Package ‘cp4p’

February 24, 2019

Type Package

Title Calibration Plot for Proteomics

Version 0.3.6

Date 2019-02-22

Author Quentin Giai Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Couté, Thomas Burger

Maintainer Quentin Giai Gianetto <quentin2g@yahoo.fr>

Description Functions to check whether a vector of p-values respects the assumptions of FDR (false discovery rate) control procedures and to compute adjusted p-values.

License GPL-3

Depends R (>= 3.2.0), MESS, graphics, stats, multtest, qvalue, limma

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2019-02-24 15:50:07 UTC

R topics documented:

cp4p-package *Introduction to the CP4P package*

Description

This package provides tools to check whether a vector of p-values respects the assumptions of classical FDR (false discovery rate) control procedures.

It is built to be easily used by non-statisticians in the context of quantitative proteomics (yet, it can be applied in other contexts).

Concretely, it allows estimating the proportion of true null hypotheses (i.e. proportion of non-differentially abundant proteins or peptides in a relative quantification experiment), as well as checking whether the p-values are adequately distributed for further FDR control.

In addition, the package allows performing an adequately chosen adaptive FDR control procedure to get adjusted p-values.

A tutorial giving a practical introduction to this package is available in the supplementary material of Giai Gianetto et al. (2016).

Details

Package:	cp4p
License:	GPL-3
Depends:	multtest, qvalue, limma, MESS, graphics, stats

This package is composed of three functions that take as input a vector of p-values resulting from multiple two-sided hypothesis testing (such as multiple t-tests for equal means for instance).

First, the function `estim.pi0` allows determining the proportion of true null hypotheses among the set of tests using eight different estimation methods proposed in literature.

Second, the function `calibration.plot` proposes an intuitive plot of the p-values, so as to visually assess their behavior and well-calibration.

Third, the function `adjust.p` allows obtaining adjusted p-values in view to perform an adaptive FDR control from a chosen level.

Two proteomic datasets named `LFQRatio2` and `LFQRatio25` allow to use these functions in a concrete framework where the proportion of non-differentially abundant proteins is known.

Author(s)

Quentin Giai Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Couté, Thomas Burger

Maintainer: Quentin Giai Gianetto <quentin2g@yahoo.fr>

References

Giai Gianetto, Q., Combes, F., Ramus, C., Bruley, C., Couté, Y., Burger, T. (2016). Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. Proteomics, 16(1), 29-32.

`adjust.p` *Adjusted p-values for adaptive FDR control*

Description

This function computes adjusted p-values for adaptive FDR control from a vector of raw (unadjusted) p-values.

Usage

```
adjust.p(p, pi0.method = 1, alpha = 0.05, nbins = 20, pz = 0.05)
```

Arguments

`p` Numeric vector of raw p-values. Raw p-values are assumed without missing values, and between 0 and 1.

`pi0.method` Numeric value between 0 and 1 corresponding to the proportion of true null hypotheses (non-differentially abundant proteins or peptides), or the name of an estimation method for this proportion among `"st.boot"`, `"st.spline"`, `"langaas"`, `"jiang"`, `"histo"`, `"pounds"`, `"abh"` or `"slim"` (see function `estim.pi0` for details). The two-stage Benjamini and Hochberg procedure (Benjamini et al. (2006)) is also available according to an expected FDR given by the `alpha` parameter (write `pi0.method="bky"`). Default is 1 (classical Benjamini and Hochberg procedure (1995) is performed in this case).

`alpha` A nominal type I error rate used for estimating the proportion of true null hypotheses (non-differentially abundant proteins or peptides) in the two-stage Benjamini and Hochberg procedure (used only if `pi0.method="bky"`). Default is 0.05.

`nbins` Number of bins. Parameter used for the `"jiang"` and `"histo"` methods in `estim.pi0`. Default is 20.

`pz` P-value threshold such as p-values below are associated to false null hypotheses. Used for the `"slim"` method in `estim.pi0`. Default is 0.05.

Details

The procedure uses an estimation of the proportion of true null hypotheses (non-differentially abundant proteins or peptides), the value or the name of which is precised in input. Next, this estimation is multiplied by the adjusted p-values of the Benjamini and Hochberg procedure (1995) to obtain the final adjusted p-values (see section 3 in Craiu and Sun (2008) for details).

