Package 'ClockstaRG'

July 29, 2014

Type Package

Title ClockstaRG: An implementation of ClockstaR for genomic data

Version 0.1

Date 2014-07-28

Author Sebastian Duchene

Maintainer Sebastian Duchene <sebastian.duchene@sydney.edu.au>

Description ClockstaRG is an implementation of ClockstaR for large data sets. Some steps of the analysis can be run in parallel. It also includes a fast clustering algorithm, CLARA.

Depends R (>= 2.15.0), ape (>= 3.0-8), cluster (>= 1.14.4), phangorn (>= 1.7-4), ClockstaR2 (>= 2.0)

License GPL (>=2)

R topics documented:

`ClockstaRG-package` *ClockstaRG: An implementation of ClockstaR for genomic data*

Description

CLockstaRG is an implementation of ClockstaR for large genomic data sets. Some of the steps in the algorithm, such as optimising the branch lengths of the trees, and estimating the tree distances can be run in parallel. This implementation also includes an fast clustering algorithm, CLARA, for large data sets.

Details

Package:	ClockstaRG
Type:	Package
Version:	1.0
Date:	2014-07-28
License:	GPL (<= 2)

Author(s)

Sebastian Duchene

Maintainer: Sebastian Duchene <sebastian.duchene@sydney.edu.au>

References

Duchêne, S, Molak, M., and Ho, SYW. "ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis." Bioinformatics (2014) 30 (7):1017-1019

See Also

github.com/sebastianduchene/clockstar github.com/sebastianduchene/nelsi

`boot.clara` *boot.clara performs a parametric bootstrap of the sBSDmin distances among trees. For every bootstrap replicate it calculates the cluster width (Wk) or the silhouette width (Sk), depending on whether run.clara.sk, or run.clara.sil is used. See the tutorial for details and instructions.*

Description

boot.clara performs a parametric bootstrap of the sBSDmin distances among trees. For every bootstrap replicate it calculates the cluster width (Wk) or the silhouette width (Sk), depending on whether run.clara.sk, or run.clara.sil is used. See the tutorial for details and instructions.

Usage

```
boot.clara(clus.matrix.name, out.boot.name = "out_boot.txt", nboot = 10, k.range = NULL, boot.te
```

Arguments

clus.matrix.name	The name of the file with the matrix with the MDS points, obtained with run.mds
out.boot.name	Name of the output file.
nboot	Number of bootstrap replicates to perform for every value of 'k'.
k.range	Range of values of 'k'
boot.temp.name	Name of temporary files of the bootstrap data. This is deleted after runing CLARA for the range of 'k' specified.
FUNboot	The function to use. run.clara.wk for cluster width, or run.clara.sil for silhouette width.

Details

Please see the tutorial for instructions.

Value

This function does not return a value to the console. It writes a text file with the statistic, Wk or Sk, for the bootstrap replicates

Author(s)

Sebastian Duchene

References

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

---

fill.matrix *fill.matirx fills the NA values in the pairwise sBSDmin matrix obtained with fold.sbsd*

Description

fill.matirx fills the NA values in the pairwise sBSDmin matrix obtained with fold.sbsd

Usage

```
fill.matrix(matrix.name, new.matrix.name = NULL)
```

Arguments

matrix.name	The name of the text file of the matrix obtained with fold.sbsd.
new.matrix.name	The file name of the matrix with filled NA values.

Details

Please see the tutorial for instructions on how to use this function.

Value

This function does not return a value to the prompt. The output is text file. Please see the tutorial for instructions on how to use this function.

Note

None. Please see the tutorial for instructions on how to use this function.

Author(s)

Sebastian Duchene

`fold.sbsd` *fold.sbsd formats the sBSDmin distances calculated with get.sbsd to be used in clustering algorithms.*

Description

fold.sbsd folds the output of get.sbsd into a pairwise matrix that can be used in clustering algorithms.

