

Package ‘lemma’

January 2, 2012

Version 1.3-1

Date 2010-03-23

Title Laplace approximated EM Microarray Analysis

Author Haim Bar, Elizabeth Schifano.

Maintainer Haim Bar <hyb2@cornell.edu>

Description LEMMA is used to detect “nonnull genes” - genes for which the average response in treatment group 1 is significantly different from the average response in group 2, in normalized microarray data. LEMMA is an implementation of an approximate EM algorithm to estimate the parameters in the assumed linear model in Bar, Booth, Schifano, Wells (2009).

License GPL (>= 2)

URL <http://www.stat.cornell.edu/lemma>

Repository CRAN

Date/Publication 2010-04-08 06:07:53

R topics documented:

apoi	2
colon	2
ecoli	3
LEMMA	4
lemma	6
lemmaPlots	8
printTopGenes	9
simdata	10

Index	12
--------------	-----------

apoi

The APO-AI data set (Callow et al., 2000)

Description

The ApoA1 experiment (Callow et al., 2000) used gene targeting in embryonic stem cells to produce mice lacking apolipoprotein A-1, a gene known to play a critical role in high density lipoprotein (HDL) cholesterol levels. Originally, 5,600 expressed sequence tags (EST) were selected. For this data set, we used the data and normalization method provided with the LIMMA R package (Smyth, 2005), which consists of 5,548 ESTs, from 8 control (wild type "black six") mice and 8 "knockout" (lacking ApoA1) mice. Common reference RNA was obtained by pooling RNA from the control mice, and was used to perform expression profiling for all 16 mice. The response of interest is the log₂ fluorescence ratio (with respect to the common reference).

Usage

apoi

Format

A data frame containing 5,548 rows and 18 columns. The first column contains the gene IDs, the second column contains the gene names. Columns 3-11 contain the responses for the 8 mice in the control group, and columns 12-18 contain the responses for the 8 mice in the treatment group.

References

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res*, 10 2022-9.

colon

The colon cancer data set (Alon et al., 1999)

Description

The data analyzed by Alon et al. (1999) consists of 2000 ESTs in 40 tumor and 22 normal colon tissue samples. Of the 40 patients involved in the study, 22 supplied both tumor and normal tissue samples. The normalized data was downloaded from <http://microarray.princeton.edu/oncology/affydata/index.html> and was further transformed by taking the log₂ of the expression data.

Usage

colon

Format

A data frame containing 2,000 rows and 64 columns. The first column contains the gene IDs, the second column contains the gene names. Columns 3-24 contain the responses for the 22 control samples, and columns 25-64 contain the responses for the 40 treatment samples.

References

Alon, Barkai, Notterman, Gish, Ybarra, Mack and Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96 6745-6750.

ecoli

The Escherichia coli data (Hung et al., 2002)

Description

The Leucine-responsive regulatory protein (Lrp) data analyzed by Hung et al. consists of 2,758 genes, with 4 replicates in each treatment group. The experimental design for the Affymetrix GeneChip experiments is as follows: four GeneChips were hybridized with biotin-labeled RNA pools 1-3, 4-6, 7-9, and 10-12 prepared from *lrp+* cells, and four GeneChips were hybridized with biotin-labeled RNA pools 1-3, 4-6, 7-9, and 10-12 prepared from *lrp-* cells, respectively. One average difference measurement for each gene probe set on each GeneChip was obtained for subsequent data processing and analysis. The normalized data was downloaded from <http://www.jbc.org/cgi/content/full/277/43/40309/DC1> and was further transformed by taking the \log_2 of the responses.

Usage

ecoli

Format

A data frame containing 2,758 rows and 10 columns. The first column contains the gene IDs, the second column contains the gene names. Columns 3-6 contain the responses for the 4 control samples, and columns 7-10 contain the responses for the 4 treatment samples.

References

Hung, S. P., Baldi, P. & Hatfield, G. W. (2002). Global gene expression profiling in *Escherichia coli* K12. The effects of leucine-responsive regulatory protein. *J Biol Chem* 277, 40309-40323

Description

LEMMA fits a linear mixed model to normalized microarray data. See the complete LEMMA paper (on the lemma web site) which contains the underlying model and the theory.

