

# USER MANUAL for CCPM Version 4.4

## Table of Contents

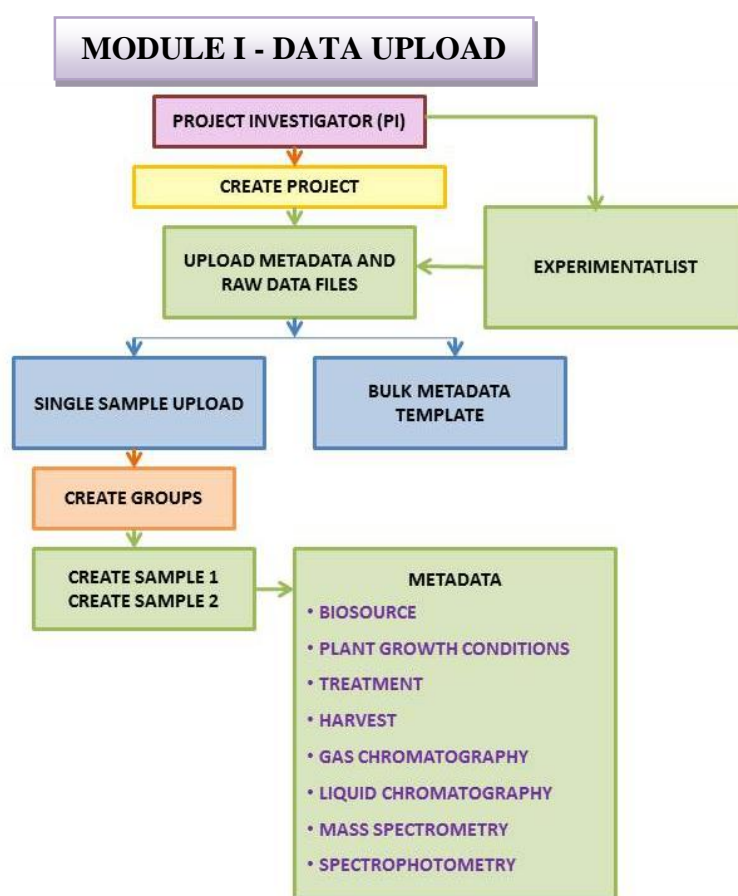
Getting started .....	3
LC/GC-MS Overview .....	3
NMR Overview .....	5
Registration, Login and Forgot password .....	7
Description of the roles .....	7
How to Create a New Project? .....	7
How to join a project .....	7
Projects .....	8
All Unpublished Projects .....	8
My Tasks .....	8
Module I - Data upload .....	9
Hierarchy of the project .....	9
Creation of Groups and samples with in each group .....	9
To upload raw data files (individual or Bulk raw file upload) .....	10
Module II - Data pre-processing, comparative analysis and visualization .....	11
LC/GC-MS .....	12
NMR .....	17
Module III – Filtration, Pre-treatment and Statistical analysis .....	18
Filtration .....	18
Pre-treatment .....	19
Statistical analysis .....	21
Module IV – KEGG Connectivity .....	33
Tools .....	36
1. Text Mzml Help page content .....	36
Requirements: .....	36
Usage: .....	36
2. NMR (Metabolite identification Module) .....	36
News .....	39
Events .....	39
People .....	39

Documentation.....	39
Forum.....	39

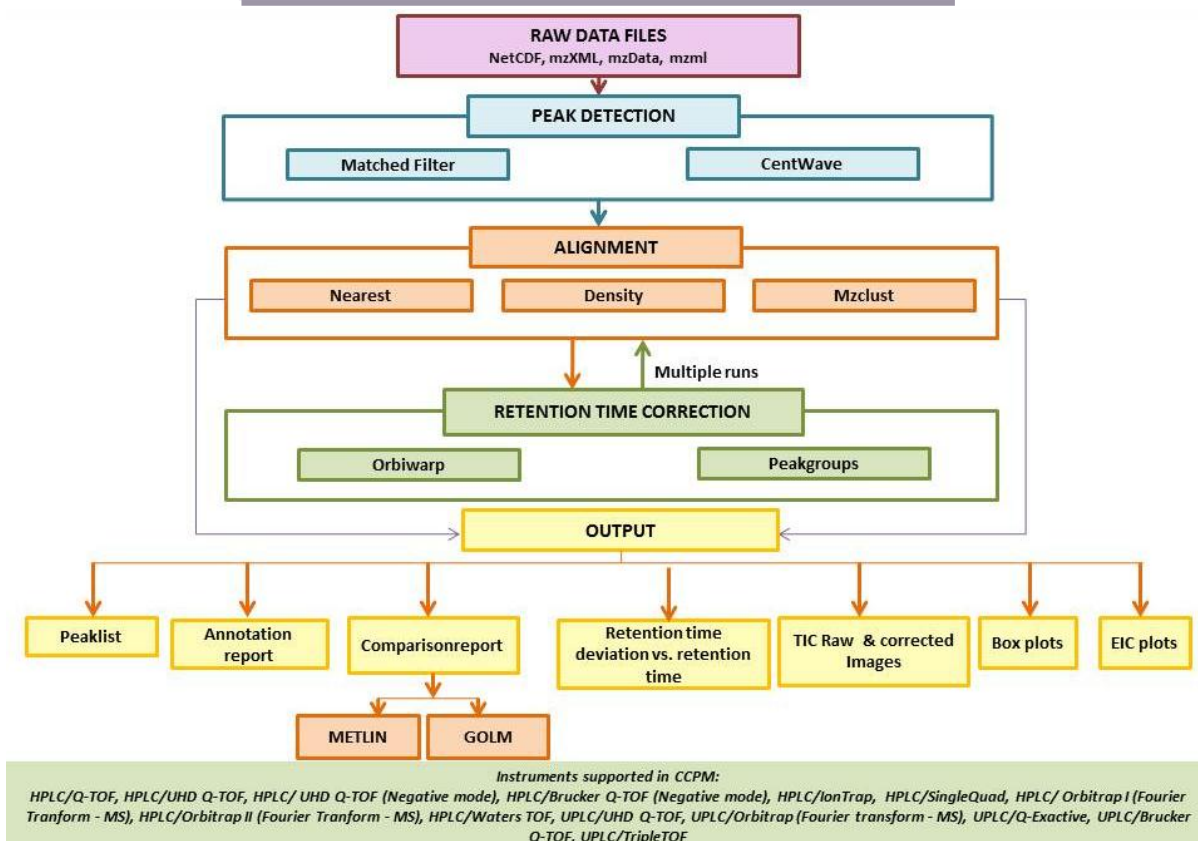
## Getting started

### LC/GC-MS Overview

The CCPM portal can be used to upload and analyse metabolomics data. The raw data from GC-MS and LC-MS can be uploaded and analysed in the pipeline. The pipeline consists of three modules: Module I- Data upload, Module II-Data pre-processing, and Module III -Pre-treatment and Statistical analysis. Links to METLIN and GOLM databases are provided for metabolite identification. The user also has an option to upload the pre-processed data from other software and carry out Pre-treatment and statistical analysis in our portal.

