Package ‘QuaternaryProd’

April 12, 2018

Type Package

Title Computes the Quaternary Dot Product Scoring Statistic for Signed and Unsigned Causal Graphs

Version 1.6.0

Date 2015-10-22

Description QuaternaryProd is an R package that performs causal reasoning on biological networks, including publicly available networks such as STRINGdb. QuaternaryProd is an open-sorce alternative to commercial products such as Quiagen and Inginuity pathway analysis. For a given a set of differentially expressed genes, QuaternaryProd computes the significance of upstream regulators in the network by performing causal reasoning using the Quaternary Dot Product Scoring Statistic (Quaternary Statistic), Ternary Dot product Scoring Statistic (Ternary Statistic) and Fisher's exact test. The Quaternary Statistic handles signed, unsigned and ambiguous edges in the network. Ambiguity arises when the direction of causality is unknown, or when the source node (e.g., a protein) has edges with conflicting signs for the same target gene. On the other hand, the Ternary Statistic provides causal reasoning using the signed and unambiguous edges only. The Vignette provides more details on the Quaternary Statistic and illustrates an example of how to perform causal reasoning using STRINGdb.

License GPL (>=3)

biocViews GraphAndNetwork, GeneExpression, Transcription

Depends R (>= 3.2.0), Rcpp (>= 0.11.3), dplyr, rlist

Suggests knitr

LinkingTo Rcpp

LazyData true

VignetteBuilder knitr

RoxygenNote 6.0.1

NeedsCompilation yes

Author Carl Tony Fakhry [cre, aut],
Ping Chen [ths],
Kourosh Zarringhalam [aut, ths]

Maintainer Carl Tony Fakhry <cfakhry@cs.umb.edu>

R topics documented:

`QuaternaryProd-package` *Computes the Quaternary Dot Product Scoring Statistic for Signed and Unsigned Causal Graphs*

Description

QuaternaryProd is an R package that performs causal reasoning on biological networks, including publicly available networks such as STRINGdb. QuaternaryProd is an open-source alternative to commercial products such as Quiagen and Inginuity pathway analysis. For a given a set of differentially expressed genes, QuaternaryProd computes the significance of upstream regulators in the network by performing causal reasoning using the Quaternary Dot Product Scoring Statistic (Quaternary Statistic), Ternary Dot product Scoring Statistic (Ternary Statistic) and Fisher's exact test. The Quaternary Statistic handles signed, unsigned and ambiguous edges in the network. Ambiguity arises when the direction of causality is unknown, or when the source node (e.g., a protein) has edges with conflicting signs for the same target gene. On the other hand, the Ternary Statistic provides causal reasoning using the signed and unambiguous edges only. The Vignette provides more details on the Quaternary Statistic and illustrates an example of how to perform causal reasoning using STRINGdb.

Details

Package:	QuaternaryProd
Type:	Package
Version:	1.5.1
Date:	2015-10-22
License:	GPL (>= 2)

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

Maintainer: Carl Tony Fakhry <cfakhry@cs.umb.edu>

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

`QP_Pmf` *Computes the probability mass function of the scores.*

Description

This function computes the probability mass function for the Quaternary Dot Product Scoring Statistic for signed causal graphs. This includes scores with probabilities strictly greater than zero.

Usage

```
QP_Pmf(q_p, q_m, q_z, q_r, n_p, n_m, n_z, epsilon = 1e-16)
```

Arguments

`q_p`	Expected number of positive predictions.
`q_m`	Expected number of negative predictions.
`q_z`	Expected number of nil predictions.
`q_r`	Expected number of regulated predictions.
`n_p`	Number of positive predictions from experiments.
`n_m`	Number of negative predictions from experiments.
`n_z`	Number of nil predictions from experiments.
`epsilon`	parameter for thresholding probabilities of matrices. Default value is 1e-16.

Details

This function computes the probability for each score in the support of the distribution. The returned value is a vector of probabilities where the returned vector has names set equal to the corresponding scores.

Setting epsilon to zero will compute the probability mass function without ignoring any matrices with probabilities smaller than epsilon*D_max (D_max is the numerator associated with the matrix of highest probability for the given constraints). The default value of 1e-16 is experimentally validated to be a very reasonable threshold. Setting the threshold to higher values which are smaller than 1 will lead to understimating the probabilities of each score since more tables will be ignored.

Value

Vector of probabilities for scores in the support.

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

See Also

QP_Pvalue, QP_Support

Examples

```
# Compute the probability mass function of the Quaternary Dot
# Product Scoring Statistic for the given table margins.
pmf <- QP_Pmf(50,50,50,0,50,50,50)
```

QP_Probability *Computes the probability of a score.*

Description

This function computes the probability of a score in the Quaternary Dot Product scoring distribution.

