GSRI

October 5, 2010

`GSRI-package` *Gene Set Regulation Index (GSRI)*

Description

This package estimates the number of differentially expressed genes in gene sets with the Gene Set Regulation Index (GSRI).

Details

Package:	GSRI
Type:	Package
Version:	0.99.5
Date:	2010-04-20
License:	GPL-2
LazyLoad:	yes

This method calculates the number of differentially expressed genes in a gene set. It does not require a cut-off value for the distinction between regulated and unregulated genes. The approach is based on the fact that p-values obtained from a statistical test are uniformly distributed under the null hypothesis and accumulated around zero for the alternative hypothesis.

The Gene Set Regulation Index (GSRI) is the 5%-quantile of the distribution of the estimated number of differentially expressed genes obtained from bootstrapping the group samples. It indicates, that with a probability of 95% more than GSRI genes in the gene set are differentially expressed. This index can also be employed to test the hypothesis whether at least one gene in a set is regulated and to compare and rank the regulation of different gene sets, see Bartholom<c3><a9> et al. 2009.

Author(s)

Kilian Bartholome, Julian Gehring

Maintainer: Julian Gehring <julian.gehring@fdm.uni-freiburg.de>

References

Kilian Bartholom<c3><a9>, Clemens Kreutz, and Jens Timmer: Estimation of gene induction enables a relevance-based ranking of gene sets, Journal of Computational Biology: A Journal

of Computational Molecular Cell Biology 16, no. 7 (July 2009): 959-967. http://www.liebertonline.com/doi/abs/10.1089/cmb.2008.0226

Functions of the following packages are used in this package:

Korbinian Strimmer (2009). fdrtool: Estimation and Control of (Local) False Discovery Rates. R package version 1.2.6. http://CRAN.R-project.org/package=fdrtool

R. Gentleman, V. Carey, W. Huber and F. Hahne. genefilter: methods for filtering genes from microarray experiments. R package version 1.28.2.

See Also

gsri

gsriFromFile

Examples

```
## Simulate expression data for a gene set of:
## 100 genes, 40 samples (20 treatment, 20 control)
## and 30 regulated genes
expdata <- matrix(rnorm(4000,mean=0),nrow=100,ncol=40)
expdata[1:30,1:20] <- rnorm(600,mean=1)
data <- data.frame(expdata)
phenotype <- c(rep(0,20),rep(1,20))
geneSetName <- "Test Gene Set"

## Estimate the number of differentially expressed genes
res <- gsri(data, phenotype, geneSetName)


## Read expression data, phenotypes and gene sets from files
## Input files can be found in the vignette directory of this package
dataFileName <- system.file("extdata", "data.gct", package="GSRI")
phenotypeFileName <- system.file("extdata", "phenotype.cls", package="GSRI")
geneSetFileName <- system.file("extdata", "geneSets.gmt", package="GSRI")

res2 <- gsriFromFile(dataFileName, phenotypeFileName, geneSetFileName)
```

`gsri` *Gene Set Regulation Index (GSRI)*

Description

Estimates the number of differentially expressed genes for a single gene set.

Usage

```
gsri(data, phenotype, geneSetName,
     useGrenander = FALSE, plotResults = TRUE, writeResults = FALSE,
     nBootstraps = 100, test = "ttest", testArgs = NULL, alpha = 0.05)
```

Arguments

`data`	A data frame or matrix of size n x m containing the gene expression dataset with rows representing genes and columns samples.
`phenotype`	Vector of size m containing the phenotypes of the samples. If not already a factorial variable it will be internally converted to a factor.
`geneSetName`	Character string with the name of the gene set. Used for identification of the data set, has no influence on the calculation.
`useGrenander`	Logical indicating whether the grenander estimate from the **fdrtool** package should be used additionally (default: FALSE). For details see the Notes section below.
`plotResults`	Logical indicating whether the results should be plotted (default: FALSE). The plot shows the cumulative density function of calculated p-values, the fit of the uniform distribution, the estimated number of regulated genes and the estimated GSRI for each gene set.
`writeResults`	Logical indicating whether a list of the estimated regulated genes for each gene set should be written to a text file in the working directory (default: FALSE). The file will be named 'GeneSet_#geneSetName#_data.txt', with #geneSetName# taken from the 'geneSetName' argument.
`nBootstraps`	Number of bootstrap samples to be drawn (default: 100)
`test`	Character string or function name to specify the statistical test to calculate p-values for effect between groups. Groups are defined by the 'phenotype' argument. In this package both a t-test (default: "ttest") and an F-test ("ftest") between groups are implemented and can be chosen with the according character string. A user-defined function can also be passed as an argument in order to specify own test statistics. For details see the Notes section below.
`testArgs`	Optional arguments passed to the function 'test' if specified (default: NULL)
`alpha`	Significance level for bootstrap (default: 0.05). The resulting GSRI will be the (1-alpha)*100% confidence interval. If a vector of values is given, GSRI will be calculated for all values and passed to the output argument. Plots and file outputs will only contain GSRI for the first value in the vector to simplify output.

