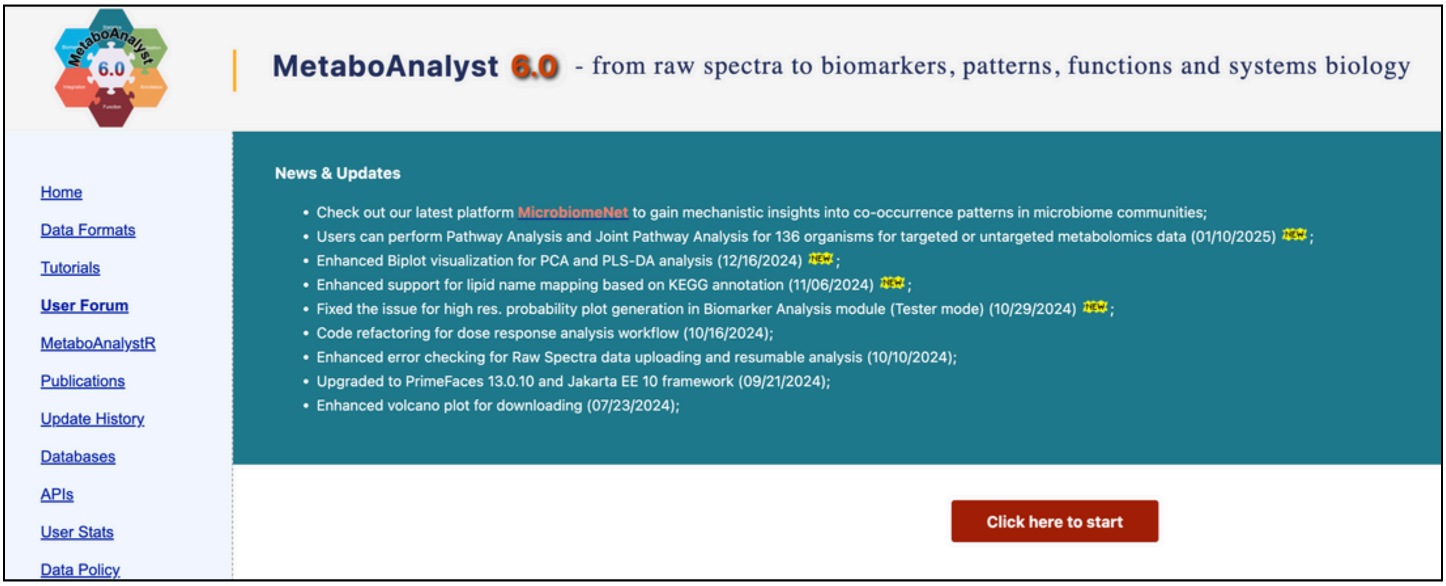
Data analysis of mass spectrometry-based metabolomics data using MetaboAnalyst

28632 LM Regulat Sci & Toxic 21 Cent
Dr. Ralf Weber, Ossama Edbali

In this assignment we will use MetaboAnalyst to analyse a metabolomics dataset (see Metabolomics_MS_SFPM.csv in Canvas page).
Before carrying out statistical analysis in MetaboAnalyst, we will first revisit some basic data visualization approaches using Microsoft Excel.

1. Load the Metabolomics_MS_SFPM.csv into Microsoft Excel .
2. Assess the structure of the data.
    a. Samples are stored in rows (how many?).
    b. Variables are in columns (how many?)
    c. Visualize the values of the rows and columns. Can you detect any interesting patterns (e.g. differences between the different classes?)