Usage

```
QP_Probability(score, q_p, q_m, q_z, q_r, n_p, n_m, n_z, epsilon = 1e-16)
```

Arguments

score	The score for which the probability will be computed.
q_p	Expected number of positive predictions.
q_m	Expected number of negative predictions.
q_z	Expected number of nil predictions.
q_r	Expected number of regulated predictions.
n_p	Number of positive predictions from experiments.
n_m	Number of negative predictions from experiments.
n_z	Number of nil predictions from experiments.
epsilon	Threshold for probabilities of matrices. Default value is 1e-16.

Details

Setting epsilon to zero will compute the probability mass function without ignoring any matrices with probabilities smaller than epsilon*D_max (D_max is the numerator associated with the matrix of highest probability for the given constraints). The default value of 1e-16 is experimentally validated to be a very reasonable threshold. Setting the threshold to higher values which are smaller than 1 will lead to understimating the probabilities of each score since more tables will be ignored.

For computing p-values, the user is advised to use the p-value function which is optimized for such purposes.

Value

This function returns a numerical value, where the numerical value is the probability of the score.

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

See Also

QP_Pmf, QP_Pvalue, QP_SigPvalue

Examples

```
# Computing The probability of score 50
# for the given table margins.
prob <- QP_Probability(0,50,50,50,0,50,50,50)
```

QP_Pvalue *Computes the p-value of a score.*

Description

This function computes the right sided p-value for the Quaternary Dot Product Scoring Statistic.

Usage

```
QP_Pvalue(score, q_p, q_m, q_z, q_r, n_p, n_m, n_z, epsilon = 1e-16)
```

Arguments

`score`	The score for which the p-value will be computed.
`q_p`	Expected number of positive predictions.
`q_m`	Expected number of negative predictions.
`q_z`	Expected number of nil predictions.
`q_r`	Expected number of regulated predictions.
`n_p`	Number of positive predictions from experiments.
`n_m`	Number of negative predictions from experiments.
`n_z`	Number of nil predictions from experiments.
`epsilon`	Threshold for probabilities of matrices. Default value is 1e-16.

Details

Setting epsilon to zero will compute the probability mass function without ignoring any matrices with probabilities smaller than epsilon*D_max (D_max is the numerator associated with the matrix of highest probability for the given constraints). The default value of 1e-16 is experimentally validated to be a very reasonable threshold. Setting the threshold to higher values which are smaller than 1 will lead to understimating the probabilities of each score since more tables will be ignored.

Value

This function returns a numerical value, where the numerical value is the p-value of the score.

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

See Also

`QP_SigPvalue`

Examples

```
# Computing The p-value of score 50
# for the given table margins.
pval <- QP_Pvalue(50,50,50,50,0,50,50,50)
```

QP_SigPvalue *Computes the p-value for a statistically significant score.*

Description

This function computes the right sided p-value for the Quaternary Dot Product Scoring Statistic for statistically significant scores.

Usage

```
QP_SigPvalue(score, q_p, q_m, q_z, q_r, n_p, n_m, n_z, epsilon = 1e-16, sig_level = 0.05)
```

Arguments

score	The score for which the p-value will be computed.
q_p	Expected number of positive predictions.
q_m	Expected number of negative predictions.
q_z	Expected number of nil predictions.
q_r	Expected number of regulated predictions.
n_p	Number of positive predictions from experiments.
n_m	Number of negative predictions from experiments.
n_z	Number of nil predictions from experiments.
epsilon	Threshold for probabilities of matrices. Default value is 1e-16.
sig_level	Significance level of test hypothesis. Default value is 0.05.

Details

Setting epsilon to zero will compute the probability mass function without ignoring any matrices with probabilities smaller than epsilon*D_max (D_max is the numerator associated with the matrix of highest probability for the given constraints). The default value of 1e-16 is experimentally validated to be a very reasonable threshold. Setting the threshold to higher values which are smaller than 1 will lead to understimating the probabilities of each score since more tables will be ignored. If the score is not statistically significant, then a value of -1 will be returned.

Value

This function returns a numerical value, where the numerical value is the p-value of a score if the score is statistically significant otherwise it returns -1.

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

See Also

QP_Pvalue

Examples

```
# Computing The p-value of score 50
# for the given table margins.
pval <- QP_SigPvalue(50,50,50,50,0,50,50,50)
```

QP_Support *Computes the support for the scores.*

Description

This function computes the support of the Quaternary Dot Product Scoring distribution for signed causal graphs. This includes all scores which have probabilities strictly greater than 0.

Usage

```
QP_Support(q_p, q_m, q_z, q_r, n_p, n_m, n_z)
```

Arguments

q_p	Expected number of positive predictions.
q_m	Expected number of negative predictions.
q_z	Expected number of nil predictions.
q_r	Expected number of regulated predictions.
n_p	Number of positive predictions from experiments.
n_m	Number of negative predictions from experiments.
n_z	Number of nil predictions from experiments.

