ctc

October 25, 2011

`hc2Newick` *Convert hclust objects to Newick format files*

Description

Convert hclust objects to Newick format files.

Usage

```
hc2Newick(hc, flat=TRUE)
```

Arguments

`hc`	a `hclust` object (as returned by the function `hclust` in the package `stats`)
`flat`	a boolean (see section value).

Value

If `flat=TRUE` the result is a string (that you can write in a file).

If `flat=FALSE` the result is a list (of lists). Each list is consituted of the elements `left`, `right` and `dist`.

Author(s)

Laurent (laurent@cbs.dtu.dk)

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

Examples

```
data(USArrests)
h = hclust(dist(USArrests))
write(hc2Newick(h),file='hclust.newick')
```

`hclust2treeview` *Hierarchical clustering and treeview export*

Description

This function compute hierachical clustering with function hcluster and export cluster to treeview files format.

Usage

```
hclust2treeview(x,file="cluster.cdt",method = "euclidean",link = "complete",keep
```

Arguments

`x`	numeric matrix or a data frame or an object of class "exprSet".
`file`	File name of export file.
`method`	the distance measure to be used. This must be one of `"euclidean"`, `"maximum"`, `"manhattan"`, `"canberra"` `"binary"` `"pearson"`, `"correlation"` or `"spearman"`. Any unambiguous substring can be given.
`link`	the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of `"ward"`, `"single"`, `"complete"`, `"average"`, `"mcquitty"`, `"median"` or `"centroid"`.
`keep.hclust`	if TRUE: function returns a list of 2 objects of class hclust

Details

This function producte all 3 files needed by treeview, with extentions: cdt, gtr, atr.

Value

if keep.hclust = FALSE, function return 1. else function returns 2 objects of class hclust, first: hierachical clustering by rows, second: hierarchical clustering by columns

Author(s)

Antoine Lucas, http://mulcyber.toulouse.inra.fr/projects/amap/

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

See Also

`hclust`

Examples

```
data(USArrests)
hclust2treeview(USArrests,file="cluster.cdt")
```

`r2cluster` *Write to Cluster file format*

Description

Converting data to Cluster format

Usage

```
r2cluster(data,labels=FALSE,colname="ACC",description=FALSE,
          file="cluster.txt",dec='.')
```

Arguments

`file`	the path of the file
`data`	a matrix (or data frame) which provides the data to put into the file
`labels`	a logical value indicating whether we use the frist column as labels (ACC column in cluster file)
`colname`	a character string indicating what kind of objects are in each row. YORF, MCLID, CLID, ACC can be used: see details.
`description`	a logical value indicating whether we use the second column as description (NAME column for cluster file)
`dec`	the character used in the file for decimal points

Details

Software *Cluster*, made by *M. Eisen* needs formatted input data like:

```
ACC     NAME    GWEIGHT GORDER  V3      V4      V5
EWEIGHT                         1       1       1
gbk01   Gene1   1       1       0.9     0.4     1.4
gbk02   Gene2   1       2       0.6     0.2     0.2
gbk03   Gene3   1       3       1.6     1.1     0.9
gbk04   Gene4   1       4       0.4     1       1
```

First field of first line (i.e "ACC") is a special field, that tells program what kind of objects are in each row.

Four special values are defined with web link (when visualize with TreeView):

- YORF `http://genome-www.stanford.edu/cgi-bin/dbrun/SacchDB?find+Locus+%22UNIQID%22`
- MCLID `http://genome.rtc.riken.go.jp/cgi-bin/getseq?g+R+UNIQID`
- CLID `http://genome-www4.stanford.edu/cgi-bin/SMD/source/sourceRes\discretionary{-}{}{}ult?op\discretionary{-}{}{}tion=CloneID&criteria1=IMAGE:UNIQID&choice=cDNA`
- ACC `http://genome-www4.stanford.edu/cgi-bin/SMD/source/sourceRes\discretionary{-}{}{}ult?op\discretionary{-}{}{}tion=Number&criteria=UNIQID&choice=cDNA`

Line begining with `EWEIGHT` gives weights for each column (variable). Column `GWEIGHT` gives weights for each line (individuals).

