Package 'sitePath'

October 16, 2019

Type Package

Title Detection of sites with fixation of amino acid substitutions in protein evolution

Version 1.0.3

Author Chengyang Ji, Aiping Wu

Maintainer Chengyang Ji <chengyang.ji12@alumni.xjtlu.edu.cn>

Description The package does hierarchical search for fixation events given multiple sequence alignment and phylogenetic tree. These fixation events can be specific to a phylogenetic lineages or shared by multiple lineages.

License MIT + file LICENSE

Depends R (>= 3.6.0)

Imports ape, seqinr, Rcpp, methods, graphics, utils, stats

Suggests testthat, knitr, rmarkdown, BiocStyle

LinkingTo Rcpp

RoxygenNote 6.1.1

Encoding UTF-8

VignetteBuilder knitr

URL https://wuaipinglab.github.io/sitePath/

BugReports https://github.com/wuaipinglab/sitePath/issues

biocViews Alignment, MultipleSequenceAlignment, Software

git_url https://git.bioconductor.org/packages/sitePath

git_branch RELEASE_3_9

git_last_commit 837018a

git_last_commit_date 2019-06-25

Date/Publication 2019-10-15

R topics documented:

---

addMSA *Prepare data for sitePath analysis*

Description

sitePath requires both tree and sequence alignment to do the analysis. addMSA wraps read.alignment function in seqinr package and helps match names in tree and sequence alignment. Either provide the file path to an alignment file and its format or an alignment object from the return of read.alignment function. If both the file path and alignment object are given, the function will use the sequence in the alignment file.

Usage

```
addMSA(tree, msaPath = "", msaFormat = "", alignment = NULL)
```

Arguments

tree	a phylo object. This commonly can be from tree paring function in ape or ggtree. All the tip.label should be found in the sequence alignment
msaPath	The file path to the multiple sequence alignment file
msaFormat	The format of the multiple sequence alignment file
alignment	an alignment object. This commonly can be from sequence parsing function in the seqinr package. Sequence names in the alignment should include all tip.label in the tree

Value

addMSA returns a phylo object with matched multiple sequence alignment

Examples

```
data(zikv_tree)
msaPath <- system.file("extdata", "ZIKV.fasta", package = "sitePath")
addMSA(zikv_tree, msaPath = msaPath, msaFormat = "fasta")
```

extractTips *Extract sitePath for a single site*

Description

Retrieve the name of the tips involved in the fixation

Usage

```
## S3 method for class 'fixationSites'
extractTips(x, site, select = 1, ...)

## S3 method for class 'multiFixationSites'
extractTips(x, site, select = 1, ...)
```

Arguments

x	A fixationSites or a multiFixationSites object.
site	A site predicted to experience fixation.
select	For a site, there theoretically might be more than one fixation on different lineages. You may use this argument to extract for a specific fixation of a site. The default is the first fixation of the site.
...	Other arguments

Value

The name of the tips involved in the fixation

findSites *Finding sites with variation*

Description

Single nucleotide polymorphism (SNP) in the whole package refers to variation of amino acid. findSNPsite will try to find SNP in the multiple sequence alignment. A reference sequence and gap character may be specified to number the site. This is irrelevant to the intended analysis but might be helpful to evaluate the performance of fixationSites.

After finding the lineagePath of a phylogenetic tree, fixationSites uses the result to find those sites that show fixation on some, if not all, of the lineages. Parallel evolution is relatively common in RNA virus. There is chance that some site be fixed in one lineage but does not show fixation because of different sequence context.

After finding the lineagePath of a phylogenetic tree, multiFixationSites uses the result to find those sites that show multiple fixations on some, if not all, of the lineages.

Usage

```
SNPsites(tree, reference = NULL, gapChar = "-", minSNP = NULL)

## S3 method for class 'lineagePath'
fixationSites(paths, reference = NULL,
  gapChar = "-", tolerance = 0.01, minEffectiveSize = NULL,
  extendedSearch = TRUE, ...)

## S3 method for class 'lineagePath'
multiFixationSites(paths, reference = NULL,
  gapChar = "-", minEffectiveSize = NULL, extendedSearch = TRUE, ...)
```

