Package ‘sitePath’

April 15, 2020

Type Package

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

Version 1.2.3

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 7.1.0

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_10

git_last_commit 86e44d4

git_last_commit_date 2020-04-01

Date/Publication 2020-04-14

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 parsing 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 info for fixation of a single site*

Description

The result of `fixationSites` contains all the possible sites with fixation mutation. The function `extractTips` retrieves the name of the tips involved in the fixation.

The function `extractSite` can be used to extract the fixation info of a single site.

Usage

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

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

## S3 method for class 'sitePath'
extractTips(x, select = 1, ...)

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

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

Arguments

`x`	A `fixationSites` or a `multiFixationSites` or a `sitePath` 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 function `extractTips` returns the name of the tips involved in the fixation.

The function `extractSite` returns a `sitePath` object

Examples

```
data(zikv_tree_reduced)
data(zikv_align_reduced)
tree <- addMSA(zikv_tree_reduced, alignment = zikv_align_reduced)
mutations <- fixationSites(lineagePath(tree))
extractTips(mutations, 139)
extractSite(mutations, 139)
```

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 random sampling on the original tree and applies the method used in `fixationSites` to each sampled tree and summarize the results from all the samples.

Usage

```
SNPsites(tree, minSNP = NULL)

## S3 method for class 'lineagePath'
fixationSites(
  paths,
  minEffectiveSize = NULL,
  searchDepth = 1,
  method = c("compare", "insert", "delete"),
  ...
)

## S3 method for class 'lineagePath'
multiFixationSites(
  paths,
  samplingSize = NULL,
  samplingTimes = 100,
  minEffectiveSize = 0,
  searchDepth = 1,
  method = c("compare", "insert", "delete"),
  ...
)
```

Arguments

`tree`	The return from `addMSA` function
`minSNP`	Minimum number of amino acid variation to be a SNP
`paths`	a `lineagePath` object returned from `lineagePath` function
`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.

`searchDepth` The function uses heuristic search but the termination of the search cannot be intrinsically decided. `searchDepth` is needed to tell the search when to stop.

`method` The strategy for predicting the fixation. The basic approach is entropy minimization and can be achieved by adding or removing fixation point, or by comparing the two.

`...` further arguments passed to or from other methods.

`samplingSize` The number of tips sampled for each round of resampling. It shoud be at least 10th and at most nine 10ths of the tree tips.

`samplingTimes` The total times of random sampling to do. It should be greater than 100.

Value

`SNPsite` returns a list of qualified SNP site

`fixationSites` returns a list of fixation mutations with names of the tips involved.

`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))
```

`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, showTips = FALSE, ...)
```

Arguments

`x`	A `lineagePath` object
`y`	Whether plot the nodes from the `extendedSearch` in `fixationSites`
`showTips`	Whether to plot the tip labels. The default is `FALSE`.
`...`	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))
```

`plot.sitePath` *Plot the fixation mutation*

Description

Visualize the `sitePath` object which is the basic unit of the result of `fixationSites` and `multiFixationSites`.

Usage

```
## S3 method for class 'sitePath'
plot(x, y = NULL, showTips = FALSE, ...)
```

Arguments

`x`	A `sitePath` object
`y`	A `sitePath` object can have more than one fixation path. This is to select which path to plot. The default is NULL which will plot all the paths.
`showTips`	Whether to plot the tip labels. The default is `FALSE`.
`...`	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).

See Also

`plotSingleSite`

Examples

```
data(zikv_align_reduced)
data(zikv_tree_reduced)
tree <- addMSA(zikv_tree_reduced, alignment = zikv_align_reduced)
paths <- lineagePath(tree)
fixations <- fixationSites(paths)
plot(fixations[[1]])
```

`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, showTips = FALSE, ...)

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

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

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.
`showTips`	Whether to plot the tip labels. The default is `FALSE`.
`...`	Arguments in `plot.phylo` functions and other arguments.
`select`	Select which fixation path in to plot. The default is NULL which will plot all the fixations.

Value

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

See Also

plot.sitePath

Examples

```
data(zikv_tree)
data(zikv_align)
tree <- addMSA(zikv_tree, alignment = zikv_align)
paths <- lineagePath(tree)
plotSingleSite(paths, 139)
fixations <- fixationSites(paths)
plotSingleSite(fixations, 139)
## 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' (actually the least percentage of 'major SNP') as a threshold against number of output lineagePath. This plot is intended to give user a rought view about how many lineages they could expect from the 'similarity' threshold in the function lineagePath. The number of lineagePath is preferably 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 user-defined maxPath. The zero lineagePath result will also be excluded.

Usage

```
similarityMatrix(tree)

sneakPeek(tree, step = 10, 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. The default is 10.
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. 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)
```

`setSiteNumbering` *Set site numbering to the reference sequence*

Description

A reference sequence can be used to define a global site numbering scheme for multiple sequence alignment. The gap in the reference will be skipped so the site ignored in numbering.

Usage

```
## S3 method for class 'phylo'
setSiteNumbering(x, reference = NULL, gapChar = "-", ...)

## S3 method for class 'lineagePath'
setSiteNumbering(x, reference = NULL, gapChar = "-", ...)
```

Arguments

`x` The object to set site numbering. It could be a `phylo` object after `addMSA` or a `lineagePath` object. The function for `fixaitonSites` and `multiFixationSites` will be added in later version.

`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.

`...` further arguments passed to or from other methods.

Value

A `phylo` object with site numbering mapped to reference sequence

Examples

```
data(zikv_tree)
msaPath <- system.file('extdata', 'ZIKV.fasta', package = 'sitePath')
tree <- addMSA(zikv_tree, msaPath = msaPath, msaFormat = 'fasta')
setSiteNumbering(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 finding 'major SNPs' which usually accumulate along the evolutionary pathways. 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 in groupTips. If not provided, the mean similarity substract standard deviation of all sequences will be used. And for lineagePath, this decides how minor SNPs are to remove. If provided as fraction between 0 and 1, then the minimum number of SNP will be total tips times similariy. If provided as integer greater than 1, the minimum number will be similariy. The default similariy is 0.1 for lineagePath.
simMatrix	A diagonal matrix of similarities for each pair of sequences. This parameter will not have effect in the function lineagePath.
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)
```

`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