Metabolomic Data Analysis with MetaboAnalyst 4.0

Name: guest8221339122704546959

December 5, 2019

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 Concentration Data

The concentration data 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 85 (samples) by 138 (compounds) 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 (KNN), 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

Zero or missing variables were replaced with a small value: 358.5

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 variables. The None option is only for less than 5000 features. Over that, if you choose None, the IQR filter will still be applied. In addition, the maximum allowed number of variables is ***10000***

No data filtering was performed.

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

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
160711-19	109	29	136
160711-23	112	26	136
160711-35	118	20	136
160711-37	114	24	136
160711-39	117	21	136
160711-43	116	22	136
160711-49	115	23	136
160711-52	122	16	136
160711-58	116	22	136
160711-67	112	26	136
160711-68	116	22	136
160711-72	121	17	136
160711-73	117	21	136
160711-74	123	15	136
160711-84	112	26	136
160711-86	117	21	136
160711-94	109	29	136
160711-95	105	33	136
160711-96	98	40	136
160711-10	113	25	136
160711-11	105	33	136
160711-12	110	28	136
160711-13	113	25	136
160711-14	108	30	136
160711-16	112	26	136
160711-21	111	27	136
160711-22	114	24	136
160711-27	123	15	136
160711-28	118	20	136
160711-38	113	25	136
160711-41	117	21	136
160711-42	114	24	136
160711-47	121	17	136
160711-56	123	15	136
160711-59	119	19	136
160711-64	111	27	136
160711-66	114	24	136
160711-69	115	23	136
160711-70	119	19	136
160711-75	116	22	136
160711-85	117	21	136
160711-87	115	23	136
160711-92	135	3	136
160711-102	115	23	136
160711-15	109	29	136
160711-18	112	26	136
160711-24	117	21	136
160711-25	111	27	136
160711-29	115	23	136
160711-40	113	25	136
160711-44	113	25	136
160711-51	119	19	136
160711-53	129	9	136
160711-54	114	24	136
160711-57	120	18	136
160711-62	114	24	136
160711-63	112	26	136
160711-65	118	20	136
160711-71	118	20	136
160711-76	117	21	136
160711-78	116	22	136
160711-89	110	28	136
160711-91	101	37	136
160711-93	111	27	136
160711-101	130	8	136
160711-17	135	3	136
160711-20	134	4	136
160711-26	137	1	136
160711-32	137	1	136
160711-33	136	2	136
160711-34	136	2	136
160711-36	136	2	136
160711-45	137	1	136
160711-46	136	2	136
160711-48	136	2	136
160711-50	136	2	136
160711-55	136	2	136
160711-77	137	1	136
160711-79	138	0	136
160711-80	136	2	136
160711-81	136	2	136
160711-82	136	2	136
160711-83	136	2	136
160711-88	102	36	136
160711-90	100	38	136

1.2 Data Normalization

The data is stored as a table with one sample per row and one variable (bin/peak/metabolite) per column. The normalization procedures implemented below are grouped into four categories. Sample specific normalization allows users to manually adjust concentrations based on biological inputs (i.e. volume, mass); row-wise normalization allows general-purpose adjustment for differences among samples; data transformation and scaling are two different approaches to make features more comparable. You can use one or combine both to achieve better results.

The normalization consists of the following options:

1. Row-wise procedures:
   - Sample specific normalization (i.e. normalize by dry weight, volume)
   - Normalization by the sum
   - Normalization by the sample median
   - Normalization by a reference sample (probabilistic quotient normalization)[3]
   - Normalization by a pooled or average sample from a particular group
   - Normalization by a reference feature (i.e. creatinine, internal control)
   - Quantile normalization
2. Data transformation :
   - Generalized log transformation (glog 2)
   - Cube root transformation
3. Data scaling:
   - Mean centering (mean-centered only)
   - Auto scaling (mean-centered and divided by standard deviation of each variable)
   - Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)
   - Range scaling (mean-centered and divided by the value range of each variable)

Figure 1 shows the effects before and after normalization.