Arguments

tree	The return from addMSA function
reference	Name of reference for site numbering. The name has to be one of the sequences' name. The default uses the intrinsic alignment numbering
gapChar	The character to indicate gap. The numbering will skip the gapChar for the reference sequence.
minSNP	Minimum number of amino acid variation to be a SNP
paths	a lineagePath object returned from lineagePath function
tolerance	A vector of two integers to specify maximum amino acid variation before/after mutation. Otherwise the mutation will not be counted into the return. If more than one number is given, the ancestral takes the first and descendant takes the second as the maximum. If only given one number, it's the maximum for both ancestral and descendant. The default is 0.01
minEffectiveSize	A vector of two integers to specifiy minimum tree tips involved before/after mutation. Otherwise the mutation will not be counted into the return. If more than one number is given, the ancestral takes the first and descendant takes the second as the minimum. If only given one number, it's the minimum for both ancestral and descendant.
extendedSearch	Whether to extend the search. The terminal of each lineagePath is a cluster of tips. To look for the fixation mutation in the cluster, the common ancestral node of farthest tips (at least two) will be the new terminal search point.
...	further arguments passed to or from other methods.

Value

SNPsite returns a list of qualified SNP site

fixationSites returns a list of mutations with names of the tips involved. The name of each list element is the discovered mutation. A mutation has two vectors of tip names: 'from' before the fixation and 'to' after the fixation.

multiFixationSites returns sites with multiple fixations.

Examples

```
data("zikv_tree_reduced")
data("zikv_align_reduced")
tree <- addMSA(zikv_tree_reduced, alignment = zikv_align_reduced)
```

```
SNPsites(tree)
fixationSites(
    lineagePath(tree),
    tolerance = c(1, 1),
    minEffectiveSize = c(50, 50)
)
data(h3n2_tree_reduced)
data(h3n2_align_reduced)
tree <- addMSA(h3n2_tree_reduced, alignment = h3n2_align_reduced)
multiFixationSites(lineagePath(tree))
```

`h3n2_align` *Multiple sequence alignment of H3N2's HA protein*

Description

The raw protein sequences were downloaded from NCBI database.

Usage

```
data(h3n2_align)
```

Format

a `alignment` object

`h3n2_align_reduced` *Truncated data for runnable example*

Description

This is a truncated version of `h3n2_align`

Usage

```
data(h3n2_align_reduced)
```

Format

a `alignment` object

h3n2_tree — *Phylogenetic tree of H3N2's HA protein*

Description

Tree was built from h3n2_align using RAxML with default settings.

Usage

```
data(h3n2_tree)
```

Format

a phylo object

h3n2_tree_reduced — *Truncated data for runnable example*

Description

This is a truncated version of h3n2_tree

Usage

```
data(h3n2_tree_reduced)
```

Format

a phylo object

plot.lineagePath — *Visualize phylogenetic lineages*

Description

Visualize lineagePath object. A tree diagram will be plotted and paths are black solid line while the trimmed nodes and tips will use grey dashed line.

Usage

```
## S3 method for class 'lineagePath'
plot(x, y = TRUE, ...)
```

Arguments

x	A lineagePath object
y	Whether plot the nodes from the extendedSearch in fixationSites
...	Arguments in plot.phylo functions.

Value

The function only makes plot and returns no value (It behaviors like the generic plot function).

Examples

```
data("zikv_tree")
data("zikv_align")
tree <- addMSA(zikv_tree, alignment = zikv_align)
plot(lineagePath(tree, 0.996))
```

---

plotSingleSite *Color the tree by a single site*

Description

For lineagePath, the tree will be colored according to the amino acid of the site. The color scheme tries to assign distinguishable color for each amino acid.

For fixationSites, it will color the ancestral tips in red, descendant tips in blue and excluded tips in grey.

For multiFixationSites, it will color the tips which have their site fixed. The color will use the same amino acid color scheme as plotSingleSite.lineagePath

Usage

```
## S3 method for class 'lineagePath'
plotSingleSite(x, site, showPath = FALSE,
  reference = NULL, gapChar = "-", ...)

## S3 method for class 'fixationSites'
plotSingleSite(x, site, ...)

## S3 method for class 'multiFixationSites'
plotSingleSite(x, site, ...)
```

Arguments

x	A fixationSites object from fixationSites or the return from addMSA function.
site	One of the mutations in the fixationSites object. It should be from the names of the object. Or an integer to indicate a site could be provide. The numbering is consistent with the reference defined at fixationSites.
showPath	If plot the lineage result from lineagePath.
reference	Name of reference for site numbering. The name has to be one of the sequences' name. The default uses the intrinsic alignment numbering.
gapChar	The character to indicate gap. The numbering will skip the gapChar for the reference sequence.
...	Arguments in plot.phylo functions and other arguments.

Value

The function only makes plot and returns no value (It behaviors like the generic plot function).

