LipidOne 2.1 functions user guide

This document outlines the functionalities implemented in the new version, LipidOne 2.1. In this release, the features have been organized into four categories: Monovariate Statistical Analysis, Multivariate Statistical Analysis, Clustering Analysis, and Lipid System Biology. This guide adheres to the same organizational structure. For a comprehensive description of the entire LipidOne project, please refer to the following scientific publications:

- Husam B.R. Alabed, Dorotea Frongia Mancini, Sandra Buratta, et al. LipidOne 2.0: Unveiling Hidden Biological Insights in Lipidomic Data with a new web bioinformatics tool. (PREPRINT) Authorea. March 25, 2024. DOI: https://doi.org/10.22541/au.171135699.95749738/v1
- Roberto Maria Pellegrino, Matteo Giulietti, Husam B R Alabed, Sandra Buratta, Lorena Urbanelli, Francesco Piva, Carla Emiliani, LipidOne: user-friendly lipidomic data analysis tool for a deeper interpretation in a systems biology scenario, Bioinformatics, Volume 38, Issue 6, March 2022, Pages 1767–1769, https://doi.org/10.1093/bioinformatics/btab867

The following publications provide examples of applications of our tool in the field of lipidomics:

- Alabed HBR, Del Grosso A, Bellani V, Urbanelli L, Carpi S, De Sarlo M, Bertocci L, Colagiorgio L, Buratta S, Scaccini L, Frongia Mancini D, Tonazzini I, Cecchini M, Emiliani C, Pellegrino RM. Untargeted Lipidomic Approach for Studying Different Nervous System Tissues of the Murine Model of Krabbe Disease. Biomolecules. 2023 Oct 23;13(10):1562. doi: 10.3390/biom13101562. PMID: 37892244; PMCID: PMC10605133.
- Alabed HBR, Pellegrino RM, Buratta S, Lema Fernandez AG, La Starza R, Urbanelli L, Mecucci C, Emiliani C, Gorello P. Metabolic Profiling as an Approach to Differentiate T-Cell Acute Lymphoblastic Leukemia Cell Lines Belonging to the Same Genetic Subgroup. Int J Mol Sci. 2024 Mar 31;25(7):3921. doi: 10.3390/ijms25073921. PMID: 38612731; PMCID: PMC11011837.

Content

Monovariate statistical Analysis

t.test/ANOVA

Univariate statistical analysis in LipidOne is based on t-test or ANOVA. It is used to identify lipids whose averages across sample groups show statistically significant differences. This technique helps determine which lipids may be potential biomarkers of interest by highlighting statistically significant variations in the lipidomic dataset. The user can select the level of significance, choosing between p-value or False Discovery Rate (FDR), to control the false positive rate.

How it works

The user selects the significance level (p-value or FDR*) to determine which lipids are statistically significant. The statistical test (t-test: used to compare the mean of two groups or ANOVA: used to compare the mean of three or more groups) is applied to each lipid to calculate the p-value. Next, the -log10 of the p-value is plotted on the y-axis to facilitate visualization of the results.

**) False Discovery Rate (FDR): This is a correction methodology for multiple tests that controls the proportion of false positives (incorrect findings) among all significant findings. Using FDR is particularly important in omics analysis where many tests are performed simultaneously, thus reducing the probability of obtaining falsely significant results*.

The user can set the significance threshold to highlight lipids that show statistically significant differences between sample groups.

This process helps focus attention on the lipids of greatest biological or clinical interest, facilitating further analysis and interpretation.

As the graph reads.

- The graph of the univariate statistical analysis represents each lipid as a point, with the y-axis showing the -log10 of the p-value obtained from the statistical test (t-test or ANOVA).
- **Red Dashed Line**: Indicates the boundary of significance chosen by the user. Points above this line are considered significant.
- **Blue and Red Dots**: Significant lipids are shown in blue, while nonsignificant lipids are shown in red. The labels of the significant dots show the names of the lipids.

Bar Chart

In LipidOne, the Bar Chart is a visual tool used to compare averages of lipid concentrations between different groups of samples. Depending on the contexts, many versions of Bar Chart are available in LipidOne but they are always grouped bars. It is particularly useful for visualizing statistically significant differences between groups being compared. Because of its immediacy, this type of chart is ideal for highlighting specific patterns and clearly communicating differences in the data.

How it works

The Bar Chart groups the averages of the concentrations of each lipid for different groups of samples. Each bar represents the mean of a lipid for a specific group; the error bars show the experimental error associated with the mean (also called the standard error of the mean). Asterisks above the bars indicate the statistical significance of the differences between the groups, based on statistical tests (t-test or ANOVA). The colors of the bars represent the different groups being compared, making it easier to visually identify the differences between the groups.