The adjusted p-values of the Benjamini and Hochberg procedure (1995) and of the two-stage Benjamini and Hochberg procedure (Benjamini et al. (2006)) are computed using the R package `multtest` (Pollard et al. (2005)).

Value

A list composed of :

pi0 The proportion of true null hypotheses (non-differentially abundant proteins or peptides) used to adjust p-values.

adjp A matrix of raw and adjusted p-values with rows corresponding to each test. First column corresponds to raw p-values and second column to adjusted p-values.

Author(s)

Quentin Giai Gianetto <quentin2g@yahoo.fr>

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289-300, 1995.

Y. Benjamini, A.M. Krieger, and D.Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93(3):491-507, 2006.

R.V. Craiu and L. Sun. Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. Statistica Sinica, 18:861-879, 2008.

K.S. Pollard, S. Dudoit and M.J. van der Laan. Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer. 2005.

See Also

estim.pi0, calibration.plot

Examples

```
#get p-values
data(LFQRatio2)
p=LFQRatio2[,7]

#adjust p-values by estimating the proportion of true null hypotheses
#with the "pounds" method.
res_pounds=adjust.p(p, pi0.method = "pounds")

#proportion of true null hypotheses with the "pounds" method.
res_pounds$pi0

#plot ajusted p-values in function of raw p-values
plot(res_pounds$adjp)

#adjust p-values by estimating the proportion of true null hypotheses
#using the two-stage Benjamini and Hochberg procedure with a FDR of 0.1.
res_bky=adjust.p(p, pi0.method = "bky", alpha = 0.1)
```

```
#proportion of true null hypotheses with the two-stage BH procedure.
res_bky$pi0

#plot adjusted p-values in function of raw p-values
plot(res_bky$adjp)

#compare the two-stage Benjamini and Hochberg procedure
#with the "pounds" method
plot(res_pounds$adjp[,2],res_bky$adjp[,2])
```

`calibration.plot` *Displaying the "Calibration Plot" of a vector of p-values.*

Description

From a proteomics viewpoint, this function displays a graph (the "Calibration Plot") which allows to visually assess the compliance of a differential abundance analysis with FDR control procedure assumptions.

From a statistical viewpoint, this function performs a plot of the cumulative distribution function of 1-p-values. It allows checking whether p-values respect several assumptions of FDR control procedures.

Usage

```
calibration.plot(p, pi0.method = "pounds", nbins = 20, pz = 0.05)
```

Arguments

`p` Numeric vector of raw p-values. Raw p-values are assumed without missing values, and between 0 and 1.

`pi0.method` Numeric value between 0 and 1 corresponding to the proportion of true null hypotheses if known by the user, or the name of an estimation method among `"st.boot"`, `"st.spline"`, `"langaas"`, `"jiang"`, `"histo"`, `"pounds"`, `"abh"` or `"slim"` (see function `estim.pi0` for details). Default is `"pounds"`. If `pi0.method="ALL"`, a plot allowing the comparison of the eight estimation methods is displayed.

`nbins` Number of bins. Parameter used for the `"jiang"` and `"histo"` methods in `estim.pi0`. Default is 20.

`pz` P-value threshold such as p-values below are associated to false null hypotheses. Used for the `"slim"` method in `estim.pi0`. Default is 0.05.

Details

This function provides a graph which displays the cumulative distribution function of 1-p-values as a function of 1-p-values (black curve) as advocated by Schweder and Spjotvoll (1982).

The blue straight line has a slope equals to the proportion of true null hypotheses (estimated by `estim.pi0`) that is recalled in the caption of the plot. It is close to the black curve for small 1-pvalues if the p-values are independently and uniformly distributed under the null hypothesis.

In addition, two other measures are given in the caption of the graphic. Each has a color that matches that of various areas of the plot and should be carefully consider to assess the well-calibration of p-values (see Giai Gianetto et al. (2016) for details).

The first measure corresponds to one minus the ratio between the green area and the grey area (referred to as "differentially abundant protein concentration"). The closer to 100% this measure is, the smaller the false nondiscovery rate is expected.

