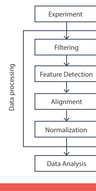
Creative Proteomics

# BIOINFOMATIC ANALYSIS FOR METABOLOMICS

## Data Processing and Normalization

The basics of data processing is to convert the original data file into a representation to help easily access the characteristics of each observed ion. These characteristics include ion retention time and m/z time, as well as ion intensity measurements in each raw data file. In addition to these basic features, data processing can also extract other information, such as the isotope distribution of ions.

Common Data Processing Pipeline




## Univariate Analysis

Metabolomic data are usually multi-dimensional, with the number of features (peaks, metabolites) ranging from several dozen to hundreds or even thousands. The features of acquired data represent snapshots of biochemical profiles of each organism. The majority of these features are expected to be within normal physiological range, while some may fluctuate dramatically due to the change in physiological conditions. Identifying these 'key' features is the first step to find potential biomarkers and unveil the underlying biological function.

### Fold Change Analysis

Fold change (FC) is a measure that describes the degree of quantitative change between the final value and the original value. FC can be used to analyze gene expression data in proteomics and metabolomics to measure changes under different conditions.

FC analysis can be easily understood by biologists.

The disadvantage of using the FC method is that it has a bias and may lose differentially expressed genes with a large difference (YX) but a small ratio (X / Y), resulting in high deletion under high intensity rate.

### T-test

T-test can be used to determine whether two datasets are significantly different from each other.

The one-sample t-test is used to test whether the difference between a sample average and a known overall average is significant.

Two-sample t-test is used to test whether the difference between the average of two samples and the population represented by each is significant. Paired-sample t test measures the difference of the data obtained by two groups of subjects that are matched or the data obtained by the same group of subjects under different conditions. The purpose is to eliminate the influence of confounding factors.

### Analysis of Variance

Analysis of variance (ANOVA) is a collection of statistical models widely used to analyze the variation of the individual value from the mean value of the group, such as "variation" among and between groups. The observed variance in a particular variable is partitioned into components attributable to different sources of variation.

ANOVAs are very useful for comparing three or more groups (or variables) for statistical significance. It is conceptually similar to multiple two-sample t-tests, but is more conservative that results in less type I error, and is therefore suited to a wide range of practical problems.

### Correlation Analysis

Correlation analysis is a simple and useful univariate method to test whether two variables are related.

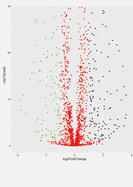
We can perform: 1. Identification of features similar to a known biomarker; 2. Identification of features following a particular pattern.

Supported similarity measures include: Euclidean distance, Pearson's correlation, Spearman's rank correlation, and Kendall's t-test.

### Volcano Plot

The volcano chart is a scatter chart used to quickly find changes in a large data set composed of complex data.

Volcano plots display both noise-level-standardized and unstandardized signals concerning differential expression of mRNA levels. Regularized test statistic and joint filtering have an intuitive geometric interpretation in a volcano plot, and its advantage over the double filter criterion of genes can be easily understood. As a scattering plot, the volcano plot can incorporate other external information, such as gene annotation, to aid the hypothesis generating process concerning a disease or phenotype.

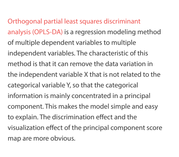



## Multivariate Analysis

Metabolomic data are usually composed of dozens of features (peaks, compounds, etc.). Many features change as a function of time, phenotype or different experimental conditions. Multivariate data analysis is desired for analyzing metabolomic data. MVA includes a lot of techniques, such as PCA, multivariate ANOVA, multivariate regression analysis, factor analysis and discriminant analysis.

### Principal Component Analysis

Principal component analysis (PCA) is a broadly used statistical method that uses an orthogonal transformation to convert a set of observations of conceivably correlated variables into a set of values of linearly uncorrelated variables called principal components. This is an unsupervised statistical analysis approach that is probably the most widely used statistical tool in metabolomics studies. PCA is mostly used as a tool in exploratory data analysis and for making predictive models.




### PLS-DA/OPLS-DA

Partial least squares discriminant analysis (PLS-DA) is a supervised multivariate statistical analysis method. It combines the regression model between metabolite changes and experimental grouping while reducing dimensionality, and uses a certain discriminant threshold to discriminant analysis of the regression results. Compared with PCA, PLS-DA analysis can further show the differences between groups.

Orthogonal partial least squares discriminant analysis (OPLS-DA) is a regression modeling method of multiple dependent variables to multiple independent variables. The characteristic of this method is that it can remove the data variation in the independent variable X that is not related to the categorical variable Y, so that the categorical information is mainly concentrated in a principal component. This makes the model simple and easy to explain. The discrimination effect and the visualization effect of the principal component score map are more obvious.

#### Comparison

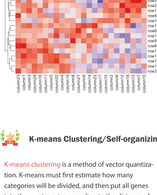
OPLS-DA can filter changes that are independent of experimental conditions. Therefore, OPLS-DA can better reflect sample differences related to experimental conditions than PLS-DA and can make the separation of samples between groups better.

Generally, PLS-DA is often used to compare two or more groups, while OPLS-DA is usually used to compare two groups. In addition, OPLS-DA is more accurate than PLS-DA in screening differential metabolites. The VIP value generated by OPLS-DA is generally used to screen differential metabolites.

## Clustering Analysis

### Dendrogram Analysis

A dendrogram is a tree diagram widely used to illustrate the arrangement of the clusters produced by hierarchical clustering. The hierarchical clustering algorithms begin with each object in individual clusters. At every step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects cannot be separated.



### Heatmap Analysis

A heatmap is a graphical representation of statistical data where the individual values contained in a matrix are represented by colors. Heatmap is suitable for displaying the differences between multiple variables, showing whether there are variables that are similar to each other, and detecting whether there is any correlation between each other.

### K-means Clustering/Self-organizing Map

K-means clustering is a method of vector quantization. K-means must first estimate how many categories will be divided, and then put all genes into these categories according to the distance of similarity. K-means calculation is much smaller and more efficient than hierarchical clustering.

Self-organizing feature map (SOM) is a data matrix and visualization method based on neural network. Each object in the data set is processed one at a time. The nearest center point is determined and updated.

#### Comparison

Unlike K-means, there is a topological order between the center points of the SOM. While updating a center point, the neighboring center points will also be updated until the set threshold is reached or the center point no longer changes significantly. Finally, a series of center points are obtained which implicitly define multiple clusters, and the objects closest to this center point are classified into the same cluster.

SOM emphasizes the proximity relationship between the center points of clusters, and the correlation between adjacent clusters is stronger. SOM is often used to visualize network data or gene expression data.

**Other BioInformatics Analysis We Offer:**

- Classification and Feature Selection
- Enrichment Analysis
- Pathway Analysis
- Biomarker Analysis

Email: info@creative-proteomics.com
Tel: 1-631-275-3058


Creative Proteomics