Details

This function calculates the number of differentially expressed genes for a single gene set, with data and gene set taken from the workspace.

From bootstrapping the group samples the (1-alpha)*100% quantile of the distribution of the estimated number of differentially expressed genes is obtained. The Gene Set Regulation Index (GSRI) is defined as the 5% quantile and indicates, that with a probability of 95% more than GSRI genes in the gene set are differentially expressed.

This index can be employed to test the hypothesis whether at least one gene in a set is regulated and to compare and rank the regulation of different gene sets.

Value

A list with components:

`geneSet`	Name of gene set
`percRegGenes`	Estimated percentage of differentially expressed genes

`percRegGenesSd`
: Estimated standard deviation of the percentage of differentially expressed genes

`numRegGenes`
: Estimated number of differentially expressed genes

`numRegGenesSd`
: Estimated standard deviation of the number of differentially expressed genes

`gsri`
: Gene Set Regulation Index

`nGenes`
: Number of genes in gene set

Note

Usage of the Grenander estimate is based on the assumption about the concave shape of the cumulative density distribution. It reduces the variance, i.e. makes the approach more stable especially for small gene sets. On the downside the number of significant genes is overestimated for few regulated genes. A conservative solution of this trade-off would be to rank the gene-sets with and without Grenander estimate and to choose the lowest rank for each gene-set. Please note that the Grenander estimate does not allow duplicates in the p-values. If this occurs in a data set, an error message will be printed and the analysis should be repeated without the Grenander estimate.

With the t-test and the F-test, two widely used statistical tests are available in this package. To allow fast computation this package uses the implementations from the **genefilter** package.

It is also possible to apply user-defined tests with this method. In this case the function has to be called by function(data, phenotype, testArgs). 'data' and 'phenotype' are of class 'matrix' and 'factor', respectively. 'testArgs' will only be passed to the function if it is defined. In general all methods that compute p-values are suitable. The function has to return a vector with one p-value per gene. For details on how to use your own test functions please refer to the vignette of this package.

Author(s)

Kilian Bartholom<c3><a9>, Julian Gehring

See Also

`gsriFromFile`

`GSRI`

Examples

```
## Simulate expression data for a gene set of:
## 100 genes, 40 samples (20 treatment, 20 control)
## and 30 regulated genes
expdata <- matrix(rnorm(4000,mean=0),nrow=100,ncol=40)
expdata[1:30,1:20] <- rnorm(600,mean=1)
data <- data.frame(expdata)
phenotype <- c(rep(0,20),rep(1,20))
geneSetName <- "Test Gene Set"

## Estimate the number of differentially expressed genes
res <- gsri(data, phenotype, geneSetName)
```

`gsriFromFile` *Gene Set Regulation Index (GSRI)*

Description

Estimates the number of differentially expressed genes for a list of gene sets, with data and gene sets read from files.

Usage

```
gsriFromFile(dataFileName, phenotypeFileName, geneSetFileName,
             useGrenander = FALSE, plotResults = FALSE,
             writeResults = FALSE, writeSummary = FALSE,
             minGeneSetSize = 10, nBootstraps = 100, test = "ttest",
             testArgs = NULL, alpha = 0.05, verbose = TRUE)
```