Value

Integer vector of support.

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

Examples

```
# Compute the support of the Quaternary Dot Product Scoring distribution with the given margins.
QP_Support(50,50,50,0,50,50,50)
```

`RunCRE_HSAStringDB` *This function runs a causal relation engine by computing the Quaternary Dot Product Scoring Statistic, Ternary Dot Product Scoring Statistic or the Enrichment test over the Homo Sapien STRINGdb causal network.*

Description

This function runs a causal relation engine by computing the Quaternary Dot Product Scoring Statistic, Ternary Dot Product Scoring Statistic or the Enrichment test over the Homo Sapien STRINGdb causal network.

Usage

```
RunCRE_HSAStringDB(gene_expression_data, method = "Quaternary",
                    fc.thresh = log2(1.3), pval.thresh = 0.05,
                   only.significant.pvalues = FALSE, significance.level = 0.05)
```

Arguments

`gene_expression_data`
: A data frame for gene expression data. The `gene_expression_data` data frame must have three columns `entrez`, `fc` and `pvalue`. `entrez` denotes the entrez id of a given gene, `fc` denotes the fold change of a gene, and `pvalue` denotes the p-value. The `entrez` column must be of type integer or character, and the `fc` and `pvalue` columns must be numeric values.

`method`
: Choose one of `Quaternary`, `Ternary` or `Enrichment`. Default is `Quaternary`.

`fc.thresh`
: Threshold for fold change in `gene_expression_data` data frame. Any row in gene_expression_data with abosolute value of `fc` smaller than `fc.thresh` will be ignored. Default value is `fc.thresh = log2(1.3)`.

`pval.thresh`
: Threshold for p-values in `gene_expression_data` data frame. All rows in `gene_expression_data` with p-values greater than `pval.thresh` will be ingnored. Default value is `pval.thresh = 0.05`.

only.significant.pvalues
: If only.significant.pvalues = TRUE then only p-values for statistically significant regulators are computed otherwise uncomputed p-values are set to -1. The default value is only.significant.pvalues = FALSE.

significance.level
: When only.significant.pvalues = TRUE, only p-values which are less than or equal to significance.level are computed. The default value is significance.level = 0.05.

Value

This function returns a data frame containing parameters concerning the method used. The p-values of each of the regulators is also computed, and the data frame is in increasing order of p-values of the goodness of fit score for the given regulators. The column names of the data frame are:

- uid The regulator in the STRINGdb network.
- symbol Symbol of the regulator.
- regulation Direction of regulation of the regulator.
- correct.pred Number of correct predictions in gene_expression_data when compared to predictions made by the network.
- incorrect.pred Number of incorrect predictions in gene_expression_data when compared to predictions made by the network.
- score The number of correct predictions minus the number of incorrect predictions.
- total.reachable Total Number of children of the given regulator.
- significant.reachable Number of children of the given regulator that are also present in gene_expression_data.
- total.ambiguous Total number of children of the given regulator which are regulated by the given regulator without knowing the direction of regulation.
- significant.ambiguous Total number of children of the given regulator which are regulated by the given regulator without knowing the direction of regulation and are also present in gene_expression_data.
- unknown Number of target nodes in the STRINGdb causal network which do not interact with the given regulator.
- pvalue P-value of the score computed according to the selected method.

Author(s)

Carl Tony Fakhry, Ping Chen and Kourosh Zarringhalam

References

Carl Tony Fakhry, Parul Choudhary, Alex Gutteridge, Ben Sidders, Ping Chen, Daniel Ziemek, and Kourosh Zarringhalam. Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics, 17:318, 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1181-8.

Franceschini, A (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. In:'Nucleic Acids Res. 2013 Jan;41(Database issue):D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29'.

Examples

```
# Get gene expression data
e2f3 <- system.file("extdata", "e2f3_sig.txt", package = "QuaternaryProd")
e2f3 <- read.table(e2f3, sep = "\t", header = TRUE, stringsAsFactors = FALSE)

# Rename column names appropriately and remove duplicated entrez ids
names(e2f3) <- c("entrez", "pvalue", "fc")
e2f3 <- e2f3[!duplicated(e2f3$entrez),]

# Compute the Quaternary Dot Product Scoring statistic for statistically significant
# regulators in the STRINGdb network
quaternary_results <- RunCRE_HSAStringDB(e2f3, method = "Quaternary",
                             fc.thresh = log2(1.3), pval.thresh = 0.05,
                             only.significant.pvalues = TRUE)
# Get FDR corrected p-values
quaternary_results["qvalue"] <- p.adjust(quaternary_results$pvalue, method = "fdr")
quaternary_results[1:4, c("uid","symbol","regulation","pvalue","qvalue")]
```

Index