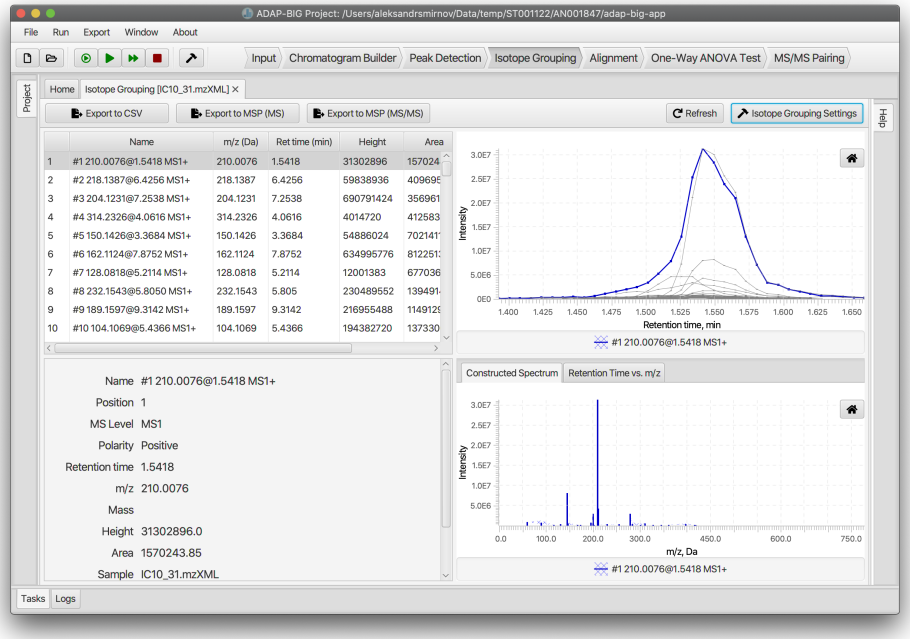

Figure 3.4: Ion Peak Grouper Step visualization.

3.5 Alignment Step

Visualization of the *Alignment* step is identical to the alignment step of the GC-MS workflow. See Section 2.6 for details.

3.6 One-Way ANOVA Test

Visualization of the *One-Way ANOVA* test is identical to that of the GC-MS workflow. See Section 2.7 for details.

3.7 MS/MS Pairing Step

The *MS/MS Pairing* step finds the MS/MS spectra and pairs them with the aligned components. For MS/MS spectra to be paired with a component, they should satisfy two conditions: (*i*) the precursor m/z value of an MS/MS spectrum should be contained in the pseudo-spectrum of the component; (*ii*) the retention time of an MS/MS spectrum should be within the half-height retention time range (i.e. the range corresponding to the upper half of the elution profile) of the component. Users can adjust parameters of the MS/MS Pairing step either by clicking button *MS/MS Pairing Settings* or by selecting menu *Run/Settings...*.

m/z tolerance is used when matching the precursor m/z value of an MS/MS spectrum to m/z values of the pseudo-spectrum of a component.

