Package 'phylotools'

December 10, 2017

Type Package

Title Phylogenetic Tools for Eco-Phylogenetics

Version 0.2.2

Date 2017-12-08

Author Jinlong Zhang [aut, cre],
Nancai Pei [ctb],
Xiangcheng Mi [ctb]

Maintainer Jinlong Zhang <jinlongzhang01@gmail.com>

Description A collection of tools for building RAxML supermatrix using PHYLIP or aligned FASTA files. These functions will be useful for building large phylogenies using multiple markers.

Depends ape

Suggests vegan

License GPL-2

LazyLoad yes

URL https://github.com/helixcn/phylotools

NeedsCompilation no

Repository CRAN

Date/Publication 2017-12-10 12:34:27 UTC

R topics documented:

phylotools-package 2
clean.fasta.name 2
dat2fasta 4
dat2phylip 5
get.fasta.name 6
get.phylip.name 7
read.fasta 8
read.phylip 9
rename.fasta 10

rm.sequence.fasta 11
split_dat 13
sub.taxa.label 14
supermat 15

Index **18**

`phylotools-package` *Phylogenetic tools for building PHYLIP supermatrix and more*

Description

A collection of a few functions for handling DNA-barcoding sequences, building PHYLIP supermatrix for RAxML etc.

Details

Package:	phylotools
Type:	Package
Version:	0.2.2
Date:	2017-12-09
License:	GLP-2
LazyLoad:	yes

Author(s)

Jinlong Zhang

Maintainer: Jinlong Zhang <jinlongzhang01@gmail.com>

`clean.fasta.name` *Clean the name of a fasta file*

Description

Cleaning the names of sequences for a fasta file. The punctuation characters and the white space will be replaced with "_".

Usage

```
clean.fasta.name(infile = NULL, outfile = "name_cleaned.fasta")
```

Arguments

infile	character string representing the name of the fasta file.
outfile	Character string representing the file name to be generated.

Details

Punctuation characters and white space will be replaced by "_". More information can be found at regex.

Value

This is a subroutine without a return value. A fasta file with all the names of sequences renamed will be saved to the working directory.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

read.fasta

Examples

```
cat(
  ">seq_1*66",  "--TTACAAATTGACTTATTATA",
  ">seq_2()r",  "GATTACAAATTGACTTATTATA",
  ">seq_3:test",  "GATTACAAATTGACTTATTATA",
  ">seq_588",  "GATTACAAATTGACTTATTATA",
  ">seq_8$$yu",  "GATTACAAATTGACTTATTATA",
  ">seq_10", "---TACAAATTGAATTATTATA",
  file = "matk.fasta", sep = "\n")

clean.fasta.name(infile = "matk.fasta")
get.fasta.name("name_cleaned.fasta")

# Delete file
unlink("matk.fasta")
unlink("name_cleaned.fasta")
```

dat2fasta *Convert and Save sequence data frame to fasta file*

Description

Convert and Save sequence data frame to fasta file.

Usage

```
dat2fasta(dat, outfile = "out.fasta")
```

Arguments

dat	data frame by read.phylip or read.fasta
outfile	a character string, representing the name of the fasta file to be generated

Details

The column of the data frame must be: 1. seq.name, 2. seq.text, represent the name of the sequences, the content of the sequence, eg. ATCGGGAAC.

Value

This is a routine without return value.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

read.fasta,read.phylip

Examples

```
cat(
">seq_2", "GTCTTATAAGAAAGAATAAGAAAG--AAATACAAA-------AAAAAAGA",
">seq_3", "GTCTTATAAGAAAGAAATAGAAAAGTAAAAAAAAA-------AAAAAAAG",
">seq_5", "GACATAAGACATAAAATAGAATACTCAATCAGAAACCAACCCATAAAAAC",
">seq_8", "ATTCCAAAATAAAATACAAAAAGAAAAAACTAGAAAGTTTTTTTTCTTTG",
">seq_9", "ATTCTTTGTTCTTTTTTTTCTTTAATCTTTAAATAAACCTTTTTTTTTTA",
file = "trn1.fasta", sep = "\n")

res <- read.fasta("trn1.fasta")
```

```
dat2fasta(res)
unlink("trn1.fasta")
unlink("out.fasta")
```

---

dat2phylip *Conver the data frame to sequential PHYLIP format file*

Description

Convert and save a data frame to sequential PHYLIP file.

Usage

```
dat2phylip(dat, outfile = "out.phy")
```

Arguments

dat	the data frame returned by read.phylip, read.fasta.
outfile	character string represents the phylip file to be generated.

