COVAIN Toolbox User's Manual

Xiaoliang Sun

2012.03.15

Revised on 2014.02.12

Table of contents

General Information ---- *2*

Introduction ---- **2**

Availability ---- **2**

Requirements ---- **2**

Release notes ---- **2**

Installation and launch ---- **3**

Citation ---- **3**

Contact ---- **4**

Nomenclatures ---- **4**

Operation guide ---- *5*

Data loading ---- **5**

Data formats ---- 5

Data header names ---- 5

Data contents ---- 7

Data preview ---- 7

Default options for data preprocessing and data analysis ---- **7**

Data preprocessing ---- **8**

Missing value imputation ---- 8

Outlier adjustment ---- 8

Sample normalization ---- 8

Data transformation ---- 9

Data filtration and visualization ---- 10

Evaluation ---- 10

ANOVA ---- 11

Error bar plot ---- 13

Data combination ---- 14

Data analysis ---- **15**

Multivariate statistics ---- 15

Conditions selection ---- 15

PCA and ICA ---- 15

Correlation ---- 16

Bi-clustering ---- 16

Time series analysis ---- 17

Time order selection ---- 17

Correlation ---- 17

Clustering---------- 17
Granger causation analysis---------- 17
Permutation entropy ---------- 18
Network analysis---------- 19
Network inference ---------- 19
Network visualization ---------- 19
Network properties---------- 20
KEGG mapping ---------- 20
Inverse Jacobian analysis ---------- 21
mzGroupAnalyzer and Pathway Viewer---------- 22
Results saving---------- 25

General Information

Introduction

The COVAIN Toolbox is statistical data analysis software that runs under Matlab enviroment. Though COVAIN was initially designed for metabolomics data processing, it can analyze other types of omics data, such as proteomics and transcriptomics. COVAIN gets its name from one of its functionalities: *COVA*riance *IN*verse engineering. The design principle is to put most common data analysis methods – including preprocessing, uni- and multi-variate statistics, time-series analysis and network properties – into one software with a full graphical user interface (GUI) support, thus making data analysis more convenient for both biologists and bioinformatician.

Availability

The latest COVAIN Toolbox can be downloaded from Dept.MoSys@Uni-Vienna homepage at:
http://www.univie.ac.at/mosys/software.html

Requirements

Matlab software and its Statistics Toolbox are required. COVAIN has been well tested under Windows XP/Vista/7 with Matlab 2009a or newer; however, most functionality could work with an older version. Microsoft Excel is required for .xls and .xlsx files operation. There are, a few incompatibility problems with Mac/Linux due to Excel-like files reading and writing, which will be solved in the future. Only for mzGroupAnalyzer and Matlab itself's network visualization that the Matlab Bioinformatics Toolbox is needed.

Release notes

2014.02.12:

- Add mzGroupAnalyzer and its associated Pathway Viewer for inferring pathways between unknown m/z features from mass spectrum data.
- Add Ridge Regression method (also known as Tikhonov Regularization) in differential Jacobian inference module to alleviate ill-conditioned problems.
- Display names for each figure, thus reduces confusion between many co-existing figures.
- Display the percentage of variances in the PCA plot.
- KEGG mapping is temporally disabled and will be recovered soon.

Installation and launch

The installation has two options. One way is adding COVAIN folder permanently into the Matlab path and the other way is every time, go to the COVAIN folder in the Matlab “Current Folder”, illustrated by Figure 1.1. The second way is recommended because: 1) The COVAIN toolbox does not have automatically uninstall tool which means if a new version is added to Matlab path, it does not remove the older versions from the path; 2) Sometimes the user is reminded by Matlab not to have privileges to save the path. For the first way, the user can set default Matlab starting folder as COVAIN folder, see Figure 1.2; for the second way, if the user cannot save path, one possible method is to revise permissions of the pathdef.m in Matlab root\toolbox\local\ folder, see Figure 1.3.

Finally, to launch the COVAIN toolbox, input capitalized letters “COVAIN” in Matlab command window (see Figure 1.1) and the toolbox GUI will be opened. If inputting small letters “covain” instead, the user can still open the toolbox but get a warning.

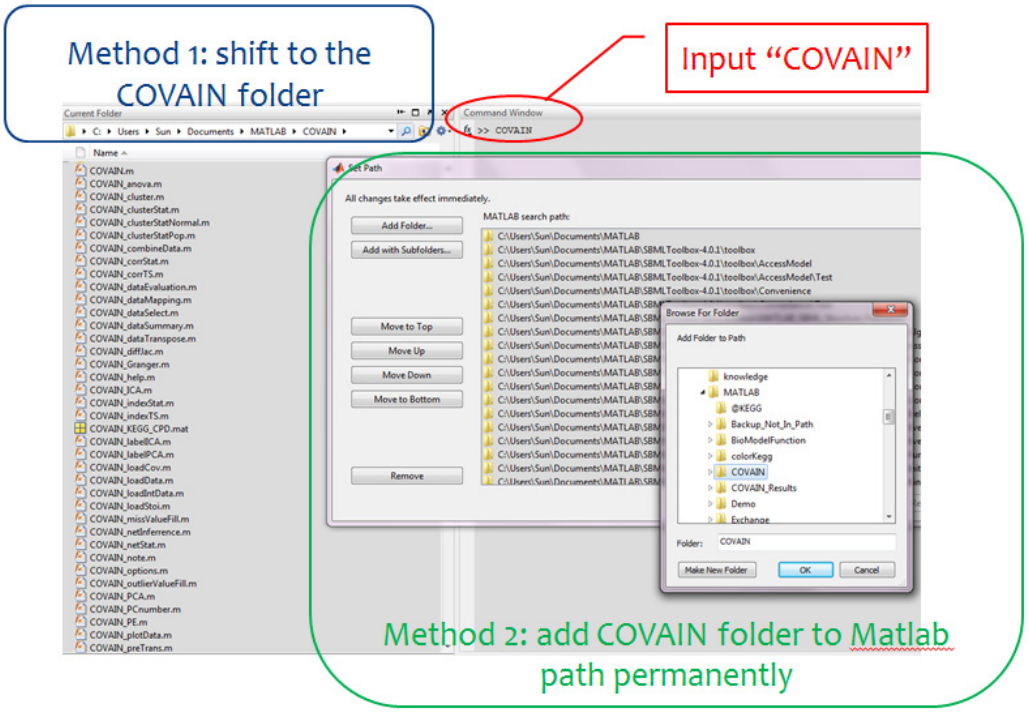


Figure 1.1: Two methods to install and launch the COVAIN toolbox

Citation

Xiaoliang Sun and Wolfram Weckwerth, *COVAIN: a toolbox for uni- and multivariate statistics, timeseries and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data*, Metabolomics (12 February 2012), pp. 1-13, doi:10.1007/s11306-012-0399-3

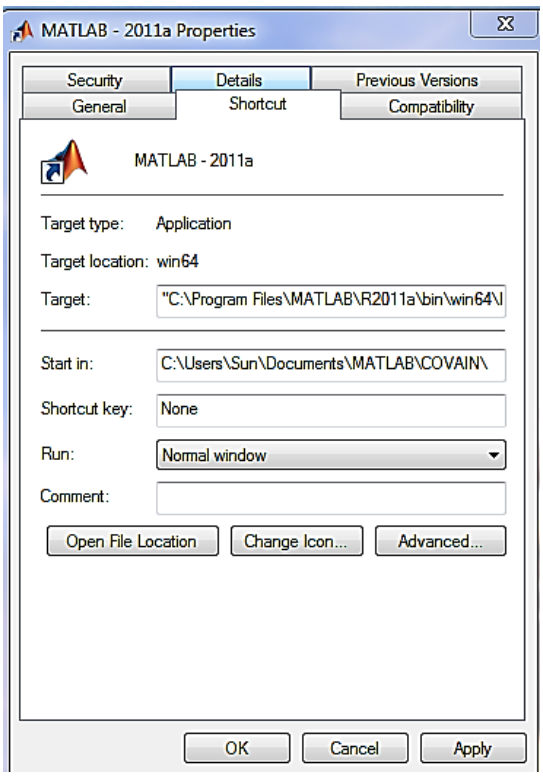
Contact

If you have problems, suggestions, or special requirements on the COVAIN toolbox, please contact us:

Dr. Xiaoliang Sun or Prof. Dr. Wolfram Weckwerth
(xiaoliang.sun, wolfram.weckwerth) @ univie.ac.at
Department of Molecular Systems Biology (MoSys),
University of Vienna,
Althanstr. 14, 1090 Vienna, Austria

Nomenclatures

“Variables” mean the measured substances such as metabolites, proteins or genes, etc;
“Conditions” mean experimental settings such as control and treatments.



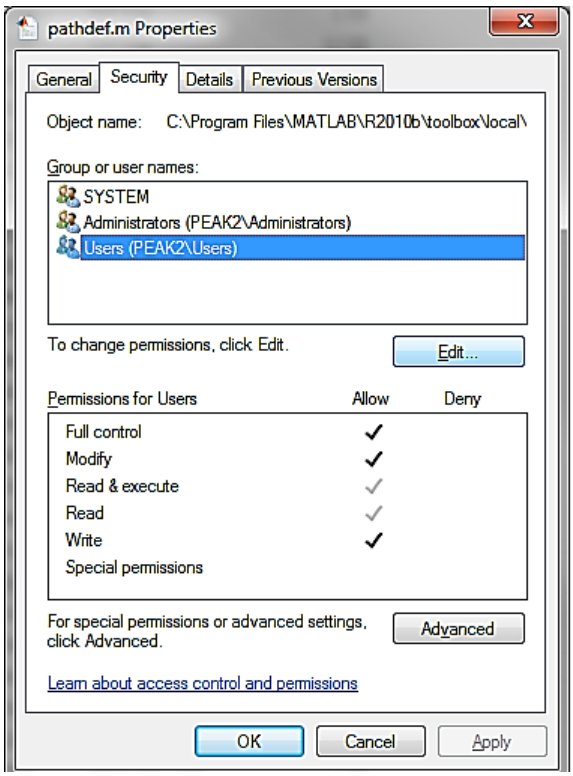


Figure 1.2 (left): Set the default starting folder. Firstly, find the shortcut of Matlab icon, then right click and find the “Start in” option. Finally input the COVAIN folder such as “C:\Users\Name\Documents\MATLAB\COVAIN\"

Figure 1.3 (right): Change the pathdef.m permission properties. The pathdef.m is located in Matlab root\toolbox\local\ folder. Add “Modify” and “Write” or “Full control”.

Operation guide

To start data analysis, user needs to load data first, because not only results but options are both associated with each dataset. Clicking on buttons without loading any data will receive errors. A workflow of COVAIN is shown in Figure 2.1.