How to interpret the result

The Bar Chart allows a quick assessment of differences in lipid levels between groups. Here is how to interpret the main elements of the chart:

- **Bars and Colors:** Each color represents a specific group. The heights of the bars show the average concentration of each lipid for that group.
- **Error Bars:** The error bars above each bar represent the experimental error, indicating the variability of the data.
- **Asterisks:** The asterisks above the bars indicate the statistical significance of the differences between the groups:
    - Absence of asterisks: differences not statistically significant
    - * : Significant difference with p-value < 0.05
    - ** : Highly significant difference with p-value < 0.01
    - *** : Highly significant difference with p-value < 0.001

The units on the y-axis represent the unit of measurement set by the user when loading data.

Biomarker Discovery

The "Biomarker Discovery" table is an essential tool for identifying lipids that can serve as candidate biomarkers. These biomarkers are selected based on statistical and performance criteria, providing a list of lipids that show significant differences between sample groups and have potential biological or clinical relevance. The goal is to help researchers focus on lipids of interest for further study or diagnostic applications.

How it works

The table lists the top 20 candidate lipids as biomarkers, evaluated according to different metrics:

- **Variable:** Name of the lipid.
- **p_value:** The p-value obtained from t.test, indicating the significance of the difference between the groups. Lower p.values indicate higher significance.
- **AUC (Area Under the Curve):** Value of the area under the ROC (Receiver Operating Characteristic) curve, which measures the ability of the lipid to distinguish between the two groups being compared. An AUC of 1 indicates perfect classification ability.
- **Cohen's d** is a measure of effect size used in statistics to quantify the difference between two groups. It is expressed in terms of "standard deviations" and is calculated as the difference between the averages of the two groups, divided by the combined standard deviation of the groups. This measure provides insight into how much one mean differs from the other in terms of standard deviations**.** Cohen's d quantifies the magnitude of the effect in terms that are independent of the unit of measurement of the original data, making it a useful indicator for comparing results across different studies. For example, if Cohen's d is equal to 2, it means that the two group averages are "2 standard deviations" away from each other.

  Interpretation of Cohen's d:

  - d = 0.2: Small effect.
  - d = 0.5: Average effect.
  - d = 0.8: Large effect.
  - d > 2: Indicates a very large effect size, suggesting a substantial difference between the groups.
- Power **(%): The** power of the statistical test, indicating the probability of detecting a true difference between the groups assuming a significance level $p < 0.001$. Higher power values indicate higher test reliability.

How to interpret the result

The table provides a summary of the main statistical characteristics of candidate lipid biomarkers:

- **Variable:** Identifies the specific lipid.
- **p_value:** Lipids with very low p-values (e.g., < 0.001) are considered highly significant. This means that the observed differences between the groups are not due to chance.
- **AUC:** An AUC close to 1 indicates that the lipid has excellent ability to discriminate between groups. An AUC of 0.5 indicates no discriminating ability.
- **Cohen's d:** Values above 2 suggest that the lipid has a substantial difference between groups, making it a good candidate as a biomarker.
- **Power (%):** Indicates the reliability of the statistical test. A value close to 100% is ideal, indicating a low probability of type II (false negative) error.
-

Volcano Plot

The Volcano Plot in LipidOne is a visual tool used to identify lipids that show significant changes between two experimental conditions. It is particularly useful for combining statistical significance (p-value) with magnitude of change (fold change*) in a single graphical representation. This type of graph is ideal for highlighting which lipids are significantly regulated differently between two groups, helping to identify potential biomarkers or targets of interest.

**) Fold Change: In the context of a volcano plot, fold change represents the ratio of change in expression levels between two conditions. On the x-axis, it is plotted as the log2 fold change, which allows for easy visualization of both increases and decreases in expression on a symmetrical scale.*

How it works

The Volcano Plot represents each lipid as a dot on a two-dimensional graph. The x-axis shows the fold change (relative change between two conditions), while the y-axis shows the -log10 of the p-value obtained from a statistical test (e.g., t-test or ANOVA). Users can set a significance level that affects the coloring of lipids in the graph:

- **Gray Dots:** Insignificant lipids (non-GIS) that do not exceed the significance threshold set by the user.
- **Red dots:** Lipids overexpressed (UP) in the experimental condition compared with the control.
- **Blue dots:** Lipids underexpressed (DOWN) in the experimental condition compared with the control.

The horizontal dashed line represents the significance threshold of the user's chosen p-value, while the vertical lines (if present) may indicate relevant fold change thresholds.

