Metabolomic Data Analysis with MetaboAnalyst 3.0

User ID: guest57235513848816868 79

April 11, 2016

1 Data Processing and Normalization

1.1 Reading and Processing the Raw Data

MetaboAnalyst accepts a variety of data types generated in metabolomic studies, including compound concentration data, binned NMR/MS spectra data, NMR/MS peak list data, as well as MS spectra (NetCDF, mzXML, mzDATA). Users need to specify the data types when uploading their data in order for MetaboAnalyst to select the correct algorithm to process them. Table 1 summarizes the result of the data processing steps.

1.1.1 Reading Peak Intensity Table

The peak intensity table should be uploaded in comma separated values (.csv) format. Samples can be in rows or columns, with class labels immediately following the sample IDs.

Samples are in columns and features in rows. The uploaded file is in comma separated values (.csv) format. The uploaded data file contains 55 (samples) by 359 (peaks(mz/rt)) data matrix.

1.1.2 Data Integrity Check

Before data analysis, a data integrity check is performed to make sure that all the necessary information has been collected. The class labels must be present and contain only two classes. If samples are paired, the class label must be from -n/2 to -1 for one group, and 1 to n/2 for the other group (n is the sample number and must be an even number). Class labels with same absolute value are assumed to be pairs. Compound concentration or peak intensity values should all be non-negative numbers. By default, all missing values, zeros and negative values will be replaced by the half of the minimum positive value found within the data (see next section)

1.1.3 Missing value imputations

Too many zeroes or missing values will cause difficulties for downstream analysis. MetaboAnalyst offers several different methods for this purpose. The default method replaces all the missing and zero values with a small values (the half of the minimum positive values in the original data) assuming to be the detection limit. The assumption of this approach is that most missing values are caused by low abundance metabolites (i.e.below the detection limit). In addition, since zero values may cause problem for data normalization (i.e. log), they are also replaced with this small value. User can also specify other methods, such as replace by mean/median, or use K-Nearest Neighbours, Probabilistic PCA (PPCA), Bayesian PCA (BPCA) method, Singular Value Decomposition (SVD) method to impute the missing values [1]. Please choose the one that is the most appropriate for your data.

[1] Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. *pcaMethods: a bioconductor package, providing PCA methods for incomplete data.*, Bioinformatics 2007 23(9):1164-1167

0 variables were removed for threshold 50 percent. Missing variables were imputated using KNN

1.1.4 Data Filtering

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step can usually improves the results. Data filter is strongly recommended for datasets with large number of variables (> 250) datasets contain much noise (i.e.chemometrics data). Filtering can usually improve your results[2].

For data with number of variables < 250, this step will reduce 5% of variables; For variable number between 250 and 500, 10% of variables will be removed; For variable number bwteen 500 and 1000, 25% of variables will be removed; And 40% of variabled will be removed for data with over 1000 varaibles.

No data filtering was performed.

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
Ctrl_BKS_2202	359	0	359
Ctrl_BKS_2203	359	0	359
Ctrl_BKS_2206	359	0	359
Ctrl_BKS_2211	359	0	359
HF_BKS_2214	359	0	359
HF_BKS_2216	359	0	359
HF_BKS_2217	359	0	359
HF_BKS_2218	359	0	359
Ctrl_BKS_db+_2225	359	0	359
Ctrl_BKS_db+_2226	359	0	359
Ctrl_BKS_db+_2227	359	0	359
Ctrl_BKS_db+_2231	359	0	359
HF_BKS_db+_2239	359	0	359
HF_BKS_db+_2241	359	0	359
HF_BKS_db+_2242	359	0	359
HF_BKS_db+_2243	359	0	359
Ctrl_BL6_2249	359	0	359
Ctrl_BL6_2250	359	0	359
Ctrl_BL6_2251	359	0	359
Ctrl_BL6_2252	359	0	359
HF_BL6_2260	359	0	359
HF_BL6_2261	359	0	359
HF_BL6_2262	359	0	359
HF_BL6_2266	359	0	359
Ctrl_BL6_db+_2273	359	0	359
Ctrl_BL6_db+_2274	359	0	359
Ctrl_BL6_db+_2275	359	0	359
Ctrl_BL6_db+_2279	359	0	359
HF_BL6_db+_2286	359	0	359
HF_BL6_db+_2287	359	0	359
HF_BL6_db+_2288	359	0	359
HF_BL6_db+_2289	359	0	359
Ctrl_BTBR_2292	359	0	359
Ctrl_BTBR_2293	359	0	359
Ctrl_BTBR_2297	359	0	359
Ctrl_BTBR_2298	359	0	359
HF_BTBR_2306	359	0	359
HF_BTBR_2307	359	0	359
HF_BTBR_2308	359	0	359
HF_BTBR_2309	359	0	359
Ctrl_BTBR_ob+_2316	359	0	359
Ctrl_BTBR_ob+_2317	359	0	359
Ctrl_BTBR_ob+_2318	359	0	359
Ctrl_BTBR_ob+_2320	359	0	359
HF_BTBR_ob+_2329	359	0	359
HF_BTBR_ob+_2331	359	0	359
HF_BTBR_ob+_2333	359	0	359
HF_BTBR_ob+_2334	359	0	359
Test Pool_Negatve_01	359	0	359
Test Pool_Negatve_02	359	0	359
Test Pool_Negatve_03	359	0	359
Test Pool_Negatve_04	359	0	359
Test Pool_Negatve_05	359	0	359
Test Pool_Negatve_06	359	0	359
Test Pool_Negatve_07	359	0	359

[2]Hackstadt AJ, Hess AM.*Filtering for increased power for microarray data analysis*, BMC Bioinformatics. 2009; 10: 11.

Error occured during normalization of your data .... Fail to proceed. Please check if the data format you uploaded is correct. For further information, please contact `jeff.xia@mcgill.ca` with a description of the problem. Your feedback is highly appreciated!

2 Statistical and Machine Learning Data Analysis

MetaboAnalyst offers a variety of methods commonly used in metabolomic data analyses. They include:

1. Univariate analysis methods:
    - Fold Change Analysis
    - T-tests
    - Volcano Plot
    - One-way ANOVA and post-hoc analysis
    - Correlation analysis
2. Multivariate analysis methods:
    - Principal Component Analysis (PCA)
    - Partial Least Squares - Discriminant Analysis (PLS-DA)
3. Robust Feature Selection Methods in microarray studies
    - Significance Analysis of Microarray (SAM)
    - Empirical Bayesian Analysis of Microarray (EBAM)
4. Clustering Analysis
    - Hierarchical Clustering
        - Dendrogram
        - Heatmap
    - Partitional Clustering
        - K-means Clustering
        - Self-Organizing Map (SOM)
5. Supervised Classification and Feature Selection methods
    - Random Forest
    - Support Vector Machine (SVM)

`Please note: some advanced methods are available only for two-group sample analyais.`

3 Data Annotation

Please be advised that MetaboAnalyst also supports metabolomic data annotation. For NMR, MS, or GC-MS peak list data, users can perform peak identification by searching the corresponding libraries. For compound concentration data, users can perform metabolite set enrichment analysis and metabolic pathway analysis.

The report was generated on Mon Apr 11 23:04:21 2016 with R version 3.2.2 (2015-08-14). Thank you for using MetaboAnalyst! For suggestions and feedback please contact Jeff Xia (*jeff.xia@mcgill.ca*).