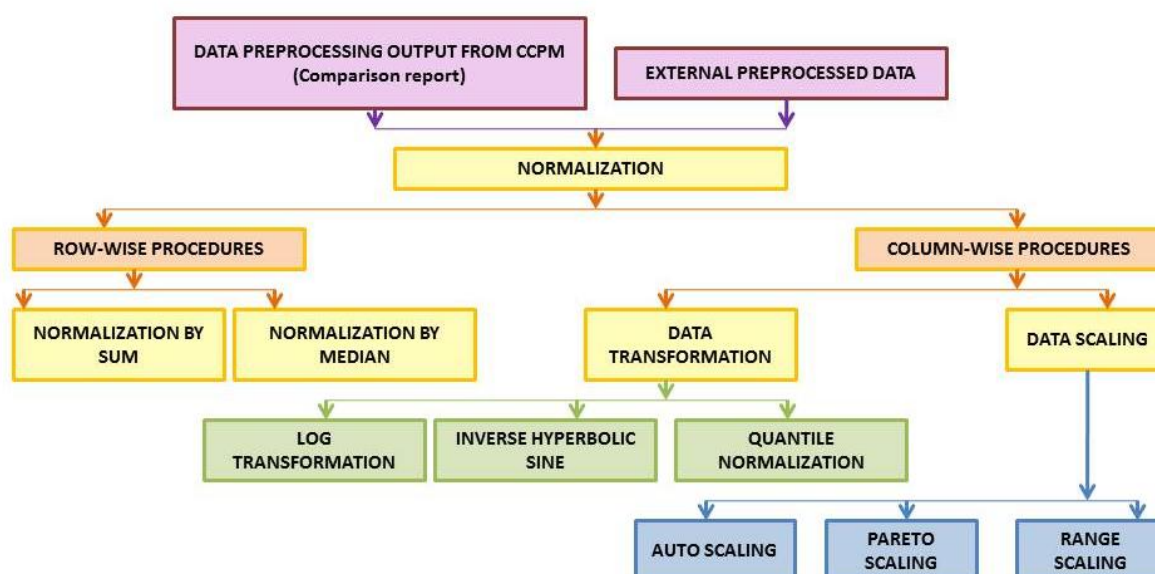


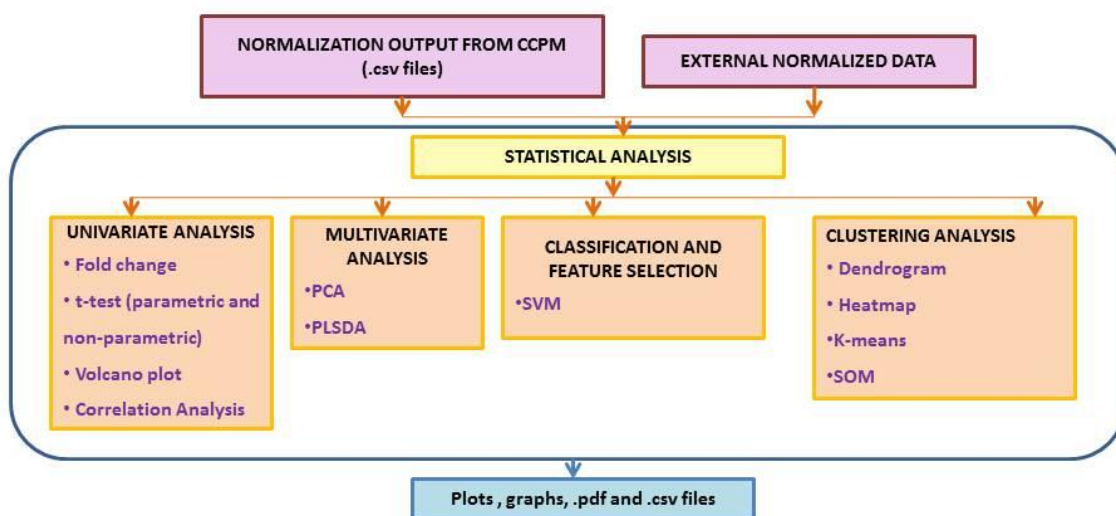
## MODULE II - DATA PRE-PROCESSING



## MODULE III – PRE-TREATMENT AND STATISTICAL ANALYSIS

### PRE-TREATMENT

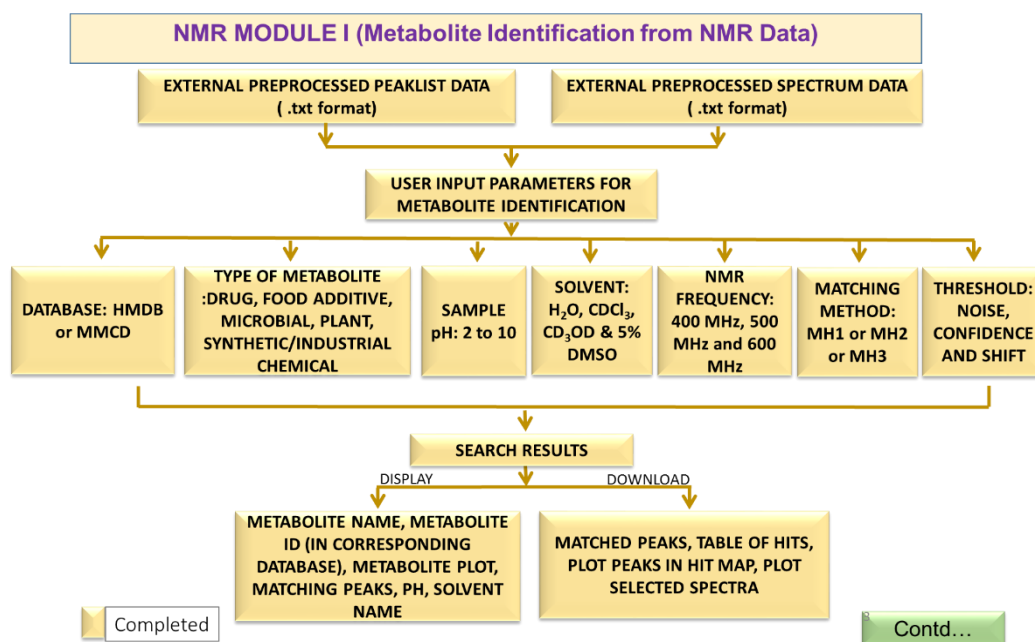


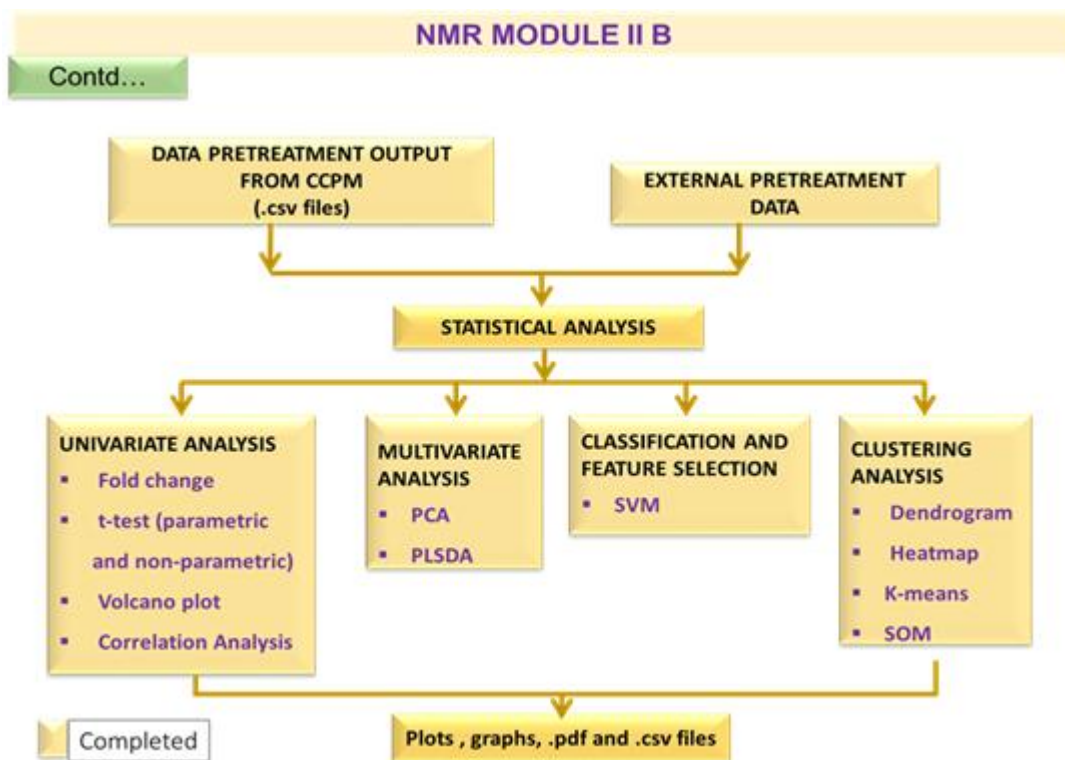
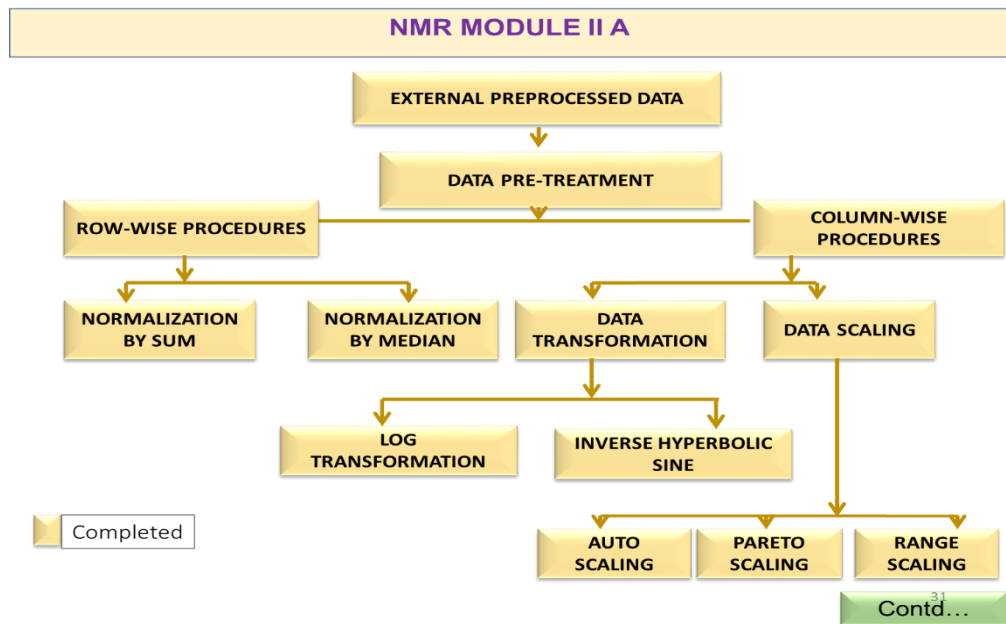


## NMR Overview

CCPM NMR platform offers the metabolite identification and comprehensive statistical analysis of various plant metabolomics projects. The NMR project workflow is divided into two modules:

1. **NMR Module I** - Metabolite identification from NMR data and
2. **NMR Module II**- **A**: Data Pre-treatment and **B**: Statistical Analysis.





## Registration, Login and Forgot password

In order to create projects, the user should first register in the portal to obtain a username and password. Registration is free. Once the user registers, he will get a verification mail in the registered mail ID. Upon clicking the link in the mail, the user will be directed to the CCPM portal. In case when the user could not remember the password later on, click on the **lost password** button. Enter the e-mail ID provided during registration process. The password will be sent via e-mail.

## Description of the roles

The user can choose any of the roles given below:

1. **Principal Investigator (PI)** have access to add a new project by clicking on “+New Project” button. They can also grant roles to the members, who want to join the project.
2. **Experimentalist**: To be an experimentalist, user should join the project first. Once their role is accepted by the PI/Co-ordinator, they can create groups and add their corresponding samples in the selected project.
3. **Co-investigator**: The Co-investigator should first join the project. He can perform analyses in the project that he has enrolled in and can view the results. However, he cannot upload/modify any data in the project.

## How to Create a New Project?

**Users registered as PI can ONLY add a new project.** The newly added project can be viewed by all registered members of the project.

Step 1: Login using user ID and password.

Step 2: Click on “+ New project” button.

Step 3: Fill in the Project form.

Step 4: Click on "Submit" button.

## How to join a project

The PIs have to first create a Project and grant roles to the users who want to sign up for the project. The users have to register in the portal and communicate the username/e-mail registered in CCPM portal to the PI offline. PIs can add the users and grant roles to them by clicking on the “Add user” button under the “People” tab of the project. PI can also edit the roles of the members at any point of time by clicking on the “edit” button that appears along the corresponding name.

Once the role is granted by the PI of the project, user can access the information and perform tasks as per the role. All project users can join either as an Experimentalist/Co-Investigator. Only Experimentalists can upload data by clicking "**Metadata**" button.

## Projects

Users having access (described in Description of roles section) can create/add projects using **+New Project**. All the projects created will be listed under **My Projects** and **All Projects**. The projects under them are sub-categorised as **Published** and **Unpublished projects** based on their status.

All open access projects are listed under **All Projects**. All projects to which the user has rights are listed under **My Projects**.

1. **Summary**: Project details can be viewed by clicking the project of interest. The project details can be edited by clicking on the "**edit project**" button. The project can be published by clicking on the "**Publish**" button.
2. **People**: Displays the details about the people working in the project. **Add User**: Visible only to the PI of the project. PI can grant/deny/block roles of signed up users.
3. **Tasks/Results**: Pre-processing, Pre-treatment and Statistical Analysis results of the project are displayed under this tab.
4. Click on **Metadata** button to get Groups and samples level information of the project.

## All Unpublished Projects

All ongoing projects are listed here. Only project details are available for public view. The PI has the option of delisting (not for public view) by clicking on **Edit project** and unchecking the box against '**List project for public view**'.

**Edit project**: Visible only to the PI of the project. PI can edit the details as needed.

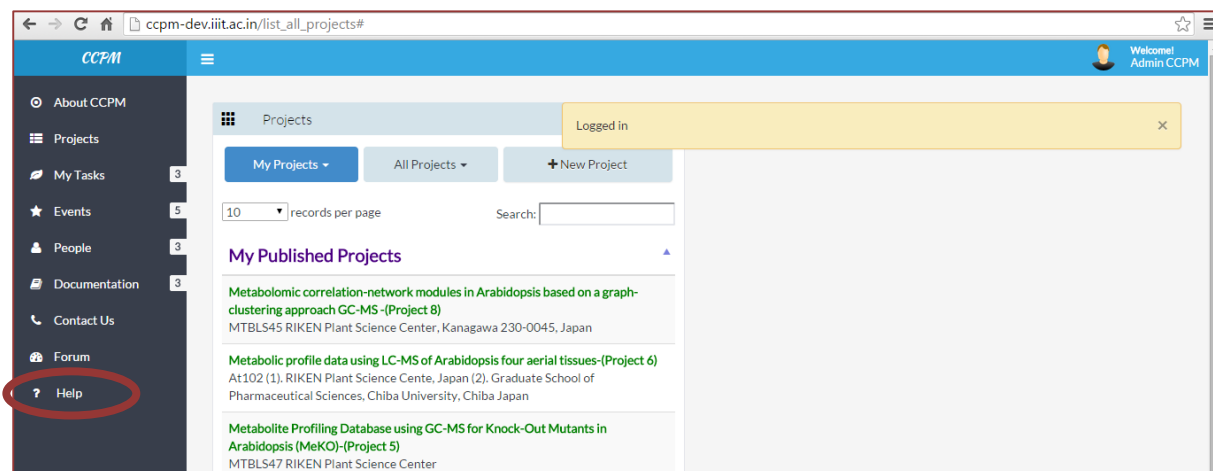
**Publish**: Visible only to the PI of the project. The PI can publish the project when all the experimentalists are done with uploading the data and performing the tasks. After publishing, the project will now appear under "**My and All Published projects**".

## My Tasks

This menu on the left side panel has 3 sub-menu options: **Pre-processing**, **Pre-treatment** and **Statistical analysis**. By clicking on each of this option, the user can view all the results for the tasks performed by him in all the projects that he has joined. The user can also perform the Pre-treatment and statistical analysis tasks for the available results. However, adding a new task (pre-processing/Pre-treatment) and external data analysis (i.e. external data Pre-treatment and external statistical analysis) cannot be performed through this menu option. To perform that, the user has to click on the project that he has enrolled in and proceed through the "**Tasks/Results**" tab option.



Page-wise help is available for all the pages in the portal.



## Module I - Data upload

### Hierarchy of the project

For a given project, there can be many groups. And for a given group, there can be many samples. In other words, sample is a subset of a group and group is a sub set of a project

Project > Group > Sample

### Definitions of Terminologies

1. **Groups:** Group is a sub set of project. A group can contain one or more samples with some commonality. For e.g., all samples derived from the same tissue of the plant or all samples treated with the same compound or controls etc. Groups can be created by the PI or the experimentalist.
2. **Samples:** A portion of material that is taken for testing. Sample is a subset of group. Only that PI or experimentalist who created the respective group can add samples to it.
3. **Replicate:** Replicates can be biological or technical. Biological replicates are parallel measurements of biologically distinct samples that capture random biological variation. Multiple samples taken within each combination of time, location, and any other controlled variables. The purpose of collecting replicate samples is to obtain precision (i.e., the spatial and temporal variations, plus variance introduced by sampling and analytical procedures). Technical replicates are usually controls to make sure that there is no error in the experiment. The metadata for the replicates are usually the same.
4. **Aliquot:** A small representative portion of the sample. It could be a smaller portion from the original sample or made by proportional dilution of the sample with inactive ingredients. The metadata for the aliquot also usually remains the same.