3) How to interpret the result

The Volcano Plot allows rapid identification of lipids that show significant changes between two conditions. Here is how to interpret the main elements of the plot:

- **Dots and colors:**
    - **Red (UP):** Lipids with positive fold change and significant p-value, indicating overexpression in the experimental condition.
    - **Blue (DOWN):** Lipids with negative fold change and significant p-value, indicating underexpression in the experimental condition.
    - **Gray (non-GIS):** Lipids with nonsignificant p-value, independent of fold change.
- **Axes:**
    - **x-axis (Fold Change):** Measures the amount of change in each lipid between the two conditions. Positive values indicate an increase, while negative values indicate a decrease.
    - **Y-axis (-log10 p-value):** Measures statistical significance. Higher values indicate greater significance.

Pie Chart

The LipidOne "Pie Chart" is a visual tool used to represent the percentage composition of lipid categories and classes within a selected experimental group. This type of chart makes it possible to quickly visualize the distribution and relative abundance of different lipid categories and classes, helping to identify which lipid groups are predominant in the sample analyzed.

How it works

The graph is a combination of two concentric doughnut graphs:

- **Inner Part:** Represents major lipid categories, such as GPL (Glycerophospholipids), SP (Sphingolipids), GL (Glycerolipids), etc. Each section shows the percentage of the lipid category to the total.
- **Outer Part:** Represents the specific lipid classes within the categories. Each section shows the percentage of the lipid class to the total. The outer sections are segments of the inner sections, making the detailed distribution within each category visible.

The numerical values indicate the percentage of each lipid category and class present in the selected experimental group.

How to interpret the result

The "Pie Chart" allows a clear visualization of the relative proportions of lipid categories and classes. Here is how to interpret the main elements of the chart:

- **Lipid Categories (Inner Part):** Each section of the inner donut represents a lipid category, and its size reflects its percentage proportion to the total.
- **Lipid Classes (Outer Part):** Each section of the outer donut represents a lipid class within a category, and its size reflects its percentage proportion to the total.

The percentages make it easy to understand the lipid composition of the sample. This type of visualization is useful for quickly identifying dominant lipid categories and classes, as well as highlighting differences in lipid composition between different experimental groups.

Summary Table

In LipidOne, the function "Summary Table" is used to provide a comparative overview of lipid classes among several user-selected experimental groups. This table summarizes the averages of the amounts of each lipid class, along with the statistical significance of the differences between the groups. It is a useful tool for identifying significant variations in lipid class levels between experimental conditions, helping to focus attention on lipid classes that might be of biological or clinical interest.

How it works

The table presents the following information for each lipid class:

- **LipidClass:** Name of the lipid class.
- **Count:** Number of lipids belonging to that class.
- **HbS:** Mean and standard error of the amount of the lipid class in the HbS group (or another user-selected experimental group).
- **WT:** Mean and standard error of the amount of lipid class in the WT group (or another user-selected control group).
- **p-value:** P-value obtained from the statistical test comparing the means between the two groups. It indicates the significance of the difference.
- **Sign:** Indicates significance of differences with symbols (* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$).

How to interpret the result

The table provides a detailed comparison of the average amounts of lipid classes among the experimental groups. Here is how to interpret the main elements of the table:

- **LipidClass and Count:** Identify the lipid class and the number of lipids in it.
- **HbS and WT:** Show the averages and standard errors of the lipid amounts in the two groups. These values indicate how much each lipid class averaged in the experimental groups.
- **p-value:** A low p-value (typically less than 0.05) indicates that the difference between the mean lipid amounts in the two groups is statistically significant. This suggests that the observed variation is not due to chance.
- **Sign:** Significance symbols (*, **, ***) provide a quick indication of the statistical significance of differences. More asterisks indicate greater significance.

Lipid Categories

The "Lipid Categories" function is used in LipidOne to visualize the distribution of lipid categories among different experimental groups. It is presented in two versions: absolute concentration and normalized (percent) concentration. These stacked bar graphs provide a clear representation of the relative and absolute amounts of different lipid categories, facilitating comparison between two or more experimental groups.

How it works

The function generates two types of stacked bar graphs:

- **Absolute Concentration:** Shows the total amount of each lipid category (in ng/ml) in the experimental groups. The bars represent the sum of lipid amounts for each category within each group.
- **Normalized Concentration (Percent):** Shows the percentage proportion of each lipid category to the total lipids in each group. The bars represent the percentage of each lipid category to the total, facilitating comparison of relative compositions between groups.

Each lipid category is represented by a different color, allowing easy visual identification of differences.

