# Package 'clusterStab'

September 24, 2012

**Title** Compute cluster stability scores for microarray data

**Version** 1.28.0

**Author** James W. MacDonald, Debashis Ghosh, Mark Smolkin

**Description** This package can be used to estimate the number of clusters in a set of microarray data, as well as test the stability of these clusters.

**Maintainer** James W. MacDonald <jmacdon@med.umich.edu>

**License** Artistic-2.0

**Depends** Biobase (>= 1.4.22), R (>= 1.9.0), methods

**Suggests** fibroEset, genefilter

**LazyLoad** True

**biocViews** Clustering

## R topics documented:



---

benhur *A Function to Estimate the Number of Clusters in Microarray Data*

---

### Description

This function estimates the number of clusters in e.g., microarray data using an iterative process proposed by Asa Ben-Hur.

## Usage

```
## S4 method for signature 'ExpressionSet'
benhur(object, freq, upper, seednum = NULL,
linkmeth = "average", distmeth = "euclidean", iterations = 100)
## S4 method for signature 'matrix'
benhur(object, freq, upper, seednum = NULL, linkmeth
= "average", distmeth = "euclidean", iterations = 100)
```

## Arguments

| | |
|---|---|
| object | Either a matrix or ExpressionSet |
| freq | The proportion of samples to use. This should be somewhere between 0.6 - 0.8 for best results. |
| upper | The upper limit for number of clusters. |
| seednum | A value to pass to set.seed, which will allow for exact reproducibility at a later date. |
| linkmeth | Linkage method to pass to hclust. Valid values include "average", "centroid", "ward", "single", "mcquitty", or "median". |
| distmeth | The distance method to use. Valid values include "euclidean" and "pearson" where pearson implies 1-pearson correlation. |
| iterations | The number of iterations to use. The default of 100 is a reasonable number. |

## Details

This function may be used to estimate the number of true clusters that exist in a set of microarray data. This estimate can be used to as input for clusterComp to estimate the stability of the clusters.

The primary output from this function is a set of histograms that show for each cluster size how often similar clusters are formed from subsets of the data. As the number of clusters increases, the pairwise similarity of cluster membership will decrease. The basic idea is to choose the histogram corresponding to the largest number of clusters in which the majority of the data in the histogram is concentrated at or near 1.

If overlay is set to TRUE, an additional CDF plot will be produced. This can be used in conjunction with the histograms to determine at which cluster number the data are no longer concentrated at or near 1.

## Value

The output from this function is an object of class benhur. See the benhur-class man page for more information.

## Author(s)

Originally written by Mark Smolkin <marksmolkin@hotmail.com> further modifications by James W. MacDonald <jmacdon@med.umich.edu>

## References

A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing, 2002. Smolkin, M. and Ghosh, D. (2003). Cluster stability scores for microarray data in cancer studies . BMC Bioinformatics 4, 36 - 42.

## Examples

```
data(sample.ExpressionSet)
tmp <- benhur(sample.ExpressionSet, 0.7, 5)
hist(tmp)
ecdf(tmp)
```

---

BenHur-class *Class "BenHur", a class for estimating clusters in microarray data, and methods for visualizing them.*

---

## Description

A specialized class representation used for estimating clusters in microarray data.

## Objects from the Class

Objects are usually created by a call to benhur, although technically a new object can also be created by a call to new("BenHur",...). However, this second method is usually not worth the work required.

## Slots

**jaccards:** Object of class "list", containing the jaccard vectors; these indicate the proportion of pairwise similarity between clusters formed from subsets of the data.

**size:** Object of class "vector", only used for plotting.

**iterations:** Object of class "vector", containing the number of iterations. Defaults to 100.

**freq:** Object of class "vector", containing the proportion of the data used for subsampling.

## Methods

**ecdf** signature(x = "BenHur"): Plot an empirical CDF. This can be used to help determine the number of clusters in the data. The most likely (e.g., most stable number) of clusters will have a CDF that is concentrated at or near one. See vignette for more information.

**hist** signature(x = "BenHur"): Plot histograms for all clusters tested. The most likely (e.g., most stable number) of clusters will have a histogram in which the data are clustered at or near one. See vignette for more information.

**show** signature(object = "BenHur"): Gives a nice summary.

## Author(s)

James W. MacDonald <jmacdon@med.umich.edu>

## References

A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing, 2002. Smolkin, M. and Ghosh, D. (2003). Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 4, 36 - 42.

clusterComp *Estimate Microarray Cluster Stability*

## Description

This function estimates the stability of clustering solutions using microarray data. Currently only agglomerative hierarchical clustering is supported.

## Usage

```
## S4 method for signature 'ExpressionSet'
clusterComp(object, cl, seednum = NULL, B = 100,
sub.frac = 0.8, method = "ave", distmeth = "euclidean", adj.score = FALSE)
## S4 method for signature 'matrix'
clusterComp(object, cl, seednum = NULL, B = 100,
sub.frac = 0.8, method = "ave", distmeth = "euclidean", adj.score = FALSE)
```

## Arguments

| | |
|---|---|
| object | Either a matrix or ExpressionSet |
| cl | The number of clusters. This may be estimated using benhur |
| seednum | A value to pass to set.seed, which will allow for exact reproducibility at a later date. |
| B | The number of permutations. |
| sub.frac | The proportion of genes to use in each subsample. This value should be in the range of 0.75 - 0.85 for best results |
| method | The linkage method to pass to hclust. Valid values include "average", "centroid", "ward", "single", "mcquitty", or "median". |
| distmeth | The distance method to use. Valid values include "euclidean" and "pearson", where pearson implies 1-pearson correlation. |
| adj.score | Boolean. Should the stability scores be adjusted for cluster size? Defaults to FALSE. |

## Details

This function estimates the stability of a clustering solution by repeatedly subsampling the data and comparing the cluster membership of the subsamples to the original clusters.

## Value

The output from this function is an object of class clusterComp. See the clusterComp-class man page for more information.

## Author(s)

James W. MacDonald <jmacdon@med.umich.edu>

## References

A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing, 2002. Smolkin, M. and Ghosh, D. (2003). Cluster stability scores for microarray data in cancer studies . BMC Bioinformatics 4, 36 - 42.

## Examples

```
data(sample.ExpressionSet)
clusterComp(sample.ExpressionSet, 3)
```

---

`ClusterComp-class` *Class "ClusterComp" a class for testing the stability of clusters in microarray data*

---

## Description

A specialized class representation used for testing the stability of clusters in microarray data.

## Objects from the Class

Objects are usually created by a call to `clusterComp`, although technically objects can be created by calls of the form `new("ClusterComp", ...)`. However, the latter is probably not worth doing.

## Slots

`clusters`**:** Object of class `"vector"` showing the cluster membership for each sample when using all the data.

`percent`**:** Object of class `"vector"` containing the percentage of subsamples that resulted in the same class membership for all samples.

`freq`**:** Object of class `"vector"` containing the subsampling percentage used. Defaults to 0.8.

`clusternum`**:** Object of class `"vector"` containing the number of clusters tested.

`iterations`**:** Object of class `"vector"` containing the number of iterations performed. Defaults to 100.

`method`**:** Object of class `"vector"` containing the agglomerative method used. Options include "average", "centroid", "ward", "single", "mcquitty", or "median".

## Methods

**show** `signature(object = "ClusterComp")`: Give a nice summary of results.

## Author(s)

James W. MacDonald <jmacdon@med.umich.edu>

## References

A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing, 2002. Smolkin, M. and Ghosh, D. (2003). Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 4, 36 - 42.

# Index