[3] Dieterle F, Ross A, Schlotterbeck G, Senn H. *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics*, 2006, Anal Chem 78 (13);4281 - 4290

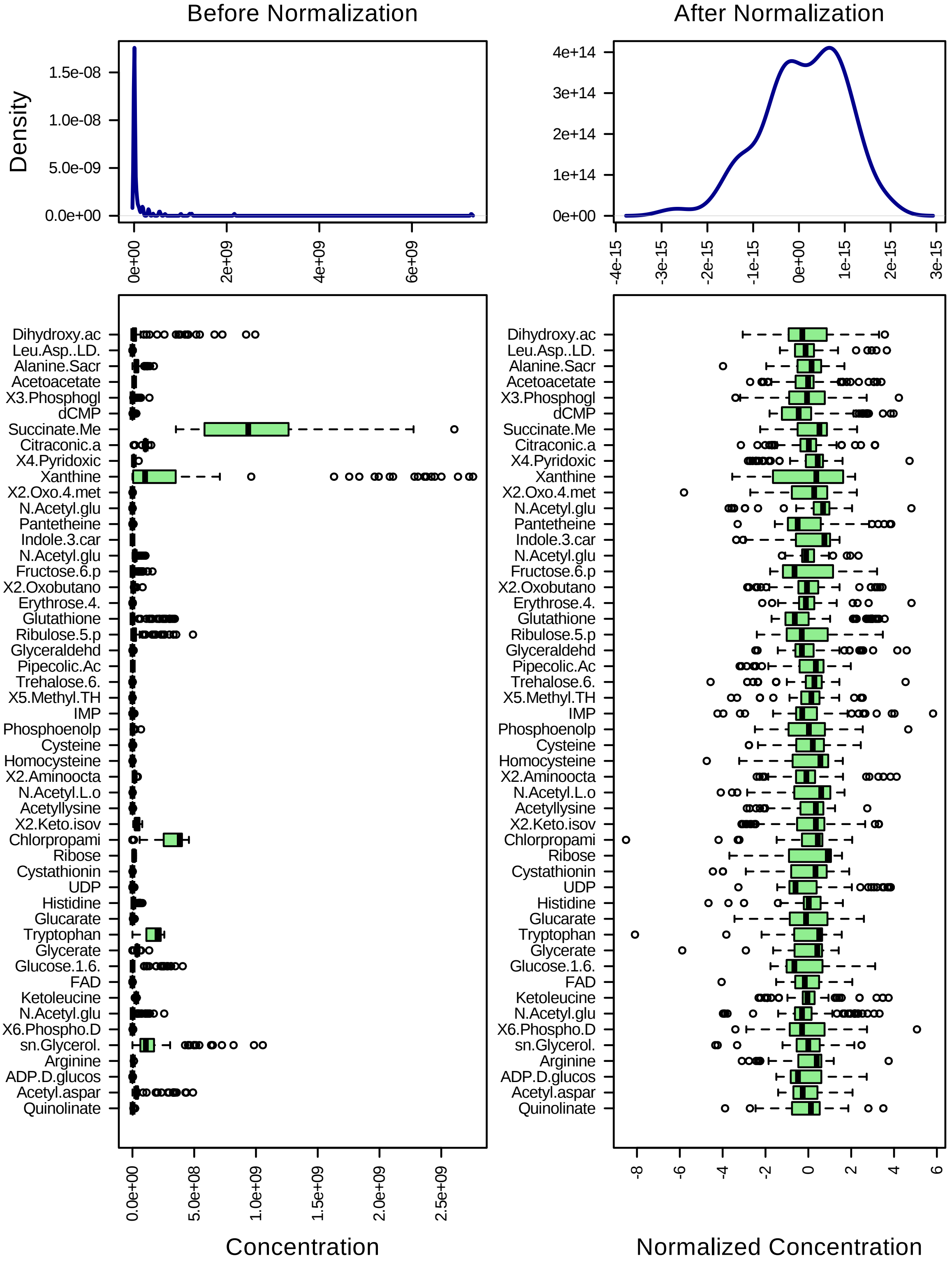

Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 50 features due to space limit. The density plots are based on all samples. Selected methods : Row-wise normalization: Quantile Normalization; Data transformation: Log Normalization; Data scaling: Pareto Scaling.

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 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"stat\", FALSE)"
 [2] "mSet<-Read.TextData(mSet, \"Replacing_with_your_file_path\", \"colu\", \"disc\");"
 [3] "mSet<-SanityCheckData(mSet)"
 [4] "mSet<-ReplaceMin(mSet);"
 [5] "mSet<-PreparePrenormData(mSet)"
 [6] "mSet<-Normalization(mSet, \"QuantileNorm\", \"LogNorm\", \"ParetoNorm\", ratio=FALSE, ratioNum=
 [7] "mSet<-PlotNormSummary(mSet, \"norm_0_\", \"png\", 72, width=NA)"
 [8] "mSet<-PlotSampleNormSummary(mSet, \"snorm_0_\", \"png\", 72, width=NA)"
 [9] "mSet<-SaveTransformedData(mSet)"
[10] "mSet<-PreparePDFReport(mSet, \"guest8221339122704546959\")\n"
```

The report was generated on Thu Dec 5 20:13:32 2019 with R version 3.6.1 (2019-07-05).