Package: clusterRepro (via r-universe)

September 16, 2024

Version 0.9

Date 2018-10-10

Title Reproducibility of Gene Expression Clusters

Author Amy Kapp <Amy_Kapp@hotmail.com> and Rob Tibshirani

<tibs@stanford.edu>

Maintainer Rob Tibshirani <tibs@stanford.edu>

Depends R (>= 2.2.0)

Description This is a function for validating microarray clusters via reproducibility, based on the paper referenced below.

URL https://www.ncbi.nlm.nih.gov/pubmed/16613834.

License GPL-2

NeedsCompilation no

Date/Publication 2018-10-15 18:20:14 UTC

Repository https://tibshirani.r-universe.dev

RemoteUrl https://github.com/cran/clusterRepro

RemoteRef HEAD

RemoteSha 1346e72549bd00792055d1080b1194d2144adbdb

Contents

---

clusterRepro   *Gene expression clusters reproducibility and validation*

Description

Validate gene expression clusters by determining whether or not they are reproducible

Usage

```
clusterRepro(Centroids, New.data, Number.of.permutations)
```

Arguments

Centroids   The matrix of centroids with annotated rows. The labeled rows are either genes
                  (for gene clusters) or samples (for sample clusters) and the columns are the
                  centroids.

New.data   The matrix of gene expression data (with annotated rows) independent of the
                  dataset used to form the centroids. For gene clusters, the rows are samples and
                  the columns are genes. For sample clusters, the rows are genes and the columns
                  are samples.

Number.of.permutations
                  The number of times the centroids will be permuted to generate the null distri-
                  bution.

Details

This function looks for gene expression clusters found in one dataset in another independent dataset.
The centroids from the first dataset are used to classify the independent data and the correspond-
ing in-group proportions (IGPs) are computed. These in-group proportions are compared to null
distributions of in-group proportions to produce p-values. The IGP null distributions are gener-
ated by repeatedly permuting the centroids within the box aligned with the principal components,
classifying the independent data, and calculating the corresponding IGPs.

Value

Actual.Size   The number of columns of New.data assigned to each centroid.

Actual.IGP   The in-group proportions of the groups formed when New.data is classified us-
                  ing Centroids.

p.value   The p-values for each of the groups represented by the centroids.

Number   The number of permutations used to compute the corresponding p-value.

Author(s)

Amy Kapp and Robert Tibshirani

References

Amy Kapp and Robert Tibshirani. Are clusters in one dataset present in another dataset? To be published.

See Also

[IGP.clusterRepro](#).

Examples

```
### Generate centroids with annotated rows
Centroids <- matrix(rnorm(30, sd = 10), 10)
rownames(Centroids) <- letters[1:nrow(Centroids)]

### Generate data with annotated rows
Data <- cbind(matrix(rep(Centroids[,1], 10), 10),
matrix(rep(Centroids[,2], 15), 10), matrix(rep(Centroids[,3], 20), 10))
Data <- Data + matrix(rnorm(length(Data), sd = 10), nrow(Data))
rownames(Data) <- letters[1:nrow(Data)]

### Classify the data and calculate the corresponding in-group
### proportions and group size
Result <- clusterRepro(Centroids, Data, Number.of.permutations = 1)
Result$Actual.IGP
Result$Actual.Size

### Generate null distributions and compare to actual in-group proportions to obtain p-values
Result2 <- clusterRepro(Centroids, Data, Number.of.permutations = 1000)

### If the number of rows in the centroid matrix does not match the
### number of rows in the data matrix and the row labels are unique, this
### function will only use the rows that the two matrices have in common.
Data <- matrix(rnorm(200), 20)
rownames(Data) <- letters[(nrow(Data)+6):7]
Result <- IGP.clusterRepro(Data, Centroids)
Result2 <- clusterRepro(Centroids, Data, Number.of.permutations = 1000)
```

IGP.clusterRepro	*In-group proportion calculation*

Description

This function classifies gene expression (microarray) data using centroids and calculates the in-group proportions of the resulting groups.

Usage

```
IGP.clusterRepro(Data, Centroids)
```

Arguments

Data
: The matrix of gene expression data with annotated rows. For gene clusters, the labeled rows are samples and the columns are genes. For sample clusters, the labeled rows are genes and the columns are samples.

Centroids
: The matrix of centroids with annotated rows. The labeled rows are either genes (for gene clusters) or samples (for sample clusters) and the columns are the centroids.

Details

The Pearson's centered correlation coefficient between each datum and each centroid is calculated. The datum is then classified to the group whose centroid had the highest correlation with the datum. The in-group proportion is defined to be the proportion of data in a group whose nearest neighbors (Pearson's centered correlation) are also classified to the same group.

Value

Class
: The data classification made using the centroids

IGP
: The in-group proportions for the groups found in the data

Size
: The number of data classified to each of the groups

Author(s)

Amy Kapp and Robert Tibshirani

References

Amy Kapp and Robert Tibshirani. Are clusters in one dataset present in another dataset? To be published.

Examples

```
### Make centroid matrix with annotated rows
C <- matrix(rnorm(30), 10)
rownames(C) <- letters[1:nrow(C)]

### Make data matrix with annotated rows
D <- matrix(rnorm(100), 10)
rownames(D) <- letters[1:nrow(C)]

### Classify data and calculate in-group proportions
Result <- IGP.clusterRepro(D, C)

### If the number of rows in the centroid matrix does not match the
### number of rows in the data matrix and the row labels are unique, this
### function will only use the rows that the two matrices have in common.
D <- matrix(rnorm(200), 20)
rownames(D) <- letters[(nrow(D)+6):7]
Result <- IGP.clusterRepro(D, C)
```

Index