The second measure corresponds to the total red area observed on the graph (referred to as "uniformity underestimation"). The smaller this measure is, the more the proportion of true null hypotheses is expected to be not under-estimated (so as to get a conservative p-value adjustment).

Supplementary theoretical justifications on these measures can be found in the tutorial available in the supplementary material of Giai Gianetto et al. (2016).

Value

A list composed of :

`pi0` Numeric value corresponding to the proportion of true null hypotheses (non-differentially abundant proteins or peptides) used for the plot. Numeric vector if `pi0.method="ALL"`.

`h1.concentration` Numeric value corresponding to one minus the ratio between the green area and the grey area. NULL if `pi0.method="ALL"`.

`unif.under` Numeric value corresponding to the total red area observed on the graph (multiplied by 100). NULL if `pi0.method="ALL"`.

Author(s)

Quentin Giai Gianetto <quentin2g@yahoo.fr>

References

Giai Gianetto, Q., Combes, F., Ramus, C., Bruley, C., Couté, Y., Burger, T. (2016). Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. Proteomics, 16(1), 29-32.

Schweder, T., Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. Biometrika, 69(3), 493-502.

See Also

`estim.pi0`

Examples

```
#get p-values
data(LFQRatio25)
p=LFQRatio25[,7]

#Plot straight lines whose slopes correspond to different estimates of
#the proportion of true null hypotheses
r=calibration.plot(p, pi0.method="ALL")
r$pi0

#Plot of the graph with the "pounds" method (default)
r=calibration.plot(p)
#Estimate of the proportion of true null hypotheses
r$pi0
#Estimate of the differentially abundant protein concentration
#(the closer to one, the better)
r$h1.concentration
#Estimate of the "uniformity underestimation" quantity
#(If null, pi0 is not underestimated.)
r$unif.under

#Plot of the graph using the "slim" method
r=calibration.plot(p, pi0.method="slim")
```

`estim.pi0` *Estimation of the proportion of true null hypotheses*

Description

From a proteomics viewpoint, this function estimates the global proportion of proteins (resp. of peptides) that are non differentially abundant from the tested protein list (resp. from the tested peptide list). This proportion is later used as a correcting factor to compute the adjusted p-values, that are in turn used to tune a threshold according to a desired false discovery rate.

From a statistical viewpoint, this function allows estimating the proportion of true null hypotheses (pi0) from a vector of raw p-values following eight different estimation methods from the literature.

Usage

```
estim.pi0(p, pi0.method = "ALL", nbins = 20, pz = 0.05)
```

Arguments

`p`	Numeric vector of raw p-values. Raw p-values are assumed without missing values, and between 0 and 1.
`pi0.method`	Name of an estimation method for the proportion of true null hypotheses among `"st.boot"`, `"st.spline"`, `"langaas"`, `"jiang"`, `"histo"`, `"pounds"`, `"abh"` or `"slim"`. Default is `"ALL"`: all the eight estimation methods are performed simultaneously.

`nbins` Number of bins. Parameter used for the `"jiang"` and `"histo"` methods. Default is 20.

`pz` P-value threshold such as p-values below are associated to false null hypotheses. Used for the `"slim"` method. Wang, Tuominen and Tsai (2011) suggest to take a value between 0.01 and 0.1. Default is 0.05.

Details

This function allows to estimate the proportion of true null hypotheses following different estimation methods :

`"abh"` the least slope method proposed in Benjamini and Hochberg (2000).

`"st.spline"` the smoother method described in Storey and Tibshirani (2003).
The `qvalue` function of R package `qvalue` with default tuning is used (Storey (2015)).

`"st.boot"` the bootstrap method described in Storey et al. (2004).
The `qvalue` function of R package `qvalue` with default tuning is used (Storey (2015)).

`"langaas"` the method described in Langaas, Ferkingstad and Lindqvist (2005) using a convex decreasing density estimate for p-values. The `convest` function of R package `limma` with default tuning is used (Ritchie et al. (2015)).

`"histo"` the histogram method described in Nettleton, Hwang, Caldo and Wise (2006).

`"pounds"` the conservative estimate described in Pounds and Cheng (2006).

`"jiang"` the average estimate method described in Jiang and Doerge (2008).