How to interpret the result

Here is how to interpret the main elements of the graphs:

- **Absolute Concentration:**

This graph shows the total amount of each lipid category in the compared experimental groups. The y-axis represents the concentration expressed in the unit of measurement indicated by the user, while the x-axis represents the experimental groups. The stacked bars indicate the absolute amounts of the lipid categories, allowing you to see which categories are most abundant in each group.

- **Normalized Concentration (Percentage):**

This graph shows the percentage distribution of each lipid category in the experimental groups. The y-axis represents the percentage, while the x-axis represents the experimental groups. The stacked bars indicate the proportion of each lipid category to the total lipids, facilitating comparison of relative compositions among the groups.

Multivariate Statistical Analysis

Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised statistical technique used to reduce data complexity while retaining as much original information as possible. In the field of lipidomics, PCA helps identify patterns and visualize differences between samples based on lipid profiles. It can reveal similar sample groups, general trends and major variations, facilitating the interpretation of complex data.

How PCA works

PCA works by transforming the original data into new variables called principal components. These components are linear combinations of the original variables (lipids) and are ordered by importance: the first component captures the most variability in the data, the second component captures the second most variability, and so on. This transformation process reduces the dimensionality of the data, allowing key differences and similarities between samples to be better visualized and interpreted. The LipidOne user can obtain graphs for the first two components or pairwise test up to five principal components.

Before applying PCA, the data are automatically normalized between samples, and the user can decide whether to apply autoscaling or pareto scaling. With Autoscaling, each variable (lipid) is centered (subtracted from the mean) and scaled (divided by the standard deviation). This makes all variables comparable in terms of units of variation. Pareto Scaling is a technique similar to autoscaling, but the variables are divided by the square root of the standard deviation. This method reduces the influence of variables with large variance while still maintaining some original structure of the data.

How to read PCA graphs

PCA graphs include the Score plot, Loading plot and Scree plot.

- **Score Plot**: Shows the samples in a two-dimensional space defined by the first two principal components. Each point represents a sample, and their relative position indicates similarity or difference in lipid profiles. Nearby samples have similar profiles, while distant samples have larger differences. In lipidOne, if there are more than 4 samples, 95% confidence ellipsoids are drawn around the groups of samples to indicate variability and separation of groups with some level of confidence. If the ellipsoids are clearly separated, it means that the lipid profile is clearly different between the groups, on the contrary if the overlap between the ellipsoids is partial or total, the differences between groups of samples are minor or zero.
- **Loading Plot**: Shows the variables (lipids) that contribute to the principal components. Indicates which lipids most influence the separation observed in the score plot. Variables with high loadings on a principal component are those that explain the most variability in the data in that direction. The more extreme the loadings in the direction where the scores are most distant, the more these variables differ and explain the variance in the data.
- **Scree Plot**: A bar plot that shows the percentage contribution of each principal component to the total variability of the data. In LipidOne, the contribution of the first five principal components is displayed. This scree plot helps determine how many principal components are needed to explain a significant portion of the variability in the data.

Partial Least Squares Discriminant Analysis

Partial Least Squares Discriminant Analysis (PLS-DA) is a supervised statistical technique used to classify samples into predefined groups based on predictor variables. In the context of lipidomics, PLS-DA helps identify which lipids discriminate best among different groups of samples. Not only does this technique reduce data complexity, but it also improves the interpretability of differences between groups, allowing potential biomarkers to be identified and biological variation to be better understood.

How it works

PLS-DA combines aspects of PCA with linear regression, finding new components that maximize the separation between predefined groups. Each new component is a linear combination of the original variables (lipids) and is chosen to maximize covariance between variables and sample classes. This process reduces the dimensionality of the data, allowing better visualization and interpretation of differences between groups.

Similar to PCA, also for PLS-DA the user can select a pair from the first five components by scaling the data with no method, autoscaling and Pareto scaling.

How to read the graphs

The graphs produced by LipidOne for PLS-DA include score plot, loading plot, Cross Validation, VIP and Confusion Matrix.

Scree plot and **Loading plot are** interpreted as explained for PCA.

- **Cross Validation:** This graph shows the performance of the model as a function of the number of components. The variables shown are Accuracy (model accuracy), R2Y (explanation of variability in response data), and Q2Y (prediction of variability in response data). In general, a model with high values for all three metrics can be considered a good predictive model.
- Variable Importance Graph (**VIP**): Shows the importance of each lipid (variable) in discriminating between groups. The 20 most important variables are those with higher values, as indicated by the graph.
- **Confusion Matrix**: This graph shows the performance of the classification model, comparing the predicted classes with the actual classes. Each cell shows the number of samples classified correctly or incorrectly.