Details

The output will be in sequential PHYLIP format.

Value

This is a subroutine, there is no return value.

Note

The names of the sequences should not contain white space or Punctuation characters. See regex for more details.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

dat2fasta, read.fasta, read.phylip

Examples

```
  cat(
  ">seq_2", "GTCTTATAAGAAAGAATAAGAAAG--AAATACAAA-------AAAAAAGA",
  ">seq_3", "GTCTTATAAGAAAGAAATAGAAAAGTAAAAAAAAA-------AAAAAAAG",
  ">seq_5", "GACATAAGACATAAAATAGAATACTCAATCAGAAACCAACCCATAAAAAC",
  ">seq_8", "ATTCCAAAATAAAATACAAAAAGAAAAAACTAGAAAGTTTTTTTTCTTTG",
  ">seq_9", "ATTCTTTGTTCTTTTTTTTCTTTAATCTTTAAATAAACCTTTTTTTTTTA",
  file = "trn1.fasta", sep = "\n")

res <- read.fasta("trn1.fasta")
dat2phylip(res)
unlink("trn1.fasta")
unlink("out.phy")
```

`get.fasta.name` *get the names of all the sequences of fasta file*

Description

get the names of all the sequences of a fasta file, and perform cleaning of the names of the sequences

Usage

```
get.fasta.name(infile, clean_name = FALSE)
```

Arguments

`infile`	character string representing the name of the fasta file.
`clean_name`	logical, representing cleaning of the names will be performed.

Value

a character vector containing the names of the sequences

Note

Punctuation characters and white space be replaced by "_". Definition of Punctuation characters can be found at `regex`.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

read.fasta, regex

Examples

```
  cat(
  ">seq_2", "GTCTTATAAGAAAGAATAAGAAAG--AAATACAAA-------AAAAAAGA",
  ">seq_3", "GTCTTATAAGAAAGAAATAGAAAAGTAAAAAAAAA-------AAAAAAAG",
  ">seq_5", "GACATAAGACATAAAATAGAATACTCAATCAGAAACCAACCCATAAAAAC",
  ">seq_8", "ATTCCAAAATAAAATACAAAAAGAAAAAACTAGAAAGTTTTTTTTCTTTG",
  ">seq_9", "ATTCTTTGTTCTTTTTTTTCTTTAATCTTTAAATAAACCTTTTTTTTTTA",
  file = "trn1.fasta", sep = "\n")
get.fasta.name("trn1.fasta")
unlink("trn1.fasta")
```

---

get.phylip.name *get the names of sequences from a PHYLIP file*

Description

get the names of sequences from a PHYLIP file.

Usage

```
get.phylip.name(infile, clean_name = FALSE)
```

Arguments

infile	character representing the name or path of the phylip file.
clean_name	logical, representing cleaning of the names will be performed.

Details

Punctuation characters and white space be replaced by "_". Definition of Punctuation characters can be found at regex.

Value

a character vector of the names of the sequences

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

See Also

read.phylip, regex

Examples

```
  cat("6 22",
  "seq_1    --TTACAAATTGACTTATTATA",
  "seq_2    GATTACAAATTGACTTATTATA",
  "seq_3    GATTACAAATTGACTTATTATA",
  "seq_5    GATTACAAATTGACTTATTATA",
  "seq_8    GATTACAAATTGACTTATTATA",
  "seq_10   ---TACAAATTGAATTATTATA",
  file = "matk.phy", sep = "\n")
get.phylip.name("matk.phy")
unlink("matk.phy")
```

`read.fasta` *Read FASTA file*

Description

Read and convert the fasta file to data frame

Usage

```
read.fasta(file = NULL, clean_name = FALSE)
```

Arguments

`file`	character string representing the name of the fasta file.
`clean_name`	logical, representing cleaning of the names will be performed. Punctuation characters and white space be replaced by "_" . See `regex` for more details.

Details

In this function, names of the sequences are identified by ">", and all the lines before next ">" will be concatenated.

Value

a data frame with two columns: (1) seq.name, the names for all the sequences. (2) seq.text, the raw sequence data.

Note

Punctuation characters and white space in the names of the sequences will be replaced by "_".