<http://www.stat.cornell.edu/lemma/docs/LEMMAsummary.pdf> contains a short summary of the model.

This version supports either two treatment groups or within-group analysis (for example, in paired tests). It provides either a two-way classification (null and nonnull genes), or a three-way classification: null genes, for which statistically there is no difference in expression between the two treatment groups; nonnull group 1 - genes that are significantly more expressed in treatment group 1 than in treatment group 2; and nonnull group 2 - genes that are significantly more expressed in treatment group 2 than in treatment group 1.

The program runs on both Windows and Linux.

Details

The input should consist of a data frame with G rows, and have the following structure:

- `geneid` - the first column must contain a vector of G gene IDs
- `genename` - the second column must contain a vector of G gene names
- `Y1` - columns 3,...,(2+n1) contain normalized data for $n1$ replicates of subjects from treatment group 1. The columns must be named `Y1_m` where $m=1,\dots,n1$.
- `Y2` - columns (n2+2),...,(2+n1+n2) contain normalized data for $n2$ replicates of subjects from treatment group 2. The columns must be named `Y2_m` where $m=1,\dots,n2$. Note that these columns should only be used in the two-groups case. In the case of within-group analysis, only `Y1_m` columns should be in the data frame (as well as the gene id and gene name columns.)

In this version $n1$ and $n2$ do not have to be the same, but all the rows in `Y1` have to have $n1$ elements, and all the rows in `Y2` have to have $n2$ elements. The program also uses the following variables when the user invokes the `lemma` function: `outdir`, `locfdrcutoff`, `fdrcutoff`, `topgenes`, `titletext`, `mgq`, `tol`, `maxIts`, `modes`, `plots`, `saveascsv`, `ErrVarEst`.

All of the parameter estimates, plots, and gene lists will be saved under the user-specified `outdir` directory. In particular, this directory will contain the following files:

- `log.txt` - reporting the total number of genes, sample sizes, $mean(d_g)$, $sd(d_g)$, $mean(m_g)$, $sd(m_g)$, estimates of the shape and scale parameters of the assumed inverse gamma prior for the error variance. It also contains the mean and variance of the fitted error variance distribution (they should be close to the sample mean and variance based on the observed m_g). Estimates for τ , ψ , σ_ψ^2 , p_1 and p_2 , as well as standard errors for all the parameters are also included in this log file. Also logged are the number of nonnulls genes detected using the user-provided local `fdr` and the `FDR` thresholds. Any convergence problems in the EM algorithm are reported in this file.

- resultsRR.txt - contains a list of genes sorted by their posterior null probability. This file also contains the estimated posterior probabilities for a gene being more expressed in treatment group 1 than in treatment group 2 (and vice versa). It also contains the gene effect ($d_g - \tau$).
- resultsFDR.txt - contains a list of genes sorted by their BH-adjusted p-values. The file also contains the gene effect ($d_g - \tau$), and the sign of the gene effect which can be used to determine if a (nonnull) gene is more expressed in treatment group 1 than in treatment group 2 (or vice versa).
- results.csv - contains a list of all the genes, their BH-adjusted p-values, the test statistic (d_g), and the posterior probabilities.
- AllData.RData - contains the following elements: dg, mg, n1, n2, f, G, RRfdr0, RRfdr1, RRfdr2, alpha_hat, beta_hat, sig2eb, tau, psi, sig2psi, p0, p1, p2, pBH0

Note:

- $f = n_1 + n_2 - 2$ (the degrees of freedom for the mean square error, m_g)
- sig2eb = a vector of length G containing the posterior mode of the error variances, given m_g
- pBH0 = a vector of length G containing the BH-adjusted p-value of genes
- RRfdr0 = a vector of length G containing the posterior probabilities of genes being null
- RRfdr1 = a vector of length G containing the posterior probabilities of genes being nonnull and more expressed in treatment group 1 than in treatment group 2
- RRfdr2 = a vector of length G containing the posterior probabilities of genes being nonnull and more expressed in treatment group 2 than in treatment group 1

Author(s)

Bar, H.Y. <hyb2@cornell.edu>, Schifano, E.D. <eds27@cornell.edu>

References

Bar, H.Y., Booth, J.G., Schifano, E.D., Wells, M.T., (2010). Laplace approximated EM Microarray Analysis: an empirical Bayes approach for comparative microarray experiments.
<http://www.stat.cornell.edu/lemma/docs/lemma.pdf>

See Also

Read [lemma](#) to see how to execute the program.

