Package 'MACER'

December 3, 2022

Type Package

Title Molecular Acquisition, Cleaning, and Evaluation in R 'MACER'

Version 0.2.1

Maintainer Robert G Young <rgyoung6@gmail.com>

Description To assist biological researchers in assembling taxonomically and marker focused molecular sequence data sets. 'MACER' accepts a list of genera as a user input and uses NCBI-GenBank and BOLD as resources to download and assemble molecular sequence datasets. These datasets are then assembled by marker, aligned, trimmed, and cleaned. The use of this package allows the publication of specific parameters to ensure reproducibility. The 'MACER' package has four core functions and an example run through using all of these functions can be found in the associated repository <https://github.com/rgyoung6/MACER_example>.

License GPL-2 | GPL-3

Encoding UTF-8

URL <https://github.com/rgyoung6/MACER>

Depends

Imports rentrez, ape, httr, stats, utils, ggplot2, parallel, pbapply, grDevices, png

Language en-GB

RoxygenNote 7.2.1

NeedsCompilation no

Author Robert G Young [aut, cre, cph] (<https://orcid.org/0000-0002-6731-2506>), Rekkab Gill [aut], Daniel Gillis [aut], Robert H Hanner [aut, cph]

Repository CRAN

Date/Publication 2022-12-02 23:30:02 UTC

R topics documented:

align_to_ref 2
auto_seq_download 4
barcode_clean 5
create_fastas 7

Index **10**

`align_to_ref` *Align and Trim MSA Against a Reference*

Description

Takes a FASTA file with target sequences and aligns them against a reference sequence submitted to the program. The output is an aligned fasta file that is trimmed to the length of the reference sequence. Sequences without full coverage (records having sequences with leading or trailing gaps) are removed. Records with characters other than IUPAC are also removed. Finally, internal gaps are removed from the sequence based on the submitted multiple sequence alignment percent coverage of the character position as provided in the pigl argument supplied by the user.

Usage

```
align_to_ref(
  data_folder = NULL,
  ref_seq_file = NULL,
  MAFFT_loc = NULL,
  output_file = NULL,
  pigl = 0.95,
  op = 1.53
)
```

Arguments

`data_folder`	This variable can be used to provide a location for the file containing all of the fasta files wanting to be aligned. The default value is set to NULL where the program will prompt the user to select the folder through point-and-click.
`ref_seq_file`	This variable can be used to provide a location for the reference sequence file. The default value is set to NULL where the program will prompt the user to select the folder through point-and-click.
`MAFFT_loc`	This variable can be used to provide a location for the MAFFT program. The default value is set to NULL where the program will prompt the user to select the folder through point-and-click.
`output_file`	This variable can be used to set the location of the output files from the program. The default value is set to NULL where the program will place the output files in the same location as the target files.

`pigl` This is the percent internal gap loop argument. This provides a percent that will remove records causing internal gaps if more than the percent value assigned to this argument is reached. If this value is set to 0 then internal gaps are not removed. The default for this value is 0.95.

`op` This is the gap opening penalty for the use of MAFFT. The higher the value the larger penalty in the alignment. The default for this value is set to 1.53 which is the default value in the MAFFT program. For alignment of highly conserved regions where no gaps are expected this should be set to a much higher number and 10 is recommended for coding regions like the COI-5P.

Details

User Input: 1. A file folder location with the fasta files that need to be aligned and trimmed using the supplied reference sequence. Please note that any and all fasta files (named *.fas) in this folder will be analyzed. 2. A reference sequence file with a sequence or MSA with all sequences having the same length. 3. The location of the MAFFT executable file <https://mafft.cbrc.jp/alignment/software/>

Value

Output: 1. In the submitted file folder location there will be a log file titled MAFFT_log. 2. The sequence output files from this script are placed into two subfolders. These folders are in the submitted file location where the fasta files of interest are located. The two folders created are MAFFT and MAFFT_trimmed. In the MAFFT folder there will be files with name of the files in the submitted file folder appended with "_MAFFT". The MAFFT_trimmed file will contain files with the same naming convention as the files in the submitted folder and appended with "_MAFFT_trimmed".

Author(s)

Robert G. Young

References

<https://github.com/rgyoung6/MACER> Young RG, Gill R, Gillis D, Hanner RH (2021) Molecular Acquisition, Cleaning and Evaluation in R (MACER) - A tool to assemble molecular marker datasets from BOLD and GenBank. Biodiversity Data Journal 9: e71378. <https://doi.org/10.3897/BDJ.9.e71378>

See Also

auto_seq_download() create_fastas() barcode_clean()

Examples

```
## Not run:
align_to_ref(pigl=0.75)
align_to_ref(pigl=0.95, op=10)
align_to_ref(pigl=0)

## End(Not run)
```

auto_seq_download *Automatic Sequence Download*

Description

Takes a list of genera, as supplied by the user, and searches and downloads molecular sequence data from BOLD and Genbank.