Correlation Analysis

Correlation analysis calculates the correlation coefficient between each pair of variables, indicating how much variation in one variable is associated with variation in another. It is a technique used to assess the strength and direction of the relationships among two or more variables. This tool is useful for understanding how lipids interact with each other and how samples cluster based on their lipid profiles, helping to identify patterns of co-variation and significant associations.

How it works

LipidOne users can generate a correlation matrix by choosing to correlate both variables (Lipid Classes, Lipid Molecular Species, or Lipid Building Blocks) and samples. Correlation coefficients vary between -1 and 1, where:

- **1:** Indicates perfect positive correlation.
- **0:** Indicates no correlation.
- **-1**: Indicates perfect negative correlation.

The resulting matrices are visually represented via heatmaps, where colors indicate the strength and direction of correlations.

How to interpret the result

The result of the correlation analysis is represented by a heatmap showing the correlation matrix. The following two examples illustrate the interpretation of the graphs:

- **Sample Correlation Matrix:**

This graph shows the correlation between different samples. Each cell represents the correlation coefficient between two samples. Darker colors indicate stronger (positive) correlations, while lighter colors indicate weaker correlations.

- **Correlation Matrix of Lipid Classes:**

This graph shows the correlation between different lipid classes. Each cell represents the correlation coefficient between two lipid classes. The colours range from blue (strong positive correlation) to red (strong negative correlation).

Using these graphs, LipidOne users can easily identify significant correlation patterns among lipids or between samples, helping to interpret lipid interactions and similarities between lipid profiles of samples.

Clustering Analysis

Heat Map

The Heat Map is a powerful visual tool used to represent and interpret large complex datasets, such as lipidomic data. In LipidOne, a Heat Map allows users to visualize the relative abundance of lipids in different samples, identifying patterns of variation and potential biomarkers. Users can select between 2 and 200 lipids to include in the graph, sorted by statistical significance (by t-test or ANOVA). This representation facilitates analysis of differences between sample groups and understanding of relationships between variables. Additionally, the Heat Map includes hierarchical clustering with dendrograms for both samples and variables (lipids), enhancing the ability to identify clusters and relationships within the data.

How it works

The Heat Map combines data visualization with clustering techniques to highlight significant patterns. Before constructing the heat map, either a t-test or ANOVA (depending on the number of groups included in the analysis) is performed to assess the statistical significance of the lipids. The lipids are then organized in descending order of their p-values, and the top n lipids (a number chosen by the user, between 2 and 200) are selected. With this selection, the heat map is constructed, including hierarchical clustering of both samples and lipids:

- **Correlation**: Measures similarity based on correlation between lipid profiles.
- **Canberra**: A distance measure that is more sensitive to small values, making it useful for data with a wide range of values.
- **Minkowski**: A generalization of Euclidean and Manhattan distances.
- **Manhattan**: Calculates the sum of the absolute differences between the coordinates of the points.
- **Euclidean**: The "linear" or "direct" distance between two points in a multidimensional space.
- **Binary**: Uses binary variables to calculate distance, suitable for data with dichotomous values.

Users select the most appropriate distance algorithm for their data, and the Heat Map visualizes how the samples and lipids cluster based on the chosen distance measure. This process allows flexibility in analysis and detailed visualization of patterns of variation in lipidomic data.

How to read the graphs

The Heat Map presents the data in a grid where rows represent lipids and columns represent samples. The values are coded in red and blue, with different intensities indicating the relative abundance of the lipids.

Clustering and **Dendrograms**: Both lipids and samples are clustered, and the resulting dendrograms show the similarity relationships between them. This helps to identify groups of lipids with similar patterns of variation between samples and groups of samples with similar lipid profiles.

Group Colors: The first row of the graph shows the colors of the groups, allowing you to see at a glance whether the samples have clustered according to the predefined groups, depending on the distance algorithm used.

Dendrogram

A dendrogram is a visual tool used to represent similarity relationships between samples or variables in a dataset. LipidOne users can use dendrogram function to identify groups (clusters) of samples with similar lipid profiles. This type of graph helps visualize how samples cluster based on their lipid characteristics, facilitating the identification of meaningful patterns.

How it works

The user can select the algorithm for measuring the distance matrix by choosing from several options, including: **Correlation, Euclidean, Maximum, Manhattan, Canberra, Binary, Minkowski**

Each algorithm calculates the distances between samples in different ways, affecting the structure of the dendrogram.

In the resulting dendrogram:

- **Colored Samples:** Samples are colored according to the group they belong to (e.g., HOM or WT).
- **Colored Dendrogram** Branches**:** The dendrogram branches take on different colors depending on the clusters formed from the first nodes.

