|  |  |
| --- | --- |
| SelBestSeq {regPhylo} | R Documentation |

## Select the sequences based on the geographic proximity with a focal area and sequence length, and export sequences (fasta format) and metadata for the selected gene regions

### Description

This function selects and exports the sequences for a list of selected gene regions for all the species present in the input data table. The selection of the sequence is optimised considering the most geographically proximate sequence to the RefPoint (for instance the New Zealand centroid in our case study) for a given species and gene region, as the first priority. If any geographic information is available or if multiple sequences are equally distants from the RefPoint, the function then also consider the length of the sequence. The selection based on geographic proximity can also be disabled.

### Usage

SelBestSeq(input = NULL, output = NULL, RefPoint = NULL,

perReposit = NULL, Alignment = NULL, MaxSeq = NULL,

gene.list = NULL, SeqChoice = NULL)

### Arguments

|  |  |
| --- | --- |
| input | a data table coming from the function Select.DNA() or SpeciesGeneMat.Bl.R (the data table should have 29 columns). |
| output | the name of the output table (including th path if necessary, but the folder must be created first). |
| RefPoint | a vector with geographic coordinates with longitude and latitude in decimal degrees WGS84 of the geographic reference point of the focal area. (e.g. the reference point for New Zealand, is 174.7976 degrees longitude, and -41.3355 degrees latitude). If RefPoint is NULL then the selection based on geographic proximity is disabled, and only the selection based on the length of the sequence is performed. |
| perReposit | if the name of the repository is provided (should be the same name as used in the function 'Congr.NCBI.BOLD.perReposit.R' option 'perReposit') it implies that several sequences are coming from a personal repository and the function will use the ID of the sequence provided by the field 'Db\_xref' as a unique id for the sequence (the function uses the first element (using ';' as separator) present in the column 'Db\_xref'), and otherwise leave it blank. |
| Alignment | if Alignment='T' then fasta files of the sequence of the selected gene regions are exported (requiring the seqinr R package). |
| MaxSeq | Maximum number of best sequences to be exported for a species when multiple sequences are available for a gene region, if 'ALL' all the sequences available for that gene region are exported. |
| gene.list | a vector of selected gene regions (gene names have to be consistent with the header of the table with the suffix '\_SpDNA\_Mat.txt' exported by the function SpeciesGeneMat\_Bl.R). For example Cytochrome c oxydase subunit 1, should have to be written 'co1' and not 'COI' or 'COX1'. |
| SeqChoice | select the longest sequence after removing ambiguous nucleotides (option SeqChoice='Longest') or the sequence with the length closest to the median sequence length (option SeqChoice='Median'). |

### Details

The choice based on the sequence length can be made using the longest sequence (option SeqChoice='Longest') or a sequence with a length closest to the median length of the sequence length distribution for a given species and a given gene region (option SeqChoice='Median') (See Antonelli et al. 2017, Systematic Biology DOI: 10.1093/sysbio/syw066 for a similar approach). The sequence length is computed after discarding the potential indels ('-') and the ambiguous nucleotides.

### Value

The function returns a table with all the information about the selected sequences including 3 additional columns ('SeqLengthNoIndelsNoAmbig' = Sequence length after removing the indels and ambiguous nucleotides, 'Dist2NZ' = Great circle distance in kilometers from the RefPoint, 'dist2med' = distance to the median of the sequence length distribution), and the alignment of the gene region in fasta format.

### References

Antonelli et al. 2017, DOI: 10.1093/sysbio/syw066

### Examples

# Load the data table exported by the function Select.DNA

data(Seq.DF4) # the table is called "Seq.DF5" and constitute the

# sixth object of the list "Seq.DF4".

Seq.DF5 = Seq.DF4$Seq.DF5

# Run the function and export all the sequences in the alignment for

# each species and gene region.

## Not run:

Seq.DF6=SelBestSeq(input = Seq.DF5, output = "Alig\_Seq.DF5.All",

RefPoint = cbind(174.7976, -41.3355), perReposit = "My.Rep",

Alignment = T, MaxSeq = "ALL", gene.list = c("co1", "16srrna"),

SeqChoice = "Median")

dim(Seq.DF6) # 10 sequences reported.

# Run the function and export the best (the most proximal to the focal

# area e.i. NZ) sequences in the alignment for each species and gene region,

# using the sequence the with a median sequence length for each gene region and species.

Seq.DF7=SelBestSeq(input = Seq.DF5, output = "Alig\_Seq.DF5.Best",

RefPoint = cbind(174.7976, -41.3355), perReposit = "My.Rep",

Alignment = T, MaxSeq =1, gene.list = c("co1", "16srrna"),

SeqChoice = "Median")

dim(Seq.DF7) # 2 sequences reported.

# Run the function and export the two most proximal sequences of the focal area (e.i. NZ)

# in the alignment for each species and gene region, using the longuest sequence per

# gene region and species (Notice that there is only one 16srrna sequence

# in the pool of sequences anyway).

Seq.DF8=SelBestSeq(input = Seq.DF5, output = "Alig\_Seq.DF5.Best",

RefPoint = cbind(174.7976, -41.3355), perReposit = "My.Rep",

Alignment = T, MaxSeq =2, gene.list = c("co1", "16srrna"),

SeqChoice = "Longest")

dim(Seq.DF8) # 3 sequences reported, 2 CO1 seqeunces and the only 16srrna

# sequence available in the pool.

# Run the function while the selection based on the geographic proximity is disabled,

# the two longuest sequences per gene region and species are retained

# (Notice that there is only one 16srrna sequence in the pool of sequences anyway).

Seq.DF9=SelBestSeq(input = Seq.DF5, output = "Alig\_Seq.DF5.Best",

perReposit = "My.Rep", Alignment = T, MaxSeq = 2,

gene.list = c("co1", "16srrna"), SeqChoice = "Longest")

dim(Seq.DF9) # 3 sequences reported, 2 CO1 sequences and the only 16srrna

# sequence available in the pool, but the geographic selection has been disabled.

## End(Not run)