Note

Cluster is a program made by *M. Eisen* that performs hierarchical clustering, K-means and SOM.

Cluster is copyrighted. To get or have information about *Cluster*: http://rana.lbl.gov/EisenSoftware.htm

Author(s)

Antoine Lucas, http://antoinelucas.free.fr/ctc

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

See Also

xcluster, r2xcluster, hclust

Examples

```
#    Create data
set.seed(1)
m <- matrix(rep(1,3*24),ncol=3)
m[9:16,3] <- 3 ; m[17:24,] <- 3    #create 3 groups
m <- m+rnorm(24*3,0,0.5)           #add noise
m <- floor(10*m)/10                #just one digits

r2cluster(m)
```

---

r2gtr *Write to gtr, atr, cdt file format*

Description

Write data frame and hclust object to gtr atr, cdt files (Xcluster or Cluster output). Visualisation of cluster can be done with tools like treeview

Usage

```
r2gtr(hr,file="cluster.gtr",distance=hr$dist.method,dec='.',digits=5)
r2atr(hc,file="cluster.atr",distance=hc$dist.method,dec='.',digits=5)
r2cdt(hr,hc,data,labels=FALSE,description=FALSE,file="cluster.cdt",dec='.')
```

Arguments

file	the path of the file
data	a matrix (or data frame) which provides the data to put into the file
hr,hc	objects of class hclust (rows and columns)

distance	The distance measure used. This must be one of ‘"euclidean"’, ‘"maximum"’, ‘"manhattan"’, ‘"canberra"’ or ‘"binary"’. Any unambiguous substring can be given.
digits	number digits for precision
labels	a logical value indicating whether we use the frist column as labels (NAME column for cluster file)
description	a logical value indicating whether we use the second column as description (DESCRIPTION column for cluster file)
dec	the character used in the file for decimal points

Details

Function hclust2treeview compute hierarchical clustering and export to all files at once.

Author(s)

Antoine Lucas, http://antoinelucas.free.fr/ctc

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

See Also

r2xcluster, xcluster2r,hclust,hcluster

Examples

```
#    Create data
set.seed(1)
m <- matrix(rep(1,3*24),ncol=3)
m[9:16,3] <- 3 ; m[17:24,] <- 3    #create 3 groups
m <- m+rnorm(24*3,0,0.5)           #add noise
m <- floor(10*m)/10                #just one digits

# use library stats
# Cluster columns
hc <- hclust(dist(t(m)))
# Cluster rows
hr <- hclust(dist(m))

# Export files
r2atr(hc,file="cluster.atr")
r2gtr(hr,file="cluster.gtr")
r2cdt(hr,hc,m ,file="cluster.cdt")
```

`r2xcluster` *Write to Xcluster file format*

Description

Converting data to Xcluster format

Usage

```
r2xcluster(data,labels=FALSE,description=FALSE,file="xcluster.txt")
```

Arguments

`file`	the path of the file
`data`	a matrix (or data frame) which provides the data to put into the file
`labels`	a logical value indicating whether we use the frist column as labels (NAME column for xcluster file)
`description`	a logical value indicating whether we use the second column as description (DESCRIPTION column for cluster file)

Details

Software *Xcluster*, made by *G. Sherlock* needs formatted input data like:

```
NAME    DESCRIPTION     GWEIGHT V2      V3      V4
EWEIGHT                         1       1       1
gbk01   Gene1           1       0.9     0.4     1.4
gbk02   Gene2           1       0.6     0.2     0.2
gbk03   Gene3           1       1.6     1.1     0.9
gbk04   Gene4           1       0.4     1       1
```

Line begining with `EWEIGHT` gives weights for each column (variable). Column `GWEIGHT` gives weights for each line (individuals).

Note

Xcluster is a C program made by *Gavin Sherlock* that performs hierarchical clustering, K-means and SOM.