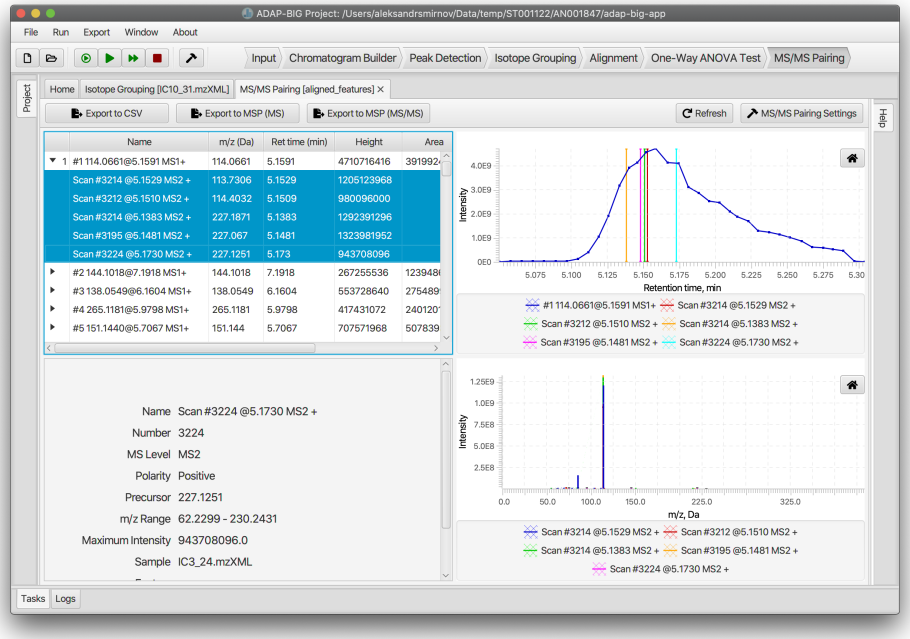

Figure 3.5: MS/MS Pairing Step visualization.

Intensity factor threshold is used to filter out low-intensity MS/MS spectra. First all MS/MS spectra that satisfy the precursor and retention time requirements are assigned to a component. Then the standard deviation of their intensities is calculated. Finally, if the intensity of an MS/MS spectrum does not exceeds that standard deviated multiplied by the *Intensity factor threshold*, then that MS/MS spectrum is removed from the component. Users can use value 0 to keep all MS/MS spectra.

The visualization of the *MS/MS Pairing* step consists of four parts (Figure 3.5). In the top-left corner, there located a table of components. Users can expand each component to see the paired MS/MS spectra. The bottom-left corner contains a panel with detailed information about the selected component or MS/MS spectrum. The fields displayed on that panel change in accordance to whether the selected element is a component or an MS/MS spectrum.

The top-right corner contains a figure that displays an elution profiles of a component (blue curve) and its paired MS/MS spectra (vertical lines). Users must select all MS/MS spectra in the component table to display them on this figure. Finally, the figure in the bottom-right corner displays the pseudo-spectrum of the selected component.

3.8 Export feature tables and mass spectra

Users have three options to export the processing results. They can export a feature table with retention times, m/z values, intensities, and areas for each feature. Next, they can export constructed pseudo-spectra into an MSP file. Finally, they can export MS/MS spectra

Workflow Step	CSV	MSP		Content
		MS	MS/MS	
Input	–	–	–	None
Chromatogram Builder	✓	–	–	Extracted-ion chromatograms (one file per sample)
Peak Detection	✓	–	–	EIC peaks (one file per sample)
Ion Peak Grouper	✓	✓	–	Components and their pseudo-spectra (one file per sample)
Alignment	✓	✓	–	Aligned components and their spectra (single file)
One-way ANOVA Test	✓	✓	–	Aligned components with p-values and their spectra (single file)
MS/MS ANOVA Test	✓	✓	✓	Aligned components with p-values, their spectra, and paired MS/MS spectra (single file)

Table 3.1: Content exported by buttons *Export to CSV*, *Export to MSP (MS)*, and *Export to MSP (MS/MS)* for each workflow step.

into an MSP file. This can be done either by clicking buttons *Export to CSV*, *Export to MSP (MS)*, *Export to MSP (MS/MS)*, or by selecting the corresponding items in the *Export* menu.

Notice that the content of the exported files depends on the workflow step currently selected in ADAP-BIG. I.e., if the current workflow step is *Chromatogram Builder*, then ADAP-BIG will export a table of extracted-ion chromatograms into a separate CSV file for each sample. Also, it will export nothing into MSP files because the chromatograms contains no spectra. Similarly, if the current workflow step is *Peak Detection*, then ADAP-BIG will only export a table of detected EIC peaks. If the current workflow step is *Ion Peak Grouper*, then ADAP-BIG will export a table of components into a CSV file and their pseudo-spectra a MSP file. Table 3.1 described the exported content for all workflow steps. Notice that *Export to MSP (MS/MS)* is not used in the GC-MS workflow.

Chapter 4

Advanced Functionality

4.1 Running individual steps in the command line

ADAP-BIG workflow steps can be executed not only through the ADAP-BIG application, but also by running individual workflow steps from the command line. In order to do it, users should have Java 9+ installed on their local machine and download the ADAP-BIG jar files available at the GitHub Release page.