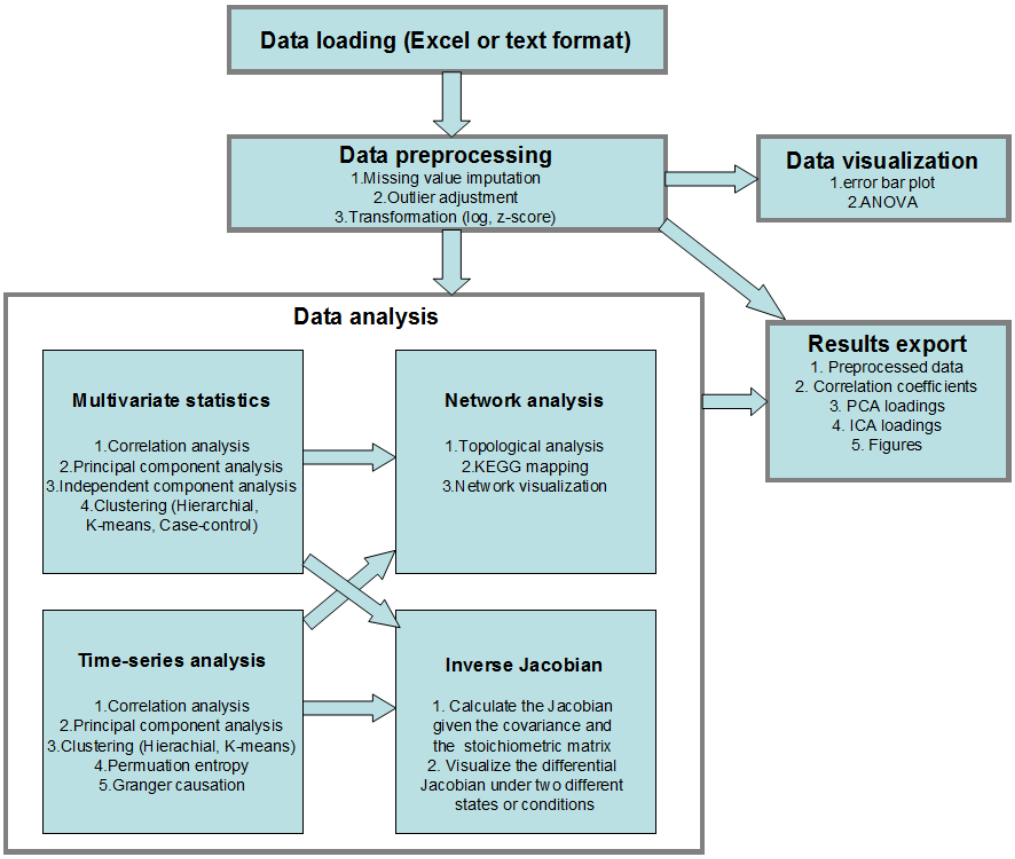


Figure 2.1: Illustration of COVAIN workflow: data loading, preprocessing to analysis, visualization and results exporting

Data loading

Data formats

The COVAIN toolbox supports loading Excel Spreadsheets (.xls or .xlsx) or tab-separated text files.

Data header names

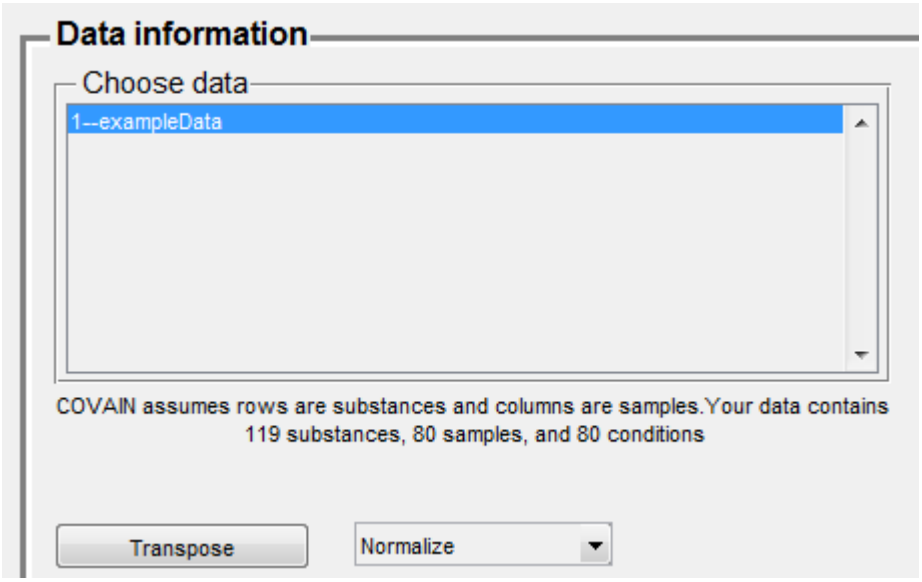
The first row header should be names of conditions or treatments, and all replicates under one condition/treatment should use the same name, such as A, A, A (one condition with three replicates), B, B, B (the other condition with three replicates)… If the replicates have different

names such as A1, A2, A3, B1, B2, B3 …, they will be regarded as different conditions. The first column header should be the names of variables (such as metabolites, proteins, m/z values, etc).

One data can only have one row header and one column header. User can use “selecting regions” function when opening Excel sheets to avoid excessive header problems. If the data does not contain header names, the toolbox will automatically name them.

If the data is arranged in a transposed way, i.e., row header is variable names and column header is condition names, it can be transposed by the button “Transpose”. See Figure 2.2.

A.



B.

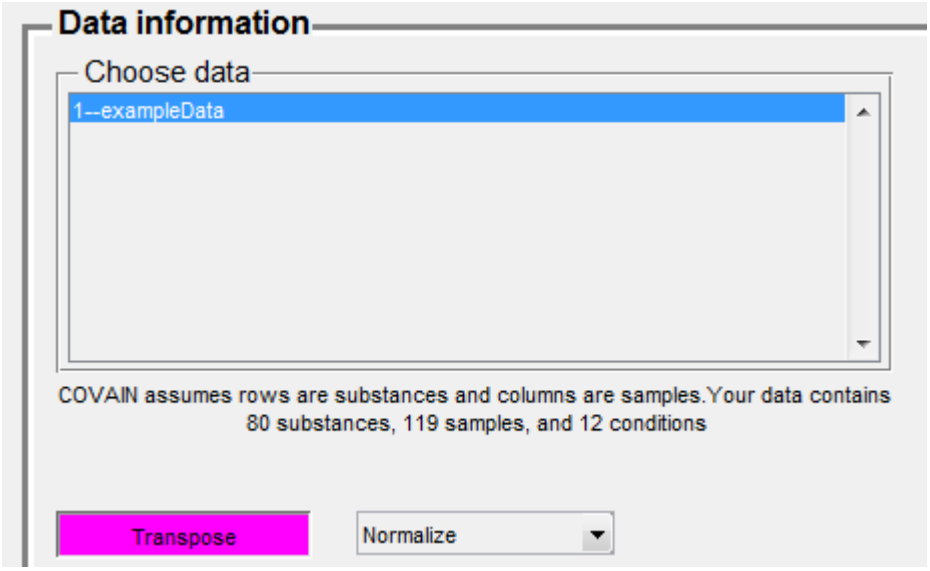


Figure 2.2: COVAIN assumes rows are substances (variables) and columns are samples. The number of variables and samples as well as conditions will be shown under the “Choose data” panel after loading the data. User can therefore decide if the data needs to be transposed. If doing the transposition, the “Transpose” button will be highlighted with fuchsia color indicating this action has been executed. The “Transpose” action can be cancelled.

Data contents

The blanks or none-number characters in the data are regarded as missing values and will be replaced by zeros.

Data preview

After loading the data, the data name (the file name) will be shown in the “Choose data” panel with a number prefix 1--, indicating it is the first dataset, and so on for further datasets (See Figure 2.2). The data is previewed in the central table. The default is mean value preview, i.e., mean value under one condition for each variable. Sometimes if the variable names are too long, or computer screen is too small, it may be favorable to use “transpose preview”. See Figure 2.3.

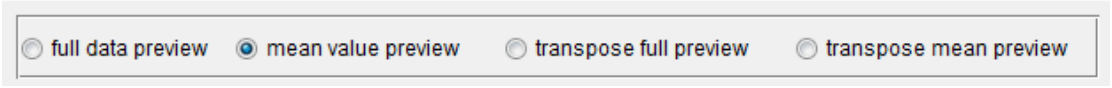


Figure 2.3: The four data preview radio buttons.

Default options for data preprocessing and data analysis

Before doing any preprocessing or analysis, it would be good to know the default options. Click the button “Options” at the right side and you will see Options window, like Figure 2.4. All options can be changed or edited. Wrong input (such as a negative p-value) will be ignored. The options are updated by clicking the “OK” button.

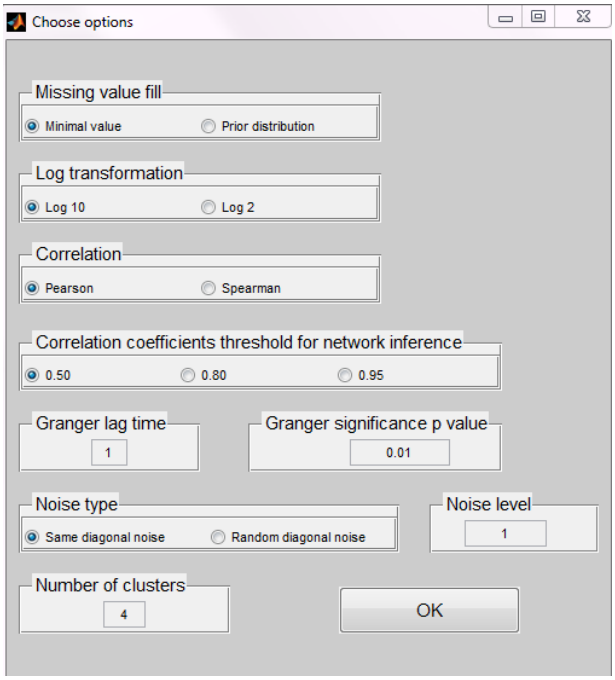


Figure 2.4: Options selection window.

Data preprocessing

Missing value imputation

The missing value information will be shown after loading the data (Figure 2.5). The default option is using the half of the minimal value of all samples of the specified variable to fill the missing values (of the same variable) that are not detected by instruments. It is also optional to use prior distribution to estimate the missing values. The strategy is based on the assumption that measurements are normal-distributed. Detailed algorithm can be referred to the paper. This action can be cancelled. For many further data analysis, missing values imputation is necessary.