Principal component analysis



MetaboAnalyst 6.0 - from raw spectra to biomarkers, patterns, functions and systems biology

News & Updates

- Check out our latest platform MicrobiomeNet to gain mechanistic insights into co-occurrence patterns in microbiome communities;
- Users can perform Pathway Analysis and Joint Pathway Analysis for 136 organisms for targeted or untargeted metabolomics data (01/10/2025) NEW;
- Enhanced Biplot visualization for PCA and PLS-DA analysis (12/16/2024) NEW;
- Enhanced support for lipid name mapping based on KEGG annotation (11/06/2024) NEW;
- Fixed the issue for high res. probability plot generation in Biomarker Analysis module (Tester mode) (10/29/2024) NEW;
- Code refactoring for dose response analysis workflow (10/16/2024);
- Enhanced error checking for Raw Spectra data uploading and resumable analysis (10/10/2024);
- Upgraded to PrimeFaces 13.0.10 and Jakarta EE 10 framework (09/21/2024);
- Enhanced volcano plot for downloading (07/23/2024);

1. Go to www.metaboanalyst.ca
2. Click the red button ***"Click here to start".***
3. Select ***"Statistical Analysis [one factor]"***.
4. In the first panel ***"A plain text file (.txt or .csv)"*** do the following:
    a. Data type: Peak intensities.
    b. Format: Samples in rows (unpaired).
    c. Data file: Metabolomics_MS_SFPM.csv
5. Click "Submit".

6. Inspect the results of the data integrity check. Do they confirm your observations made in Excel?
7. Click on ***“Missing values”***, impute the missing values using the KNN algorithm (“feature-wise“), and click “Process”.
8. In the ***“Data filtering”*** step make all three filters as 0%. The purpose of data filtering is to identify and remove unreliable columns or variables from the data table. Click “Submit” then “Proceed”.
9. In the data ***“Normalization”*** window select ***“Normalization by sum”***, ***“Log transformation”*** and ***“Mean centering”***. Click on ***“Normalize”*** followed by ***“View Results”*** to compare the peak intensities before and after normalization.
10. Click “Proceed”.
11. Under ***“Chemometrics Analysis”*** click on ***“Principal Component Analysis (PCA)”***. How would you judge the quality and biological meaning of the data.
    a. Are all the samples for each biological class clustered together?
    b. Can a separation between the biological classes be observed?
    c. Are there any outlier samples?
    d. How many components are required to capture 60% of the variance?
    e. Hint: use the overview window to determine which combinations of PCs show interesting patterns. If yes, which metabolites seem to be most related to this separation?
12. Repeat the previous analysis, with different pre-processing approaches and visualize them using PCA. To do this, click on “Data check” under “Processing” in the left sidebar.
13. Finally, we will study the influence that outliers can have on a PCA model.
14. (In Excel) Multiply column BAF (i.e., feature 320.23461) of rows 127, 128 by 1000.
15. Exit MetaboAnalyst and reprocess the data and study the PCA output. Can you explain the difference? Note that PC1 now explains more variance compared to analysis of the original data. This shows that more explained variance does not equal a better model!

Univariate data analysis

1. Exit MetaboAnalyst and Repeat steps 1 - 6 from the previous section using the unmodified data file.
2. Click “Proceed“ and do not click “Missing Values”.
3. Skip the data filtering step (i.e. make all steps 0%) and click “Submit” followed by “Proceed”.
4. Skip the normalisation steps (i.e. all None).
5. Perform a ***“Fold change (FC) analysis”***.
   This analysis studies the ratio in average peak intensity between the two groups of samples.
   a. Which peaks to show large FC-values?
   b. ***Note***: *The FC plot displays log2(FC) rather than the FC values themselves. This is to ensure that the FC are symmetrically distributed around zero. If FC is defined as group1/group2 positive log2(FC)-values correspond to peaks where group1 has a larger average intensity, while negative log2(FC)-values indicate peaks where group2 has the largest intensity.*
6. Perform a ***“T-test”*** from the left sidebar under Statistics.
   a. Which peaks show a significant difference between the two groups of samples?
7. Download the analysis report and assess which (if any) metabolites are significant (Select ***“Download”*** from the left sidebar and select ***“Generate Report”***).
8. A significant peak intensity (between groups) does not necessarily correspond to a biologically meaningful difference. Especially for large sample sizes very minute differences in peak intensity may be marked as significant. Therefore, the p-values of the statistical test are often studied in combination with the fold-changes in a **“Volcano plot”.** Study the volcano plot for further interpretation of the results.