Xcluster is copyrighted. To get or have information about *Xcluster*: http://genome-www.stanford.edu/~sherlock/cluster.html

Author(s)

Antoine Lucas, http://antoinelucas.free.fr/ctc

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

See Also

xcluster, xcluster2r, hclust, hcluster

Examples

```
##    Create data
set.seed(1)
m <- matrix(rep(1,3*24),ncol=3)
m[9:16,3] <- 3 ; m[17:24,] <- 3    #create 3 groups
m <- m+rnorm(24*3,0,0.5)           #add noise
m <- floor(10*m)/10                #just one digits

r2xcluster(m)

## And once you have Xcluster program:

## Not run:
  system('Xcluster -f xcluster.txt -e 0 -p 0 -s 0 -l 0')
  h <- xcluster2r('xcluster.gtr')
  plot(h,hang=-1)

## End(Not run)
```

---

read.eisen *Read expression data from a file formatted for Eisen clustering*

Description

The input for Eisen-clustering is a slight variation of a tab delimited file. This method reads the expression data from such files as a matrix and provides optional additional information on the experiments as attributes.

Usage

```
read.eisen(file,sep="\t",dec=".", format.check = TRUE)
```

Arguments

file	The relative or absolute path to the file to be read, as internally forwarded to the read.table function.
sep	Separator of fields, passed on to read.table.
dec	Passed on to read.table. This is particulary helpful for the interpretation of data from localised spreadsheet programs.
format.check	TRUE or FLASE: to disable file format check.

Details

The software of Michael Eisen and its plain tab separated format for the presentation of gene expression data prior to their clustering is supported by many hard- and software providers, both as an input for their tools and as resulting from the analysis and normalisation of the chip images. To be able to read and write this format, the Bioconductor suite is enabled to easily reanalyse or extend older experiments that might have been analysed with the Eisen tools before.

Value

A numerical matrix is returned. It is a complete analogue of the Eisen-format, except the descriptions, weights and other information being passed to attributes. The first row will be the column names, the first column will be the respective row name. A second row that has a first empty field is referred to via the attribute "second.row". A column NAME is stored in the attribute "NAME".

Author(s)

Steffen Moeller

References

Michael Eisen Lab http://rana.lbl.gov/

Michael Hoon's Cluster 3.0 http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

Eisen M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. /Proc. Natl. Acad. Sci. USA /, 95:14863-14868.

De Hoon M.J.L., S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. Bioinformatics *20* (9): 1453–1454 (2004).

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

`xcluster` *Hierarchical clustering*

Description

Performs a hierarchical cluster analysis on a set of dissimilarities (this function launch an external program: Xcluster).

Usage

```
xcluster(data,distance="euclidean",clean=FALSE,tmp.in="tmp.txt",tmp.out="tmp.gtr
```

Arguments

`data`	a matrix (or data frame) which provides the data to analyze
`distance`	The distance measure used with *Xcluster*. This must be one of `"euclidean"`, `"pearson"` or `"notcenteredpearson"`. Any unambiguous substring can be given.
`clean`	a logical value indicating whether you want the true distances (`clean=FALSE`), or you want a clean dendrogram
`tmp.in, tmp.out`	temporary files for Xcluster

Details

Available distance measures are (written for two vectors x and y):

- Euclidean: Usual square distance between the two vectors (2 norm).
- Pearson: $1 - \text{cor}(x, y)$
- Pearson not centered: $1 - \frac{\sum_i x_i y_i}{\left(\sum_i x_i^2 \sum_i y_i^2\right)^{1/2}}$

Xcluster does not use usual agglomerative methods (single, average, complete), but compute the distance between each groups' barycenter for the distance between two groups.

This have a problem for this kind of data:

A	0	0
B	0	1
C	0.9	0.5

Ie: a triangular in $\mathbf{R}^2$, the distance between A and B is larger than the distance between the group A,B and C (with euclidean distance).

For that case it can be useful to use `clean=TRUE` and that mean that you must not consider A and B as a group without C.