How to interpret the result

The dendrogram allows us to visualize the hierarchy of similarity relationships between samples. Here is how to interpret the main elements of the graph:

- **Vertical Distance:** The length of vertical branches represents the distance or dissimilarity between samples. Samples with a common branch near the base of the dendrogram are more similar to each other.
- **Colors of Samples:** The colors of the samples indicate which group they belong to (e.g., red for HOM and blue for WT). This allows you to quickly see which samples belong to which groups.
- **Branch Colors:** Colored branches indicate distinct clusters. For example, if samples from one group form a separate cluster with a specific color, it suggests that those samples are more similar to each other than samples from other groups.

In the case presented, the group and branch colorations suggest a clear separation between the HOM and WT groups, indicating that the samples in each group form distinct clusters. However, this result is not guaranteed in every analysis and may vary depending on the dataset and distance algorithm selected.

K-Means Clustering

K-means clustering is an unsupervised clustering technique used to group samples based on their similar characteristics. In the context of lipidomics, this technique allows the identification of groups of samples with similar lipid profiles, facilitating pattern identification and visualization of separation between groups of samples.

How it works

K-means clustering is a "partitional" clustering algorithm. It works by dividing the data into a specified number of clusters (k), which can be set by the user between 2 and 5. The algorithm begins by randomly selecting k centroids, one for each cluster. Each sample is then assigned to the nearest centroid, and the centroids are updated iteratively until convergence is achieved, minimizing the sum of the squared distances within each cluster. The K-means algorithm assigns each sample to the nearest cluster. Colored ellipses represent the distribution of samples within each cluster and are drawn only if a cluster consists of at least four points.

How to interpret the result

The result of the K-means analysis is displayed in a clustering graph. Each point in the graph represents a sample, with colors indicating the different clusters. The centroids of the clusters are represented in the graph with different geometric shapes, helping to visualize the center of each cluster. The colored ellipses show the distribution of samples within each cluster. Labels on the dots identify the samples. This makes it easy to observe how the samples are grouped and to identify any sharp separations between clusters.

Lipid System Biology

Lipid Pathway

The Lipid Pathway function of LipidOne is used to analyze lipid metabolism pathways and their alterations under different conditions. This function helps identify active pathways and changes in the lipid profile that occur due to various biological processes or treatments. This analysis can provide insights into metabolic changes and signaling pathways involving lipids, which are critical to understanding diseases, cellular responses, and the activation or deactivation of certain biochemical pathways. When uploading the dataset, users have the option of selecting a model organism from Homo sapiens, Mus musculus, Rattus norvegicus, Arabidopsis thaliana, Saccharomyces cerevisiae, Drosophila melanogaster, Danio rerio, and Caenorhabditis elegans. The results of the analysis are then relative to the selected organism.

How It Works

The method follows a statistical approach to identify significant changes in the lipid metabolism pathways of the selected model organism. For a detailed description of the applied algorithm, see the reference at the bottom of the page. The main steps are:

- **Selection of Control Group and Experimental Group**: Collection of data on lipid species concentrations from experiments comparing two different sample groups, one considered as the reference and one as experimental.
- **Calculation of Weights**: Calculation of weights for each reaction based on the concentration ratios of lipid products and reactants. These weights reflect the "shifts" in lipid metabolism of the experimental group relative to the control. The weight is the ratio of the concentrations of the product and reactant. Reactions and associated enzymes/genes are specific to each model organism selected by the user during dataset loading.
- **Scoring of Reactions**: Each reaction in the pathway is scored according to the calculated weights and their significance by performing a t-test to compare the mean scores of reactions between the two selected sample groups. The t-test p-values are then converted to Z-scores.
- **Pathway Selection**: Pathways with p-values below the threshold set by the user are considered significantly active.

How to Interpret Graphs.

The analysis produces three types of visual outputs:

- **Network**: Shows lipid classes (nodes) and their transformations (arcs). The arcs are colored and weighted according to the significance of the changes:
  - **Arc Color**: Red arcs indicate negative changes, while blue arcs indicate positive changes in lipid transformations.
  - **Arc Thickness**: Thicker arcs represent higher significance (lower p-values).
- **Gene Involvement**: Shows t-scores of significant lipid transformations, with blue bars for positive changes and red bars for negative changes. The height of the bars indicates the magnitude of change.
- **Interaction Table**: Summarizes lipid transformations with columns for "From" and "To" lipid variables, associated genes/enzymes, Z-scores, and activation status (positive or negative).