Use [lemmaPlots](#) to produce diagnostics plots.

Use [printTopGenes](#) to produce a list of genes sorted by their adjusted p-values or by their posterior null probabilities.

Examples

```
## Not run:
lemma(apoi,titletext="APO-AI, Callow et al (2000)",outdir="OUT/apoi",
      plots=F)
lemmaPlots("OUT/apoi",mgq=0.99, titletext="APO-AI (Callow et al., 2000)")
```

```
lemma(simdata,titletext="Simulated data",outdir="OUT/simdata")

# Similarly, if the user wants to use the 2-way classification:
lemma(apoai,titletext="APO-AI, Callow et al (2000)",outdir="OUT/apoai",
      modes=2, plots=F)
lemmaPlots("OUT/apoai",mgq=0.99,titletext="APO-AI (Callow et al., 2000)",
          modes=2)

## End(Not run)
```

 lemma

LEMMA - the main program

Description

This program implements the EM algorithm for the LEMMA model. LEMMA fits a linear mixed-model for normalized microarray data analysis. See the reference to the LEMMA paper, which contains the underlying model and the theory. For a short summary of the model, see

<http://www.stat.cornell.edu/lemma/docs/LEMMAsummary.pdf>.

This version supports two treatment groups and either a two-way classification (null and nonnull genes), or a three-way classification: null genes, for which statistically there is no difference in expression between the two treatment groups; nonnull group 1 - genes that are significantly more expressed in treatment group 1 than in treatment group 2; and nonnull group 2 - genes that are significantly more expressed in treatment group 2 than in treatment group 1. This version also supports within-group analysis (for example, in paired experiments).

The program runs on both Windows and Linux.

Usage

```
lemma(dataframe, locfdrcutoff=0.2, fdrcutoff=0.2, mgq=1,
      titletext="", outdir="OUT", topgenes="nonnull", tol=1e-6,
      maxIts=50000, modes=3, plots=TRUE, saveascsv=TRUE,
      ErrVarEst="MLE")
```

Arguments

dataframe	The data frame file containing the normalized data (the first column contains the gene IDs, the second column contains the gene names, the next n1 columns contain the normalized data for treatment group 1 (Y1), and in the two-treatment case, the last n2 columns contain the normalized data for treatment group 2 (Y2).
locfdrcutoff	the local fdr cutoff value for detecting nonnull genes. Default=0.2.
fdrcutoff	the level for the FDR procedure for detecting nonnull genes. Default=0.2.
mgq	The quantile used for eliminating genes with extreme m_g values. Default=1.
titletext	The title to be used in the output files.
outdir	the output directory (use / as a directory separator, even on Windows). Default=OUT.

topgenes	the number of genes to print (sorted in ascending order by local-fdr or adjusted p-value). Use 'all' to print all the genes, or 'nonnull' to print only the genes that are declared nonnull with the given local fdr or FDR cut-off values. Default='nonnull'.
tol	The tolerance level for determining the convergence of the EM algorithm. Default=1e-6.
maxIts	The maximum number of EM iterations. Default=50000.
modes	the number of assumed (null + nonnull) groups. Can be either 2 or 3 (default=3).
plots	logical - if TRUE, produce the diagnostics plots in ps and pdf formats (default=TRUE).
saveascsv	logical - if TRUE, produce the comma-separated file with all the genes and their adjusted p-values and posterior probabilities (default=TRUE).
ErrVarEst	The estimation method for the error variance parameters. Can be either MLE for maximum likelihood estimation, or MM for Method of Moments estimation (default=MLE).