Usage

```
fold.sbsd(trees.file, comps.file, out.name = "test.fold.txt", method = "lite", comps.range = NUL
```

Arguments

`trees.file`	The file with the gene trees, as produced in optim.trees.g
`comps.file`	The file with the sBSDmin distances among pairs of trees.
`out.name`	The name of the output file.
`method`	There are two methods available: 'memory' and 'lite'. The 'memory' method loads all the sBSDmin distances into RAM, while the 'lite' method reads one at a time. The 'memory' method is faster, but with 'lite' it is possible to overcome memory limitations from very large files.
`comps.range`	A sequence of integers. This is the range of sBSDmin distances to read and format into the pairwise matrix. Use this for very large files that are difficult to read into RAM.

Details

None. See the tutorial for instructions and deatails on some of the settings.

Value

This function does not return a value to the prompt. Instead, it writes a text file with the pairwise distance matrix. Note that it only prints the below diagonal values of the matrix, and the rest is filled with NA. To complete these values in the matrix use fill.matrix. This is a requirement of some clusteing algorithms.

Note

See the tutorial for instructions and deatails on some of the settings.

Author(s)

Sebastian Duchene

`get.gap` *Get gap statistic.*

Description

get.gap obtains the gap statistic for a range of Wk and bootstrap replicates.

Usage

```
get.gap(true.data, boot.data)
```

Arguments

`true.data`	The data with Wk for a range of k. Note that this is not the file name, but the data loaded as a matrix or data fram in R.
`boot.data`	The data with Wk for bootstrap replicates for a range of k. Note that this is not the file name, but the data loaded as a matrix or data fram in R.

Details

Please see the tutorial for instructions.

Value

A matrix with the Gap statistic and the standard errors (SE). The first column is the Gap, and the second is the SE.

Note

Please see the tutorial and the documentation for 'clusGap' function of the 'cluster' package for more details.

Author(s)

Sebastian Duchene

References

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. _Journal of the Royal Statistical Society B_, *63*, 411-423.

Tibshirani, R., Walther, G. and Hastie, T. (2000). Estimating the number of clusters in a dataset via the Gap statistic. Technical Report. Stanford.

Per Broberg (2006). SAGx: Statistical Analysis of the GeneChip. R package version 1.9.7. <URL: http://home.swipnet.se/pibroberg/expression_hemsida1.html>

get.sbsd	*get.sbsd calculates the sBSDmin distance for pairs of trees. It is similar to the function bsd.dist in ClockstaR2, but it can handle data sets with many trees.*

Description

get.sbsd calculates the sBSDmin distance for pairs of trees. It is similar to the function bsd.dist in ClockstaR2, but it can handle data sets with many trees.

Usage

```
get.sbsd(trees.file, comps.file, method = "lite", range.comps = NULL, out.file = "sbsd.txt")
```

Arguments

trees.file	The name of the file with the gene trees. This is a newick with a list of the gene trees with branch lengths. The trees can include names. See the tutorial and example files for details.
comps.file	The name of the file with the tree comparissons.
method	There are two methods available: 'memory' and 'lite'. Use 'memory' to load all the trees in RAM. Use 'lite' to read the trees one at a time, so it can be parallelised over different machines or processors. The method 'memory' is faster, i the analysis is not run in parallel.
range.comps	This is useful for the 'lite' method. In this case it is possible to run several instances of the analysis, each for a range of sBSDmin distances among trees.
out.file	The name of the file to write the output.

Details

Please see the tutorial for instructions.

Value

This function does not return a value to the prompt. It saves the output to a text file.

Note

NONE

Author(s)

Sebastian Duchene

See Also

NONE

`make.tree.comps` *This function creates a file with the names of the pairwise comparissons for all trees.*

Description

This function creates a file with the names of the pairwise comparissons for all trees.

Usage

```
make.tree.comps(trees.file, tree.comps = "treecomps.txt")
```

Arguments

`trees.file`	Name of file with the trees
`tree.comps`	Name of file to write the list of tree comparissons

Details

None. See the tutorial for instructions on how to run.

Value

This function does not return values on the command prompt. It creates a text file with the tree comparissons. See the tutorial for instructions on how to run.

Note

See the tutorial for instructions on how to run.