Outlier adjustment

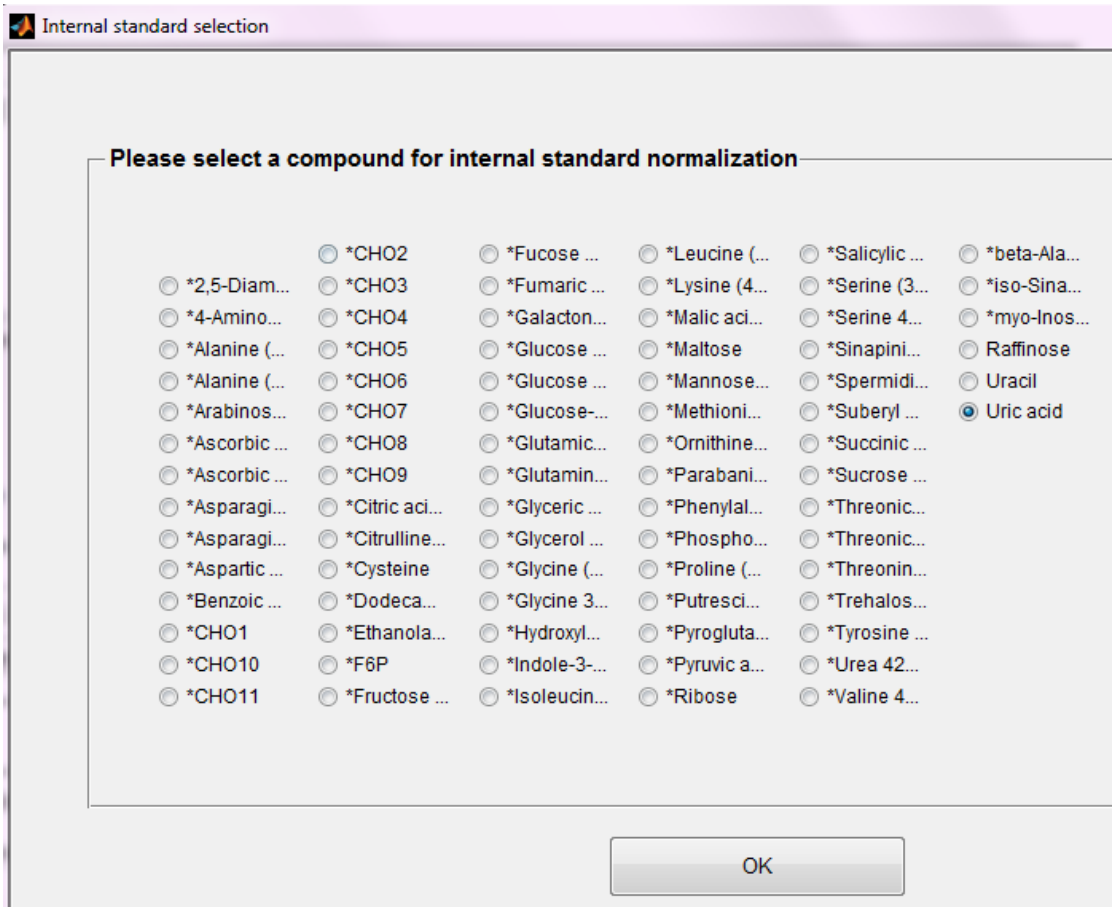
The outlier information will be shown after loading the data (Figure 2.5). The outliers are defined as measurements outside of two standard deviations of mean values for each condition of each compound. The outlier adjustment firstly proposes a prior distribution of the rest of the data and then randomly samples values from this distribution to fill outliers. This action is optional and can be cancelled.

Figure 2.5: Fill missing values and adjust outliers. After executing actions, the background color of the buttons are highlighted in fuchsia color.

Sample normalization

Note the sample normalization is not related with statistical data normalization. It applies in metabolomics or proteomics raw data and uses internal standard or fresh weight to calibrate the data. For internal standard, a window will be launched to let user select one variable (Figure 2.6A); for fresh weight, one way is to append a row in the data file using the “Fresh Weight” as variable name, thus the toolbox reads the fresh weight values; the other way is to input fresh weight values for each sample in a series of new windows (Figure 2.6B).

A.



B.

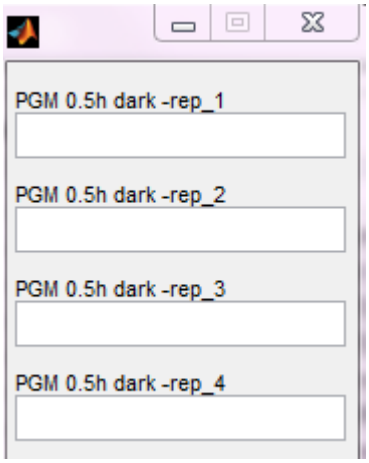


Figure 2.6: Two sample normalization methods. A. Internal standard. B. Fresh weight.

Data transformation

COVAIN applies three transformations: Log (Log_{10} or Log_2), Z score and 0-1 normalization as shown in Figure 2.7. The 0-1 normalization means, for a vector X (x_1, x_2, x_3, …, x_n), the

transformed vector will be $(x_i - x_{min})/(x_{max} - x_{min})$, i = 1,2, 3, …, n, where x_{max} and x_{min} are maximal and minimal value of *X*, respectively, thus all variables are scaled into [0,1] range.

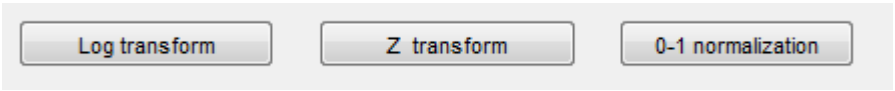


Figure 2.7: The three data normalization buttons.

Data filtration and visualization

Evaluation

Click on the "Evaluation" button and you will see a window like Figure 2.8. The upper panel shows the reproducibility report, i.e., how the replicates are similar to each other. The similarity criterion is Pearson's correlation coefficients. Generally, a close to 1 value indicates good reproducibility (So please look at values instead of colors). The bottom labels are arranged as R1 vs R2, R1 vs R3, …, denoting pairs between Replicate 1 with 2, Replicate 1 with 3, etc. It is optional to remove some replicates by inputting their indices in the right edit box, for example, "1", or "1:2", or "1,3,4". Click "OK" and a new dataset will be imported.

The lower panel shows some statistical features of all variables and enable user to select variables that satisfy the specified criterions. There are three criterion to filter out variables: 1) SNR (Signal-to-noise ratio), defined as mean value divided by standard deviation of a variable under each condition; 2) Level, the mean value of a variable under each condition; 3) Change, the relative change of mean value of one condition (such as treatment) to a reference condition (such as control). In the future, more criterions will be introduced.

User can choose a subset of variables by applying "Top", "Bottom" or "Middle" (list boxes under the "Choose" options) to select some percentage (input boxes under "Percentage %" options) and at the same time satisfying one or more criterions (check boxes under the "AND" options), over one reference condition or all conditions (list boxes under the "Conditions" options).

User can use "View" button preview the selection effect before accepting. For overlapping part of two or three criterions, a Venn diagram[1] will be shown in the right side (two criterions) or as a new window (three criterions). Finally, clicking "Accept" will import the selected variables as a new dataset, and closing the figure will not change the current data.

[1] A third-party script from Matlab Central, vennX.m, was modified and applied. The original file can be downloaded from: http://www.mathworks.com/matlabcentral/fileexchange/6116