Value

An object of class **hclust** which describes the tree produced by the clustering process. The object is a list with components:

`merge`	an $n-1$ by 2 matrix. Row i of `merge` describes the merging of clusters at step i of the clustering. If an element j in the row is negative, then observation $-j$ was merged at this stage. If j is positive then the merge was with the cluster formed at the (earlier) stage j of the algorithm. Thus negative entries in `merge` indicate agglomerations of singletons, and positive entries indicate agglomerations of non-singletons.
`height`	a set of $n-1$ non-decreasing real values. The clustering *height*: that is, the value of the criterion associated with the clustering `method` for the particular agglomeration.
`order`	a vector giving the permutation of the original observations suitable for plotting, in the sense that a cluster plot using this ordering and matrix `merge` will not have crossings of the branches.
`labels`	labels for each of the objects being clustered.
`call`	the call which produced the result.
`method`	the cluster method that has been used.
`dist.method`	the distance that has been used to create `d` (only returned if the distance object has a `"method"` attribute).

Note

Xcluster is a C program made by *Gavin Sherlock* that performs hierarchical clustering, K-means and SOM.

Xcluster is copyrighted. To get or have information about *Xcluster*: `http://genome-www.stanford.edu/~sherlock/cluster.html`

Author(s)

Antoine Lucas, http://mulcyber.toulouse.inra.fr/projects/amap/

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

See Also

r2xcluster, xcluster2r,hclust, hcluster

Examples

```
#    Create data
set.seed(1)
m <- matrix(rep(1,3*24),ncol=3)
m[9:16,3] <- 3 ; m[17:24,] <- 3    #create 3 groups
m <- m+rnorm(24*3,0,0.5)           #add noise
m <- floor(10*m)/10                #just one digits


# And once you have Xcluster program:
#
#h <- xcluster(m)
#
#plot(h)
```

---

xcluster2r *Importing Xcluster/Cluster output*

Description

Converting Xcluster/Cluster output (.gtr or .atr) to R hclust file

Usage

```
xcluster2r(file,distance="euclidean",labels=FALSE,fast=FALSE,clean=FALSE,
           dec='.')
```

Arguments

file	the path of a *Xcluster/Cluster* file (.gtr or .atr)
distance	The distance measure used with *Xcluster/Cluster*. This must be one of "euclidean", "pearson" or "notcenteredpearson". Any unambiguous substring can be given.
labels	a logical value indicating whether we use labels values (in the .cdt file) or not.
fast	a logical value indicating whether we reorganize data like R (Fast=FALSE) or we let them like *Xcluster/Cluster* did
clean	a logical value indicating whether you want the true distances (clean=FALSE), or you want a clean dendrogram (see details below).
dec	the character used in the file for decimal points

Details

See **xcluster** for more details.

Value

An object of class **hclust** which describes the tree produced by the clustering process.

Note

Xcluster is a C program made by *Gavin Sherlock* that performs hierarchical clustering, K-means and SOM.

Xcluster is copyrighted. To get or have information about *Xcluster*: http://genome-www.stanford.edu/~sherlock/cluster.html

Cluster is a program made by *Michael Eisen* that performs hierarchical clustering, K-means and SOM.

Cluster is copyrighted. To get or have information about *Cluster*: http://rana.lbl.gov/EisenSoftware.htm

Author(s)

Antoine Lucas, http://mulcyber.toulouse.inra.fr/projects/amap/

References

Antoine Lucas and Sylvain Jasson, *Using amap and ctc Packages for Huge Clustering*, R News, 2006, vol 6, issue 5 pages 58-60.

See Also

xcluster, r2xcluster, hclust, hcluster

Examples

```
#    Create data
set.seed(1)
m <- matrix(rep(1,3*24),ncol=3)
m[9:16,3] <- 3 ; m[17:24,] <- 3    #create 3 groups
m <- m+rnorm(24*3,0,0.5)           #add noise
m <- floor(10*m)/10                #just one digits

r2xcluster(m)


# And once you have Xcluster program:
#
#system('Xcluster -f xcluster.txt -e 0 -p 0 -s 0 -l 0')
#h <- xcluster2r('xcluster.gtr')
#plot(h,hang=-1)
```

Index