### Creation of Groups and samples with in each group

## 1. Manual or Single Meta-Data upload

2. Choose a projects and click on the “Metadata” button to upload metadata for single sample.
3. Click on **Unpublished Project** of **My Projects**
4. Click on **Metadata** button which will lead to the page where groups and samples can be added.
5. Click on “**Create new group**” button, fill the form and click “**submit**” to create a new group. Samples can be added by clicking on the “**add new sample**” button on the right panel of that group and submitting the details. The metadata for the sample can be entered by clicking “**Edit**” under each individual field (BIO, PGC etc.) below. Replicate and aliquot can be added by clicking on the “**Replicate**” or “**aliquot**” button on the right panel.

## 2. Automated / Bulk Meta-Data upload (via comma separated values (.CSV file))

Click on the “**Bulk upload**” button. Meta data for multiple groups and its samples in a project can be uploaded as a single .csv file by the user. The corresponding file can be uploaded by clicking the button **Choose File** followed by clicking **Upload Bulk Data**.

1. **Bulk Upload template:** By clicking on this button, the User can download an empty template for Bulk groups/samples data upload. The downloaded empty template bulk metadata form should be filled and submitted by clicking on **Bulk Upload** button and **Choose File** followed by clicking **Upload Bulk Data**. The groups and samples with their filled up columns should be populated under the project.

## To upload raw data files (individual or Bulk raw file upload)

User can also upload Multiple Raw files with a group simultaneously by using **Bulk Upload button**. When user clicks on the group the right hand side of the page displays group information here user can locate the **bulk upload** button.

The user can also upload the single raw data files by clicking on **Raw Data file Upload**.

Raw data files with following file extensions are supported for Data pre-processing by this portal.

- i) .netCDF
- ii) .mzML
- iii) .mzXML
- iv) .mzDATA

## Instruments supported for raw Data upload by the portal

-- High Performance Liquid Chromatography (HPLC)

- HPLC/Q-TOF
- HPLC/UHD Q-TOF
- HPLC/ UHD Q-TOF (Negative mode)
- HPLC/Brucker Q-TOF (Negative mode)
- HPLC/Ion Trap
- HPLC/Single Quad
- HPLC/ Orbitrap I (Fourier Transform - MS)
- HPLC/Orbitrap II (Fourier Transform - MS)
- HPLC/Waters TOF
  
- **Ultra Performance Liquid Chromatography (UPLC)**
- UPLC/UHD Q-TOF
- UPLC/Orbitrap (Fourier transform - MS)
- UPLC/Q-Executive
- UPLC/Brucker Q-TOF
- UPLC/Triple TOF
  
- **Gas Chromatography (GC)**
- GC/TOF
- GC/Single Quad

### **Data upload from Phenotypic Barcode reader**

Phenotypic data from the barcode reader can be uploaded (excel file) by clicking on the “**Phenotype samples**” button. This button is displayed to the person who created the group. By uploading the corresponding file, the values will automatically populate into the phenotype forms of the samples. The user also has the option of downloading the “**Phenotype template**” by clicking this button. Then the values can be filled in and uploaded by clicking on the “**Phenotype samples**” button.

## **Module II - Data pre-processing, comparative analysis and visualization**

## LC/GC-MS

1. Click on the “Pre-processing” menu under **Tasks/Results** on the left panel. User can start a new pre-processing task by clicking on the button “+New Pre-Processing Task”
2. Give a study name (optional). Let the study name be precise and distinct to avoid confusion for later tasks.
3. “Select a single project/cross project”: Desired option to be selected by the user as accord to the samples under study.
4. "Select a project" Desired project to be selected by the user for single project option and select another project for cross project study.
5. “Select two or more distinct sets of Samples”:

---- Set 1:

----- “Select a group”: Select the first group in which samples have to be pre-processed. The raw data files already uploaded in the data capture would be considered.

----- “Select an instrument”: Select the instrument using which raw data files of the samples of interest have been generated.

---- Set 2:

----- “Select a group”: Select the second group in which samples have to be pre-processed for comparison. The raw data files already uploaded in the data capture for these samples would be considered.

----- “Select an instrument”: Select the instrument using which raw data files of the samples of interest have been generated.

....

---- Set N:

----- “Select a group”: Select the N<sup>th</sup> group in which samples have to be pre-processed for comparison. The raw data files already uploaded in the data capture for these samples would be considered.

----- “Select an instrument”: Select the instrument using which raw data files of the samples of interest have been generated.

Samples under the selected groups (for both sets) are displayed below.

6. Depending on the requirement, user can select a few samples (by clicking on each sample code) or all samples (by clicking on “Select All” button), for both the sets.
7. The user has an option to change instrument’s default parameters (scroll down for more information).
8. Click “Submit” button to submit a job.

## Explanation of parameters:

Peak Detection: This is to detect peaks using different algorithms (given below)

----- CentWave: It is one of the methods used for peak detection. This algorithm is mostly used for high resolution LC/fTOF, OrbiTrap, FTICRg-MS data in centroid mode. Due to the fact that peak centroids are used, a binning step is not necessary.

----- Ppm (Parts per million): The ppm parameter has to be set according to the machine accuracy, e.g. ppm=25 and should fall in the range 2.5 to 100.

----- Min peak width: The peak width range (0.6 to 10) has to be set according to the chromatographic peak width range, e.g. minimum peak width = 20 seconds for HPLC and minimum peak width = 5 seconds for UPLC chromatography.

----- Max peak width: The peak width range (1.4 to 60) has to be set according to the chromatographic peak width range, e.g. maximum peak width = 50 seconds for HPLC and maximum peak width = 12 seconds for UPLC chromatography.

----- Matched filter: This method uses Gaussian model peak width as an integral part of the peak detection algorithm. Model peak width affects the signal to noise ratio. It is specified as full width at half maximum (fwhm).

----- Fwhm (full width at half maximum): Gaussian model peak width. Depending on the type of chromatography, the correct model peak width can be quite different. One means of determining the peak width is to fit the Gaussian function to one or more peaks in representative samples produced with your experimental protocol. It should fall in the range 3 to 30.

----- Step: Step size should fall in the range 0.1 to 0.5.

Alignment: After peak identification, peaks representing the same analyte across samples must be placed into groups. That is accomplished with the alignment methods (given below). This algorithm processes the peak lists in order of increasing mass.

----- Density: Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

----- Mzwid: Width of overlapping m/z slices to use for creating peak density chromatograms and grouping peaks across samples.

----- Minfrac: Minimum fraction of samples necessary in at least one of the sample groups for it to be a valid group.

----- Bw: Band width (standard deviation or half width at half maximum) of Gaussian smoothing kernel to apply to the peak density chromatogram.

----- MzClust: Runs high resolution alignment on single spectra samples stored in a given xcmsSet.

----- Nearest: Group peaks together across samples by creating a master peak list and assigning corresponding peaks from all samples.

Retention time correction: After matching peaks into groups, we (CCPM) used those groups to identify and correct correlated drifts in retention time from run to run. The aligned peaks are then used for a second pass of peak grouping which will be more accurate than the first. The whole process can be repeated in an iterative fashion.

----- Obiwar: Calculate retention time deviations for each sample.

----- Peak Groups: These two methods use “well behaved” peak groups to calculate retention time deviations for every time point of each sample. Use smoothed deviations to align retention times.

----- Smooth: Options available are

-----loess: For non-linear alignment

----- Linear: For linear alignment

### **Instruments supported for Data pre-processing by the portal**

#### **-- High Performance Liquid Chromatography (HPLC)**

----- HPLC/Q-TOF

----- HPLC/UHD Q-TOF

----- HPLC/ UHD Q-TOF (Negative mode)

----- HPLC/Brucker Q-TOF (Negative mode)

----- HPLC/Ion Trap

----- HPLC/Single Quad

----- HPLC/ Orbitrap I (Fourier Transform - MS)

----- HPLC/Orbitrap II (Fourier Transform - MS)

----- HPLC/Waters TOF

#### **-- Ultra Performance Liquid Chromatography (UPLC)**

----- UPLC/UHD Q-TOF

----- UPLC/Orbitrap (Fourier transform - MS)

----- UPLC/Q-Executive

----- UPLC/Brucker Q-TOF

----- UPLC/Triple TOF

#### **-- Gas Chromatography (GC)**

----- GC/TOF

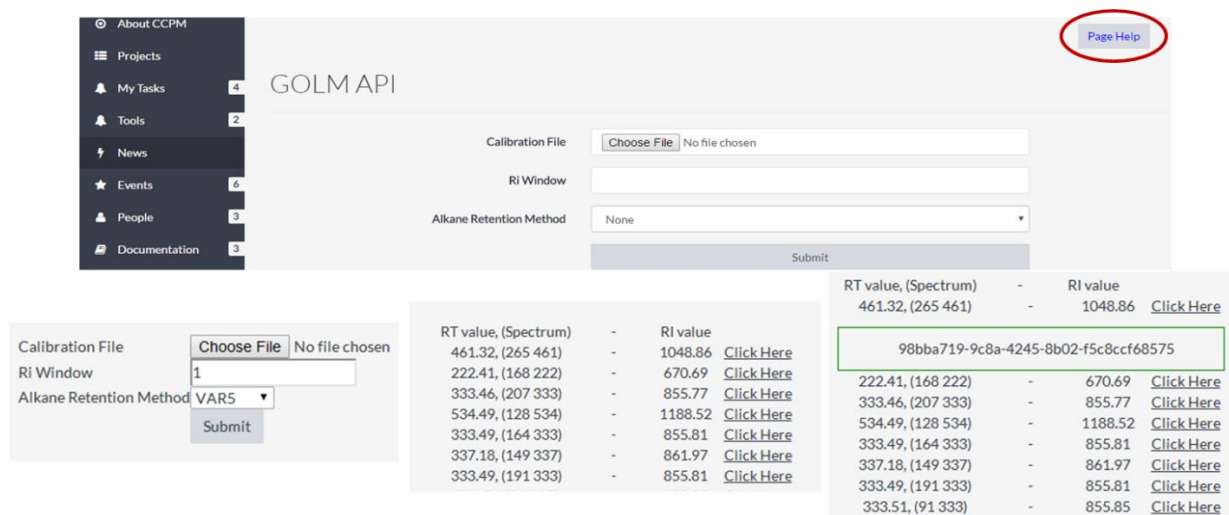
----- GC/Single Quad

### **Data Pre-processing results**

The results can be viewed under different menu options displayed on the left hand side panel:

----**Parameters**: gives the parameters with which the pre-processing task was done including the names of instrument, groups and samples used for pre-processing.

----**Pre-processed files:** Peaklist.csv, Diffreport and Annotation files are available for download individually or together in Results.zip. METLIN and GOLM links are also provided in this page for individual peak annotation.



The screenshot shows the GOLM API web interface. On the left is a sidebar with navigation links: About CCPM, Projects, My Tasks (4), Tools (2), News, Events (6), People (3), and Documentation (3). The main area contains the GOLM API form with fields for Calibration File (Choose File), RI Window (1), and Alkane Retention Method (VAR5). A red circle highlights the 'Page Help' link in the top right corner. Below the form, a table displays results for the selected parameters. The table has two columns: 'RT value, (Spectrum)' and 'RI value'. The first row shows '461.32, (265 461)' and '1048.86' with a 'Click Here' link. The second row shows '98bba719-9c8a-4245-8b02-f5c8ccf68575'. The third row shows '222.41, (168 222)' and '670.69' with a 'Click Here' link. The fourth row shows '333.46, (207 333)' and '855.77' with a 'Click Here' link. The fifth row shows '534.49, (128 534)' and '1188.52' with a 'Click Here' link. The sixth row shows '333.49, (164 333)' and '855.81' with a 'Click Here' link. The seventh row shows '337.18, (149 337)' and '861.97' with a 'Click Here' link. The eighth row shows '333.49, (191 333)' and '855.81' with a 'Click Here' link. The ninth row shows '333.51, (91 333)' and '855.85' with a 'Click Here' link.