These visualizations provide a comprehensive view of alterations in lipid metabolism, helping researchers understand how specific pathways and lipid classes are affected under different experimental conditions.

References

An Nguyen, Simon A Rudge, Qifeng Zhang, Michael JO Wakelam, Using lipidomics analysis to determine signalling and metabolic changes in cells, Current Opinion in Biotechnology, Volume 43, 2017, Pages 96-103, ISSN 0958-1669, https://doi.org/10.1016/j.copbio.2016.10.003

Proteins Interaction Network

The "Proteins Interaction Network" function of LipidOne is designed to visualize and analyze the interactions between proteins in a specific dataset. In this context, the function focuses on the proteins predicted by the previous function, "Lipid Pathway." The main goal is to identify and understand the relationships and functional interactions between proteins, highlighting how these proteins work together within biological networks. This analysis utilizes the STRING database to provide comprehensive interaction data. Understanding these interactions can help discover new biological connections and potential therapeutic targets.

How it works

1. **Input**: The proteins of interest are those identified by the "Lipid Pathway" function. The user can select between two types of interactions: "full" (all possible interactions) and "physical" (only direct physical interactions between proteins).
2. **Significance threshold**: The user can set a significance threshold value, varying between 0 and 999, that filters interactions based on their reliability or statistical significance. This parameter allows the network shown to be refined to include only the most relevant interactions.
3. **Node coloring**: Proteins in the network are colored according to their activation state, as determined in the previous lipid pathway analysis. Blue nodes represent "activated" proteins, while red nodes represent "deactivated" proteins.
4. **Graph generation**: Using the STRINGdb library, the graph of protein interactions is generated and displayed. This graph shows the proteins as nodes and their interactions as lines connecting the nodes.

As the graph reads.

- **Nodes (Proteins)**: Each node in the graph represents a protein. Proteins colored blue are those that are activated, while those in red are deactivated.
- **Lines (Interactions)**: Lines between nodes represent interactions between proteins. The density and number of lines can indicate the complexity of the network and the centrality of the proteins. Interactions are based on STRING data and can be physical or functional depending on the settings selected. The colors of the interactions have the following explanation:
    - Red: evidence of melting.
    - Green: neighborhood evidence.
    - Blue: evidence of co-occurrence.
    - Purple: experimental evidence.
    - Yellow: evidence of text mining.
    - Light blue: evidence from database.
    - Black: evidence of co-expression.

This type of analysis is crucial to better understand how proteins interact with each other and what the functional consequences of these interactions may be, providing a deeper insight into biological processes and possible therapeutic implications.

Neighboring Protein Interaction Network

Within a biological context, each protein exists within a neighboring interaction milieu with other proteins (clusters), such as those comprising a metabolic pathway. In lipidOne, the 'Neighboring Protein Interaction Network' feature enables the visualization of a network of protein interactions associated with a protein dataset previously generated by the Lipid Pathway function. This functionality leverages the STRINGdb 12.0 library.

This analysis facilitates the exploration of the intricate network of interactions among predicted proteins and their neighboring counterparts, governed by specific scoring thresholds and interaction types.

This capability is particularly valuable for elucidating the functional and structural interrelations among proteins. Notably, within a protein interaction network, proteins exhibiting a higher number of connections are considered more crucial than those with fewer connections. This principle draws upon various centrality metrics employed in network theory and bioinformatics:

- **Degree Centrality**: This metric counts the number of direct connections a protein has with other proteins in the network. A protein with many connections (high degree) is considered a "**hub**" and tends to be central in the network. These hubs are crucial because they can influence many other proteins and, consequently, many biological processes.
- **Betweenness Centrality**: Measures the number of times a protein serves as a bridge along the shortest path between two other proteins. Proteins with high betweenness centrality are important for information flow in the network.
- **Closeness Centrality**: Assesses how close a protein is to all other proteins in the network. The effect of the activation/deactivation state of a protein with high closeness centrality can "reach" other proteins more directly.

Biological Significance

1. **Crucial Role in Biological Processes**: Proteins with many connections, or hubs, often play key roles in various biological processes. For example, they may be involved in regulating metabolic pathways, cell signaling, or maintaining cell structure.
2. **Network Robustness**: Hubs contribute to the robustness of the network. Their removal can cause a significant collapse in network functionality, making them potential points of vulnerability. This concept is exploited, for example, in the study of disease, where hub proteins can be therapeutic targets.
3. **Evolution and Conservation:** Hubs are often conserved across species because of their functional importance. Studies have shown that highly connected proteins tend to be more essential for the survival of the organism.

Application Example

In a network produced by STRINGdb, the protein with the most connections is likely to be central to many biological processes. Its importance can be confirmed by studying the biological roles of the proteins to which it is connected and analyzing whether its removal leads to significant perturbations in the network.