`"slim"` the method of Wang, Tuominen and Tsai (2011) using a sliding linear model.
The default tuning suggested by Wang, Tuominen and Tsai (2011) is used.
Using their notations, lambda1 is fixed to 0.1, n to 10 and B to 100.

To take into account of right censorship on the vector of p-values, each p-value is divided by the maximum p-value present in `p`. Accordingly, the p-values of the true null hypotheses are assumed uniformly distributed between 0 and this maximum. This kind of censorship happens in proteomics when a first thresholding is performed on the fold-changes.

If you want to assume that the p-values are uniformly distributed between 0 and 1, replace `p` by `c(p,1)` when using `estim.pi0`.

Value

`pi0` Numeric value of the estimated proportion of true null hypotheses from the selected method; Numeric vector if `pi0.method="ALL"`.

Author(s)

Quentin Giai Gianetto <quentin2g@yahoo.fr>

References

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics, 25(1):60-83, 2000.

H. Jiang and R.W. Doerge. Estimating the proportion of true null hypotheses for multiple comparisons. Cancer informatics, 6:25, 2008.

M. Langaas, B.H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(4):555-572, 2005.

D. Nettleton, J.T.G. Hwang, R.A. Caldo, and R.P. Wise. Estimating the number of true null hypotheses from a histogram of p values. Journal of Agricultural, Biological, and Environmental Statistics, 11(3):337-356, 2006.

S. Pounds and C. Cheng. Robust estimation of the false discovery rate. Bioinformatics, 22(16):1979-1987, 2006.

M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi and G.K. Smyth. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43(7), pp.e47. 2015.

J.D. Storey, J.E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(1):187-205, 2004.

J.D. Storey and R. Tibshirani. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16):9440-9445, 2003.

J.D. Storey. qvalue: Q-value estimation for false discovery rate control. R package version 2.0.0, http://qvalue.princeton.edu/, http://github.com/jdstorey/qvalue. 2015.

H.-Q. Wang, L.K. Tuominen, and C.-J. Tsai. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. Bioinformatics, 27(2):225-231, 2011.

See Also

`calibration.plot`, `adjust.p`

Examples

```
#get p-values
data(LFQRatio2)
p=LFQRatio2[,7]

#estimate the proportion of true null hypotheses with different methods
r=estim.pi0(p)
r$pi0

#estimate the proportion of true null hypotheses with the "abh" method
r=estim.pi0(p, pi0.method="abh")
r$pi0

#compare with one minus the proportion of human proteins
prop_human=sum(LFQRatio2$Organism=="human")/length(LFQRatio2$Organism)
```

```
pi0_true=1-prop_human
pi0_true
```

LFQRatio2 *Dataset "LFQRatio2"*

Description

This dataset is the final outcome of a quantitative mass spectrometry-based proteomic analysis of two samples containing different concentrations of 48 human proteins (UPS1 standard from Sigma-Aldrich) within a constant yeast background (see Giai Gianetto et al. (2016) for details). It contains the abundance values of the different human and yeast proteins identified and quantified in these two conditions. The conditions A and B represent the measured abundances of proteins when respectively 10fmol and 5fmol of UPS1 human proteins were mixed with the yeast extract before mass spectrometry analyses. Three technical replicates were acquired for each condition.

To identify and quantify proteins, spectra were searched using MaxQuant (version 1.5.1.2) against the Uniprot database, the UPS database and the frequently observed contaminants database. Maximum false discovery rates were set to 0.01 at peptide and protein levels by employing a reverse database strategy.

The abundance values of the dataset were obtained from LFQ values calculated using MaxQuant from MS intensity of unique peptides (see Cox et al. (2014)). The following pre-processing steps were performed to obtain these values using the Perseus toolbox (version 1.5.1.6) available in the MaxQuant environment: log2-transformation of LFQ values, filtering of proteins with at least 3 measured values in one condition and data imputation by replacing missing values with values generated from a normal distribution (see Deeb et al. (2012)).

From a statistical viewpoint, the goal is to find which proteins are differentially abundant between the two conditions among the 1481 quantified proteins. Ideally, the 38 quantified human proteins (out of the original 48 ones) should be concluded as differentially abundant (in such a case the proportion of non-differentially abundant proteins will be pi0=1-38/1481).