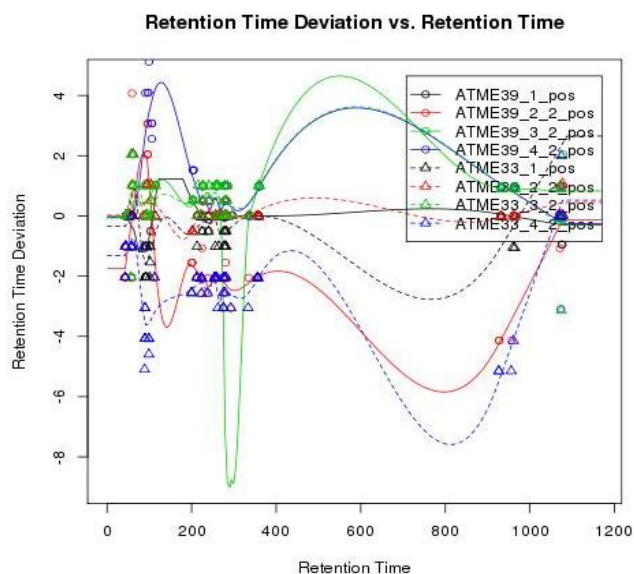
RT value, (Spectrum)	RI value	
461.32, (265 461)	1048.86	<a href="#">Click Here</a>
98bba719-9c8a-4245-8b02-f5c8ccf68575		
222.41, (168 222)	670.69	<a href="#">Click Here</a>
333.46, (207 333)	855.77	<a href="#">Click Here</a>
534.49, (128 534)	1188.52	<a href="#">Click Here</a>
333.49, (164 333)	855.81	<a href="#">Click Here</a>
337.18, (149 337)	861.97	<a href="#">Click Here</a>
333.49, (191 333)	855.81	<a href="#">Click Here</a>
333.51, (91 333)	855.85	<a href="#">Click Here</a>

The project used in above screen shot analyses GC data. Hence clicking on the Golm link takes us to the Golm website. The user can upload the Calibration file, enter RI window and Alkane retention Method to access GOLM links to metabolites. The results shown above are obtained by opting for the default calibration file, entering a RI window value (=1) and choosing Alkane Retention Method as VAR5. Page Help on the top right provides guidance in setting these parameters.

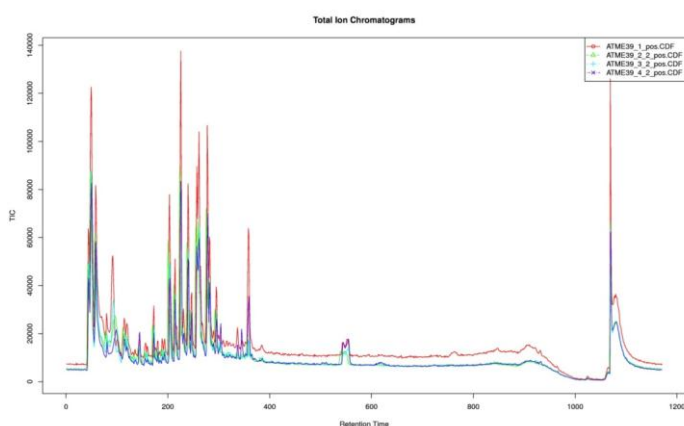
Note:-

1. In GOLM API result page the “RT value, (spectrum)” follows the format as (mz RT). So the first value in bracket is mz and second value is RT.
2. For RI window if user provides smaller value then the number of hits will be reduced so recommended value is more than 5.
3. RI calculation see [https://en.wikipedia.org/wiki/Kovats\\_retention\\_index](https://en.wikipedia.org/wiki/Kovats_retention_index)
4. A default Calibration file will be used when no file is attached.

----**Retention time:** The plot displays Retention time deviation vs. Retention time across various samples. A negative number indicates a sample was eluting before most of the others, and vice versa. An example of this plot is shown below

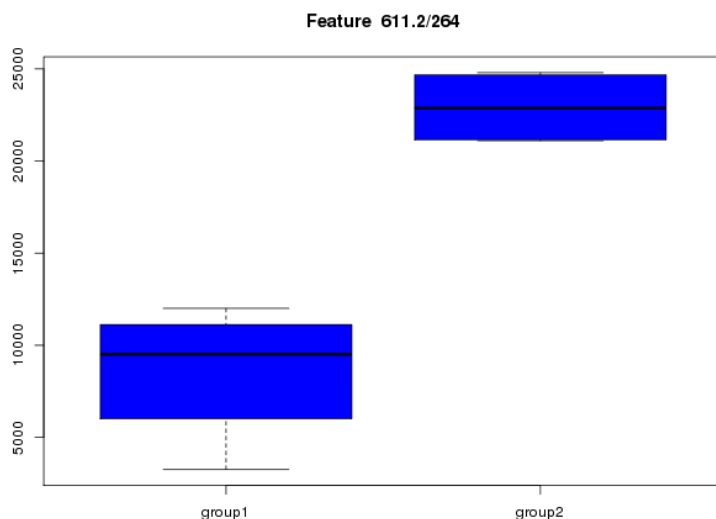


---- **TIC raw and corrected images:** The total ion current (TIC) chromatogram represents the summed intensity across the entire range of masses being detected at every point in the analysis. The range is typically several hundred mass-to-charge units or more. An example of this plot is shown below

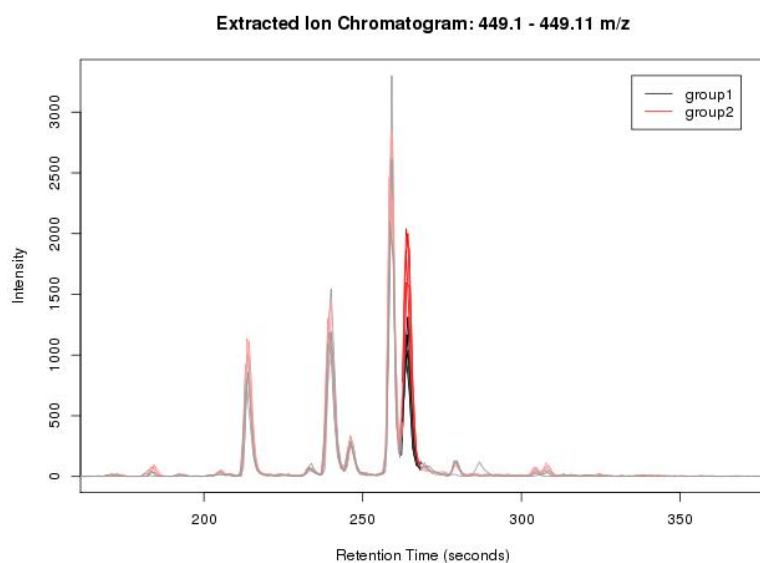


---- **Box plots:** Box-plot is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots for different classes were auto-generated for top 10 statistically significant features. An example of this plot is shown below





---**EIC plots:** In an extracted-ion chromatogram (XIC or EIC), also called a reconstructed-ion chromatogram (RIC), one or more  $m/z$  values representing one or more analytes of interest are recovered ('extracted') from the entire data set for a chromatographic run. Auto-generated extracted ion chromatograms for the top 10 (default) differentially regulated ions. Darkened lines indicate where the peaks were integrated for quantitation. An example of this plot is shown below



## NMR

### 1. Binning

Uploaded untreated raw spectral data in text format allow to bin using binning feature. Binned data use for the further pre-treatment and statistical analysis.

### NMR Instruments supported for Data pre-processing (binning) by the portal

----NMR Spectrometer – Bruker

----NMR Spectrometer - Varian/Agilent

----NMR Spectrometer – Others

The screenshot shows a web portal for NMR data pre-processing. At the top, there is a 'Study Name' input field containing 'testNMR' and a 'Preprocessed Results' button. Below this, the interface is divided into two main steps. 'STEP-I: SELECT A PROJECT' features a dropdown menu currently set to 'NMR TEST'. 'STEP-II: SELECT TWO DISTINCT SETS OF SAMPLES' includes a 'Select number of Sets' dropdown set to '2'. It then displays two columns, 'Set1' and 'Set2'. Each column has a dropdown for the experiment type (Set1 is 'HNN', Set2 is 'KLN') and another for the spectrometer (both are 'NMR Spectrometer - Bruker'). Below these, there are lists of sample files: Set1 has 'HNN\_0\_2992\_spc', 'HNN\_90\_3321\_spc', 'HNN\_60\_3211\_spc', and 'HNN\_30\_3101\_1\_spc'; Set2 has 'KNL\_0\_2552\_spc', 'KNL\_30\_2661\_spc', 'KNL\_60\_2771\_spc', and 'KNL\_90\_2881\_spc'. Each file name is on a blue button. At the bottom of each set's list are 'Select all' and 'Unselect all' buttons. A 'Submit' button is located at the bottom center. A link for 'Advanced parameters' is visible on the left side.

## Module III – Filtration, Pre-treatment and Statistical analysis

### Filtration

Filtration can be performed by clicking on the

- 1) Click on “**My Tasks**” on the left hand side bar. Select “**Pre-processing**” option.
- 2) “**Filtration**” button on the pre-processed results.

Upon clicking the “**Filtration**” button the user is navigated to a interactive data table where he can choose the type of Filtration he wants to do on the data and submit. The task submitted can be found under the “**Running Tasks**”. The status of the task is displayed (queued, assigned, running or failed). A **blue refresh** button is given next to the task name. The task can be deleted by clicking the **red** button listed under “**actions**”.

Once the task is completed, it is listed under the “**results**”. The results can be viewed by clicking the blue checked button listed under “**actions**”. The task can be deleted by clicking the **red** button listed under “**actions**”.

### **Explanation of parameters:**

User has the option to perform Filtration on the data-table based on different parameters like mz-median, p-value, fold-change, and RT-median.

### **Pre-treatment**

Pre-treatment can be performed by clicking on the

- 1) Click on “**My Tasks**” on the left hand side bar. Select “**Filtration**” option.
- 2) “**pre-treatment**” button on the filtration results.

Upon clicking the “**Pre-treatment**” button the user is navigated to a form where he can choose the type of Pre-treatment (row wise/column wise) he wants to do on the data and submit. The task submitted can be found under the “**Running Tasks**”. The status of the task is displayed (queued, assigned, running or failed). A **blue refresh** button is given next to the task name. The task can be deleted by clicking the **red** button listed under “**actions**”.

Once the task is completed, it is listed under the “**results**”. The results can be viewed by clicking the blue checked button listed under “**actions**”. The task can be deleted by clicking the **red** button listed under “**actions**”.

**External Pre-processing:** Data pre-processed by other software can be Pre-treated in the portal by clicking the option “**Upload Pre-processed results**” under “**Pre-processing**” option in the **Tasks/Results** tab of the respective project as well as by choosing **Pre-processing** under “**My Tasks**” on the left hand side bar. Select the project name, data format (if the Samples are in rows or columns) and upload the .csv file. User can click **Preprocessed File Template** button and download **Sample by Columns** or **Samples by Rows** for reference. The downloaded files can be opened in Windows with MS excel. Once uploaded, click on the **Pre-treatment** button for Pre-treatment and specify the methods to be performed. Click on the checkbox if the data is log-transformed and click on submit.

## Explanation of parameters:

User has the option to perform the following Pre-treatment methods:

### Row Wise Procedures

Row – wise aims at adjusting the variance of different samples (to normalize each sample (row) so that it is comparable to the other). User has the option of choosing

1. None :- skip the Normalization process,
2. Normalization by Sum
3. Normalization by Median: - The median is preferred to average as a statistical measure since it is more robust.
4. Normalization by reference sample (probabilistic quotient norm)
5. Normalization by a pooled sample from group
6. Normalization by reference feature.

### Column Wise Procedures

Column-wise aims at adjusting the variance of different features executed on the columns of data (intensity) across all samples. There are two components such as ‘Data transformation’ and ‘Data Scaling’.

#### A. Data Transformations

Transformations are generally applied to convert multiplicative relations into additive relations, and to make skewed distributions symmetric. User has the option of choosing ‘None’ to skip, ‘Log transformation’ or ‘Inverse hyperbolic sine (IHS)’.

1. *Log Transformation*: It is the classic solution to normalizing skewed data. It adds symmetry to positive and negative changes within a dataset as well as minimizes the impact of outlier entries. A drawback of the log transformation is that it is unable to deal with the value zero. Metabolomics data from any given sample will always possess far more metabolites in the low-abundance range than the higher range, thus necessitating a log transform to eliminate the impact of the minority high-abundance metabolites.
2. *Inverse hyperbolic sine (IHS)*: Generally possesses the same characteristics as a classical log transform, such as large value suppression. The primary difference between the IHS and log transformations is that while a standard logarithm is undefined at zero, the IHS function is fully defined and thus will not remove zeros from the data during transformation.

#### B. Data Scaling

User has the option of choosing ‘None’ to skip, ‘Auto scaling’, ‘Pareto scaling’ and ‘range scaling’.

1. *Auto Scaling*: Mean-centered (each column of the table can be achieved a mean of “0” by subtracting the column mean from each value in the column) and divided by the standard deviation of each variable. Each column of the table can be scaled so that it has unit variance by dividing each value in the column by the standard deviation of the column. It results in every feature displaying a standard deviation of one, i.e. the data is transformed to standard units.

Advantage: All metabolites are equally important.

Disadvantage: Inflation of measurement errors.

