

Package ‘BoSSA’

February 14, 2012

Type Package

Title a Bunch of Structure and Sequence Analysis

Version 1.2

Date 2009-12-28

Author Pierre Lefevre

Maintainer Pierre Lefevre <bossa.package@gmail.com>

Imports ape, SoDA

Description Sort sequence from genbank. Retrieve sequence information from genbank (designed for viruses sequences and retrieve information such as isolation date and host). BLAST sequence and accession numbers. Detect group of sequences presenting phylogeography signal. Read PDB file (protein 3D structure file).

License GPL

LazyLoad yes

Repository CRAN

Date/Publication 2010-01-26 07:48:47

R topics documented:

blast	2
distGPS	3
FastaSorter	4
grpPhyloge	5
InfoGenBank	6
read.PDB	7
TaxoGB	8

Index	9
--------------	----------

blast *use NCBI blast*

Description

This function use the BLAST URL-API and query either DNA sequences store in object of class "DNABin" or accession numbers

Usage

```
blast(X, program = "blastn", database = "nr", entrezquery = "none",
      nb = 5, oot = 35)
```

Arguments

X	either an object of class "DNABin" or a vector of accession numbers
program	specify the BLAST program to use. either "blastn" (default), "blastp", "blastx", "tblastn", "tblastx"
database	specify the database to use. See the reference section
entrezquery	add an entrez query
nb	maximal number of hit to display
oot	out of time parameter. Default correpond to a maximum of 2 minutes search

Value

the output is a list where each compartment correspond to a blast search. If the query was a "DNABin" object, there is as many compartment as sequence in the object. If the query was an accession numbers vector, there is as many compartment as accession numbers in the vector. Each blast result is a data.frame with the accession number of the hit, the definition of the hit (i.e. the name), the score and the evaluate as colomns.

Author(s)

Pierre Lefevre

References

BLAST URL-API : <http://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html>

BLAST homepage : <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Entrez query : <http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml>

See Also

[as.DNABin](#)

Examples

```
#require internet connexion
#require R4X package, available at http://r-forge.r-project.org/projects/r4x/
#blast("aj865339")
#blast("aj865339",entrezquery="viridiplantae[organism]")
```

distGPS*Compute physical distance between point using GPS coordinates*

Description

The function read a data frame with the column "nom", "lat" and "lon" (respectively, the name of the sample, the longitude in degree.decimal and latitude in degree.decimal) and compute the distance between point in kilometers.

Usage

```
distGPS(input)
```

Arguments

input	A data frame with the column "nom", "lat" and "lon" (respectively, the name of the sample, the longitude in degree.decimal and latitude in degree.decimal)
-------	--

Details

Require the SoDA package.

Value

The output is a distance matrix with row and column names corresponding to the "nom" columns of the input data frame.

Author(s)

Pierre Lefevre

Examples

```
input <- data.frame(nom=c("a", "b", "c"), lat=c(-21, 18, 12), lon=c(32, 16, -5))
distGPS(input)
```

FastaSorter	<i>Rename and get information from fasta sequences set download from GenBank</i>
-------------	--

Description

The function read a fasta alignment download from Genbank, rename the sequence with the accession number and write a file with information about the sequences.

Usage

```
FastaSorter(Y)
```

Arguments

Y The path/name of a sequence set in fasta format download from genbank

Details

This function is handfull to go through datasets download from GenBank/TaxBrowser allowing to replace the composite GenBank sequence name with the accession number.

Value

The output is (i) a sequence set in fasta format with names of sequence being replace by the accession number and (ii) a comma separated file with the information about the sequence.

Author(s)

Pierre Lefevre

Examples

```
##obtain a fasta file the GenBank  
##copy the fasta file to your current directory  
##FastaSorter("sequence.fasta")
```

`grpPhyloge`*Determine sets of sequence presenting phylogeography signal*

Description

From an alignment of sequence (fasta format) and a csv file with columns `nom` (name of the sequence), `lon` and `lat` (GPS coordinates in degree.decimals of the longitude and latitude of sampling of the sequence), the function define sets of sequence presenting phylogeography signal using a mantel test between physic and genetic distance matrices.