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

read.phylip,dat2fasta,dat2phylip,split_dat

Examples

```
cat(
">seq_2", "GTCTTATAAGAAAGAATAAGAAAG--AAATACAAA-------AAAAAAGA",
">seq_3", "GTCTTATAAGAAAGAAATAGAAAAGTAAAAAAAAA-------AAAAAAAG",
">seq_5", "GACATAAGACATAAAATAGAATACTCAATCAGAAACCAACCCATAAAAAC",
">seq_8", "ATTCCAAAATAAAATACAAAAAGAAAAAACTAGAAAGTTTTTTTTCTTTG",
">seq_9", "ATTCTTTGTTCTTTTTTTTCTTTAATCTTTAAATAAACCTTTTTTTTTTA",
file = "trn1.fasta", sep = "\n")

res <- read.fasta("trn1.fasta")
unlink("trn1.fasta")
```

---

read.phylip *read phylip file*

Description

read the phylip file, and store the sequences and their names in data frame.

Usage

```
read.phylip(infile, clean_name = TRUE)
```

Arguments

infile	character string for the name of the phylip file.
clean_name	logical, representing cleaning of the names will be performed.

Details

read.phylip accepts both interleaved and sequential phylip, the number of sequences is identified by parsing the first line of the file. Sequences and their names will be stored in a data frame.

If clean_name is TRUE, punctuation characters and white space be replaced by "_". Definition of punctuation characters can be found at regex.

Value

a data frame with two columns: (1) seq.name, the names for all the sequences; (2) seq.text, the raw sequence data.

Note

the Punctuation characters and white space in the names of the sequences will be replaced by "_".

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

See Also

read.fasta

Examples

```
  cat("6 22",
  "seq_1    --TTACAAATTGACTTATTATA",
  "seq_2    GATTACAAATTGACTTATTATA",
  "seq_3    GATTACAAATTGACTTATTATA",
  "seq_5    GATTACAAATTGACTTATTATA",
  "seq_8    GATTACAAATTGACTTATTATA",
  "seq_10   ---TACAAATTGAATTATTATA",
  file = "matk.phy", sep = "\n")

res <- read.phylip(infile = "matk.phy")
unlink("matk.phy")
```

---

rename.fasta *Rename the sequences for a fasta file*

Description

Rename the sequences within a fasta file according to a data frame supplied.

Usage

```
rename.fasta(infile = NULL, ref_table, outfile = "renamed.fasta")
```

Arguments

infile	character string containing the name of the fasta file.
ref_table	a data frame with first column for original name, second column for the new name of the sequence.
outfile	The name of the fasta file with sequences renamed.

Details

If the orginal name was not found in the ref_table, the name for the sequence will be changed into "old_name_" + orginal name.

Value

This is a subroutine without return value.

Note

Since whitespace and punctuation characters will be replaced with "_", name of a sequence might change. It is suggest to obtain the name of the sequences by calling read.fasta first, and save the data.frame to a csv file to obtain the "original" name for the sequences.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

read.fasta, split_dat

Examples

```
cat(
    ">seq_1",  "--TTACAAATTGACTTATTATA",
    ">seq_2",  "GATTACAAATTGACTTATTATA",
    ">seq_3",  "GATTACAAATTGACTTATTATA",
    ">seq_5",  "GATTACAAATTGACTTATTATA",
    ">seq_8",  "GATTACAAATTGACTTATTATA",
    ">seq_10", "---TACAAATTGAATTATTATA",
    file = "matk.fasta", sep = "\n")
old_name <- get.fasta.name("matk.fasta")
new_name <- c("Magnolia", "Ranunculus", "Carex", "Morus", "Ulmus", "Salix")
ref2 <- data.frame(old_name, new_name)
rename.fasta(infile = "matk.fasta", ref_table = ref2, outfile = "renamed.fasta")
unlink("matk.fasta")
unlink("renamed.fasta")
```

---

rm.sequence.fasta *Delete sequences from fasta file*

Description

Delete sequences from fasta file

Usage

```
rm.sequence.fasta(infile, outfile = "sequence.removed.fasta", to.rm = NULL)
```

Arguments

infile	Character string representing the name of the fasta file.
outfile	Character string representing the name of the output fasta file.
to.rm	Vector of character string containing the names of sequences to be deleted.

Details

Delete sequences from a fasta file.

Value

This is a subroutine without return value.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

read.fasta, dat2fasta

Examples

```
cat(
">seq_1",  "---TCCGCCCCCCTACTCTA",
">seq_3",  "CTCTCCGCCCCTCTACTCTA",
">seq_5",  "---TCCGCCC-TTTACTCTA",
">seq_6",  "---TCCGCCCCTCTACTCTA",
">seq_9",  "---TCCGCCC-TCTACTCTA",
">seq_12", "CTCTCCGCCC-TCTACTCTA",
file = "trn2.fasta", sep = "\n")

rm.sequence.fasta(infile = "trn2.fasta", to.rm = c("seq_1","seq_12"))

unlink("trn2.fasta")
unlink("sequence.removed.fasta")
```

split_dat *grouping the data frame containing sequences and names and generate fasta file*

Description

Splite the data frame of sequences based on the reference table of grouping.