2. *Pareto Scaling*: Mean-centered (each column of the table can be achieved a mean of “0” by subtracting the column mean from each value in the column) and divided by the square root of standard deviation of each variable) Pareto scaling is similar to auto scaling but the square root of the standard deviation is used as the scaling factor instead of standard deviation. Its normalizing effect is less intense, such that the Pre-treated data stays closer to its original values. It is less likely to blow up noisy background and reduces the importance of large fold changes compared to small ones. However, very large fold changes may still show a dominating effect.

Advantage: Stays closer to original measurement.

Disadvantage: Sensitive to large fold changes.

3. *Range Scaling*: Mean-centered ((each column of the table can be achieved a mean of “0” by subtracting the column mean from each value in the column) and divided by the range of each variable. A disadvantage of range scaling with regard to the other scaling methods tested is that only two values are used to estimate the biological range, while for the standard deviation all measurements are taken into account. This makes range scaling more sensitive to outliers.

Advantage: All metabolites equally important. Biologically related scaling

Disadvantage: Inflation of measurement errors, sensitive to outliers.

**Note:** Auto scaling and range scaling seem to perform better than the other methods with regard to the biological expectations.

## Statistical analysis

The user can start the statistical analysis by clicking on the,

1. “Pre-treatment” option in the “Tasks/Results” tab on the right hand panel of the corresponding project.
2. “Pre-treatment” option under the “My Tasks” menu option on the left hand side bar.

After Pre-treatment, user can proceed for statistical analysis by clicking the “Statistical Analysis” button next to the corresponding Pre-treated result. On clicking “Statistical analysis” button, the user is navigated to a form where he can choose the type of statistical analysis he wants to do and click on the “Perform statistical analysis” button. Statistical analysis can be done for two groups at a time.

Once the task is completed, the result is displayed. The task submitted can be found under the “Running Tasks”. The status of the task is displayed (queued, assigned, running or failed). A blue refresh button is given next to the task name. The task can be deleted by clicking the red button listed under “actions”.

**Externally Pre-treated data:** Already Normalised data can also be uploaded in the portal by clicking the “Upload Pre-treated Results” button under Pre-treatment option in Tasks/Results of the corresponding project or under My Tasks of left hand side bar. The format of the file is shown below. Once uploaded click on “Statistical Analysis” button corresponding to the uploaded file. The format of the .csv file to be uploaded is shown below.

Samples		Groups/factors		Variables															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Sample_Name	Group_Name	M101T107	M101T993	M101T979	M101T102	M104T55	M105T553	M111T45	M112T45	M113T107	M114T107	M116T417	M116T562	M116T539	M116T377	M116T452	M116T477	M116T520
2	ATME32_1_2_pos	Flower	-0.50217	-0.40059	-0.40405	-0.50771	-0.50964	-0.51423	-0.49738	-0.40144	-0.4818	-0.48809	-0.61478	-0.64955	-0.52997	-0.4886	-0.50339	-0.51519	-0.51186
3	ATME91_1_2_pos	Leaf	0.502166	0.400591	0.404045	0.507706	0.509639	0.514227	0.497376	0.401445	0.481798	0.488087	0.614782	0.649554	0.529974	0.488602	0.503392	0.515191	0.511861
4																			
5																			

The accepted format is .csv. It is essential to have ‘samples in rows’ i.e. there is a sample at each row and its features are present in various columns. Sample names should be unique with group names in the second column.

### Explanation of parameters:

User has the option to perform the following statistical methods:

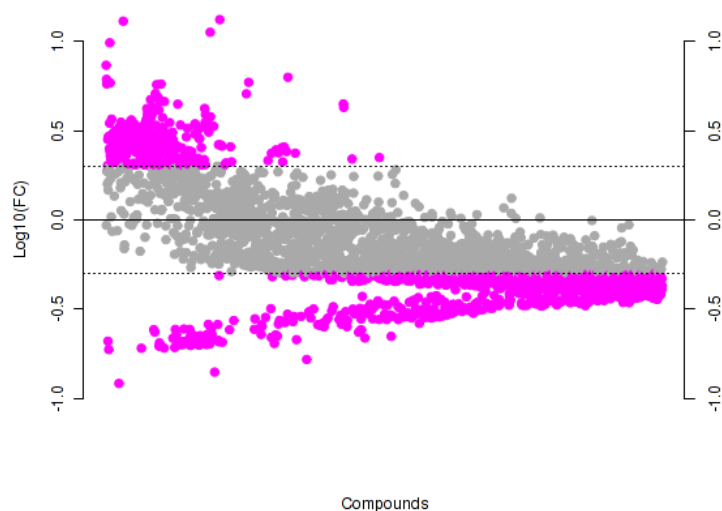
#### Univariate Analysis

Although metabolomics experiments generate multivariate data; one can employ univariate methods to test for individual metabolites that are increased or decreased significantly between different groups. Univariate methodologies are frequently used to reduce a possibly large number of measured analytes to only those that show the strongest response under the investigated conditions.

Because of their simplicity and interpretability, univariate analyses are often first used to obtain an overview or rough ranking of potentially important features before applying more sophisticated analyses. Univariate analysis examines each variable separately and does not consider the effect of multiple comparisons. Examples for such univariate approaches are: t-test, volcano plot, fold change and correlations.

### Fold Change (Applicable only for two groups)

Fold change is a measure describing how much a quantity changes going from an initial to a final value. For example, an initial value of 30 and a final value of 60 correspond to a fold change of 2, or in common terms, a two-fold increase. For example, the image looks like



### Correlation Analysis

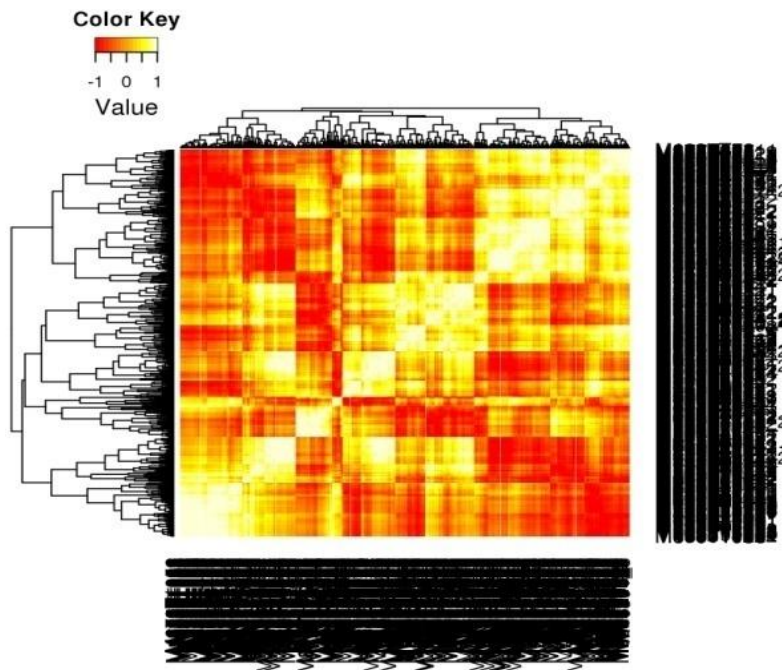
The correlation analysis is performed on normalised compound abundance levels. Compounds can then be clustered according to how closely correlated they are. Compounds with a high correlation value (i.e. close to 1) show similar abundance profiles while compounds which a high negative correlation value (i.e. close to -1) show opposing abundance profiles. It can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values, measuring the degree of correlation:

Pearson correlation coefficient is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other).

Spearman correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or -1 occurs. Spearman's coefficient, like any correlation calculation, is appropriate for both continuous and discrete variables, including ordinal variables.

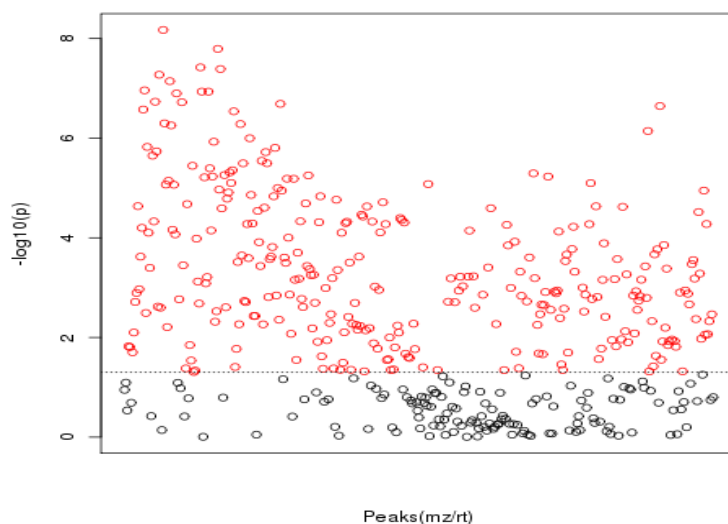
Kendall's rank correlation provides a distribution free test of independence and a measure of the strength of dependence between two variables. Kendall's rank correlation improves upon this by reflecting the strength of the dependence between the variables being compared. For example, the image looks like.

There is also availability of downloading the edge list for visualization of network in standalone Cytoscape software.



### t-test (parametric) (Applicable only for two groups)

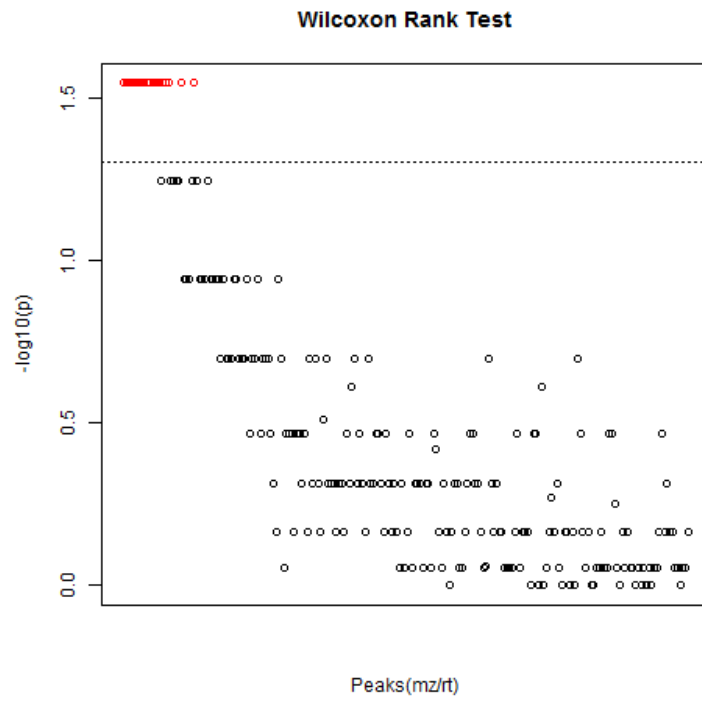
Student's t-test determines whether the means of two groups are distinct or not. This test assumes that the data is normally distributed. For example, the image looks like



### t-test (non- parametric) (Wilcoxon test) (Applicable only for two groups).

Non parametric-Wilcoxon test does not assume a normal distribution. Output is a .csv file containing the p-values for each variable. For example, the image looks like

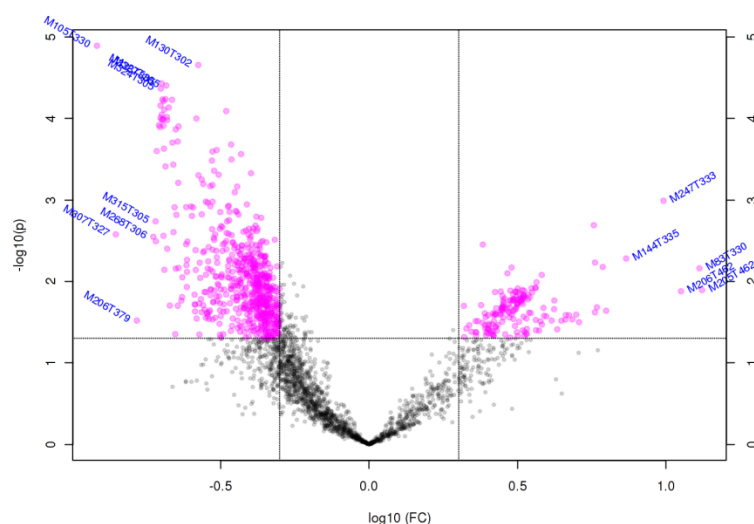




### Volcano Plot (Applicable only for two groups)

When one has a list of thousands of replicate data points between two conditions, one can identify most-meaningful changes quickly by plotting a scatter-plot like volcano plot. It plots statistical significance (e.g., a p-value from an ANOVA model) versus fold-change on the y- and x-axes, respectively. A volcano plot gives quick visual identification of data-points (genes or samples etc.) that display large-magnitude changes that are also statistically significant.