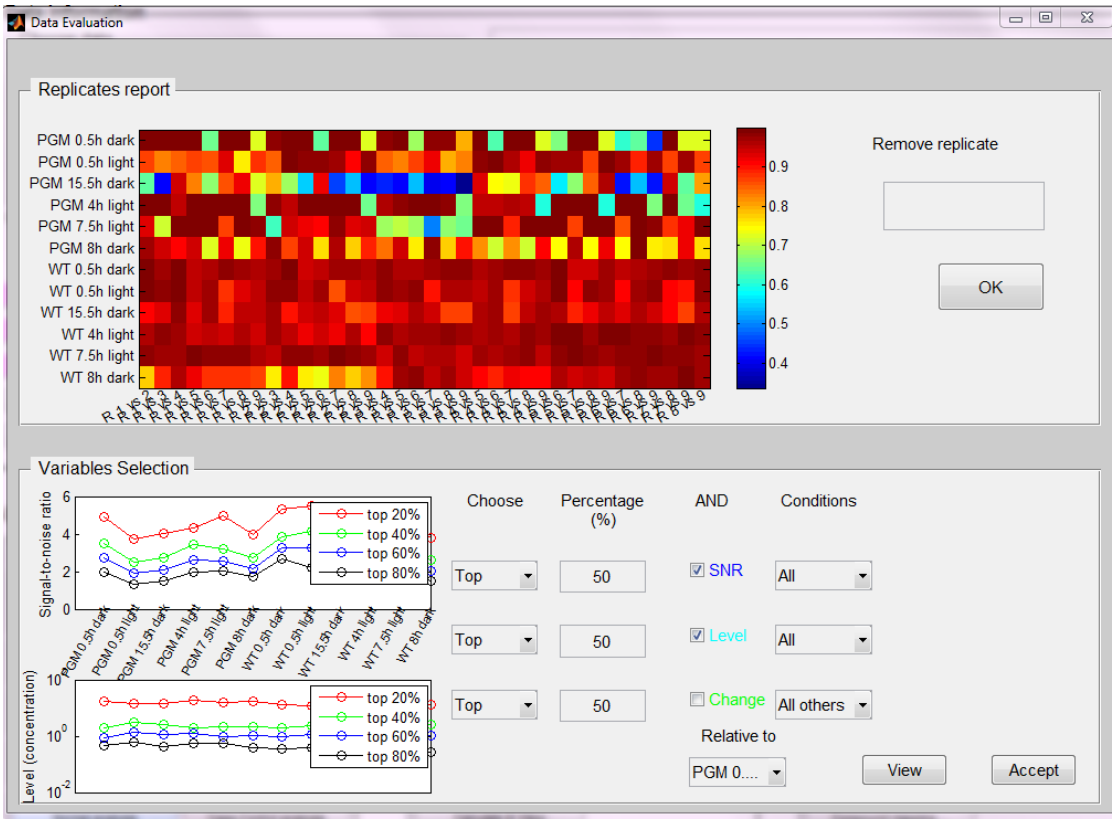


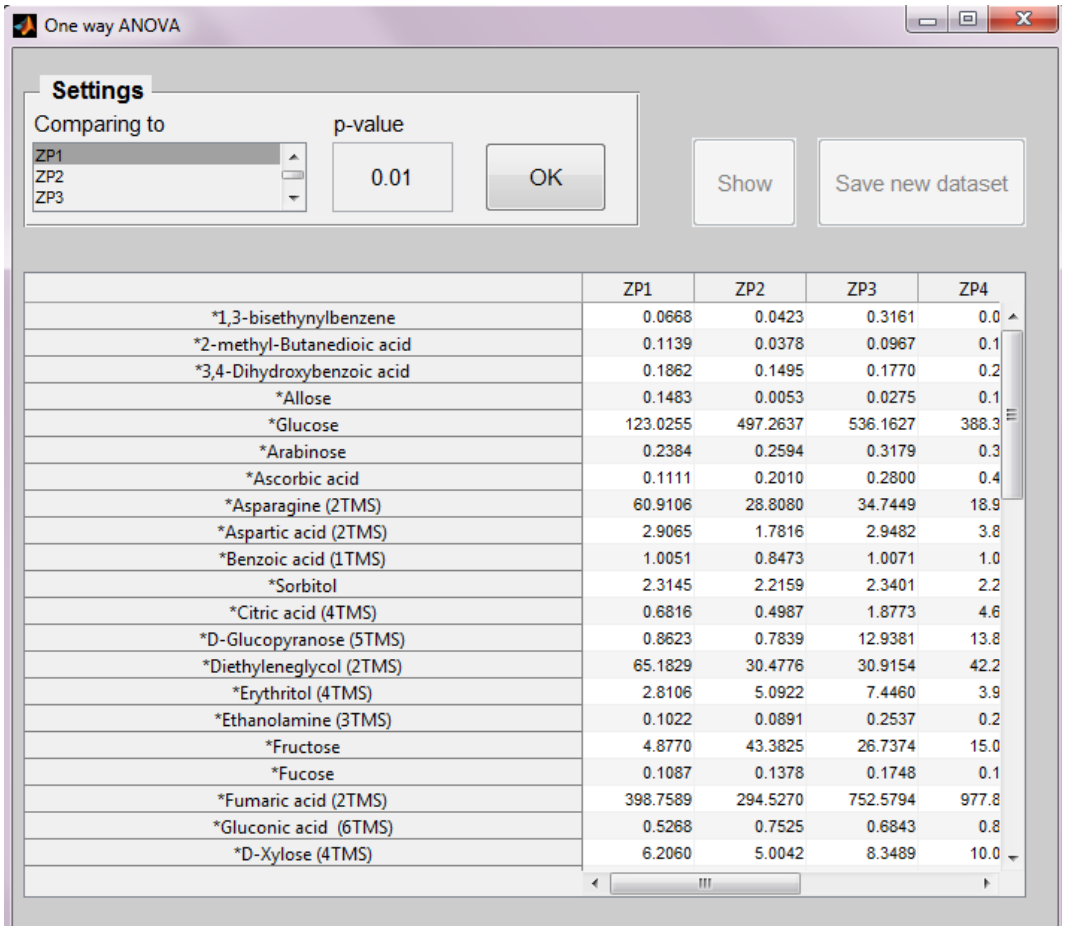
Figure 2.8: The data evaluation enables user to filter out replicates or variables. See texts for details.

ANOVA

Currently only one way ANOVA is supported. It uses Matlab Statistics Toolbox ANOVA analysis function *anova1* and multi-compare correction function *multcompare*. In the new popped out window after clicking the "ANOVA" button (Figure 2.9), user needs to define a reference condition and define a p-value. Then after clicking the "OK" button, a colored data table will be shown, with grey color denoting the reference condition, blue color denoting significant negative change (smaller than the reference condition at the specified p-value) and pink color significant positive change (larger than the reference condition at the specified p-value).

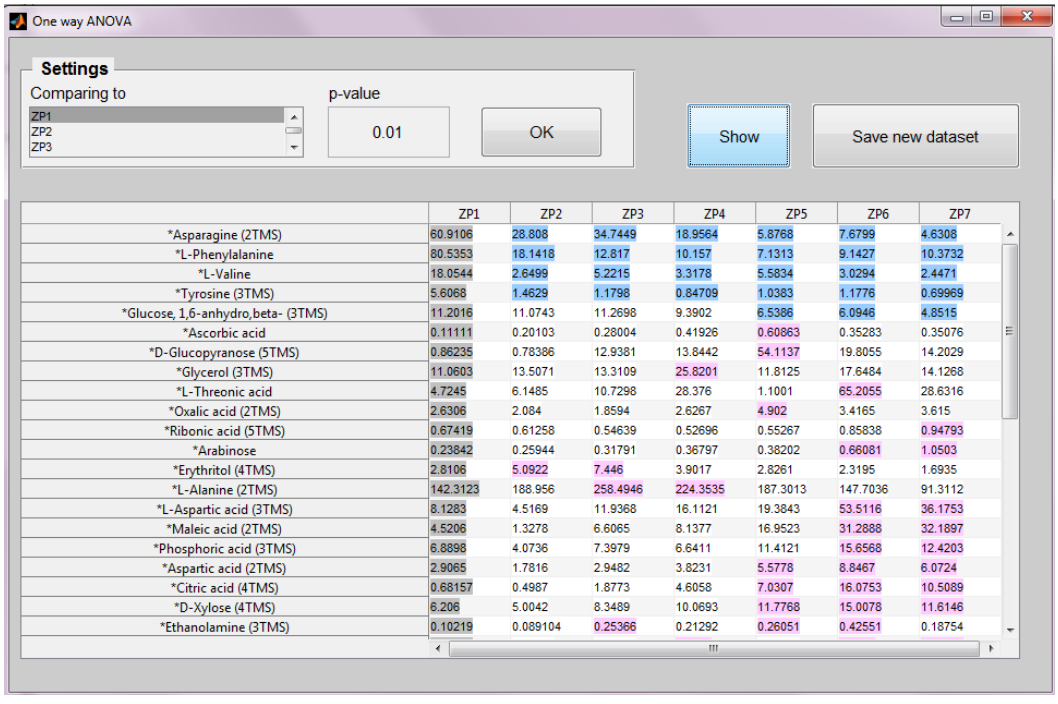
Use the "Show" button to display variables with significant changes (both negative and positive) over the reference condition: in the new table the variables are sorted from mostly negative change to mostly positive change (Figure 2.10).

Use "Save new dataset" to import the variables with significant changes over the reference condition as a new dataset. In the table, single or multi rows selections will produce box plots of one or more variables under all conditions see Figure 2.11.



	ZP1	ZP2	ZP3	ZP4
*1,3-bisethynylbenzene	0.0668	0.0423	0.3161	0.0
*2-methyl-Butanedioic acid	0.1139	0.0378	0.0967	0.1
*3,4-Dihydroxybenzoic acid	0.1862	0.1495	0.1770	0.2
*Allose	0.1483	0.0053	0.0275	0.1
*Glucose	123.0255	497.2637	536.1627	388.3
*Arabinose	0.2384	0.2594	0.3179	0.3
*Ascorbic acid	0.1111	0.2010	0.2800	0.4
*Asparagine (2TMS)	60.9106	28.8080	34.7449	18.9
*Aspartic acid (2TMS)	2.9065	1.7816	2.9482	3.8
*Benzoic acid (1TMS)	1.0051	0.8473	1.0071	1.0
*Sorbitol	2.3145	2.2159	2.3401	2.2
*Citric acid (4TMS)	0.6816	0.4987	1.8773	4.6
*D-Glucopyranose (5TMS)	0.8623	0.7839	12.9381	13.8
*Diethyleneglycol (2TMS)	65.1829	30.4776	30.9154	42.2
*Erythritol (4TMS)	2.8106	5.0922	7.4460	3.9
*Ethanolamine (3TMS)	0.1022	0.0891	0.2537	0.2
*Fructose	4.8770	43.3825	26.7374	15.0
*Fucose	0.1087	0.1378	0.1748	0.1
*Fumaric acid (2TMS)	398.7589	294.5270	752.5794	977.8
*Gluconic acid (6TMS)	0.5268	0.7525	0.6843	0.8
*D-Xylose (4TMS)	6.2060	5.0042	8.3489	10.0

Figure 2.9: The new ANOVA window.



	ZP1	ZP2	ZP3	ZP4	ZP5	ZP6	ZP7
*Asparagine (2TMS)	60.9106	28.808	34.7449	18.9564	5.8768	7.6799	4.6308
*L-Phenylalanine	80.5353	18.1418	12.817	10.157	7.1313	9.1427	10.3732
*L-Valine	18.0544	2.6499	5.2215	3.3178	5.5834	3.0294	2.4471
*Tyrosine (3TMS)	5.6068	1.4629	1.1798	0.84709	1.0383	1.1776	0.69969
*Glucose, 1,6-anhydro,beta- (3TMS)	11.2016	11.0743	11.2698	9.3902	6.5386	6.0946	4.8515
*Ascorbic acid	0.11111	0.20103	0.28004	0.41926	0.60863	0.35283	0.35076
*D-Glucopyranose (5TMS)	0.86235	0.78386	12.9381	13.8442	54.1137	19.8055	14.2029
*Glycerol (3TMS)	11.0603	13.5071	13.3109	25.8201	11.8125	17.6484	14.1268
*L-Threonic acid	4.7245	6.1485	10.7298	28.376	1.1001	65.2055	28.6316
*Oxalic acid (2TMS)	2.6306	2.084	1.8594	2.6267	4.902	3.4165	3.615
*Ribonic acid (5TMS)	0.67419	0.61258	0.54639	0.52696	0.55267	0.85838	0.94793
*Arabinose	0.23842	0.25944	0.31791	0.36797	0.38202	0.66081	1.0503
*Erythritol (4TMS)	2.8106	5.0922	7.446	3.9017	2.8261	2.3195	1.6935
*L-Alanine (2TMS)	142.3123	188.956	258.4946	224.3535	187.3013	147.7036	91.3112
*L-Aspartic acid (3TMS)	8.1283	4.5169	11.9368	16.1121	19.3843	53.5116	36.1753
*Maleic acid (2TMS)	4.5206	1.3278	6.6065	8.1377	16.9523	31.2888	32.1897
*Phosphoric acid (3TMS)	6.8898	4.0736	7.3979	6.6411	11.4121	15.6568	12.4203
*Aspartic acid (2TMS)	2.9065	1.7816	2.9482	3.8231	5.5778	8.8467	6.0724
*Citric acid (4TMS)	0.68157	0.4987	1.8773	4.6058	7.0307	16.0753	10.5089
*D-Xylose (4TMS)	6.206	5.0042	8.3489	10.0693	11.7768	15.0078	11.6146
*Ethanolamine (3TMS)	0.10219	0.089104	0.25366	0.21292	0.26051	0.42551	0.18754

Figure 2.10: The significant changes will be highlighted by blue or pink color, corresponding to negative or positive change over the reference condition.

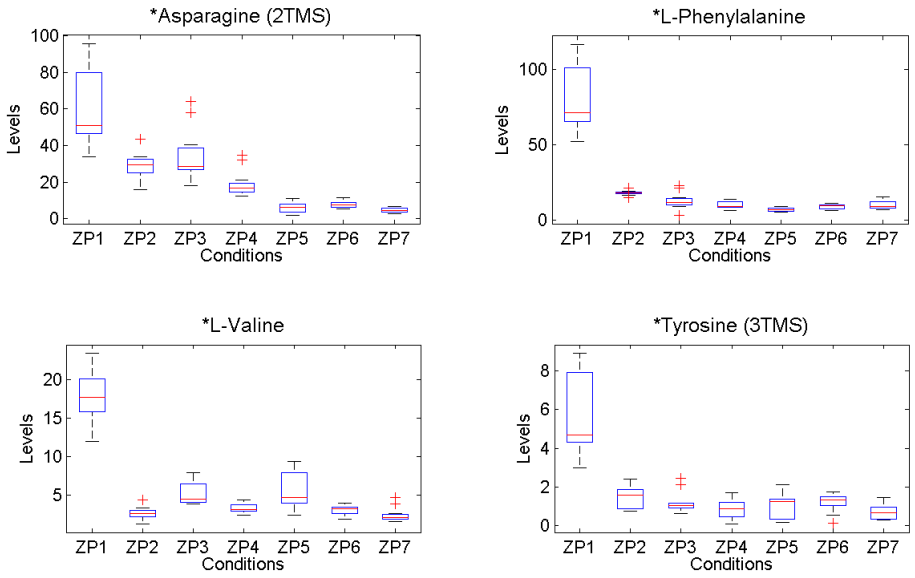


Figure 2.11: Multi rows selection of the ANOVA table produces box plot.