Usage

```
split_dat(dat, ref_table)
```

Arguments

dat	data frame generated by read.phylip or read.fasta
ref_table	data frame with first column for the name of the sequence, second column for the group the sequence belongs to.

Details

Each group of sequences will be saved to a fasta file. Sequences not included in the ref_table will be saved in "Ungrouped.fasta"

Value

This is a subroutine, there is no return value.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

http://www.genomatix.de/online_help/help/sequence_formats.html

See Also

rename.fasta

Examples

```
cat(
">seq_1",   "--TTACAAATTGACTTATTATA",
">seq_2",   "GATTACAAATTGACTTATTATA",
">seq_3",   "GATTACAAATTGACTTATTATA",
">seq_5",   "GATTACAAATTGACTTATTATA",
">seq_8",   "GATTACAAATTGACTTATTATA",
">seq_10",  "---TACAAATTGAATTATTATA",
```

```
  ">seq_11",  "--TTACAAATTGACTTATTATA",
  ">seq_12",  "GATTACAAATTGACTTATTATA",
  ">seq_13",  "GATTACAAATTGACTTATTATA",
  ">seq_15",  "GATTACAAATTGACTTATTATA",
  ">seq_16",  "GATTACAAATTGACTTATTATA",
  ">seq_17",  "---TACAAATTGAATTATTATA",
  file = "trnh.fasta", sep = "\n")

sequence_name <- get.fasta.name("trnh.fasta")
sequence_group <- c("group1","group1","group1","group1","group1",
"group2","group2","group2","group3","group3","group3","group3")
group <- data.frame(sequence_name, sequence_group)

fasta <- read.fasta("trnh.fasta")
split_dat(fasta, group)

unlink("trnh.fasta")
unlink("ungrouped.fasta")
unlink("group1.fasta")
unlink("group2.fasta")
unlink("group3.fasta")
```

`sub.taxa.label` *Substitute the tip labels of a phylogenetic tree*

Description

Substitute the tip labels of a phylogenetic tree according to a reference data table.

Usage

```
sub.taxa.label(tree, dat)
```

Arguments

`tree`	Phylogenetic tree
`dat`	A dataframe with the first column the tip labels and the second column the new names.

Value

A Phylogenetic tree with the tip labels substituted

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

See Also

read.tree

Examples

```
library(ape)
data(bird.families)
tips <- bird.families$tip.label
abr <- paste("fam",1:length(tips), sep = "")
dat <- data.frame(tips, abr)
ntree <- sub.taxa.label(bird.families, dat)
```

---

supermat *Build PHYLIP supermatrix and RAxML partition file using aligned FASTA or PHYLIP files.*

Description

Build PHYLIP supermatrix and create RAxML partition file using aligned fasta or phylip files.

Usage

```
supermat(infiles, outfile = "supermat.out.phy",
         partition.file = "gene_partition.txt")
```

Arguments

infiles	a character string vector for phylip or aligned fasta file.
outfile	the name of the PHYLIP supermatrix
partition.file	partition data summary describing the genes.

Details

Supermatrix here means a phylip file with combined aligned sequences. The missing sequences should be replaced with either "?" or "-".

Value

A list containing: (1)supermat.dat:a list containing all the data frames read by read.phylip or read.fasta (2)res.super.dat: a data frame containing the sequences and the names (3)partition.dat: summary for all the fasta or phylip files (4)partition.dat.vector: character string vector for the partition file for RAxML

Note

Punctuation characters and white space in the names of the sequences will be replaced by "_". More information can be found at regex. Type of the sequence in the RAxML partition file should be changed manually according to the manual of RAxML.