Details

The input should consist of a data frame with G rows (one row per gene), and have the following structure:

- geneid - the first column must contain a vector of G gene IDs
- genename - the second column must contain a vector of G gene names
- Y1 - columns 3,...,(2+n1) contain normalized data for $n1$ replicates of subjects from treatment group 1. The columns must be named Y1_m where $m=1,\dots,n1$.
- Y2 - columns (n2+2),..., (2+n1+n2) contain normalized data for $n2$ replicates of subjects from treatment group 2. The columns must be named Y2_m where $m=1,\dots,n2$. Note that these columns are only used in between-treatment analysis. In the case of within-group analysis, these columns should not be used.

In this version $n1$ and $n2$ do not have to be the same, but all the rows in Y1 have to have $n1$ elements, and all the rows in Y2 have to have $n2$ elements.

All of the parameter estimates, plots, and gene lists will be saved under the user-specified outdir directory. In particular, this directory will contain the following files:

- log.txt - reporting the total number of genes, sample sizes, $mean(d_g)$, $sd(d_g)$, $mean(m_g)$, $sd(m_g)$, estimates of the shape and scale parameters of the assumed inverse gamma prior for the error variance. It also contains the mean and variance of the fitted error variance distribution (they should be close to the sample mean and variance based on the observed m_g). Estimates for τ , ψ , σ_ψ^2 , p_1 and p_2 are also included in this log file, as well as the number of nonnulls genes detected using the user-provided local fdr cutoff, and the FDR threshold. Note that in this version parameter estimates are provided with their standard errors (based on the Fisher Information matrix for the error variance parameter, and the Oakes method for the mean and mixture parameters). Any convergence problems in the EM algorithm are reported in this file.

- resultsRR.txt - contains a list of genes sorted by their posterior null probability. This file also contains the estimated posterior probabilities for a gene being more expressed in treatment group 1 than in treatment group 2 (and vice versa). It also contains the gene effect ($d_g - \tau$).
- resultsFDR.txt - contains a list of genes sorted by their BH-adjusted p-values. The file also contains the gene effect ($d_g - \tau$), and the sign of the gene effect which can be used to determine if a (nonnull) gene is more expressed in treatment group 1 than in treatment group 2 (or vice versa).
- results.csv - contains a list of all the genes, their BH-adjusted p-values, the test statistic (d_g), and the posterior probabilities.
- AllData.RData - contains the following elements: dg, mg, n1, n2, f, G, RRfdr0, RRfdr1, RRfdr2, alpha_hat, beta_hat, sig2eb, tau, psi, sig2psi, p0, p1, p2, pBH0

Note:

- $f = n_1 + n_2 - 2$ (the degrees of freedom for the mean square error, m_g)
- sig2eb = a vector of length G containing the posterior mode of the error variances, given m_g
- pBH0 = a vector of length G containing the BH-adjusted p-value of genes
- RRfdr0 = a vector of length G containing the posterior probabilities of genes being null
- RRfdr1 = a vector of length G containing the posterior probabilities of genes being nonnull and more expressed in treatment group 1 than in treatment group 2
- RRfdr2 = a vector of length G containing the posterior probabilities of genes being nonnull and more expressed in treatment group 2 than in treatment group 1

Examples

```
## Not run:
lemma(apoai,titletext="APO-AI, Callow et al (2000)",outdir="OUT/apoai",
      plots=F)
lemmaPlots("OUT/apoai",mgq=0.99, titletext="APO-AI (Callow et al., 2000)")

lemma(simdata,titletext="Simulated data",outdir="OUT/simdata")

# Similarly, if the user wants to use the 2-way classification:
lemma(apoai,titletext="APO-AI, Callow et al (2000)",outdir="OUT/apoai",
      modes=2, plots=F)
lemmaPlots("OUT/apoai",mgq=0.99,titletext="APO-AI (Callow et al., 2000)",
          ErrVarEst="MM",modes=2)

## End(Not run)
```

Description

This program produces five diagnostics plots in two files (in postscript and pdf formats). The file LEMMAplots1 contains four plots: (a) the fitted distribution for the histogram of d_g , (b) the fitted distribution for the histogram of m_g , (c) and (d) are ‘volcano plots’ that show the distribution of the Benjamini-Hochberg adjusted p-values, and the posterior null probabilities, respectively. The second file contains a plot showing the values of d_g by gene No. on the array. The darker and larger the point, the smaller the adjusted p-value. The gene with the smallest p-value is highlighted in green and its full name is printed.