Usage

```
grpPhyloge(gbinfo, align, seuil = 0.1, method = "single",  
model = "raw", pairwise.deletion=TRUE)
```

Arguments

<code>gbinfo</code>	name/path of a csv file with information on name (<code>nom</code> column) and GPS coordinates of the sequence (<code>lon</code> and <code>lat</code> columns)
<code>align</code>	sequence alignment in fasta format with the same name as those specified in the <code>gbinfo</code> file
<code>seuil</code>	significativity threshold for the mantel test
<code>method</code>	clustering method used; set to "single" by default; See the <code>hclust</code> function for details
<code>model</code>	DNA evolutionnary model use to compute genetic distance. Set to "raw" by default; See <code>dist.dna</code> (ape package) for details
<code>pairwise.deletion</code>	See <code>dist.dna</code> (ape package) for details

Details

The function draw a clustering plot (with physic distance represented) with a red frame around sequences with significative mantel test for correlation between physic and genetic distance.

Value

A list of group of sequences with significative correlation between physic and genetic distance is returned.

Author(s)

Pierre Lefevre

See Also

[hclust](#), [dist.dna](#), [mantel.test](#)

InfoGenBank

Download sequence information from GenBank

Description

This function is designed to work with virus accession number. It download from genbank a gbfile (INSDSeq XML) and parse it for different information, such as the year of isolation, host, sampling location...

Usage

```
InfoGenBank(X, tsleep=3)
```

Arguments

X	a vector of accession numbers
tsleep	the time between two query to genbank. set to 3 (the unit is the second) as asked by GenBank

Details

Require the R4X package (available at <http://r-forge.r-project.org/projects/r4x/>) and internet connexion.

Value

The output is a tab separated table ready to be write on the disk. It include 18 colomns with the accession number, the organism, the isolate name, the taxonomy, the submission date of the sequence, the sampling date, the host, the host taxonomy ID in GenBank, the host family, the host genus, the host subgenus, the host name proposition in case of possible misspel, the sampling location, the GPS coordinates of the sampling location, the authors, the title, the journal in which it was published and a pubmed URL to the publication. Note that if the information is missing, the cell is left empty.

Author(s)

Pierre Lefevre

See Also

[read.GenBank](#), [TaxoGB](#)

Examples

```
#require internet connexion
#require R4X package, available at http://r-forge.r-project.org/projects/r4x/
#accnb <- c("AJ86539", "AJ865337")
#InfoGenBank(accnb)
```

read.PDB	<i>read Protein Data Bank (PDB) file</i>
----------	--

Description

Read a PDB file (structure of a protein).

Usage

```
read.PDB(X)
```

Arguments

X	The path/name of a pdb file.
---	------------------------------

Value

The output is a list of objects

header	the header of the pdb file
compound	a data frame summarizing the CMPND part of the pdb file. This include the molecule ID, the molecule name and the chain ID
atom	a data frame with the atom type, the amino acid, the amino acid number, the chain and the euclidian X, Y, Z coordnates of the atom
sequence	a list with the numbering of the amino acid and the amino acid sequence for each chain

Author(s)

Pierre Lefevre

Examples

```
##obtain a pdb file the protein data bank
##copy the pdb file to your current directory
##pdb <- read.PDB("2B4C.pdb")
```

TaxoGB

Search for taxonomy information in genbank from a text query

Description

This function used the ESearch and ESpell tool from Eutils (<http://eutils.ncbi.nlm.nih.gov/>) to obtain taxonomy information from a text query. In case no hit is obtain, a misspell search is done.

Usage

```
TaxoGB(X,tsleep=3,organism="viridiplantae")
```

Arguments

X	a vector of accession numbers
tsleep	the time between two query to genbank. set to 3 (the unit is the second) as asked by GenBank
organism	a GenBank search query on organism. set to "viridiplantae" by default. See http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml for details

Details

Require the R4X package (available at <http://r-forge.r-project.org/projects/r4x/>) and internet connexion.

Value

The output is a character string with value tab separated. It includes the query, the taxonomy ID from GenBank, the Family, species and subspecies and a proposition of correction in case of possible misspell.

Author(s)

Pierre Lefevre

See Also

[read.GenBank](#)

Examples

```
#require internet connexion
#require R4X package, available at http://r-forge.r-project.org/projects/r4x/
#TaxoGB("tomato")
#TaxoGB("tomata")
```

Index

`as.DNAbin`, 2

`blast`, 2

`dist.dna`, 5

`distGPS`, 3

`FastaSorter`, 4

`grpPhylogeo`, 5

`hclust`, 5

`InfoGenBank`, 6

`mantel.test`, 5

`read.GenBank`, 6, 8

`read.PDB`, 7

`TaxoGB`, 6, 8