Author(s)

Jinlong Zhang <jinlongzhang01@gmail.com>

References

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences, 106(44), 18621-18626.

de Queiroz, A.and Gatesy, J. (2007). The supermatrix approach to systematics. Trends in Ecology & Evolution, 22(1), 34-41.

https://github.com/stamatak/standard-RAxML

See Also

read.fasta,read.phylip,dat2phylip,

Examples

```
cat("6 22",
"seq_1    --TTACAAATTGACTTATTATA",
"seq_2    GATTACAAATTGACTTATTATA",
"seq_3    GATTACAAATTGACTTATTATA",
"seq_5    GATTACAAATTGACTTATTATA",
"seq_8    GATTACAAATTGACTTATTATA",
"seq_10   ---TACAAATTGAATTATTATA",
file = "matk.phy", sep = "\n")

cat("5 15",
"seq_1     GATTACAAATTGACT",
"seq_3     GATTACAAATTGACT",
"seq_4     GATTACAAATTGACT",
"seq_5     GATTACAAATTGACT",
"seq_8     GATTACAAATTGACT",
file = "rbcla.phy", sep = "\n")

cat("5 50",
"seq_2          GTCTTATAAGAAAGAATAAGAAAG--AAATACAAA-------AAAAAAGA",
"seq_3          GTCTTATAAGAAAGAAATAGAAAAGTAAAAAAAAA-------AAAAAAAG",
"seq_5          GACATAAGACATAAAATAGAATACTCAATCAGAAACCAACCCATAAAAAC",
"seq_8          ATTCCAAAATAAAATACAAAAAGAAAAAACTAGAAAGTTTTTTTTCTTTG",
"seq_9          ATTCTTTGTTCTTTTTTTTCTTTAATCTTTAAATAAACCTTTTTTTTTTA",
file = "trn1.phy", sep = "\n")
```

```
supermat(infiles = c("matk.phy", "rbcla.phy", "trn1.phy"))
unlink(c("matk.phy", "rbcla.phy", "trn1.phy"))
unlink(c("supermat.out.phy","gene_partition.txt"))

cat(
  ">seq_1",  "--TTACAAATTGACTTATTATA",
  ">seq_2",  "GATTACAAATTGACTTATTATA",
  ">seq_3",  "GATTACAAATTGACTTATTATA",
  ">seq_5",  "GATTACAAATTGACTTATTATA",
  ">seq_8",  "GATTACAAATTGACTTATTATA",
  ">seq_10", "---TACAAATTGAATTATTATA",
  file = "matk.fasta", sep = "\n")

cat(
  ">seq_1", "GATTACAAATTGACT",
  ">seq_3", "GATTACAAATTGACT",
  ">seq_4", "GATTACAAATTGACT",
  ">seq_5", "GATTACAAATTGACT",
  ">seq_8", "GATTACAAATTGACT",
  file = "rbcla.fasta", sep = "\n")

cat(
  ">seq_2", "GTCTTATAAGAAAGAATAAGAAAG--AAATACAAA-------AAAAAAGA",
  ">seq_3", "GTCTTATAAGAAAGAAATAGAAAAGTAAAAAAAAA-------AAAAAAAG",
  ">seq_5", "GACATAAGACATAAAATAGAATACTCAATCAGAAACCAACCCATAAAAAC",
  ">seq_8", "ATTCCAAAATAAAATACAAAAAGAAAAAACTAGAAAGTTTTTTTTCTTTG",
  ">seq_9", "ATTCTTTGTTCTTTTTTTTCTTTAATCTTTAAATAAACCTTTTTTTTTTA",
  file = "trn1.fasta", sep = "\n")

supermat(infiles = c("matk.fasta", "rbcla.fasta", "trn1.fasta"))
unlink(c("matk.fasta", "rbcla.fasta", "trn1.fasta"))

unlink(c("supermat.out.phy","gene_partition.txt"))
```

Index

∗Topic **RAxML**
    supermat, 15
∗Topic **fasta**
    clean.fasta.name, 2
    dat2fasta, 4
    get.fasta.name, 6
    read.fasta, 8
    rename.fasta, 10
    rm.sequence.fasta, 11
    split_dat, 13
    supermat, 15
∗Topic **package**
    phylotools-package, 2
∗Topic **partition**
    supermat, 15
∗Topic **phylip**
    dat2phylip, 5
    get.phylip.name, 7
    read.phylip, 9
    supermat, 15
∗Topic **supermatrix**
    supermat, 15

clean.fasta.name, 2

dat2fasta, 4, *5*, *9*, *12*
dat2phylip, 5, *9*, *16*

get.fasta.name, 6
get.phylip.name, 7

phylotools *(*phylotools-package*)*, 2
phylotools-package, 2

read.fasta, *3–5*, *7*, 8, *10–13*, *16*
read.phylip, *4, 5*, *7*, *9*, 9, *13*, *16*
read.tree, *15*
regex, *3*, *5–9*, *16*
rename.fasta, 10, *13*
rm.sequence.fasta, 11

split_dat, *9*, *11*, 13
sub.taxa.label, 14
supermat, 15