Currently, ADAP-BIG include 11 jar files representing each ADAP-BIG workflow step. Below, these jar files are listed together with their command-line arguments:

- `input.jar` imports raw data files into an ADAP-BIG project.
  `--project` path to the project folder;
  `--input` path to a raw data file;
  `--factors` (optional) path to a CSV file with factor values.
- `chromatogram-builder.jar` finds raw data points with similar m/z and constructs extracted-ion chromatograms.
  `--project` path to the project folder;
  `--filename` sample name (name of the corresponding raw data file).
- `peak-detection.jar` detects peaks in every extracted-ion chromatogram.
  `--project` path to the project folder;
  `--filename` sample name (name of the corresponding raw data file).
- `spectral-deconvolution.jar` forms components and constructs their pure fragmentation mass spectra.
  `--project` path to the project folder;
  `--filename` sample name (name of the corresponding raw data file).
- `simple-spectral-deconvolution.jar` groups similar peaks into components and constructs their pseudo-spectra.
  `--project` path to the project folder;
  `--filename` sample name (name of the corresponding raw data file).
- `alignment.jar` finds similar components across samples and aligns them
  `--project` path to the project folder.

- `significance.jar` performs the one-way ANOVA test for each aligned component with the factor groups specified during the raw data import.
  `--project` path to the project folder;
  `--type` types of features this module it applied to, can be one of `chromatogram`, `peak`, `component`, `aligned_component`;
  `--factor` (optional) name of the factor used to calculate the ANOVA test. If no factor is provided, then ANOVA is performed for all available factors.
- `adduct-search.jar`
  detects adducts by matching peaks of the pseudo-spectrum to a list of known adducts and calculates components' masses.
  `--project` path to the project folder;
  `--type` types of features this module it applied to, can be one of `chromatogram`, `peak`, `component`, `aligned_component`.
- `mass-search.jar`
  matches the computed masses to the RefMet database.
  `--project` path to the project folder;
  `--type` types of features this module it applied to, can be one of `chromatogram`, `peak`, `component`, `aligned_component`.
- `ms2-pairing.jar`
  pairs MS/MS spectra to the constructed components.
  `--project` path to the project folder;
  `--type` types of features this module it applied to, can be one of `chromatogram`, `peak`, `component`, `aligned_component`.
- `output.jar`
  exports the processing results into CSV or MSP files.
  `--project` path to the project folder;
  `--filename` (optional) sample name (name of the corresponding raw data file). If not specified, then data for all samples will be exported;
  `--type` types of features this module it applied to, can be one of `chromatogram`, `peak`, `component`, `aligned_component`;
  `--output` name of the file to export data to, this file should have either .csv or .msp extension;
  `--outputlevel` (optional) can be either ms1 (for exporting MS1 spectra) or ms2 (for exporting MS/MS spectra).

It should be noted that every jar file has a mandatory argument `--project`, which specifies path to the project folder. That folder should be already exist and contain file `settings.xml` containing parameters of the workflow steps. Users can create such file manually or export it from the ADAP-BIG application.

Then, each jar file can be executed as following:

```
java -jar input.jar --project PATH_TO_PROJECT --input RAW_DATA_FILE
```

where `java` is Java 9+ command, `input.jar` path to an ADAP-BIG jar file, `PATH_TO_PROJECT` path to the project folder, and `RAW_DATA_FILE` path to a raw data file to be imported.

```
<?xml version="1.0" encoding="utf-8" ?>
<workflow>
    <org.dulab.adapbig.input.InputModule>
        <input/>
        <factors/>
    </org.dulab.adapbig.input.InputModule>

    <org.dulab.adapbig.chromatogrambuilder.ChromatogramBuilderModule>
        <filename/>
    </org.dulab.adapbig.chromatogrambuilder.ChromatogramBuilderModule>

    <org.dulab.adapbig.peakdetection.PeakDetectorModule>
        <filename/>
    </org.dulab.adapbig.peakdetection.PeakDetectorModule>

    <org.dulab.adapbig.spectraldeconvolution.SpectralDeconvolutionModule>
        <filename/>
    </org.dulab.adapbig.spectraldeconvolution.SpectralDeconvolutionModule>

    <org.dulab.adapbig.alignment.AlignmentModule/>

    <org.dulab.adapbig.significance.SignificanceModule>
        <type value="aligned_component"/>
    </org.dulab.adapbig.significance.SignificanceModule>
</workflow>
```

Figure 4.1: Example of file `workflow.xml`.

