Package 'orthGS'

April 30, 2025

Title Orthology vs Paralogy Relationships among Glutamine Synthetase from Plants

Version 0.1.8

Description Tools to analyze and infer orthology and paralogy relationships between glutamine synthetase proteins in seed plants.

License GPL (>= 2)

Encoding UTF-8

RoxygenNote 7.3.2

Depends R (>= 2.10)

LazyData true

Imports ape, bio3d, castor, igraph, phangorn, phytools, seqinr, TreeTools

Suggests BiocManager, Biostrings, fs, knitr, muscle, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Author Elena Aledo [aut, cre, cph], Juan-Carlos Aledo [aut] (<https://orcid.org/0000-0002-3497-9945>)

Maintainer Elena Aledo <elenaaledoesteban@gmail.com>

Repository CRAN

Date/Publication 2025-04-30 08:50:02 UTC

Contents

agf 2
AngGym 3
A_selected 4
coltips 4
gapless_msa 5
getseqGS 6
madRoot 6

mapTrees 7
mltree 8
msa 9
orthG 10
orthology 10
orthP 11
sdf 12
selected_tr 13
speciesGS 13
subsetGS 14

Index **15**

`agf` *Angiosperms Gymnosperms Ferns*

Description

Angiosperms Gymnosperms Ferns

Usage

```
agf
```

Format

A dataframe with 275 rows (GS proteins) and 23 columns:

n Reference number

phylo_id Unique identification label of the protein/gen

species Species

taxon Acrogymnospermae, Angiospermae, Polypodiopsida

dna CDS sequence

prot Protein sequence

short Unique three letter identification of the species

gs GS2, GS1a or GS1b_Ang, GS1b_Gym

pI isoelectric point

factor Ferns, GS2, GS1a, GS1b_Ang, GS1b_Gym

size number of residues

CSpos position signal

prediction prediction

Lk_SP seq pep

Lk_mTP mit

Lk_cTP chl

Lk_Thylak thy

secAa amino acid at position 2

core core

dabase db

acc acc

up_id uniprot

note note

Source

It has been manually curated by the authors

`AngGym` *Angiosperms Gymnosperms*

Description

Angiosperms Gymnosperms

Usage

```
AngGym
```

Format

A dataframe with 155 rows (GS proteins) and 10 columns:

n Reference number

phylo_id Unique identification label of the protein/gen

species Species

taxon Acrogymnospermae or Angiospermae

class Angiosperms: Amborellopsida, Liliopsida, Magnoliopsida; Gymnosperms: Ginkgoopsida, Cycadopsida, Gnetopsida, Pinopsida

dna CDS sequence

prot Protein sequence

short Unique three letter identification of the species

gsLineage Either GS2, GS1a or GS1b

plant_group Primitive angiosperms, Modern angiosperms, Ginkgo-Cycadales, Gnetales, Pinacea, Conifer II

Source

It has been manually curated by the authors

A_selected *Adjacency Matrix for Orthology Graph*

Description

155 x 155 square matrix (155 GS proteins from 45 seed plant species)

Usage

```
A_selected
```

Format

A matrix with 155 rows and 155 columns

Source

It has been generated using the function orthG::mapTrees() and the reconciliation output file 'selected'. Verbigracia: orthG::mapTrees('./inst/extdata/selected') The reconciliation was carried out using RANGER-DTL with parameters D = 1, T = 10 and L = 1.

coltips *Colouring Tree Tips*

Description

Make a color vector for colouring tree tips

Usage

```
coltips(phy)
```

Arguments

phy	tree as a phylo object

Details

Each tip is given a color according to the nature of the isoform: green (GS2), blue (GS1a), brown (GS1b Gym), salmon (GS1b Ang), purple (other).

Value

a color vector as long as the number of tips

Examples

```
coltips(ape::read.tree(text = "((Bdi, Sly), (Pp, Ap));"))
```

gapless_msa *Remove Gaps in a MSA*

Description

Removes gaps in a given msa.

Usage

```
gapless_msa(msa, seqtype = 'AA', df = TRUE, sfile = FALSE)
```

Arguments

msa	input alignment.
seqtype	the nature of the sequences: 'DNA' or 'AA'.
df	logical. When TRUE msa should be a matrix, when FALSE msa should be a string giving the path to a fasta file containing the alignment.
sfile	if different to FALSE, then it should be a string indicating the path to save a fasta alignment file.

Details

It should be noted that this function does not carry out the alignment itself.

Value

an alignment without gaps in form of matrix or a file containing such an alignment in fasta format.

See Also

msa

Examples

```
## Not run: gapless_msa(msa(sequences = c("APGW", "AGWC", "CWGA"),ids = c("a", "b", "c"))$ali)
```

getseqGS *Get the GS Sequence*

Description

Provides the requested GS sequence

Usage

```
getseqGS(phylo_id, molecule = "Prot")
```

Arguments

phylo_id	the unique sequence identifier
molecule	either "Prot" or "CDS"

Details

The identifier should be one of the 'phylo_id' from data(agf).