Examples

```
data("zikv_tree")
data("zikv_align")
tree <- addMSA(zikv_tree, alignment = zikv_align)
paths <- lineagePath(tree, 0.996)
plotSingleSite(paths, 139)
## Not run:
fixations <- fixationSites(paths)
plotSingleSite(fixations, 139)

## End(Not run)
## Not run:
multiFixations <- multiFixationSites(paths)
plotSingleSite(multiFixations, 1542)

## End(Not run)
```

---

pre-assessment *Things can be done before the analysis*

Description

similarityMatrix calculates similarity between aligned sequences The similarity matrix can be used in groupTips or lineagePath

sneakPeek is intended to plot similarity as a threshold against number of output lineagePath. This plot is intended to give user a feel about how many sitePaths they should expect from the similarity threshold. The number of lineagePath should not be too many or too few. The result excludes where the number of lineagePath is greater than number of tips divided by 20 or self-defined maxPath. The zero lineagePath result will also be excluded

Usage

```
similarityMatrix(tree)

sneakPeek(tree, step = NULL, maxPath = NULL, minPath = 1,
  makePlot = FALSE)
```

Arguments

tree	The return from addMSA function
step	the similarity window for calculating and ploting. To better see the impact of threshold on path number. This is preferably specified. The default is one 50th of the difference between 1 and minimal pairwise sequence similarity.
maxPath	maximum number of path to return show in the plot. The number of path in the raw tree can be far greater than trimmed tree. To better see the impact of threshold on path number. This is preferably specified. The default is one 20th of tree tip number.

minPath	minimum number of path to return show in the plot. To better see the impact of threshold on path number. This is preferably specified. The default is 1.
makePlot	whether make a dot plot when return

Value

similarityMatrix returns a diagonal matrix of similarity between sequences

sneakPeek return the similarity threhold against number of lineagePath. There will be a simple dot plot between threshold and path number if makePlot is TRUE.

Examples

```
data("zikv_tree")
data("zikv_align")
tree <- addMSA(zikv_tree, alignment = zikv_align)
simMatrix <- similarityMatrix(tree)
sneakPeek(tree)
```

---

treemer *Topology-dependent tree trimming*

Description

groupTips uses sequence similarity to group tree tips. Members in a group are always constrained to share the same ancestral node. Similarity between two tips is derived from their multiple sequence alignment. The site will not be counted into total length if both are gap. Similarity is calculated as number of matched divided by the corrected total length. So far the detection of divergence is based on one simple rule: the miminal pairwise similarity. The two branches are decided to be divergent if the similarity is lower than the threshold. (Other more statistical approaches such as Kolmogorov-Smirnov Tests among pair-wise distance could be introduced in the future)

lineagePath finds the lineages of a phylogenetic tree providing the corresponding sequence alignment. This is done by trimming the tree to the ancestor node of tips in each group and then find the bifurcated terminals of the trimmed tree. The nodepath between root node and the bifurcated terminals is the lineages. In order to extend the search of mutational site. The lineages will tag some of its trailing nodes. Here nodes up to the ancestor of the tips with the longest nodepath are added.

Usage

```
groupTips(tree, similarity = NULL, simMatrix = NULL,
  forbidTrivial = TRUE, tipnames = TRUE)

lineagePath(tree, similarity = NULL, simMatrix = NULL,
  forbidTrivial = TRUE)
```

Arguments

tree	The return from addMSA function
similarity	Similarity threshold for tree trimming. If not provided, an average value of similarity among all sequences will be used.

simMatrix	S diagonal matrix of similarity between sequences
forbidTrivial	Does not allow trivial trimming
tipnames	If return as tipnames

Value

grouping of tips

path represent by node number

Examples

```
data("zikv_tree")
data("zikv_align")
tree <- addMSA(zikv_tree, alignment = zikv_align)
groupTips(tree, 0.996)
lineagePath(tree, 0.996)
```

---

zikv_align *Multiple sequence alignment of Zika virus polyprotein*

Description

The raw protein sequences were downloaded from ViPR database (https://www.viprbrc.org/) and aliged using MAFFT. with default settings.

Usage

```
data(zikv_align)
```

Format

a alignment object

---

zikv_align_reduced *Truncated data for runnable example*

Description

This is a truncated version of zikv_align

Usage

```
data(zikv_align_reduced)
```

Format

a alignment object

zikv_tree *Phylogenetic tree of Zika virus polyprotein*

Description

Tree was built from zikv_align using RAxML with default settings.The tip "ANK57896" was used as outgroup to root the tree.

Usage

```
data(zikv_tree)
```

Format

a phylo object

zikv_tree_reduced *Truncated data for runnable example*

Description

This is a truncated version of zikv_tree

Usage

```
data(zikv_tree_reduced)
```

Format

a phylo object

Index