4.2 Editing the `workflow.xml` file

Figure 4.1 shows an example of file `workflow.xml` used by the ADAP-BIG application to store the GC-MS and LC-MS workflows of processing raw data. However, users can build custom processing workflows by creating their own `workflow.xml` files.

The workflow description file should follow certain rules. First, it must contain the root XML element called `<workflow>`. Then, children of the `<workflow>` element must have names matching the full class names of the Java classes of ADAP-BIG workflow steps. These classes typically similar to the names of the jar files of the corresponding workflow steps. See the class names in Figure 4.1 and the jar files described in Section 4.1 for examples.

In the workflow description file, users can specify the command-line arguments listed in Section 4.1 for each workflow step. However, there are some important differences between the command-line arguments and their counterparts in the workflow description file:

- command-line argument `--project` is not specified in the file `workflow.txt` because it is added automatically by the ADAP-BIG application prior to every execution of a data processing workflow.

- arguments `--input`, `--factors`, and `--filename` correspond to self-closing XML elements `<input/>`, `<factors/>`, `<filename/>` with no content. Their values are assigned by the ADAP-BIG application to the names of raw data files and factors selected in the *New Project* window of the ADAP-BIG application.

- all other XML elements are converted into a command-line arguments. ADAP-BIG application converts XML elements of the workflow description file into command-line arguments as follows: element `<tag value="TAG_VALUE">` is converted into command-line argument `--tag TAG_VALUE`.

Below, we list several examples of command-line execution of workflow steps and their counterparts in file `workflow.txt`.

Example 1. Command-line execution of the import of a raw data file

```
java -jar significance.jar --project PATH_TO_PROJECT --input PATH_TO_FILE
```

corresponds to the following lines in the workflow description file:

```
<org.dulab.adapbig.input.InputModule>
    <input/>
</org.dulab.adapbig.input.InputModule>
```

Example 2. Command-line execution of the alignment step

```
java -jar alignment.jar --project PATH_TO_PROJECT
```

corresponds to the following line in the workflow description file:

```
<org.dulab.adapbig.alignment.AlignmentModule/>
```

Example 3. Command-line execution of the one-way ANOVA test on all aligned components

```
java -jar significance.jar --project PATH_TO_PROJECT --type aligned_component
```

corresponds to the following lines in the workflow description file:

```
<org.dulab.adapbig.significance.SignificanceModule>
    <type value="aligned_component"/>
</org.dulab.adapbig.significance.SignificanceModule>
```

4.3 Editing the `settings.xml` file

Figure 4.2 shows an example of file `settings.xml` used by the ADAP-BIG application and by individual jar files to store user-defined parameters of every workflow step. Currently, users can choose between default and unit-mass preset parameters. However, they can also create a custom settings file with their own parameters.