Error bar plot

The error bar plots[2] show the mean value of *all* variables under all conditions with one standard deviation on the error tick (Figure 2.12). Error bar plot is disabled for datasets with more than 100 variables.

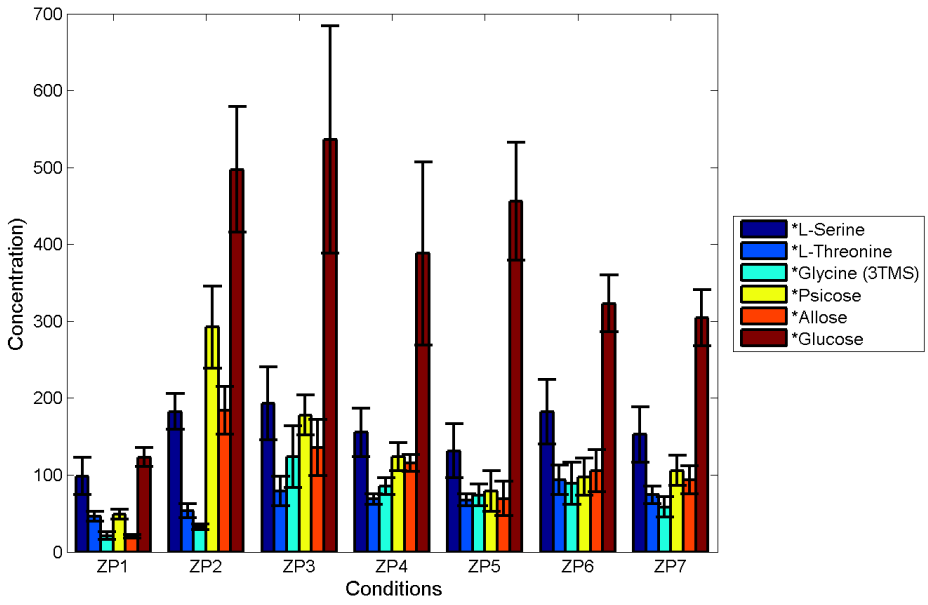
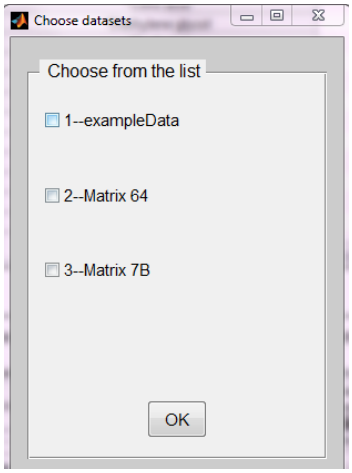


Figure 2.12 Error bar plot.

[2] Two third-party scripts from Matlab Central, barweb.m and errorbar_tick.m, were modified and applied. The original files can be downloaded from: http://www.mathworks.com/matlabcentral/fileexchange/10803, http://www.mathworks.com/matlabcentral/fileexchange/22826

Data combination

COVAIN supports combining two or more datasets into one new dataset. This button lies in the right side of the interface, just under the “Load data …” button. As the first step, user needs to select datasets to combine; then, because different datasets may have different number of replicates or conditions, user need to define which replicates or conditions are combined together. For replicates, please input their indices; for conditions, please select them from the list boxes, see an example in Figure 2.13.



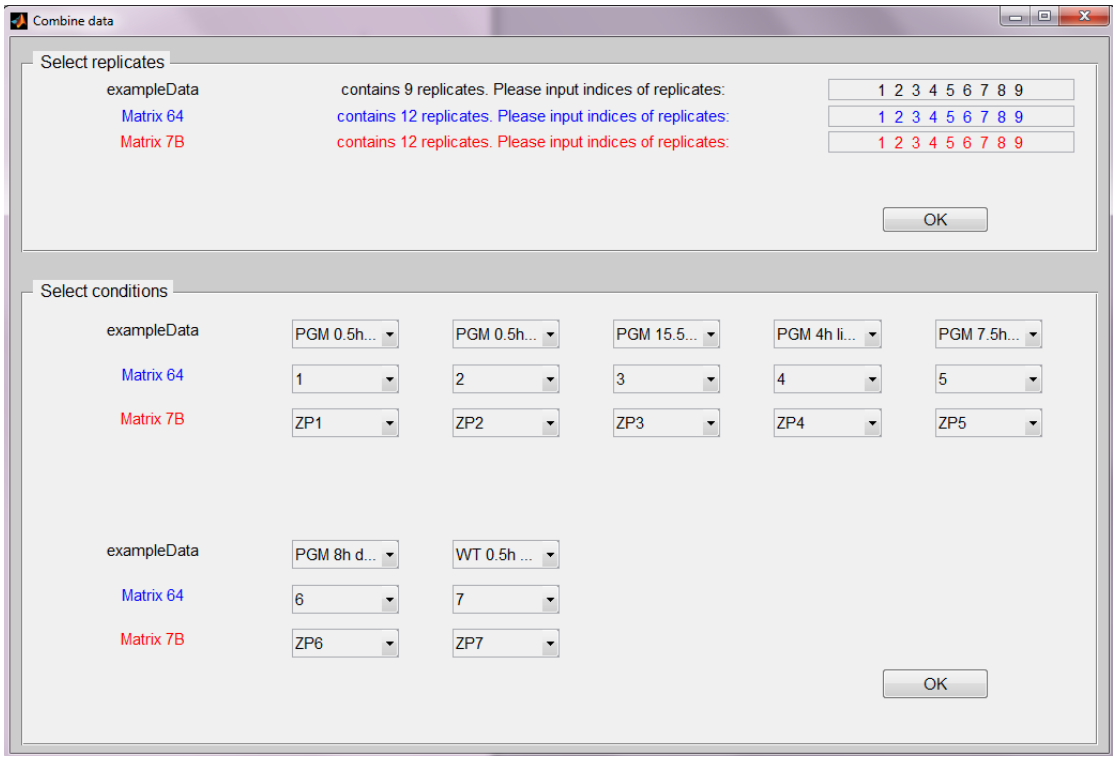


Figure 2.13: Two steps to combine datasets. First step (left figure), select datasets using check boxes; second step (bottom figure), select replicates and conditions using list boxes.

Data analysis

Multivariate statistics

Conditions selection

Before doing analysis, user needs to set which conditions to be analyzed. Please simply select “Default” if analyzing the whole data, or input the indices of conditions in the edit box (such as “1,2,3”, or “1:3”, or “1 2 3”). No selection will result a warning.

PCA and ICA

Up to five principal components (PCs) or independent components (ICs) can be displayed. Please specify the number of PCs/ICs in the edit box. Default number (2) will be used for wrong or empty input. Then click the “Plot” button and four figures will appear: the PC variance occupancy (PCA) or IC kurtosis values (ICA), loadings (PCA and ICA), score plot 2D and 3D (PCA and ICA), see an example in figure 2.14 (PCA only). The names of variables on the loading plot can be labeled by using the “Label PCA” or “Label ICA” buttons in the bottom of “Multivariate Statistics” panel.

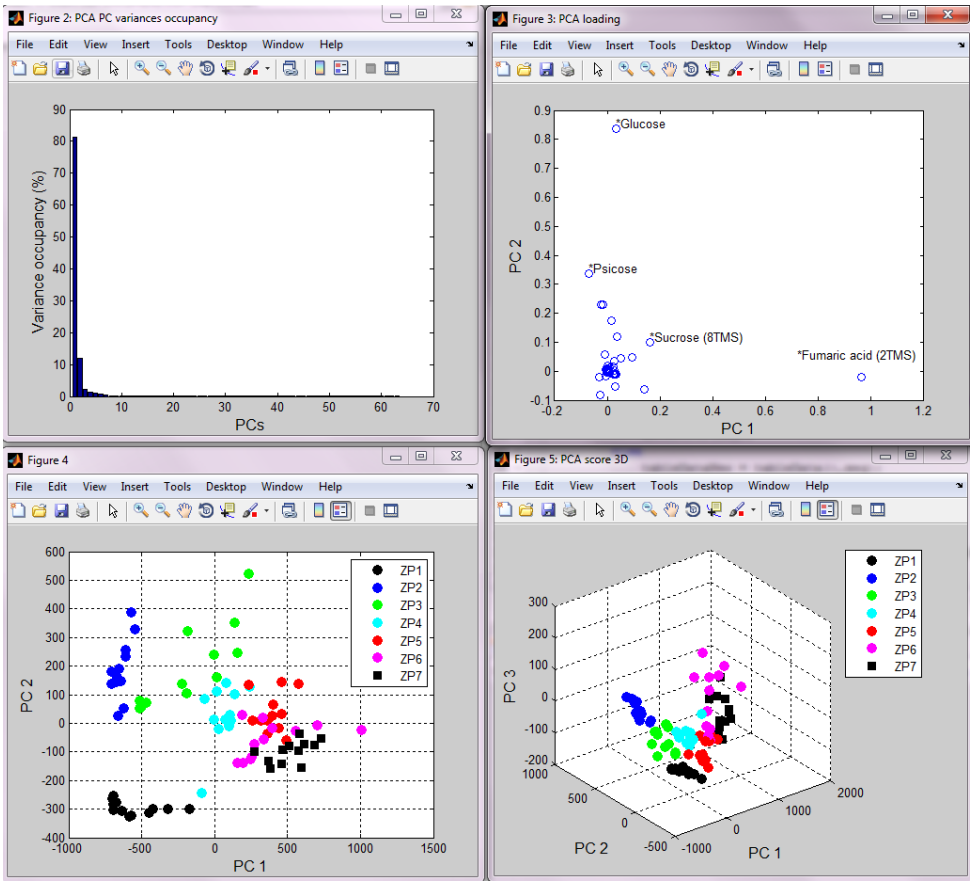


Figure 2.14: The four plots for PCA or ICA.

Correlation

The default method is Pearson's correlation coefficients, but can be changed to Spearman's method in "Options". The threshold value in "Options" does not influence the correlation coefficients calculation, but affects the inferred correlation network (See Network analysis section). A heat map denoting correlation coefficients values will be shown after clicking the "Calculate and View" button like Figure 2.15.