A volcano plot is constructed by plotting the negative log of the p-value on the y-axis (usually base 10). This results in data points with low p-values (highly significant) appearing toward the top of the plot. The x-axis is the log of the fold change between the two conditions. The log of the fold-change is used so that changes in both directions (up and down) appear equidistant from the center. Plotting points in this way results in two regions of interest in the plot: those points that are found toward the top of the plot that are far to either the left- or the right-hand side. These represent values that display large magnitude fold changes (hence being left- or right- of center) as well as high statistical significance (hence being toward the top).



## Clustering Analysis

## Dendrogram

The dendrogram is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom of the dendrogram and referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node. This can be seen in the diagram below. At each dendrogram node we have a right and left sub-branch of clustered compounds. In the following discussion, compound clusters can refer to a single compound or a group of compounds. The vertical axis is labelled distance and refers to a distance measure between compounds or compound clusters. The height of the node can be thought of as the distance value between the right and left sub-branch clusters.

If compounds are highly correlated, they will have a correlation value close to 1. Therefore, highly correlated clusters are nearer the bottom of the dendrogram. Compound clusters that are

not correlated have a correlation value of zero. Compounds that are negatively correlated, i.e. showing opposite abundance behaviour, will have a correlation value of -1.

As we move up the dendrogram, the compound clusters get bigger and the distance between compound clusters increases in value. It becomes difficult to interpret distance between compound clusters when compound clusters increase in size.

Manhattan and Euclidian distance functions are used to determine similarity.

#### Euclidean distance

This is an absolute distance between two datasets. Euclidean distances between objects, such as dots on a paper could be measured with a ruler.

#### Manhattan distance

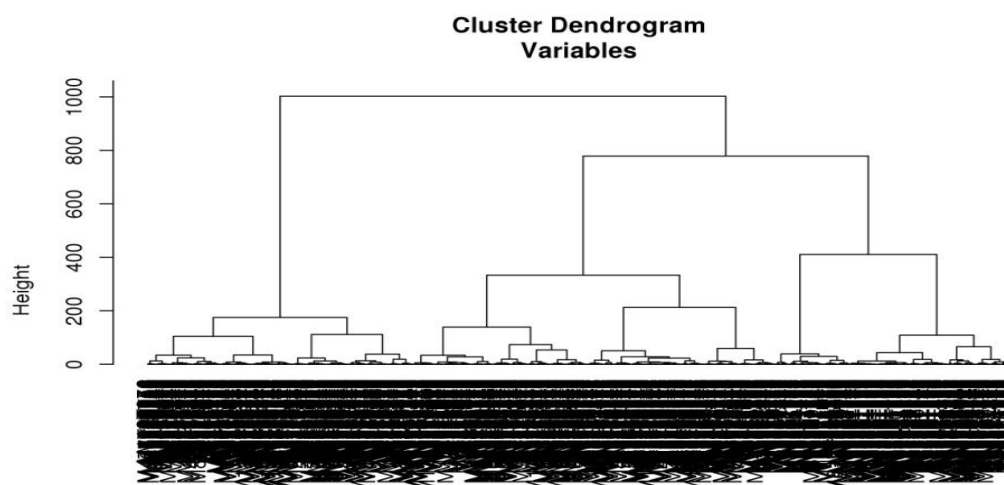
It is also possible to modify the dendrogram drawing method. Four possible options are:

Single linkage: A distance between clusters in the tree is calculated using the shortest distance between them.

Average linkage (UPGMA): A distance between clusters in the tree is calculated using average distance between them.

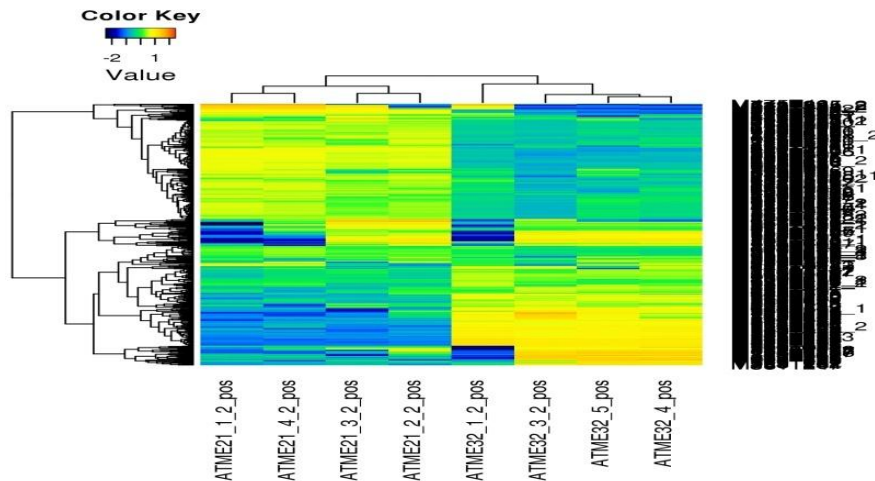
Complete linkage: A distance between clusters in the tree is calculated using the longest distance between them.

Ward: At every step of clustering two clusters that result into a minimal loss of information are combined. Information loss is measured using error sum-of-squares criterion. For example, the image looks like



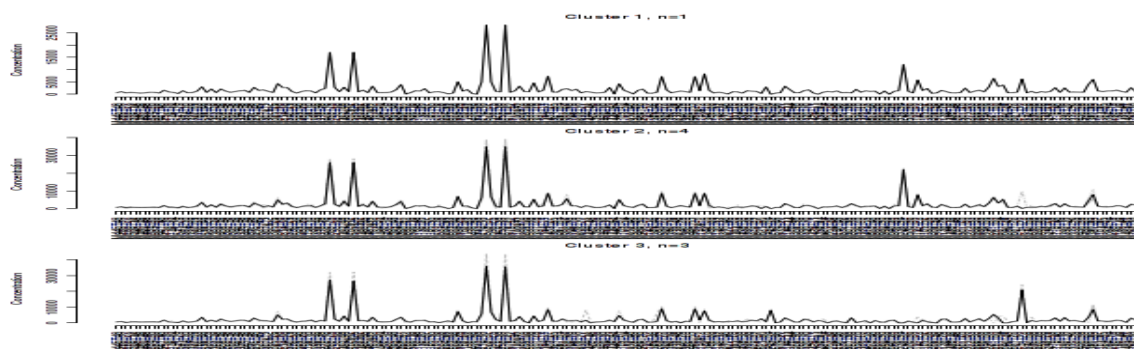
## Heatmap

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colours. The colours are from blue to red showing minimum to maximum value respectively. For example, the image looks like



## K-means

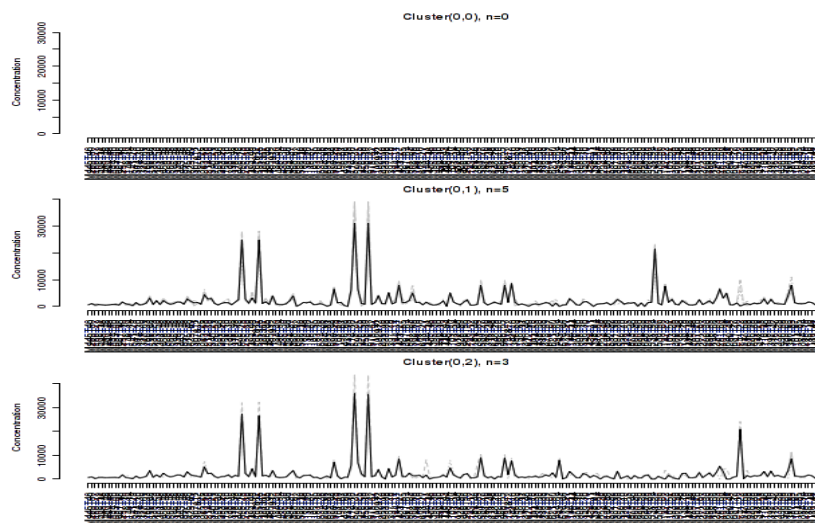
K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The user has to specify the number of centroids. The main idea is to define  $k$  centroids, one for each cluster. This results in a partitioning of the data space into Voronoi cells. For example, the image looks like



## SOM

Self-organizing map (SOM) or self-organizing feature map (SOFM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. The user hence has to specify the parameters like dimensions, initialization methods and neighbourhood.

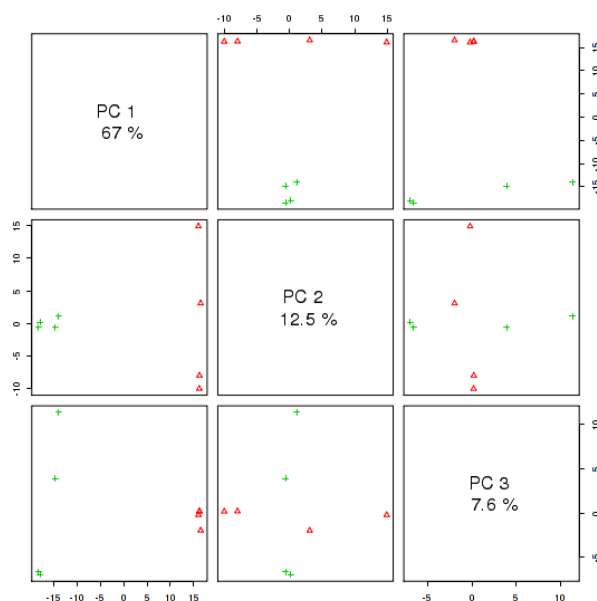
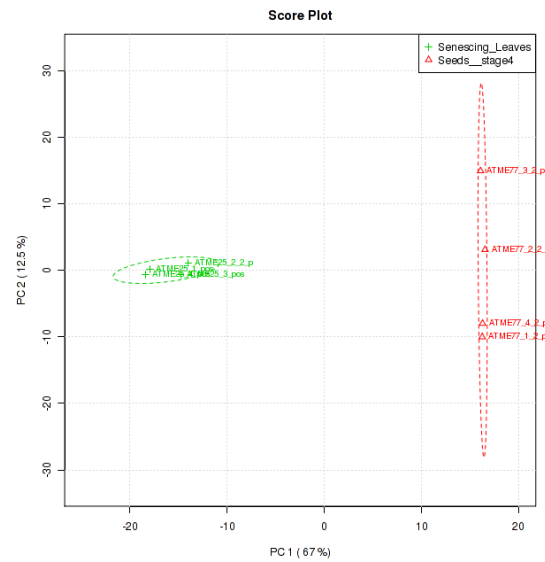
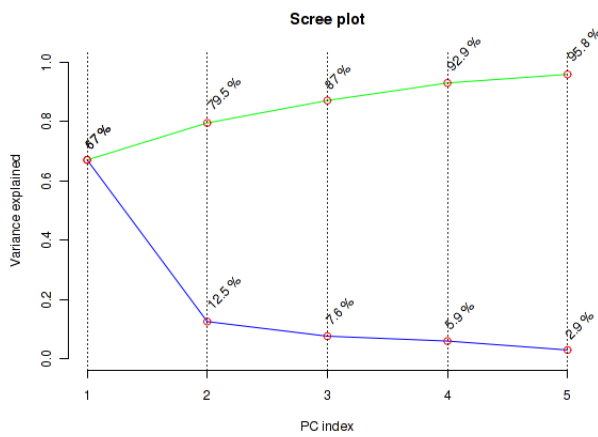
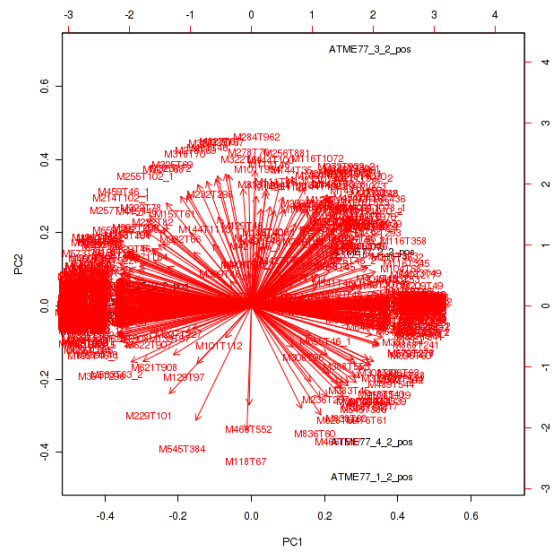
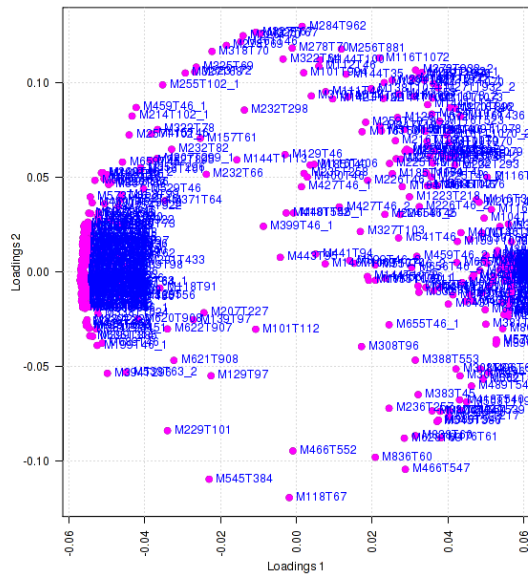
The goal of learning in the self-organizing map is to cause different parts of the network to respond similarly to certain input patterns. For example, the image looks like



