Package 'MEB'

November 28, 2021

Type Package

Title A normalization-invariant minimum enclosing ball method to detect differentially expressed genes for RNA-seq data

Version 1.8.0

Author Yan Zhou, Jiadi Zhu

Maintainer Jiadi Zhu <2160090406@email.szu.edu.cn>, Yan Zhou <zhouy1016@szu.edu.cn>

Description Identifying differentially expressed genes between the same or different species is an urgent demand for biological and medical research. For RNA-seq data, systematic technical effects and different sequencing depths are usually encountered when conducting experiments. Normalization is regarded as an essential step in the discovery of biologically important changes in expression. The present methods usually involve normalization of the data with a scaling factor, followed by detection of significant genes. However, more than one scaling factor may exist because of the complexity of real data. Consequently, methods that normalize data by a single scaling factor may deliver suboptimal performance or may not even work. The development of modern machine learning techniques has provided a new perspective regarding discrimination between differentially expressed (DE) and non-DE genes. However, in reality, the non-DE genes comprise only a small set and may contain housekeeping genes (in same species) or conserved orthologous genes (in different species). Therefore, the process of detecting DE genes can be formulated as a one-class classification problem, where only non-DE genes are observed, while DE genes are completely absent from the training data.
We transform the problem to an outlier detection problem by treating DE genes as outliers, and we propose a normalization-invariant minimum enclosing ball (NIMEB) method to construct a smallest possible ball to contain the known non-DE genes in a feature space. The genes outside the minimum enclosing ball can then be naturally considered to be DE genes. Compared with the existing methods, the proposed NIMEB method does not require data normalization, which is particularly attractive when the RNA-seq

data include more than one scaling factor. Furthermore, the NIMEB method could be easily extended to different species without normalization.

License GPL-2

Encoding UTF-8

LazyData true

Depends R (>= 3.6.0)

Suggests knitr,rmarkdown,BiocStyle

VignetteBuilder knitr

RoxygenNote 7.1.0

Imports e1071, SummarizedExperiment

biocViews DifferentialExpression, GeneExpression, Normalization, Classification, Sequencing

git_url https://git.bioconductor.org/packages/MEB

git_branch RELEASE_3_14

git_last_commit f808247

git_last_commit_date 2021-10-26

Date/Publication 2021-11-28

R topics documented:

NIMEB 2
real_data_dsp 4
real_data_sp 4
sim_data_dsp 5
sim_data_sp 5

Index **6**

`NIMEB` *Detect differential expression genes for RNA-seq data*

Description

Use a normalization-invariant minimum enclosing ball (NIMEB) method to discriminate differential expression (DE) genes in the same or different species.

Usage

```
NIMEB(countsTable, train_id, gamma, nu = 0.01, reject_rate = 0.1,
ds = FALSE)
```

Arguments

countsTable	Matrix or data.frame of short read counts for each genes in the same or different species.
train_id	A vector shows the position of housekeeping genes or conserved genes in countsTable.
gamma	A parameter needed for all kernels except linear.
nu	parameter needed for one-classification.
reject_rate	A value used in controling the scale of ball, default is 0.01.
ds	A value to show the data is for the same species or different species. If ds is FALSE, the data is the same species, else the data is the different species.

Value

list(.) A list of results, "model" represents the model of NIMEB, which could be used to discriminate a new gene, "gamma" represents the selected gamma parameters in model NIMEB, "train_error" represents the corresponding train_error when the value of gamma changed.

Examples

```
## Simulation data for the same species.
library(SummarizedExperiment)
data(sim_data_sp)
gamma <- seq(1e-06,5e-05,1e-06)
sim_model_sp <- NIMEB(countsTable = assay(sim_data_sp), train_id=1:1000,
gamma, nu = 0.01, reject_rate = 0.05, ds = FALSE)

## Real data for the same species.
data(real_data_sp)
gamma <- seq(1e-06,5e-05,1e-06)
real_model_sp <- NIMEB(countsTable = assay(real_data_sp), train_id=1:530,
gamma, nu = 0.01, reject_rate = 0.1, ds = FALSE)

## Simulation data for the different species.
data(sim_data_dsp)
gamma <- seq(1e-07,2e-05,1e-06)
sim_model_dsp <- NIMEB(countsTable = assay(sim_data_dsp), train_id=1:1000,
gamma, nu = 0.01, reject_rate = 0.1, ds = TRUE)

## Real data for the different species.
data(real_data_dsp)
gamma <- seq(5e-08,5e-07,1e-08)
real_model_dsp <- NIMEB(countsTable = assay(real_data_dsp), train_id=1:143,
gamma, nu = 0.01, reject_rate = 0.1, ds = TRUE)
```

real_data_dsp *A real dataset of genes between the different species.*

Description

This data set includes two species and 19330 genes with corresponding short read counts, in which the first 143 genes are conserved genes.

Usage

```
real_data_dsp
```

Format

A data.frame contains two species and 19330 genes.

Source

Brawand, D., Soumillon, M., Necsulea, A. and Julien, P. et al. (2011). The evolution of gene expression levels in mammalian organs. Nature, 478, 343-348.

real_data_sp *A real dataset of genes between the same species.*

Description

This data set includes two samples and each sample with five biological replicates and 16519 genes with corresponding short read counts, in which the first 530 genes are housekeeping genes.

Usage

```
real_data_sp
```

Format

A data.frame contains two samples and each sample with five biological replicates and 16519 genes.

Source

Marioni J.C., Mason C.E., et al. (2008). RNA-seq: an assessment of technical reproducibility and comparisonwith gene expression arrays. Genome Res. 18(9), 1509–1517.

sim_data_dsp *A simulation dataset of genes between the differetn species.*

Description

This data set includes two species and 18472 genes with corresponding short read counts, in which the first 1000 genes are conserved genes.

Usage

```
sim_data_dsp
```

Format

A data.frame contains two species and 18472 genes.

Source

Jiadi Zhu, Yan Zhou, Junhui Wang, Youlong Yang (2019, pending publication). A minimum enclosing ball method to detect differential expression genes for RNA-seq data.

sim_data_sp *A simulation dataset of genes between the same species.*

Description

This data set includes two samples and 10623 genes with corresponding short read counts, in which the first 1000 genes are housekeeping genes.

Usage

```
sim_data_sp
```

Format

A data.frame contains two samples and 10623 genes.

Source

Jiadi Zhu, Yan Zhou, Junhui Wang, Youlong Yang (2019, pending publication). A minimum enclosing ball method to detect differential expression genes for RNA-seq data.

Index

∗ **datasets**
real_data_dsp, 4
real_data_sp, 4
sim_data_dsp, 5
sim_data_sp, 5

NIMEB, 2

real_data_dsp, 4
real_data_sp, 4

sim_data_dsp, 5
sim_data_sp, 5