Value

The requested sequence as a character string.

Examples

```
getseqGS("Pp_GS1b_2")
```

madRoot *Find The Root of a Phylogenetic Tree Using MAD Method*

Description

Finds the root of an unrooted phylogenetic tree by minimizing the relative deviation from the molecular clock.

Usage

```
madRoot(tree, output_mode = 'phylo')
```

Arguments

tree	unrooted tree string in newick format or a tree object of class 'phylo'.
output_mode	amount of information to return. If 'phylo' (default) only the rooted tree is returned. If 'stats' also a structure with the ambiguity index, clock cv, the minimum ancestor deviation and the number of roots. If 'full' also an unrooted tree object, the index of the root branch, the branch ancestor deviations and a rooted tree object.

Details

This function is a slight modification of the code provided by Tria et al at https://www.mikrobio.uni-kiel.de/de/ag-dagan/ressourcen.

Value

a rooted tree and supplementary information if required.

Author(s)

Tria, F. D. K., Landan, G. and Dagan, T.

References

Tria, F. D. K., Landan, G. and Dagan, T. Nat. Ecol. Evol. 1, 0193 (2017).

Examples

```
## Not run: a <- msa(sequences=c("RAPGT", "KMPGT", "ESGGT"), ids = letters[1:3])$ali
rownames(a) <- letters[1:3]
tr <- mltree(a)$tree
rtr <- madRoot(tr)
## End(Not run)
```

---

mapTrees *Map Gene Tree into Species Tree*

Description

Maps a gene/protein tree into a species tree

Usage

```
mapTrees(path2rec)
```

Arguments

path2rec	path to the file containing the reconciliation output.

Details

Mapping gene tree into species tree allow to infer the sequence of events (Duplication, Speciation, Transfer).

Value

A list with three elements. The first one is a 'phylo' object where the nodelabels indicate the event: D, duplication or T transfer. If no label is shown is because the event correspond to speciation. The second element is a dataframe (the first column is the label of the internal nodes in the gene tree; the second column is the label of the internal nodes in the species tree, and the third and fourth columns label each internal node according to the inferred event). The third element of the list is an adjacency matrix: 1 when two proteins are orthologous, 0 if they are paralogous.

Examples

```
mapTrees(fs::path_package("extdata", "representatives", package = "orthGS"))
```

`mltree` *Build Up a ML Tree*

Description

Given an alignment builds an ML tree.

Usage

```
mltree(msa, df = TRUE, gapl = TRUE, model = "WAG")
```

Arguments

`msa`	input alignment.
`df`	logical. When TRUE msa should be a dataframe, when FALSE msa should be a string giving the path to a fasta file containing the alignment.
`gapl`	logical, when TRUE a gapless alignment is used.
`model`	allows to choose an amino acid models (see the function phangorn::as.pml)

Details

The function makes a NJ tree and then improve it using an optimization procedure based on ML.

Value

a ML optimized tree (and parameters)

See Also

gapless_msa

Examples

```
## Not run: a <- msa(sequences=c("RAPGT", "KMPGT", "ESGGT"), ids = letters[1:3])$ali
rownames(a) <- letters[1:3]
tr <- mltree(a)$tree
## End(Not run)
```

msa *Multiple Sequence Alignment*

Description

Aligns multiple protein, DNA or CDS sequences using inhouse software.

Usage

```
msa(sequences, ids = names(sequences), seqtype = "prot", method, sfile = FALSE)
```

Arguments

sequences	vector containing the sequences as strings.
ids	character vector containing the sequences' ids.
seqtype	it should be either "prot" of "dna" or "cds" (see details).
method	the software to be used for the alignment, as invoked in your system. For instance, "muscle3" or "clustalo".
sfile	if different to FALSE, then it should be a string indicating the path to save a fasta alignment file.

Details

Either Clustal Omega or MUSCLE must be installed, and their executable be in your system's PATH. If seqtype is set to "cds" the sequences must not contain stop codons and they will be translated using the standard code. Afterward, the amino acid alignment will be used to lead the codon alignment.

Value

Returns a list of four elements. The first one ($seq) provides the sequences analyzed, the second element ($id) returns the identifiers, the third element ($aln) provides the alignment in fasta format and the fourth element ($ali) gives the alignment in matrix format.

Examples

```
## Not run: msa(sequences = c("APGW", "AGWC", "CWGA"),
                              ids = c("a", "b", "c"))
## End(Not run)
```

orthG *Infer GS OrthoGroups Within a Set of Species*

Description

Infers GS orthogroups using tree reconciliation

Usage

```
orthG(set = "all")
```

Arguments

set	set of species of interest provided as a character vector either with the binomial or short code of the species (see data(sdf)).

Details

When set = "all", all the species in the database will be included.