Usage

```
data(LFQRatio2)
```

Format

A data frame with the 1481 identified proteins in row.

Columns `A.R1`, `A.R2` and `A.R3` correspond to the (numeric) abundance values of proteins in the three replicates of condition A.

Columns `B.R1`, `B.R2` and `B.R3` correspond to the (numeric) abundance values of proteins in the three replicates of condition B.

Column `Welch.test.pval` contains the p-values of the Welch t-test between condition A and condition B computed with the Perseus software.

Column `Organism` contains categorical values: `human` if the protein is identified as human and `yeast` otherwise.

Column `Majority.protein.IDs` contains the IDs of proteins.

References

Cox J., Hein M.Y., Luber C.A., Paron I., Nagaraj N., Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014 Sep, 13(9):2513-26.

Deeb S.J., D'Souza R.C., Cox J., Schmidt-Supprian M., Mann M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Mol Cell Proteomics. 2012 May, 11(5):77-89.

Giai Gianetto, Q., Combes, F., Ramus, C., Bruley, C., Couté, Y., Burger, T. (2016). Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. Proteomics, 16(1), 29-32.

Examples

```
data(LFQRatio2)

#p-values of the Welch t-test between condition A and condition B
p=LFQRatio2[,7]
```

`LFQRatio25` *Dataset "LFQRatio25"*

Description

This dataset is the final outcome of a quantitative mass spectrometry-based proteomic analysis of two samples containing different concentrations of 48 human proteins (UPS1 standard from Sigma-Aldrich) within a constant yeast background (see Giai Gianetto et al. (2016) for details). It contains the abundance values of the different human and yeast proteins identified and quantified in these two conditions. The conditions A and B represent the measured abundances of proteins when respectively 25fmol and 10fmol of UPS1 human proteins were mixed with the yeast extract before mass spectrometry analyses. Three technical replicates were acquired for each condition.

To identify and quantify proteins, spectra were searched using MaxQuant (version 1.5.1.2) against the Uniprot database, the UPS database and the frequently observed contaminants database. Maximum false discovery rates were set to 0.01 at peptide and protein levels by employing a reverse database strategy.

The abundance values of the dataset were obtained from LFQ values calculated using MaxQuant from MS intensity of unique peptides (see Cox et al. (2014)). The following pre-processing steps were performed to obtain these values using the Perseus toolbox (version 1.5.1.6) available in the MaxQuant environment: log2-transformation of LFQ values, filtering of proteins with at least 3 measured values in one condition and data imputation by replacing missing values with values generated from a normal distribution (see Deeb et al. (2012)).

From a statistical viewpoint, the goal is to find which proteins are differentially abundant between the two conditions among the 1472 quantified proteins. Ideally, the 46 quantified human proteins (out of the original 48 ones) should be concluded as differentially abundant (in such a case the proportion of non-differentially abundant proteins will be pi0=1-46/1472).

Usage

```
data(LFQRatio25)
```

Format

A data frame with the 1472 identified proteins in row.

Columns `A.R1`, `A.R2` and `A.R3` correspond to the (numeric) abundance values of proteins in the three replicates of condition A.

Columns `B.R1`, `B.R2` and `B.R3` correspond to the (numeric) abundance values of proteins in the three replicates of condition B.

Column `Welch.test.pval` contains the p-values of the Welch t-test between condition A and condition B computed with the Perseus software.

Column `Organism` contains categorical values: `human` if the protein is identified as human and `yeast` otherwise.

Column `Majority.protein.IDs` contains the IDs of proteins.

References

Cox J., Hein M.Y., Luber C.A., Paron I., Nagaraj N., Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014 Sep, 13(9):2513-26.

Deeb S.J., D'Souza R.C., Cox J., Schmidt-Supprian M., Mann M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. Mol Cell Proteomics. 2012 May, 11(5):77-89.

Giai Gianetto, Q., Combes, F., Ramus, C., Bruley, C., Couté, Y., Burger, T. (2016). Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying FDR control in quantitative experiments. Proteomics, 16(1), 29-32.

Examples

```
data(LFQRatio25)

#p-values of the Welch t-test between condition A and condition B
p=LFQRatio25[,7]
```

Index