## Multivariate Analysis

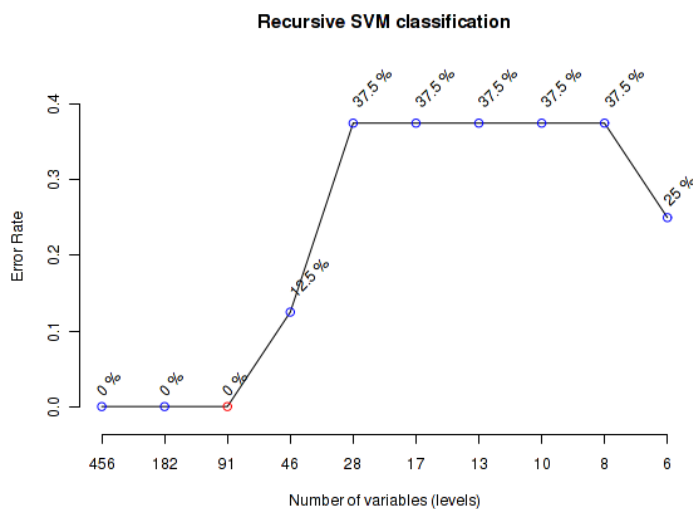
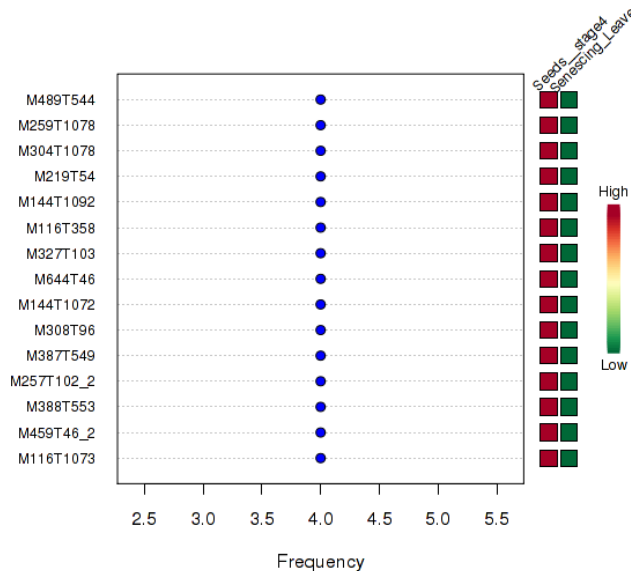
### PCA

Principal Component Analysis (PCA) is a data transformation technique which is used to reduce a multidimensional data set to a lower number of dimensions for further analysis. In PCA, a data set of interrelated variables is transformed to a new set of variables called principal components (PCs) in such a way that they are uncorrelated, and the first few of these PCs retain most of the variation present in the entire data set. The first PC is a linear combination of all the actual variables in such a way that it has the greatest amount of variation, and the second PC is a combination of the variables that have the next greatest variation in the remaining PCs. This script produces multiple plots – residual variance, scores plot labelled with sample names, scores plot labelled with group names and a loading plot. For example, the image looks like



## SVM

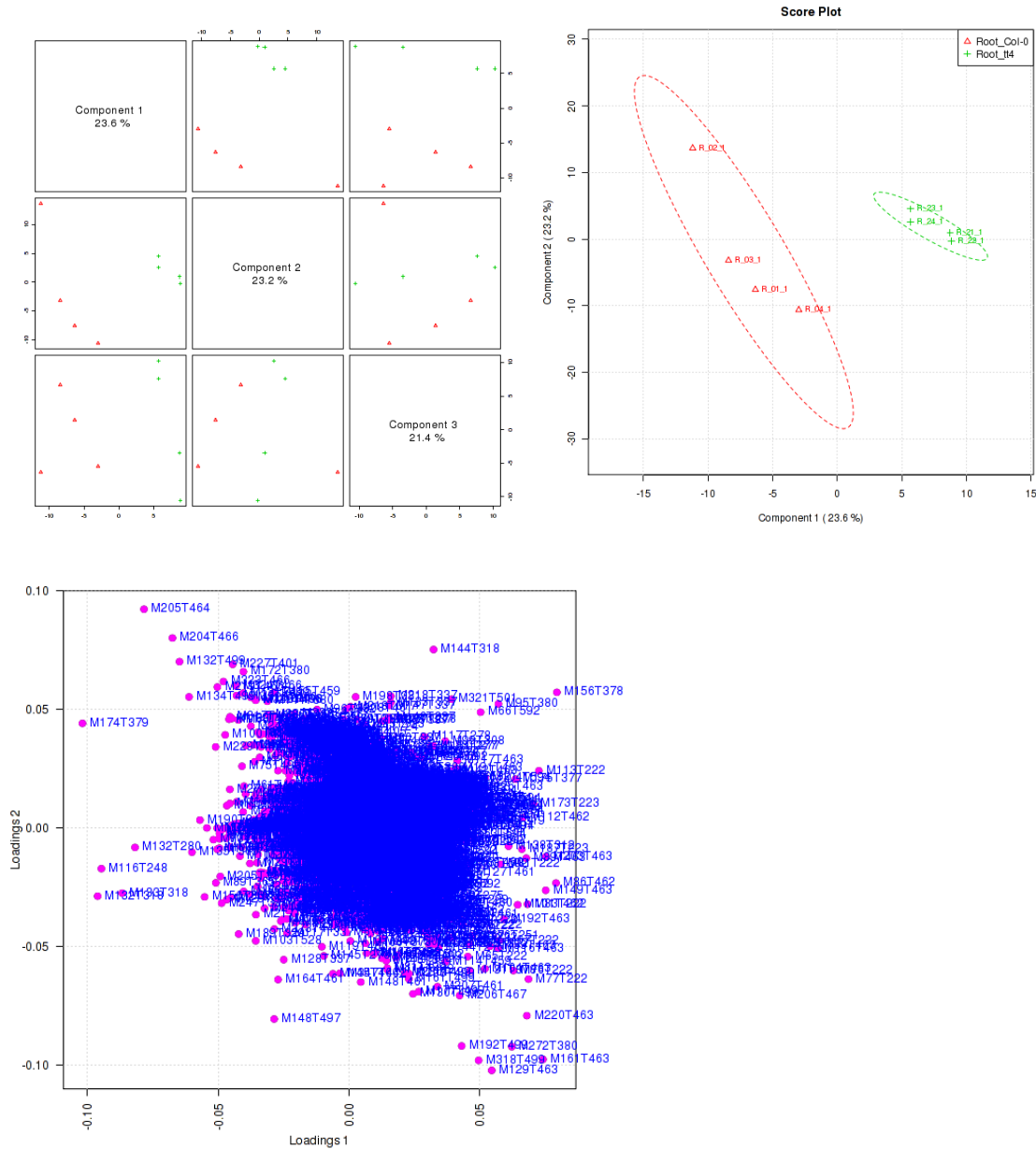
SVM is "Support Vector Machine". It is machine learning technique used for classification of the samples into various groups. It works by first training the models based on training set and then apply the model on a sample to predict its class (or group).



## PLS-DA

PLS Discriminant Analysis (PLS-DA) is performed in order to sharpen the separation between groups of observations, by hopefully rotating PCA (Principal Components Analysis) components such that a maximum separation among classes is obtained, and to understand which variables carry the class separating information.

PLS-DA consists in a classical PLS regression where the response variable is a categorical one (replaced by the set of dummy variables describing the categories) expressing the class membership of the statistical units. Therefore, PLS-DA does not allow for other response variables than the one for defining the groups of individuals. As a consequence, all measured variables play the same role with respect to the class assignment. Actually, PLS components are built by trying to find a proper compromise between two purposes: describing the set of explanatory variables and predicting the response ones.





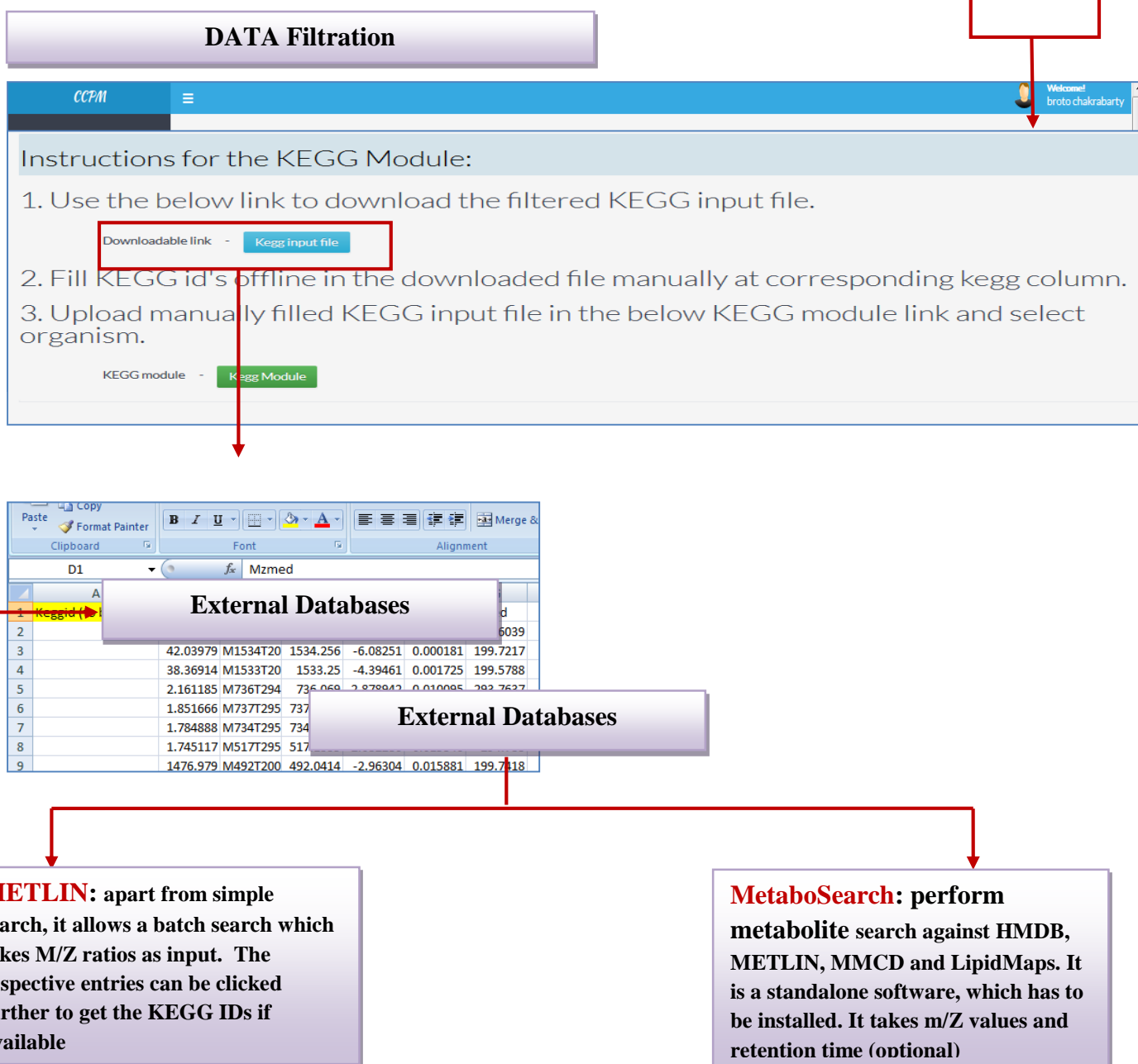
## Module IV – KEGG Connectivity

### Entry to the KEGG Module:

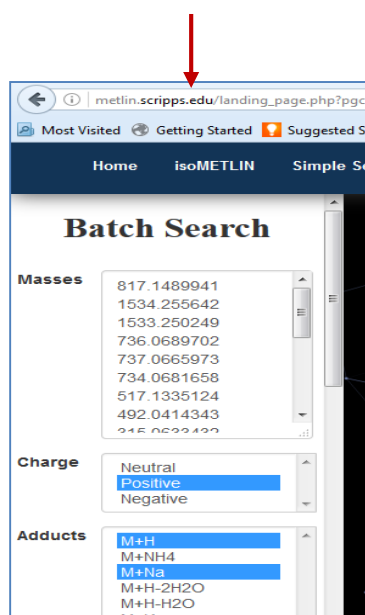
The KEGG pathway analysis can be carried out in two ways:

- 1) From the “Data Filtration” step, the file can be downloaded. The KEGG IDs are then manually inserted after querying various external databases (like METLIN, HMDB, etc.) with the **m/Z values**
- 2) Alternatively, one can directly go to the KEGG module and upload the user-defined list of compounds (optionally with numerical parameters). Here also, the KEGG compound IDs have to be fetched manually from the external databases

The detailed flowchart is given below for METLIN and MetaboSearch. Similar steps can be followed for any other database. Kindly follow the instructions for the respective databases as given in their websites



[Back To Index Page](#)



metlin.scripps.edu/landing\_page.php?pgc

Most Visited Getting Started Suggested S

Home isoMETLIN Simple Se

## Batch Search

**Masses**


817.1489941  
1534.255642  
1533.250249  
736.0689702  
737.0665973  
734.0681658  
517.1335124  
492.0414343  
245.0622422

**Charge**

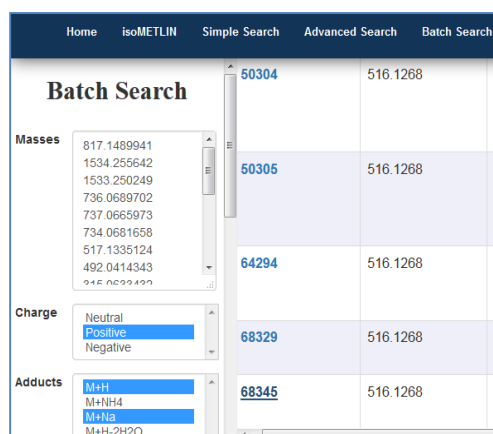
Neutral  
Positive  
Negative

**Adducts**

M+H  
M+NH4  
M+Na  
M+H-2H2O  
M+H-H2O  
M+K



	A	B
1	mz	rt
2	62.98193	31.44898
3	69.06956	50.27389
4	70.06484	29.59302
5	72.0804	34.29358
6	73.08333	34.24128
7	73.53049	23.18216
8	80.94842	27.18991
9	82.01386	23.07363
10	82.53558	23.18671
11	82.94562	27.18991
12	83.01783	23.4163



Home isoMETLIN Simple Search Advanced Search Batch Search

## Batch Search

**Masses**

817.1489941  
1534.255642  
1533.250249  
736.0689702  
737.0665973  
734.0681658  
517.1335124  
492.0414343  
245.0622422

**Charge**

Neutral  
Positive  
Negative

**Adducts**

M+H  
M+NH4  
M+Na  
M+H-2H2O

50304 516.1268

50305 516.1268

64294 516.1268

68329 516.1268

68345 516.1268

Query_ID	Query_m/z	Input_RT	Name	Formula	Exact_Mass
3	371.2282742	162.0353134	3,6,9,12,15,18,21-HEPTAOXATRICOSANE-1,23-DIOL	C16H34O9	370.2202827
4	450.3220874	265.8486871	Glycoursodeoxycholic acid;Chenodeoxyglycocholate;	C26H43NO5	449.3141235
4	450.3220874	265.8486871	Deoxycholic acid glycine conjugate;3alpha,12alpha-Di	C26H43NO5	449.3141235
KEGG ID	PubChem CID	PubChem SID	HMDB ID	Databases	dppm
-	-	-	-	MMCD	1.978340795
C05462;C05466;	670344;7822;	93353;440686;	HMDB00708;HMDB00637;	MMCD;Metlin;HMDB;LIPIDMaps	1.562168358
C04721;C05464;	375907;7293;	440459;440688	HMDB00631;HMDB04012;	MMCD;Metlin;HMDB;LIPIDMaps	1.562168358
Delta	Number of possible stereoisomers InChI String				
7.32E-04	1 InChI=1S/C16H34O9/c17-1-3-19-5-7-21-9-11-23-13-15-25-16-14-24-12-10-22-8-6-20-4-2-18/h17-18H,1-16H2				
7.02E-04	7 InChI=1S/C26H43NO5/c1-15(4-7-22)(30)27-14-23(31)(32)18-5-6-19-24-20(9-11-26(18,19)(3)25)2(10-8-17(28)12-1				
7.02E-04	6 InChI=1S/C26H43NO5/c1-15(4-9-23)(30)27-14-24(31)(32)19-7-8-20-18-6-5-16-12-17(28)10-11-25(16,2)21(18)13				
InChI Key					
GLZWNFNQMJAZGY-UHFFFAOYSA-N					
GHCZAUBVMUEKPP-AOJCKHDSN-N;GHCZAUBVMUEKPP-FJTOLMLMSA-N;					
WVULKSPCQVQLCU-FJTOLMLMSA-N;WVULKSPCQVQLCU-ODQMPPHLSA-I					
Peak No	m/z	RT	Isotopes	Adducts	Monoisotopic m/z Metabolite_group_ID
7	394.2136421	162.0353134	[2][M+1]+		371.2282742 1
8	432.312108	265.8486871	[13][M]+	[M+H-H2O]+ 449.319	450.3220874 4
8	432.312108	265.8486871	[13][M]+	[M+H-H2O]+ 449.319	450.3220874 4



METLIN ID	68345
Mass	516.126776232 <a href="#">m/z calculator</a>
Name	Isochlorogenic acid b
Synonym	
Systematic Name	
Formula	C <sub>25</sub> H <sub>24</sub> O <sub>12</sub>
CAS	14534-61-3
Purchase Option	
LMID	
KEGG	C10468
HMDB	

**Input file for KEGG**

### Pathway Analysis:

Analysis can be performed using two different inputs:

- 1) Upload a file with numeral parameters like fold change/p-values/concentrations. The file has to be a “**tab-separated**” file with KEGG compound ids in the first column and numerical parameters in the second column. **The compound IDs have to be fetched manually as shown above**
- 2) Upload a file without any numerical parameters. The file has only one column with KEGG compound ids

The appropriate templates for both the options are provided.

After uploading the file, the “**Organism Name**” has to be selected from the dropdown menu. The names are sorted in alphabetical order. For mapping to **KEGG Reference Pathways (KO)**, select “**KO**”. This will fetch the entries irrespective of the organisms.

Once the mapping is complete, the “**Fetched Pathways**” table gives a compound-wise view of the data. It tabulates the KEGG compound ids, names, the numerical parameters, the KEGG pathways and KEGG modules. The pathway and module names are clickable and will take the user to the external KEGG database to view the compound (in red) in the pathway/module maps.

On scrolling down, the “**Pathway Mapping**” table provides the pathway-wise view of the data. The KEGG pathways, pathway ids, the compounds mapped, total number of compounds mapped (which can be sorted) will be displayed. Additionally, the total list of compounds and genes present in the pathway, apart from what is given in the input file is given as an external link to the KEGG database.

### PathView Parameters

For viewing the pathway maps, the user can click on the “**View Pathway**” for each of the pathways. This will open a new tab. The following parameters are provided if the user provided numerical data along with compound ids:

**Limit:** This argument specifies the limit values for compound data when converting them to pseudo colours. Default is 1. The user can change it according to the max/min values of the numerical data

**Bin Size:** This argument specifies the number of levels or bins for compound data when converting them to pseudo colours. Default bin size is 10.

**Data has both +ve and -ve values:** This argument specifies if the numerical data has positive or negative values.

For data without numerical parameters, the compounds will be highlighted in yellow on the KEGG pathway maps. The maps can be downloaded as images

## Tools

### 1. Text Mzml Help page content

#### Requirements:

Linux 64-bit distro (preferably Ubuntu)

#### Usage:

Open Terminal

CD to the current directory

Run the program using following command `./format-convert`

A GUI window will be started. Press input button to select a text file for conversion.

Output field will be automatically filled with a filename having current time-stamp.

Press Convert to start the conversion. It may take huge amount of memory for large files.

Please contact me at [abi\\_chem@iiit.ac.in](mailto:abi_chem@iiit.ac.in) if you face any issues. Binaries for different operating systems will be made available if needed.

### 2. NMR (Metabolite identification Module)

NMR Metabolite identification tool uses the open source metabohunter scripts for semi-automatic assignment of 1D NMR spectra of metabolites. The metabolite identification interface search two major publically available NMR database (HMDB: Human Metabolome Database and MMCD: Madison Metabolomics Consortium Database) in background and display the result output from the selected database based on user input parameters. This tool for metabolite identification is based on spectra or peak lists with different search methods and with possibility for peak drift in a user defined spectral range. The NMR Module I result into assignments details displayed as Metabolite Name, Metabolite ID (in corresponding database), metabolite plot, matching peaks, while Matched peaks, table of hits, plot peaks in hit map, plot selected spectra available for download.

## Statistics

NMR Metabolite identification module in CCPM uses the manually curated HMDB and MMCD databases (there are 867 NMR spectra in HMDB and 448 in MMCD).

## Data upload for NMR metabolite identification

NMR raw data converted to text (.txt) format use to upload for metabolite identification.

**Figure 1:** Snapshot of metabolite identification tool (NMR module I)

### Details of User input parameters

**Input type:** The input data for CCPM must contain a list of peaks or a complete <sup>1</sup>H NMR spectrum. If a complete spectrum is provided as input, noise filtering and peak detection is applied.

**Input file:** The input file must have a list of peaks or a full spectrum. Data must be organized in two columns, such that each line must have only two values. The first value must be the ppm and the second value must be the height of the signal. A sample file is provided [here](#).

**Peaks list:** The input could be also pasted as a list of peaks or a spectrum. Data must be organized as described above.

**Database:** CCPM use the data extracted from two public databases, namely the Human Metabolome Database (HMDB) - v2.5 (Jan. 2011) and the Madison Metabolomics Consortium

Database (MMCD) - Jan 2011. Based on your selection, MetaboHunter use the information from one of the two databases to provide a list of potential metabolite matches.

**Type of metabolite:** When available, the metabolite type can be used as a screening criterion. The metabolite data originated from HMDB contains metabolites that can fall in one of the following classes: drug, food additive, microbial, plant, synthetic/industrial chemical.

**Sample pH:** When available, the sample pH can be used as a screening criterion. The metabolite data originated from HMDB and MMCD samples was measured on a broad range of pH values, from 2 to 10.

**Solvent:** When available, the solvent can be used as a screening criterion. The following solvents have been used to measure the metabolite spectra in HMDB and MMCD: water, CDCl<sub>3</sub>, CD<sub>3</sub>OD and 5% DMSO

**NMR frequency:** When available, the NMR frequency can be used as a screening criterion. The metabolite data originated from HMDB and MMCD was measured using the following NMR frequencies: 400 MHz, 500 MHz and 600 MHz.

**Matching method:** This allows the user to select their preferred method for metabolite identification when a query spectrum or list of peaks is provided. The available methods are:

MH1: Highest number of matched peaks,

MH2: Highest number of matched peaks with shift tolerance,

MH3: Greedy selections of metabolites with disjoint peaks and shift tolerance.

**Noise threshold:** The noise threshold must be a number above which all peaks in the input file considered as potential matches for metabolite spectra.

**Confidence threshold (percentage):** The confidence threshold must be a number greater than 0 that represents the cut-off score or percentage for matched spectra. The scoring function for all metabolite identification methods integrated in MetaboHunter is the ration between the number of matching peaks and one plus the total number of peaks for any given metabolite.

**Shift tolerance:** The shift tolerance must be a real number greater or equal to zero that represents the amount of variability (measured in ppm) allowed for each metabolite peak in the sample to match a metabolite peak in the database.

The lists of curated peaks used by the available python scripts can be downloaded here: [HMDB peaks](#), [MMCD peaks](#).

## News

Updates of the CCPM portal can be found here.

## Events

Information about the workshops can be found here.

## People

The past and the current team members, Project PIs and DBT experts list can be found here.

## Documentation

Under **Documentation** the user can find the **User manual**, **PI enrolment form** and **Security/Privacy Policy** of the portal.

## Forum

Anything and everything related to this portal can be discussed in this forum. The issues are categorized under question, feature request, bug, testimonial, and complaint& other. If your request is already posted below, then vote for it instead of posting again by using Top Ten Issues. Your issue has a better chance of getting resolved if many people vote for it.

---