Value

A list with two elements. The first one is the adjacency matrix (1 for orthologous, 0 for paralogous). The second element is an orthogroup graph.

Examples

```
orthG(set = c("Pp", "Psy", "Psm", "Ap"))
```

orthology *Infer OrthoGroups Using Tree Reconciliation*

Description

Infer orthogroups using species and gene trees reconciliation

Usage

```
orthology(trees, invoke, d = 2, t = 10, l = 1, plot = TRUE, saverec = FALSE)
```

Arguments

trees	path to a single file containing first the species tree, followed by a single gene/protein tree (see details).
invoke	character string representing the way in which the executable of RANGER-DTL (see details) is invoked.
d	cost assigned to gene duplication.
t	cost assigned to gene transfer.
l	cost assigned to gene loss.
plot	when TRUE, the orthology network graph is plotted.
saverec	path to the directory where to save the reconciliation file. If not provided the file is not saved (default)

Details

The executable of RANGER-DTL (https://compbio.engr.uconn.edu/software/RANGER-DTL) should be installed. All input trees must be expressed using the Newick format terminated by a semicolon, and they must be fully binary (fully resolved) and rooted. Species names in the species tree must be unique. E.g, E.g., (((speciesA_gene1, speciesC_gene1), speciesB_geneX), speciesC_gene2); and (((speciesA, speciesC), speciesB), speciesC); are both valid gene tree inputs and, in fact, represent the same gene tree. This gene tree contains one copy of the gene from speciesA and speciesB, and two copies from speciesC.

Value

A list with four elements. The first one is a 'phylo' object where the nodelabels indicate the event: D, duplication or T transfer. If no label is shown is because the event correspond to speciation. The second element is a dataframe (the first column is the label of the internal nodes in the gene tree; the second column is the label of the internal nodes in the species tree, and the third and fourth columns label each internal node according to the inferred event). The third element of the list is an adjacency matrix: 1 when two proteins are orthologous, 0 if they are paralogous. The last element of the list is an orthogroup graph.

Examples

```
orthology(trees = system.file("extdata", "input.trees", package = "orthGS"))
```

---

orthP *Search Orthologous of a Given Protein*

Description

Searches orthologous of a given protein within a set of selected species

Usage

```
orthP(phylo_id, set = "all")
```

Arguments

phylo_id	phylo_id of the query protein
set	set of species of interest provided as a character vector, either with the binomial or short code of the species (see details).

Details

When set = "all", the search will be carry out against all the species in the database.

Value

A list with thee elements: 1. subtree of the relevant proteins; 2. vector color; 3. phylo_ids of the orthologous found.

Examples

```
orthP(phylo_id = "Pp_GS1a", set = c("Pp", "Psy", "Psm", "Ap"))
```

---

sdf *Seed Plants and Ferns GS*

Description

155 GS proteins from 25 seed plants species and 41 GS proteins from 11 fern species

Usage

```
sdf
```

Format

A dataframe with 196 rows (GS proteins) and 7 columns:

n Reference number

Sec.Name_ Unique identification label of the protein

species Species

taxon Acrogymnospermae, Angiospermae or Polypodiopsida

short Unique three letter identification of the species

gs Either GS2, GS1a, GS1b_Gym or GS1b_Ang. Here the ferns proteins have been forced to be either GS1a or GS2

tax_group Taxonomic group

Source

It has been curated manually by the authors

selected_tr *Ultrametric Rooted Seed Plants Tree*

Description

155 GS proteins from 45 seed plants species Rooted using MAD (Minimal Ancestor Deviation)

Usage

```
selected_tr
```

Format

An phylo object

Source

It has been manually curated by the authors

speciesGS *Map Species Names*

Description

Map binomial species name to short code species name and vice versa

Usage

```
speciesGS(sp)
```

Arguments

sp	set of species of interest (either binomial or short code name)

Details

The species set should be given as a character vector (see example)

Value

A dataframe containing the information for the requested species.

Examples

```
speciesGS(c("Pinus pinaster", "Ath"))
```

subsetGS *GS Proteins Report*

Description

Assembles a report regarding the GS proteins found in the indicated subset of species

Usage

```
subsetGS(sp)
```

Arguments

sp	set of species of interest (either binomial or short code name)

Details

This function returns the protein and DNA sequences of the different isoforms found in each species, along with other relevant data.

Value

A dataframe with the information for the requested species.

Examples

```
subsetGS(c("Pinus pinaster", "Ath"))
```

Index

∗ **datasets**
A_selected, 4
agf, 2
AngGym, 3
sdf, 12
selected_tr, 13

A_selected, 4
agf, 2
AngGym, 3

coltips, 4

gapless_msa, 5
getseqGS, 6

madRoot, 6
mapTrees, 7
mltree, 8
msa, 9

orthG, 10
orthology, 10
orthP, 11

sdf, 12
selected_tr, 13
speciesGS, 13
subsetGS, 14