Data Normalization of 1H-NMR Metabolite Fingerprinting Datasets

About the Software

Data normalization in presence of unbalanced regulation

Data normalization is an essential part in NMR- and MS-based metabolomics. Done correctly it will lead to an improvement in data quality and a reduction of unwanted biases. However, the presence of unbalanced metabolic regulations, where the different cohorts under investigation do not contain approximately equal shares of up- and down-regulated features, may strongly influence data normalization and may lead to erroneous results. We recommend using the Shapiro-Wilk-Test for the detection of unbalanced regulation. In case of unbalanced regulation we recommend to use Linear Baseline Normalization, Probabilistic Quotient Normalization or Variance Stabilization Normalization in combination with variance based feature selection. We provide here an *R*-script that automatically performs feature selection followed by data normalization.

Download

Please download the R-script *here* **.**

Citation

When employing this tool please cite:

Hochrein J, Zacharias HU, Taruttis F, Samol C, Engelmann JC, Spang R, Oefner PJ & Gronwald W (2015): Data Normalization of ^{1}H-NMR Metabolite Fingerprinting Datasets in the Presence of Unbalanced Metabolite Regulation. J. Proteome Res., available online ahead of print, DOI: 10.1021/acs.jproteome.5b00192

System requirements

R-environment. It was tested with *R*-version 3.2.0

Installation instructions

There is no installation required; you can run the script within your *R*-environment.

Execution

With the script comes an text file explaining how to run the script and how to interpret the results. For a more detailed description please have a look at the corresponding publication.

Getting help

In case of trouble running the script, please read the commented *R*-code or contact the authors:

- jochen.hochrein@mol-med.uni-freiburg.de
- wolfram.gronwald@ur.de

Warranty

The Authors make no warranties expressed or implied, regarding the fitness of the software for any particular purpose. The authors claim no liability for data loss or other problems caused directly or indirectly by the software. The user is assuming the entire risk as to the software's quality and accuracy.

Running the software

Prerequisites:

Features in rows and samples in columns . Make sure that data do not contain any zeros . To obtain plots make sure that graphics are enabled for example by using Xming

To run the software:

```
source(file=" mswsd_resamp_publi.R") # this will install the necessary functions
```

to check for unbalanced regulation it will be analyzed whether total spectral areas are normally distributed (Shapiro-Wilk normality test)
use:

```
unbal_reg(my.data)
```

In case of unbalanced regulation you may want to perform normalizations without highly variable features. For this you have
to identify the amount of features to be excluded.

This is based on resampling of mswsd values
To do the resampling of the mswsd values use newly installed function

```
"resamp_mswsd(my.data)"
```

whwere ***my.data*** contains a data matrix of your metabolite data without zeros with features in rows and samples in
columns. Note, data should not be log transformed.
Results of the resampling approach will be given as a plot in PDF-formate ("*resamp_mswsd.pdf*").

From this plot identify manually the percentage of features where the mswsd values approach a stable value.
This value may then be used for subsequent data normalization for example 80 percent.
Here it is important that you do not reduce the amout of features too much so that in an extrem case only noise features remain.

Then run normalization with

```
norm.my.data<-norm_unbal(my.data,80,"VSN")
```

The first argument is your data, the second the percentage of features to be used and the third the normalization to be applied.

Available normalizations are

- linear baseline normalization based on mean values ("LBME"),
- linear baseline normalization based on median values (LBMD"),
- probabilistic quotient normalization ("PQN") and
- variance stabilization normalization ("VSN")

norm.my.data contains now the normalized data