Usage

```
lemmaPlots(outdir, mgq=0.99, titletext, modes)
```

Arguments

outdir	The directory containing the RData file with the saved estimates and data from a previous invocation of lemma(). Default='OUT'
mgq	The quantile used for plotting the m_g histogram. Genes with m_g values outside the mgq quantile are excluded from the plot. Default=0.99
titletext	The title of the plot
modes	The number of assumed (null + nonnull) groups. Can be either 2 or 3 (default=3).

Details

This function produces five diagnostics plots in two files (in postscript and pdf formats, saved in the outdir directory). The file LEMMAplots1 contains four plots:

(a) is the fitted distribution for the histogram of d_g ,

(b) is the fitted distribution for the histogram of m_g ,

(c) and (d) are ‘volcano plot’ that show the distribution of the Benjamini-Hochberg adjusted p-values, and the posterior null probabilities, respectively.

The file LEMMAplots2 contains a plot showing the values of d_g by gene number on the array. The darker and larger the point, the smaller the adjusted p-value. The gene with the smallest p-value is highlighted in green and its full name is printed.

Note - we are using ps2pdf to convert the postscript format to pdf. It is assumed to be installed on the computer.

Description

This program prints ‘topgenes’ genes sorted in ascending order by their posterior null probability, or by their adjusted p-value. Use a positive integer, or ‘all’ to print all the genes, or ‘nonnull’ to print only the genes that are declared nonnull with the given local fdr or FDR cut-off values. It is executed by lemma(), but can be executed later separately, if the user wants to print a different number of ‘topgenes’, without running the EM algorithm again.

Usage

```
printTopGenes(type, outdir, data0, data1, data2, geneid, genename,
              topgenes, titletext, cutoff, modes)
```

Arguments

type	Either ‘RR’ or ‘FDR’.
outdir	The directory with the AllData.RData file from a previous invocation of lemma(). The output files will be stored in the outdir directory (overwriting the previous copy).
data0	Enter RRfdr0 if type=‘RR’, or pBH0 if type=‘FDR’
data1	Enter RRfdr1 if type=‘RR’, or (dg-tau) if type=‘FDR’
data2	Enter (dg-tau) if type=‘RR’, or c() if type=‘FDR’
geneid	Use the geneid vector
genename	Use the genename vector
topgenes	the number of genes to print (sorted in ascending order by local-fdr or adjusted p-value). Use ‘all’ to print all the genes, or ‘nonnull’ to print only the genes that are declared nonnull with the given local fdr or FDR cut-off values. Default=‘nonnull’.
titletext	The title to be used in the output files.
cutoff	Enter locfdrcutoff if type=‘RR’, or fdrcutoff if type=‘FDR’
modes	the number of assumed (null + nonnull) groups. Can be either 2 or 3 (default=3)

simdata

Simulated data

Description

Data simulation using the LEMMA model. The true parameter values are: $\alpha = 5$, $\beta = 1/12$, $\tau = 0$, $\psi = -1$, $\sigma_\psi^2 = 0.5$, $p_1 = p_2 = 0.25$. Both groups have 6 replicates (‘subjects’), and 2,000 ‘genes’.

Usage

```
simdata
```

Format

A data frame containing 2,000 rows and 14 columns. The first column contains the gene IDs, the second column contains the gene names. Columns 3-8 contain the reponses for the 6 control samples, and columns 9-14 contain the reponses for the 6 treatment samples.

Index

*Topic **datasets**

- apoi, [2](#)
- colon, [2](#)
- ecoli, [3](#)
- simdata, [10](#)

*Topic **documentation**

- LEMMA, [4](#)
- lemma, [6](#)
- lemmaPlots, [8](#)
- printTopGenes, [9](#)

apoi, [2](#)

colon, [2](#)

ecoli, [3](#)

LEMMA, [4](#)

lemma, [5](#), [6](#)

lemmaPlots, [5](#), [8](#)

printTopGenes, [5](#), [9](#)

simdata, [10](#)