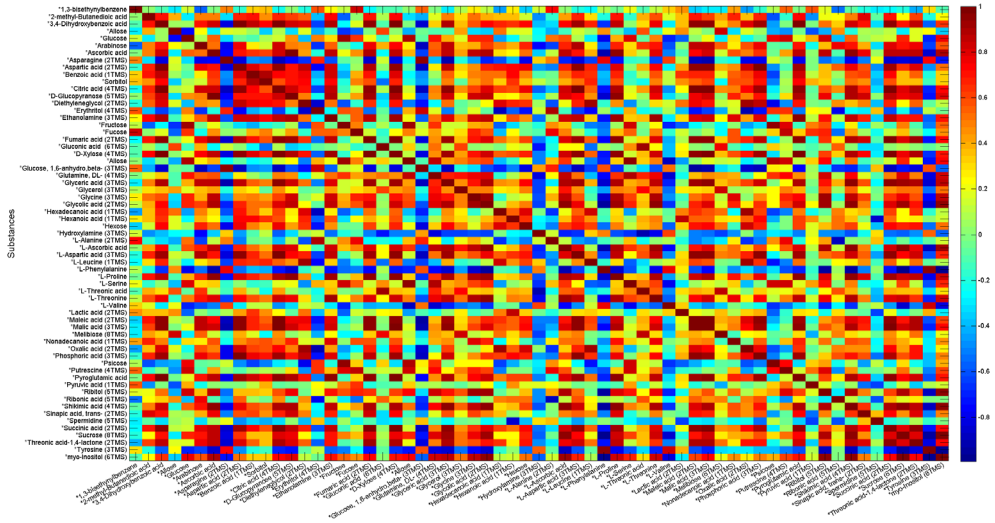


Figure 2.15: The correlation coefficients heat map with a color bar.

Bi-clustering

The bi-clustering uses average linkage of Euclidean distance between groups as the metric. The "Normal analysis" clusters all conditions and all variables (Figure 2.16 left) while the "Case-control analysis" needs to set up the control condition (Figure 2.16 right) and cluster the difference of other conditions over the control condition for all variables.

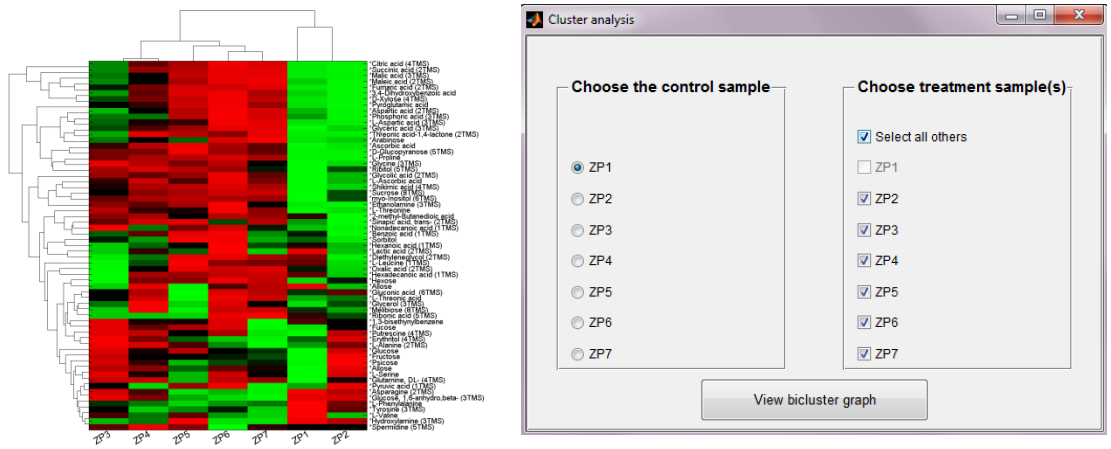


Figure 2.16: Two ways for bi-clustering.

Time series analysis

Time order selection

Before doing analysis, user needs to set up the correct time series. Please simply select “Default” if analyzing the whole data in the listed (shown in the table) time order, or input the indices of conditions in the edit box (such as “4,1,2,3, or “4,1:3”, or “4 1 2 3”). No selection will result a warning. Note: time series analysis in COVAIN needs a minimum length of four time points.

Correlation

The correlation analysis is the same as described in the Multivariate Statistics section.

Clustering

The clustering metric is similar to those described in the Multivariate Statistics section, but the produced figures are profiles plot, i.e., levels as y-axis and time points as x-axis.

Granger causation analysis

The Granger causation analysis uses the “Granger time lag” and “p-value” (both can be defined in the “Options”, the lag is limited to 2 for short time series) as optional parameters and plot three figures with information on their figure titles, and “sort” and “export” function in their bottom parts: 1) which variables causes which others; 2) which variables are caused by which other; 3) which variables pair with which others, i.e., one causes the other and vice versa.

The number of causations and the significance levels are shown in the figure title. The lists of variables and Benjamini-Hochberg corrected false discovery rate p-value are shown in the table region, such as a four cell row “Variable A causes Variable B 0.001”. User can use the bottom buttons “Sort …” to sort the table by different rules. It is possible to export the table to an Excel file.

Single or multi- sections of table cells will produce figures showing the profiles of causing variables (in red) and the being caused variables (in blue).

Figure 2.17 illustrates the Granger causation analysis usage.

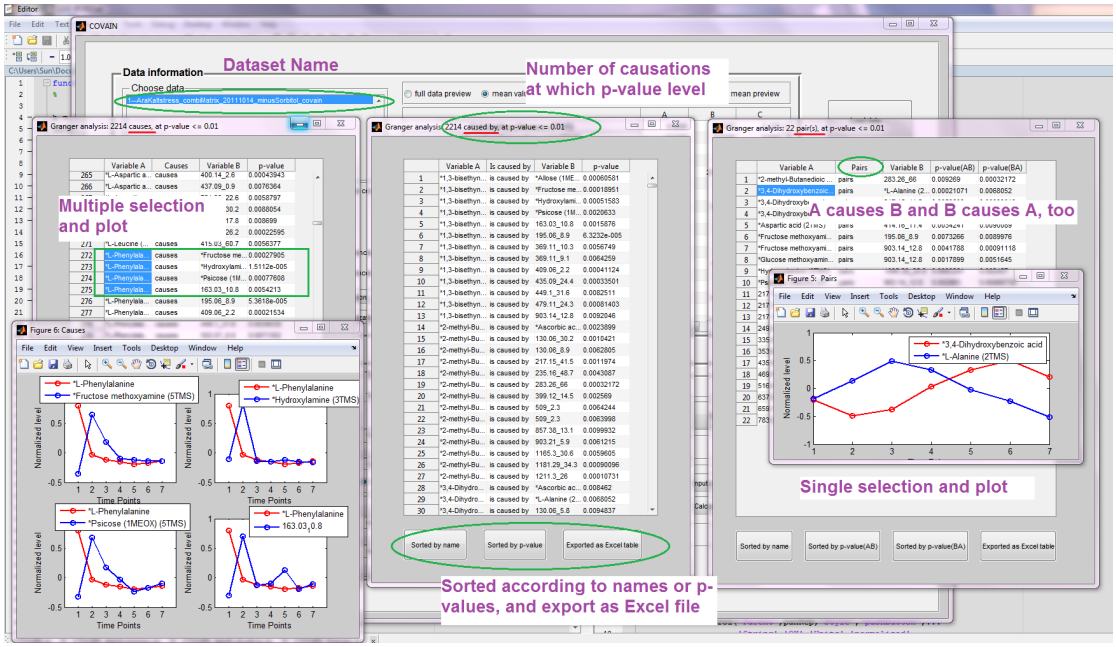


Figure 2.17: Granger causation analysis.

Permutation entropy

The permutation entropy (PE) values for all variables are shown by stem plot like Figure 2.18. The PE as a metric for complexity was investigated with transcriptomics time series data[3] but not applied in metabolomics data yet.

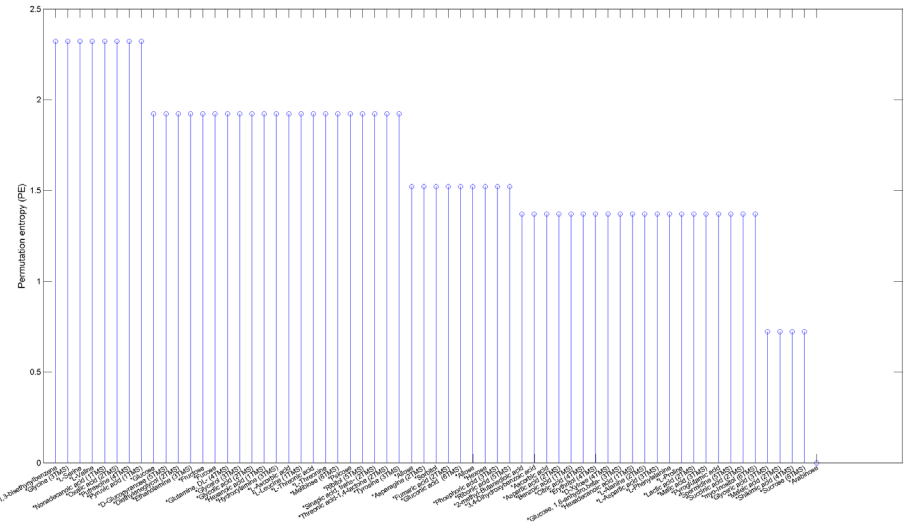


Figure 2.18: The permutation entropy analysis.

[3] Sun X, Zou Y, Nikiforova V, Kurths J, Walther D., The complexity of gene expression dynamics revealed by permutation entropy, BMC Bioinformatics. 2010 Dec 22;11:607

Network analysis

Network inference

The network inference uses either correlation coefficient or Granger causation p-value as metric. For correlation coefficient, two variables are considered connected if their pairwise absolute correlation coefficient value is higher than the threshold values; for Granger causation analysis, a connection between two variables is established if its p-value lower than the threshold. Note that, correlation network is undirected but Granger network is directed.

Network visualization

Network visualization checks if the selected inference result is available or not. If not, a warning will be shown; if yes, it uses Matlab Biograph class (in Bioinformatics Toolbox) to visualize the network as shown in Figure 2.19. For large network with over 500 interactions, visualizing in Matlab is disabled (because it would be too slow) but the network can be visualized by a good visualization software Cytoscape[4] using the .sif file if you save the results (See Save section).

