

# Package ‘clustsig’

February 14, 2012

**Version** 1.0

**Date** 2010-09-01

**Title** Significant Cluster Analysis

**Author** Douglas Whitaker <whitaker@stat.ufl.edu> and Mary Christman <mcxman@ufl.edu>

**Maintainer** Douglas Whitaker <whitaker@stat.ufl.edu>

**Depends** R (>= 2.10.0)

**Description** A complimentary package for use with hclust; simprof tests to see which (if any) clusters are statistically different. The null hypothesis is that there is no a priori group structure. See Clarke, K.R., Somerfield, P.J., and Gorley R.N. 2008. Testing of null hypothesis in exploratory community analyses: similarity profiles and biota-environment linkage. J. Exp. Mar. Biol. Ecol. 366, 56-69.

**License** GPL (>= 2)

**URL** <http://www.stat.ufl.edu/~whitaker/clustsig/>

**Repository** CRAN

**Date/Publication** 2010-09-01 18:19:49

## R topics documented:

simprof . . . . .	2
simprof.plot . . . . .	4

<b>Index</b>	<b>6</b>
--------------	----------

---

 simprof

*Similarity Profile Analysis*


---

## Description

A tool for determining the number of significant clusters produced using `hclust()` with the assumption of no a priori groups.

## Usage

```
simprof(data, num.expected=1000, num.simulated=999,
method.cluster="average", method.distance="euclidean",
method.transform="identity", alpha=0.05,
sample.orientation="row", const=0,
silent=TRUE, increment=100)
```

## Arguments

<code>data</code>	Input data in a matrix.
<code>num.expected</code>	The number of similarity profiles to generate for creating the expected distribution of the data. This value should be large.
<code>num.simulated</code>	The number of similarity profiles to generate for use in comparing the observed test statistic with its null distribution. This value should be large.
<code>method.cluster</code>	The method of clustering to use with <code>hclust</code> . Standard values from <code>hclust</code> are "ward", "single", "complete", "average", "mcquitty", "median" or "centroid".
<code>method.distance</code>	This value should be either an option to pass to the function <code>dist</code> (standard values are "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski"), or "braycurtis" for Bray-Curtis Dissimilarity. This value can also be any function which returns a "dist" object.
<code>method.transform</code>	An option to specify a transform, if any, to be applied to the data. Possible values are "identity" (no transformation), "squareroot", "log", "PA" (Presence/Absence), or any numeric value (of type "double"). This transform is applied before the adjustment constant is applied, so choose a constant accordingly.
<code>alpha</code>	The alpha level at which to reject the null hypothesis. If the null is rejected, the test continues and tests each sub-tree recursively until either all subtrees are exhausted by reaching the individual level or there are no significant distance. Due to the nature of multiple testing inherent in this process, care should be taken when choosing this alpha level.
<code>sample.orientation</code>	The orientation of the data, either "row" or "column". The practical effect of this is that the transpose will be examined if "column" is chosen.

const	The value of the constant to be used in adjusting the Bray-Curtis Dissimilarity coefficient, if any is to be used. Any positive value of "const" will be appended as a new variable to each sample, acting as a sort of "dummy species" (where that interpretation is appropriate).
silent	A logical value indicating whether anything should be printed during the code execution. If FALSE, a message will be printed every increment (see below) number of times in the main looping procedure. This was implemented because the code can take a while to run due to many permutations and its recursive nature; however, for the same reason, many messages could be printed.
increment	An integer value indicating, if silent=FALSE, one which iterations a message should be printed. (If the iteration number modulus increment equals 0, that number will be printed.)

### Value

A list object is produced with the following components:

numgroups	The number of groups which are found to be statistically significant.
significantclusters	A list of length numgroups with each element containing the sample IDs (row/column numbers in the corresponding original data) that are in each significant cluster.
pval	The merge component from the hclust results with an extra column of p-values. These p-values are for testing whether the two groups in that row are statistically different.
hclust	An object of class hclust which is just the results of running hclust on the original data.

### Author(s)

Douglas Whitaker and Mary Christman

### References

Clarke, K.R., Somerfield, P.J., and Gorley, R.N, 2008. Testing of null hypotheses in exploratory community analyses similarity profiles and biota-environment linkage. *J. Exp. Mar. Biol. Ecol.* 366, 56-69.

### See Also

[hclust](#)

### Examples

```
## Not run:
# Load the USArrests dataset included with R
# And use abbreviations of state names
# We leave out the third column because
# it is on a different scale
usarrests<-USArrests[,c(1,2,4)]
```

```
rownames(usarrests)<-state.abb
# Run simprof on the data
res <- simprof(data=usarrests,
method.distance="braycurtis")
# Graph the result
pl.color <- simprof.plot(res)

## End(Not run)
```

---

simprof.plot

*Similarity Profile Analysis Dendrogram Plotter*


---

## Description

A function to plot a dendrogram based on the results of simprof().

## Usage

```
simprof.plot(results, leafcolors=NA, plot=TRUE, fill=TRUE,
leaflab="perpendicular", siglinetype=1)
```

## Arguments

results	The object returned by simprof. However, the only parts "results" needs to have are "hclust" and "significantclusters".
leafcolors	A vector of color names/identifiers (names or hex codes); it should be the same length as "results\$significantclusters". If this isn't supplied, the <a href="#">rainbow</a> function will be used to generate enough colors (recommended). Because the colors are used in the order generated by rainbow, if there are a large number of significant clusters, they may appear to form a continuous color spectrum. If this is the case, the appropriate option is to manually supply a vector of colors to more clearly delineate different clusters.
plot	A logical option indicating whether to plot the dendrogram ("TRUE" for plot).
fill	A logical option indicating whether to color the entire subtree which comprises a significant color (as opposed to just coloring the individual leaves).
leaflab	The option from dendrogram indicating the label text orientation. Possible values are "perpendicular" (vertical), "textlike" (horizontal), or "none" (labels suppressed).
siglinetype	A numeric option indicating the type of line to use for the significant clusters (the line type chosen applies to all significant clusters). The possible values are 0=blank, 1=solid (default), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash, and "different". This can be a vector of line-types to use; in the event that there are fewer available line-types than there are significant groups, some line-types will be repeated. The selection of "different" will use try to automatically use as many different line-types as necessary to have a unique type for each significant group (the line-types will be used in the order of 6 through 1 so that solid is used last to increase clarity).

**Value**

A dendrogram is returned. If "plot=TRUE", the dendrogram is also plotted.

**Author(s)**

Douglas Whitaker and Mary Christman

**See Also**

[hclust](#), [dendrogram](#)

**Examples**

```
## Not run:  
# Load the USArrests dataset included with R  
# And use abbreviations of state names  
# We leave out the third column because  
# it is on a different scale  
usarrests<-USArrests[,c(1,2,4)]  
rownames(usarrests)<-state.abb  
# Run simprof on the data  
res <- simprof(data=usarrests,  
method.distance="braycurtis")  
# Graph the result  
pl.color <- simprof.plot(res)  
  
## End(Not run)
```

# Index

- \*Topic **cluster analysis**
  - [simprof](#), 2
- \*Topic **cluster**
  - [simprof](#), 2
  - [simprof.plot](#), 4
- \*Topic **dendrogram**
  - [simprof.plot](#), 4
- \*Topic **significant cluster**
  - [simprof](#), 2
  - [simprof.plot](#), 4
- \*Topic **similarity profile**
  - [simprof](#), 2

[dendrogram](#), 5

[dist](#), 2

[hclust](#), 2, 3, 5

[rainbow](#), 4

[simprof](#), 2

[simprof.plot](#), 4