Author(s)

Sebastian Duchene

References

NONE

`optim.trees.g` *optim.trees.g is can be used to optimise the branch lengths for a tree topology over many alignments. Please see the tutorial for instructions to use.*

Description

This function is practical for large data sets. For data sets with less than 20 genes, it may be sufficient to use the standard version of ClockstaR. Please see the tutorial for instructions on how to use.

Usage

```
optim.trees.g(data.folder, init.alin = NULL, end.alin = NULL, out.trees = "out.trees", model.tes
```

Arguments

data.folder	This is the folder with the individual alignments in fasta files and a tree topology to optimise the branch lengths.
init.alin	The number of the first alignment. This is useful for runing several instances of this function. Please see the tutorial for instructions on how to use.
end.alin	The number of the first alignment. This is useful for runing several instances of this function. Please see the tutorial for instructions on how to use.
out.trees	A character. The name of the file to write the optimised trees.
model.test	A logical. Select T for model selection. The default is F. It is recommended to leave the default because it can be very slow to run
out.models	The name of the file to write the models to.

Details

Please see the tutorial for instructions on how to use

Value

The function does not return anything to the prompt. It saves the trees with optimised branch lenghts to the file specified.

Author(s)

Sebastian Duchene

References

Duchêne, S, Molak, M., and Ho, SYW. "ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis." Bioinformatics (2014) 30 (7):1017-1019

See Also

optim.trees.interactive in ClockstaR

Examples

```
## Not run:
optim.trees.g(data_folder)

## End(Not run)
```

`run.clara.sil` *run.clara is used to run the CLARA algorithm from package 'cluster'. For values of 'k' from 2:N, it estimates the average silhouette width.*

Description

run.clara is used to run the CLARA algorithm from package 'cluster'. For values of 'k' from 2:N, it estimates the average silhouette width.

Usage

```
run.clara.sil(clus.matrix.name, out.clus.name = "out_clus_sil.txt", k.range = NULL)
```

Arguments

`clus.matrix.name`	This is the filename of the MDS of the sBSDmin distances for pairs of trees.
`out.clus.name`	The name of the out put name of the avearage cluster width (Wk) for the values of k in k.range.
`k.range`	The range of 'k' values to calculate Wk.
`...`	Other arguments passed to the 'clara' function from pacakge 'cluster'.

Details

Please see the tutorial for instructions.

Value

This function does not return a value to the prompt. It writes a text file with the average silhouette width for values of k.

Author(s)

Sebastian Duchene

References

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

run.clara.wk	*run.clara is used to run the CLARA algorithm on the MDS of the sBSDmin distances for pairs of trees. For every value of 'k' it estimates the cluster width, 'Wk'.*

Description

run.clara is a raper from the CLARA algorithm ,implemented in the package 'cluster. It can be run for a range of values of k.

Usage

```
run.clara.wk(clus.matrix.name, out.clus.name = "out_clus_wk.txt", k.range = NULL, ...)
```

Arguments

clus.matrix.name	This is the filename of the MDS of the sBSDmin distances for pairs of trees.
out.clus.name	The name of the out put name of the avearage cluster width (Wk) for the values of k in k.range.
k.range	The range of 'k' values to calculate Wk.
...	Other arguments passed to the 'clara' function from pacakge 'cluster'.

Details

Please see the tutorial for instructions.

Value

This function does not return a value to the prompt. It writes a text file with Wk for values of k.

Author(s)

Sebastian Duchene

References

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

`run.mds` *run.mds obtains a multidimensional scaling (MDS) of the pariwise sBSDmin distances between trees.*

Description

run.mds obtains a multidimensional scaling (MDS) of the pariwise sBSDmin distances between trees.

Usage

```
run.mds(matrix.name, out.mds.name = "test_mds.txt")
```

Arguments

`matrix.name`	The path and name of the file with the pairwise sBSDmin distances between trees.
`out.mds.name`	The name of the output files. Please see the tutorial for instructions.

Details

None. Please see the tutorial for instructions.

Value

This function does not return an object to the prompt. Instead, it returns two text files: The points of the MDS of the sBSDmin distances, and the eigen values.

Author(s)

Sebastian Duchene

Index