Arguments

`dataFileName`	Data file of type '.gct' containing the expression data. For details on the format see the Notes section below.
`phenotypeFileName`	Phenotype file of type '.cls' specifying the phenotypes of the samples. For details on the format see the Notes section below.
`geneSetFileName`	Gene set file of type '.gmt' containing a list of gene sets. For details on the format see the Notes section below.
`useGrenander`	Logical indicating whether the grenander estimate from the **fdrtool** package should be used additionally (default: FALSE). For details see the Notes section below.
`plotResults`	Logical indicating whether the results should be plotted (default: FALSE). The plot shows the cumulative density function of calculated p-values, the fit of the uniform distribution, the estimated number of regulated genes and the estimated GSRI for each gene set.
`writeResults`	Logical indicating whether a list of the estimated regulated genes for each gene set should be written to a text file in the working directory (default: FALSE). The file will be named 'GeneSet_#geneSetName#_data.txt', with #geneSetName# taken from the 'geneSetName' argument.
`writeSummary`	Logical indicating whether a summary of results for all gene sets should be written into a single file (default: FALSE). The file will be named '#dataFileName#_results.txt', with #dataFileName# taken from the file name of the 'dataFileName' argument.
`minGeneSetSize`	Integer specifying the minimal gene set size (default: 10). Only gene sets having at least this size will be used for estimation.
`nBootstraps`	Number of bootstrap samples to be drawn (default: 100)
`test`	Character string or function name to specify the statistical test to calculate p-values for effect between groups. Groups are defined by the 'phenotype' argument. In this package both a t-test (default: "ttest") and an F-test ("ftest")

between groups are implemented and can be chosen with the according character string. An user-defined function can also be passed as an argument in order to specify own test statistics. For details see the Notes section below.

`testArgs` Optional arguments passed to the function 'test' if specified (default: NULL)

`alpha` Significance level for bootstrap (default: 0.05). The resulting GSRI will be the (1-alpha)*100% confidence interval. If a vector of values is given, GSRI will be calculated for all values and passed to the output argument. Plots and file outputs will only contain GSRI for the first value in the vector to simplify output.

`verbose` Logical indicating whether information about data loading and gene set computation should be printed on screen (default: TRUE).

Details

This function calculates the number of differentially expressed genes for a list of gene sets, with data and gene sets read from files.

From bootstrapping the group samples the (1-alpha)*100% quantile of the distribution of the estimated number of differentially expressed genes is obtained. The Gene Set Regulation Index (GSRI) is defined as the 5% quantile and indicates, that with a probability of 95% more than GSRI genes in the gene set are differentially expressed.

This index can be employed to test the hypothesis whether at least one gene in a set is regulated and to compare and rank the regulation of different gene sets.

Value

A list with components (one entry per gene set):

`geneSet` Names of gene sets

`percRegGenes` Estimated percentage of differentially expressed genes

`percRegGenesSd` Estimated standard deviation of the percentage of differentially expressed genes

`numRegGenes` Estimated number of differentially expressed genes

`numRegGenesSd` Estimated standard deviation of the number of differentially expressed genes

`gsri` Gene Set Regulation Index

`nGenes` Number of genes in gene sets

`geneSetGenes` Names of genes in gene sets.

Note

The input files should have the format typical for those file types. Details on the specific formats can be found at http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats. For more details and example files please refer to the vignette of this package.

Usage of the Grenander estimate is based on the assumption about the concave shape of the cumulative density distribution. It reduces the variance, i.e. makes the approach more stable especially for small gene sets. On the downside the number of significant genes is overestimated for few regulated genes. A conservative solution of this trade-off would be to rank the gene-sets with and without Grenander estimate and to choose the lowest rank for each gene-set. Please note that the Grenander estimate does not allow duplicates in the p-values. If this occurs in a data set, an error message will be printed and the analysis should be repeated without the Grenander estimate.

With the t-test and the F-test, two widely used statistical tests are available in this package. To allow fast computation this package uses the implementations from the **genefilter** package.

It is also possible to apply user-defined tests with this method. In this case the function has to be called by function(data, phenotype, testArgs). 'data' and 'phenotype' are of class 'matrix' and 'factor', respectively. 'testArgs' will only be passed to the function if it is defined. In general all methods that compute p-values are suitable. The function has to return a vector with one p-value per gene. For details on how to use your own test functions please refer to the vignette of this package.

Author(s)

Kilian Bartholom<c3><a9>, Julian Gehring

See Also

gsri

GSRI

Examples

```
## Read expression data, phenotypes and gene sets from files
## Input files can be found in the vignette directory of this package
dataFileName <- system.file("extdata", "data.gct", package="GSRI")
phenotypeFileName <- system.file("extdata", "phenotype.cls", package="GSRI")
geneSetFileName <- system.file("extdata", "geneSets.gmt", package="GSRI")

res <- gsriFromFile(dataFileName, phenotypeFileName, geneSetFileName)
```

Index