In summary, proteins with many connections in a network of protein interactions are considered most important because they tend to play central roles in various biological processes, contribute to the robustness of the network, and are often evolutionarily conserved. Their analysis can provide valuable insights into cellular mechanisms and therapeutic approaches in diseases.

How it works

1. **Data preparation**: Input data, containing previously identified proteins (hits) with the Proteins Interaction Network function, are mapped to STRING IDs.
2. **Network configuration**: The user can specify two parameters: **score_threshold** and **network_type**:

- The score_threshold is a number ranging from 0 to 999. It defines the minimum confidence score to consider an interaction as significant. Low values provide networks with more interactions, conversely high values provide less complex graphs.
- The "network_type" can be "**full**" (all interactions) or "**physical**" (only physical interactions).

3. **Network generation**: A STRINGdb object is created that handles the connection to the STRING data. Then a request is sent to the STRINGdb server to obtain information about the interactions between the mapped proteins. When finished, a network graph is generated showing the proteins as nodes and the interactions as lines.

How to interpret the graph

-Nodes (circles): represent proteins, each accompanied by its synthetic name. Input proteins (hits) are highlighted with a blue halo (activated) or red halo (deactivated). By using the DOWNLOAD button, the user can access the 'interacting_proteins.csv' table, which includes the functional annotations of each protein, along with the STRING code and the sizes in kilodaltons. The first rows of the table (marked with an asterisk) are the input proteins (hits) to the function.

-Line (edge): represent protein interactions. The color of the line indicates the type of evidence of the interaction:

- Red line - indicates the presence of fusion evidence
- Green line - neighborhood evidence
- Blue line - cooccurrence evidence
- Purple line - experimental evidence
- Yellow line - textmining evidence
- Light blue line - database evidence
- Black line - coexpression evidence.

The text above the graph indicates the number of proteins and interactions displayed, and the expected interactions based on a random model, with the associated p-value.

The graph makes it possible to quickly visualize the most central proteins in the network and their connections with other proteins, facilitating the interpretation of biological interactions and the discovery of new potential functional partners. Proteins with greater connections are more important than others.

The metrics associated with each protein represented in the network (**Degree** Centrality, **Betweenness** Centrality, and **Closeness** Centrality) are reported in the table Interacting_proteins.csv, available via the DOWNLOAD button. The same table also includes data from the STRING database regarding the model organism selected by the user: protein_external_id, preferred_name, protein_size, and annotation.

Proteins Enrichment Analysis

The "Proteins Enrichment Analysis" function is used to identify and visualize functional categories, cellular components, biological processes, tissues, diseases, and cellular compartments enriched in proteins of interest. This type of analysis helps to understand the biological functions, cellular localizations, and involvement of selected proteins in various biological processes and diseases. The proteins considered in the analysis are those selected by the previous function "Neighboring Protein Interaction Network." This function is essential for understanding the biological roles of the proteins of interest and for identifying the most relevant pathways and cellular compartments in the processes studied.

How it works

1. **Data Input**: The function starts with a set of proteins of interest, obtained from the "Neighboring Protein Interaction Network" function, and includes the proteins derived from the previous Lipid Pathway analysis.
2. **Initialization**: Uses the STRINGdb package to obtain the functional enrichment data of selected proteins.
3. **Filtering and Sorting**: Filters enrichment data by specific categories (Function, Component, Process, Compartments, Diseases, Tissues) and sorts them by p-values to highlight the most significant categories. Due to the type of organism, the set thresholds, and the availability of enrichment terms, not all six categories may always generate graphs.
4. **Saving Results**: Save filtered and sorted results in separate csv files for each category. The csv files are available for download.
5. **Graphical Visualization**: Generates scatter plots (scatter plots) that display the impact and significance of enriched categories, using point sizes and colors to represent additional metadata such as BgRatio and Impact.

How to read the graphs

Scatter plots generated by the "Proteins Enrichment Analysis" function show the following features:

X-axis (Impact): Represents the impact of each category. A higher value indicates greater involvement of the proteins of interest in that particular category.

Y-axis (Significance): Represents the statistical significance of enrichment, measured as -log10 of the p-value. A higher value indicates greater significance.

Size of Points (BgRatio): They indicate the proportion of the proteins of interest to the background. Larger dots represent a greater proportion.

Colors of Points (Impact): Colors range from yellow to red, where red indicates greater impact.

Graphs provide an immediate visual representation of the most enriched and meaningful categories. Labels on the dots identify specific categories, allowing quick interpretation of results.