The settings file should follow certain rules. First, it must contain the root XML element called `<settings>`. Then, children of the `settings>` element must have names matching the full names of the Java classes of ADAP-BIG workflow steps. This rule should be followed in the file `workflow.xml` as well. Next, the children of each workflow step in settings file is converted into parameters of that workflow step as follows: XML element

```
<?xml version="1.0" encoding="utf-8" ?>
<settings>
    <org.dulab.adapbig.input.InputModule/>

    <org.dulab.adapbig.chromatogrambuilder.ChromatogramBuilderModule>
        <mz.tolerance type="double" value="0.02"/>
        <minimum.scan.span type="int" value="4"/>
        <intensity.thresh.2 type="double" value="1000.0"/>
        <min.intensity.for.start.chrom type="double" value="1000.0"/>
    </org.dulab.adapbig.chromatogrambuilder.ChromatogramBuilderModule>

    <org.dulab.adapbig.peakdetection.PeakDetectorModule>
        <snr.threshold type="double" value="10.0"/>
        <snr.estimator type="text" value="intensity_window"/>
        <min.feature.height type="double" value="1000.0"/>
        <peak.width.range type="double.range" start="0.02" end="1.0"/>
        <coef.area.ratio.tolerance type="double" value="0.0"/>
        <cwt.ret.time.range type="double.range" start="0.001" end="0.06"/>
    </org.dulab.adapbig.peakdetection.PeakDetectorModule>

    <org.dulab.adapbig.output.OutputModule>
        <force.integer.mz type="boolean" value="false"/>
    </org.dulab.adapbig.output.OutputModule>

    <org.dulab.adapbig.ms2pairing.Ms2PairingModule>
        <mz.tolerance type="double" value="0.01"/>
        <intensity.factor.threshold type="double" value="3.0"/>
    </org.dulab.adapbig.ms2pairing.Ms2PairingModule>

    <org.dulab.adapbig.spectraldeconvolution.SpectralDeconvolutionModule>
        <pref.window.width type="double" value="0.2"/>
        <ret.time.tolerance type="double" value="0.02"/>
        <min.cluster.size type="int" value="5"/>
        <adjust.apex.ret.times type="boolean" value="false"/>
    </org.dulab.adapbig.spectraldeconvolution.SpectralDeconvolutionModule>

    <org.dulab.adapbig.alignment.AlignmentModule>
        <sample.count.ratio type="double" value="0.5"/>
        <ret.time.range type="double" value="0.05"/>
        <score.tolerance type="double" value="0.75"/>
        <score.weight type="double" value="0.1"/>
        <mz.tolerance type="double" value="0.005"/>
    </org.dulab.adapbig.alignment.AlignmentModule>

    <org.dulab.adapbig.significance.SignificanceModule/>
</settings>
```

Figure 4.2: Example of file `settings.xml`.

```
<name type="TYPE" value="VALUE">
```

is converted into a parameter with name, type, and value corresponding to attributes `name`, `type`, and `value`. The tag name `name` should match one of the hard-coded parameters of the workflow step. The attribute `type` should be one of `int`, `double`, `text`, `boolean`. Finally, the attribute `value` will be converted into either integer, real number, boolean in accordance with the specified type.

Currently, there are two additional XML elements for representing a user-defined parameter, that do not follow the above described rule. First, it is possible to specify a range-of-values parameter by using the pattern

```
<name type="double.range" start="START" end="END">
```

Second, users can provide a list of adducts for the adduct-search step by using the following pattern:

```
<adducts type="adduct.list">
    <adduct name="M+2H+Na" num.molecules="1" adduct.mass="25.0038" charge="3"
            quasi.molecular.ion="false"/>
    <adduct name="M+2N+H" num.molecules="1" adduct.mass="46.9857" charge="3"
            quasi.molecular.ion="false"/>
    ...
</adducts>
```

Here, attribute `name` can take any value but is preferably a user-friendly name of an adduct. Attributes `num.molecules`, `adduct.mass`, and `charge` correspond to variables n, m, and q defined in the formula

$$mz = \frac{n \cdot M - m}{q},$$

where M is the mass of the original molecule, and mz m/z value of the adduct.

Finally, every workflow step must have a corresponding segment in the settings file. However, that segment may not have any parameters (see `InputModule` and `SignificanceModule` in Figure 4.2). In the latter case, the default parameters of that workflow step will be used.

Bibliography

[1] Owen D. Myers, Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: New algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. *Analytical Chemistry*, 89(17):8696–8703, 09 2017.

[2] Aleksandr Smirnov, Yunping Qiu, Wei Jia, Douglas I. Walker, Dean P. Jones, and Xiuxia Du. Adap-gc 4.0: Application of clustering-assisted multivariate curve resolution to spectral deconvolution of gas chromatography–mass spectrometry metabolomics data. *Analytical Chemistry*, 91(14):9069–9077, 07 2019.

[3] Wenxin Jiang, Yunping Qiu, Yan Ni, Mingming Su, Wei Jia, and Xiuxia Du. An automated data analysis pipeline for gc-tof-ms metabonomics studies. *Journal of Proteome Research*, 9(11):5974–5981, 11 2010.