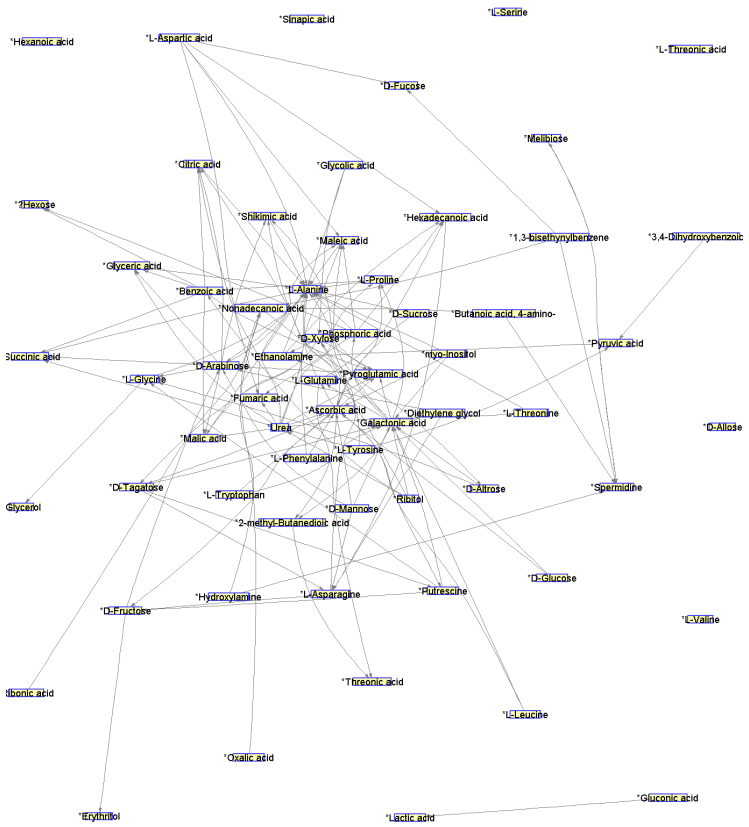


Figure 2.19: Network visualized by biograph. The names of variables have been labeled.

[4] www.cytoscape.org

Network properties

Currently this function is not supported; User can use Cytoscape plug-ins to calculate the network topological properties using the .sif file in the results (See Save section).

KEGG mapping

Currently only metabolomics data is supported, i.e., only metabolite names can be mapped to KEGG pathways. First, the metabolite names stemming from a typical metabolomics analysis are mapped against KEGG names and formulas to identify all involved pathways. In a second step a minimal interconnected graph model, i.e., each node can be reached from every other node, is defined. The resulting pathway network corresponding to the experimentally detected metabolites and a minimal set of interconnecting metabolites is exported and visualized with Cytoscape.

It is possible to select a control reference condition and color the change on other conditions using predefined colors (red for positive change and blue for negative change). The strategy is shown in Figure 2.20.

The KEGG mapping needs Matlab Bioinformatics Toolbox. It links to KEGG by a SOUP service. Note the mapping usually takes a long time; therefore, a “cancel” button is designed.

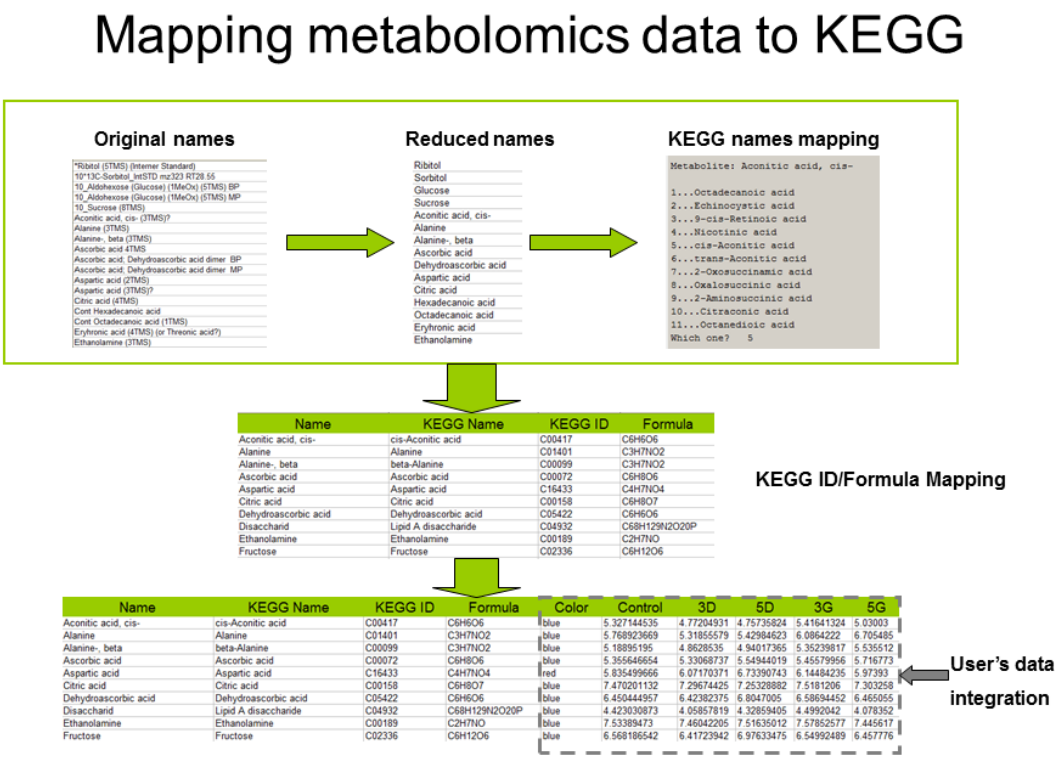


Figure 2.20: Mapping metabolomics data to KEGG pathway workflow.

Inverse Jacobian analysis

User can use the example data (data_glycolysis … and stoichiometry_glycolysis) to try this function. Please follow the steps in Figure 2.21. User can also upload their own data and the corresponding stoichiometric matrix to get the inverse Jacobian from the covariance matrix.

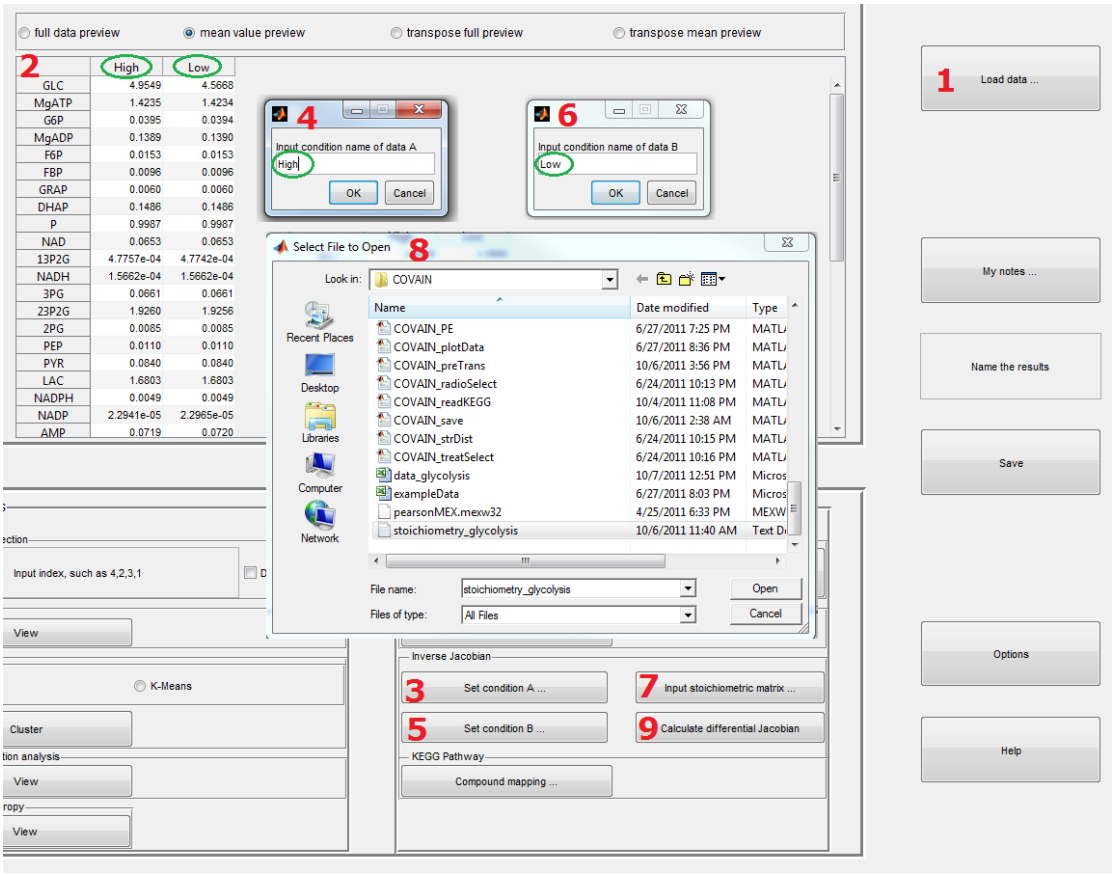


Figure 2.21: A simple tutorial on how to use the inverse Jacobian analysis on the example data. There are nine steps (labeled in red numbers) to obtain differential Jacobian from metabolomics data. In this example, the two conditions are labeled as “High” and “Low” meaning high and low glucose import, respectively.

1. Launch the COVAIN package and load data;
2. The data information will be shown, where rows are names of metabolites and columns are the names of conditions (green circles);
3. In the “Inverse Jacobian” panel, click the button “Set condition A …” and
4. Choose condition A by inputting its dataset index (See Figure 2.2A), NOT the names in Figure 2.21.
5. Similarly, click the button “Set condition B …” and
6. Choose condition B by inputting its dataset index (See Figure 2.2A), NOT the names in Figure 2.21.
7. Click the button “Input stoichiometric matrix” and
8. Load the stoichiometric matrix file;
9. Finally, click the button “Calculate differential Jacobian” and obtain the result.

mzGroupAnalyzer and Pathway Viewer

Format of inputs:

In Matlab GUI, it is difficult for drag-and-drop operations, so the format of input files need to strictly obey predefined rules. Please follow the format of the example data files. In detail:

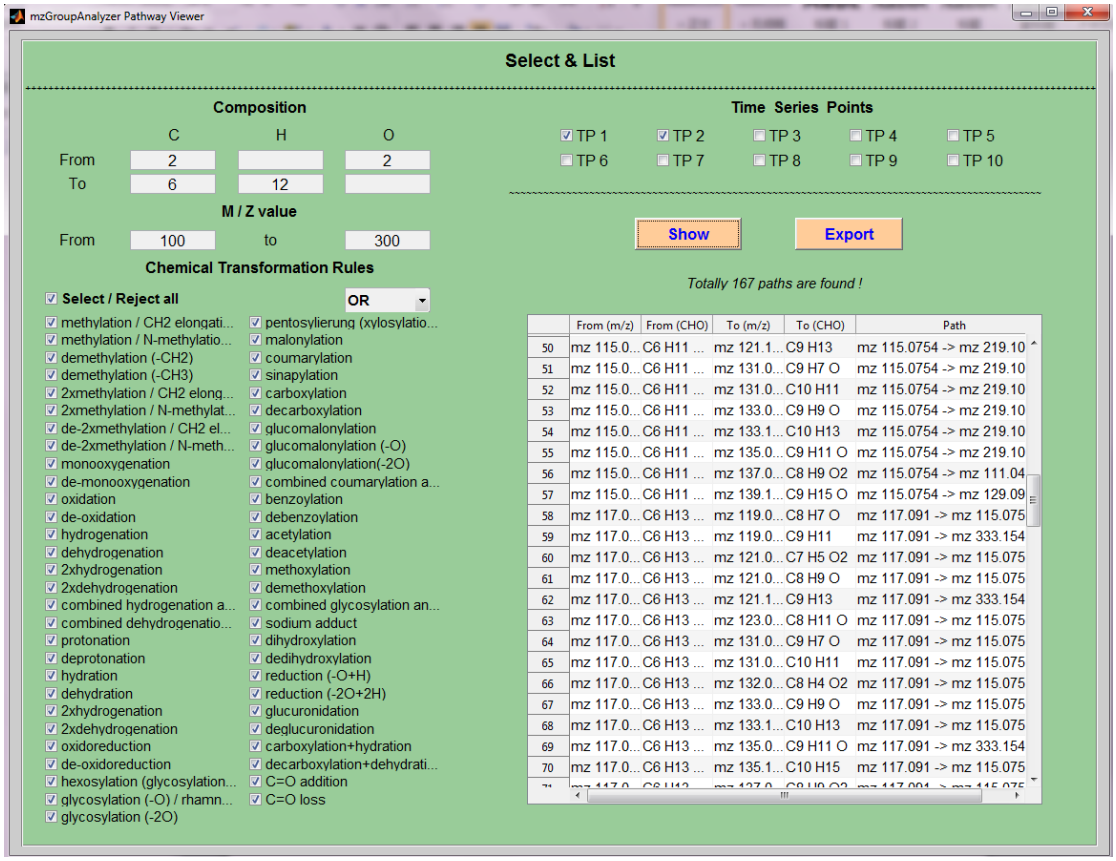
1) The m/z data Excel spreadsheet:
    a) The first row (header) must have these names: m/z, Intensity, Theo. Mass, Delta (ppm), Composition. Names are capital sensitive.
    b) The second row is left for empty. The program reads data from the third row.
    c) Any number in the composition need to closely follow characters. For example, H2 O, H2O, are correct, but H 2 O is wrong.
    d) m/z features with empty composition are removed from further processing.
2) The rule Excel spreadsheet:
    a) The first row (header) is not important. The program reads data from the second row.
    b) Chemical elements that are not appearing in any m/z's composition are removed. For example, C, H, O, N, P, S are defined in rules, but P and S do not appear in any composition in the m/z data file, so P and S will be removed.

How to use:

In the COVAIN main panel, find two mzGroupAnalyzer buttons in the Network Analysis module.

For first use, please click the first button "Load data & Analysis", and you will see three consecutive windows, representing respectively, 1) load the m/z data files (**multiple files allowed**. If they are time series, the names of files should follow correct order); 2) load the rule file (**single file**); 3) determine the name and folder to save the results. Then the program will do calculations while displaying a dynamic wait bar. Then you will see a question dialog asking if proceeding to the Pathway Viewer. You can select "Yes" to proceed, "No" not to proceed, or "Load" to open the other existing mzStruct workspace which is titled as "mzStruct.mat". If you select "Yes", you will see a new window similar to the below one.

The second button, Pathway Viewer, can be used separately if you already run the program before (hence the mzStruct workspace is available). Click this button, and you will see a window to find the workspace file "mzStruct.mat". After loading this workspace, you will see the below window. This button is especially useful if you have processed a big dataset and only need to use the Pathway Viewer.



Explanation of all parts of Pathway Viewer:

Composition: Empty input means using default minimal and maximal values (0-10000). So if you do not need this filtering option, just leave them empty.

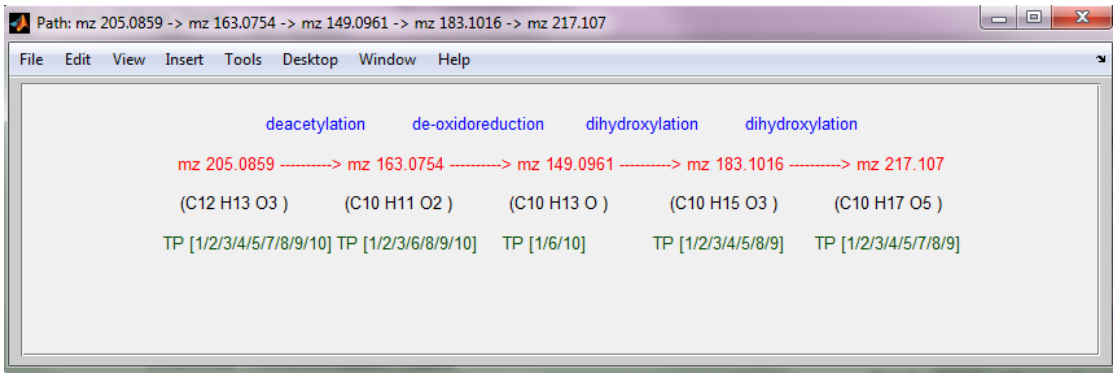
m / z value: Empty input means using default minimal and maximal values (0-1000000). If you do not need this filtering option, just leave them empty. Note: this filtering **only** applies on the first and last compounds.

Chemical rules: You can select or reject all rules by clicking the "Select/Reject all" button; you can also select only a few rules. The selected rules can be in AND relation, meaning the obtained pathways (shown in the right side) contain ALL rules, or in OR relation, meaning ANY of these rules. Very long rule names may not be completely displayed, but you can see their full names when moving the mouse around (also known as tooltip string).

Time series points: The available time points of the first compound of a pathway. Multiple selection means AND operation.

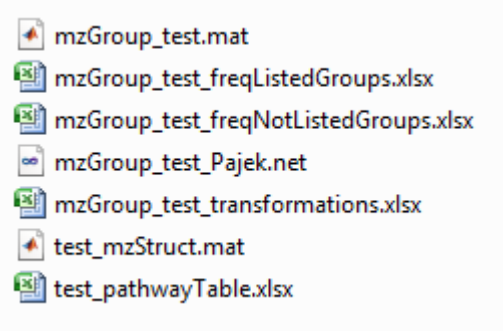
Show: All pathways satisfying the chosen criteria will be displayed in the table below. The number of pathways will be shown. You can see the pathway chain by clicking the Path column. An example is shown below, where the blue texts show the rules of each step, red denote the m/z values, black mean the composition of each m/z value and green mean the available time points of each compound (the time points are separated by "/").

Export: Save the table into an Excel file.



Outputs:

The results from a typical running may look like this:



The first five results are saved by mzGroupAnalyzer(or, using the first button) and the other two are exported by mzPathViewer(or, using the second button).

Notes:

1. Requires Matlab Bioinformatics Toolbox and Statistical Toolbox.
2. If your Matlab has Parallel Computing Toolbox, parallel computing will be used for acceleration. You can increase the number of parallel threads by modifying the default properties of this toolbox.
3. The program was tested only under Windows. Please contact me if you have problems with other OS. Most possible problems may be during the Excel file import/export.
4. The pathway searching time for big data may be quite long. My testing results are shown below (Data size is represented by number of m/z features. Empty means not available). The testing results seem lack of "scalability" because not all tasks can be run in parallel.

CPU	**Q9300**		**i5-3320M**	**i7-3770K**		**2 x Xeon E5-2690**
Property	2.5 G, 4 cores 4 threads		2.6 ~ 3.3 G, 2 cores 4 threads	3.5 ~ 3.9 G, 4 cores 8 threads		2.9 ~ 3.8 G, 8 cores 16 threads
Number of parallel threads	1	2	4	6	8	12
200 m/z		10 s	< 1 s	< 1 s	< 1 s	
500 m/z		50 s	10 s			
1800 m/z	8 h		1.6 h	1.4 h	1.2 h	1 h

(I suggest processing big data in a desktop PC or server)

Results saving

COVAIN creates default saving folder “COVAIN_Results” under the folder where each dataset is saved. However COVAIN asks the user select other place to save results, see Figure 2.22.

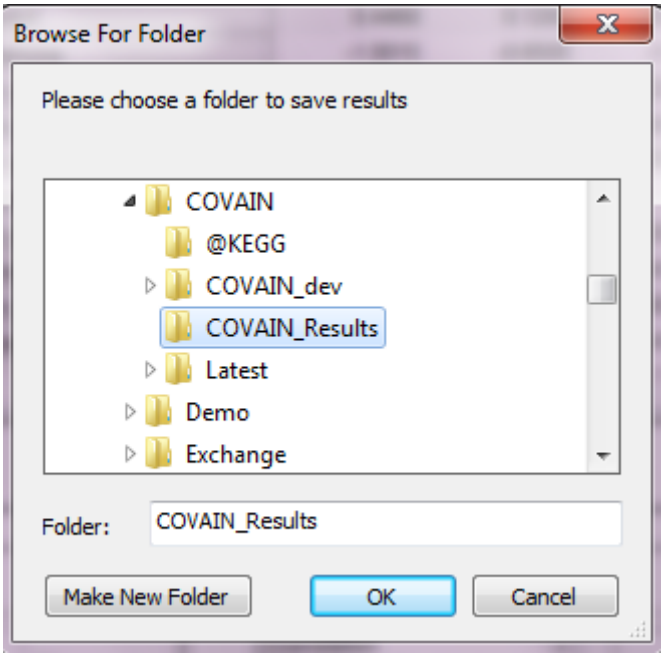


Figure 2.22: Choose the saving folder.

The saved results for each datasets: one Excel file including processed data, the PCA/ICA loadings (if done), notes (if not empty), and .sif network file (if done). The GUI data of COVAIN will be saved in a Matlab workspace which may be useful for experienced users. Please use “Name results” edit box to determine the Matlab workspace name. An example is shown in Figure 2.23.

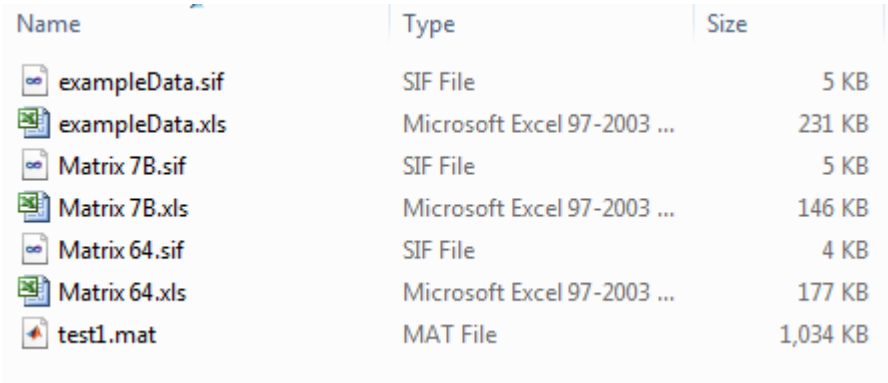


Figure 2.23: The list of saved results. Here, there are three datasets whose file names are “exampleData”, “Matrix 7B” and “Matrix 64”. For each dataset, one Excel and one .sif file are saved; for this COVAIN operation, a Matlab workspace (.mat) is saved, which contains all the analysis results of three datasets in a structure type.