GenePattern

RankNormalize Documentation

Description: Normalize the rows in a data set by rank.

Author: Pablo Tamayo, David Eby, gp-help@broadinstitute.org

Summary

Normalization is the process of reducing unwanted variation within a dataset. This module normalizes a data set by column rank using R's built-in "rank" function.

Ties in ranking are broken using the "average" method. That is, for a data set with N rows, each column will have its data points ranked from 1 (least) to N (greatest), with the score for any ties being averaged over those ties.

Example

Say you have a primitive microarray with only seven probes that you use to test two samples, getting the following results:

#1.2			
7	2		
Name	Description	Sample A	Sample B
probe1_at	gene1	1.2	7.0
probe2_at	gene2	0.8	4.4
probe3_at	gene3	3.5	3.7
probe4_at	gene4	8.4	9.0
probe5_at	gene5	0.5	0.4
probe6_at	gene6	1.2	7.0
probe7_at	gene7	1.2	6.5

Using this as input to RankNormalize with default settings produces the following output:

#1.2			
7	2		
Name	Description	Sample A	Sample B
probe1_at	gene1	4	5.5
probe2_at	gene2	2	3
probe3_at	gene3	6	2
probe4_at	gene4	7	7
probe5_at	gene5	1	1
probe6_at	gene6	4	5.5
probe7_at	gene7	4	4

The intensity values have been ranked from 1 to 7 in each column, with several ties ranked at 4 (for Sample A) and 5.5 (for Sample B).

Scaling

The scale of the results is controlled by the *scale to value* parameter. By default, for a data set with N rows, the results will be scaled to values in the interval [1, N], directly passing the row count as the *scale to value*. If the user sets this parameter, it will override that default value. For instance, the use of 1 as the *scale to value* will scale results on the interval [0, 1] (or more precisely, on the interval [1/N, N]). Using 100 will scale results to percentage values.

Parameters for Adjusting the Data Set Before Ranking

There are three additional parameters that can be used to constrain or adjust the data set before ranking:

- The *threshold* parameter sets the minimum for values in the data set. Any value below this will be increased to *threshold* before ranking.
- The *ceiling* parameter sets the maximum for values in the data set. Any value below this will be decreased to *ceiling* before ranking.
- The *shift* parameter gives an amount to adjust values in the data set. This *shift* adjustment value will be added to every value in the data set before ranking.

If more than one of these are specified, the order of precedence is *threshold*, then *ceiling*, then *shift*.

Requirements

The RankNormalize module requires R2.15.2 with the following packages:

- getopt_1.17
- optparse_0.9.5

These R packages have been bundled into a GenePattern patch and will be automatically downloaded and installed when the module is installed. R2.15.2 must be installed and configured independently.

References

R: A Language and Environment for Statistical Computing, Reference Index, Version 2.15.2 (2012-10-26), by The Core R Team.

Parameters

Name	Description
input file (required)	The dataset to be normalized in GCT or RES format.
output file name (required)	The name to be given to the output file (defaults to <input.file_basename>.NORM.<input.file_extension>)
scale to value (optional)	Result values will be scaled to this value by multiplication after normalization. Leaving this blank will give results scaled from 1 to N (where N is the number of rows).
threshold (optional)	Minimum threshold for values in the data set. Any data set value below this will be increased to the threshold value before normalization. If more than one of the *threshold*, *ceiling*, and *shift* parameters is specified, the order of precedence is *threshold*, then *ceiling,* then shift.
ceiling (optional)	Maximum ceiling for values in the data set. Any data set value above this will be decreased to the ceiling value before normalization. If more than one of the *threshold*, *ceiling*, and *shift* parameters is specified, the order of precedence is *threshold*, then *ceiling,* then *shift*.
shift (optional)	Shift all values in the data set. The shift value will be added to all data set values before normalization. If more than one of the *threshold*, *ceiling*, and *shift* parameters is specified, the order of precedence is *threshold*, then *ceiling*, then *shift*.

Input Files

1. <input file>
   A data set in GCT or RES file format to be normalized by rank.

Output Files

1. <output file name>
   The resulting normalized data set. The output format will match the input format, either GCT or RES. Any calls in a RES file will be maintained unchanged.
2. stdout.txt
   Logging output produced during the RankNormalize run.

Platform Dependencies

Module type:	Statistical Methods
CPU type:	Any
OS:	Any
Language:	R2.15.2

GenePattern Module Version Notes

Date	Version	Description
2/19/13	1	Initial version.