Usage

```
auto_seq_download(
  BOLD_database = TRUE,
  NCBI_database = TRUE,
  search_str = NULL,
  input_file = NULL,
  output_file = NULL,
  seq_min = 100,
  seq_max = 2500
)
```

Arguments

BOLD_database	TRUE is to include, FALSE is to exclude; default TRUE
NCBI_database	TRUE is to include, FALSE is to exclude; default TRUE
search_str	NULL uses the default string, anything other than NULL then that string will be used for the GenBank search; default NULL. The Default String is: (genus[ORGN]) NOT (shotgun[ALL] OR genome[ALL] OR assembled[ALL] OR microsatellite[ALL])
input_file	NULL prompts the user to indicate the location of the input file through point and click prompts, anything other than NULL then the string supplied will be used for the location; default NULL
output_file	NULL prompts the user to indicate the location of the output file through point and click prompts, anything other than NULL then the string supplied will be used for the location; default NULL
seq_min	holds the minimum length value to not flag the sequence; default 100
seq_max	holds the maximum length value to not flag the sequence; default 2500

Details

User Input: A list of genera in a text file in a single column with a new line at the end of the list.

Value

Outputs: One main folder containing three other folders. Main folder - Seq_auto_dl_TTTTTT_MMM_DD Three subfolders: 1. BOLD - Contains a file for every genus downloaded with the raw data from the BOLD system. 2. NCBI - Contains a file for every genus downloaded with the raw data from GenBank. 3. Total_tables - Contains files for the running of the function which include... A_Summary.txt - This file contains information about the downloads. A_Total_Table.tsv - A file with a single table containing the accumulated data for all genera searched.

Note

When using a custom search string for NCBI only a single genus at a time can be used.

Author(s)

Robert G. Young

References

<https://github.com/rgyoung6/MACER> Young, R. G., Gill, R., Gillis, D., Hanner, R. H. (Submitted June 2021). Molecular Acquisition, Cleaning, and Evaluation in R (MACER) - A tool to assemble molecular marker datasets from BOLD and GenBank. Biodiversity Data Journal.

See Also

create_fastas() align_to_ref() barcode_clean()

Examples

```
## Not run:
auto_seq_download()
auto_seq_download(BOLD_database = TRUE, NCBI_database = FALSE)
auto_seq_download(BOLD_database = FALSE, NCBI_database = TRUE)

## End(Not run)
```

---

barcode_clean *DNA Barcode Clean*

Description

Takes an input fasta file and identifies genus level outliers and species outliers based on the 1.5 x greater than the interquartile range. It also, if selected, checks the sequence using amino acid translation and has the option to eliminate sequences that have non-IUPAC codes. Finally, the program calculates the barcode gap for the species in the submitted dataset.

Usage

```
barcode_clean(
  AA_code = "invert",
  AGCT_only = TRUE,
  data_folder = NULL,
  outliers = TRUE,
  dist_model = "raw",
  replicates = 1000,
  replacement = TRUE,
  conf_level = 1,
  numCores = 1
)
```

Arguments

AA_code	This is the amino acid translation matrix (as implemented through ape) used to check the sequences for stop codons. The following codes are available std, vert, invert, F. The default is invert.
AGCT_only	This indicates if records with characters other than AGCT are kept, the default is TRUE. TRUE removes records with non-AGCT FALSE is accepting all IUPAC characters
data_folder	This variable can be used to provide a location for the MSA fasta files to be cleaned. The default value is set to NULL where the program will prompt the user to select the folder through point-and-click.
outliers	This is the variable to indicate if the use would like to remove suspected sequence record outliers using 1.5X the genetic distance. If set to TRUE genus and species level outliers will be removed. If FALSE this will not occur. Default TRUE.
dist_model	This is the model of nucleotide evolution that the ape program will use (see ape documentation for options. Default is "raw"
replicates	This is the number of replicates that the bootstrapping will perform. Note: more replicates will take longer. Default is 1000
replacement	This indicates that the replacement of MSA nucleotide columns will be replaced in the random resampling. Default is set to TRUE
conf_level	This is a percentage of the initial MSA nucleotide length. When set to 1 the bootstrapped resampling will have the same length as the initial MSA. Default is set to 1
numCores	This is the number of cores that the user would like to use where multithreading is available. Default is set to 1, indicating only a single thread will be used.

Details

Input: A file folder with one or more fasta files of interest

Value

Output: A single log file for the running of the function with the name A_Clean_File_YYYY-DD-TTTTTTTT. The function will also output three files for each fasta file submitted. The first is the distance matrix that was calculated and used to assess the DNA barcode gaps. This file is named the same as the input file with dist_table.dat appended to the end of the name. The second file is the total data table file which provides a table of all submitted records for each data set accompanied with the results from each section of the analysis. This file is named the same as the input fasta with data_table.dat appended to the end, Finally, a fasta file with all outliers and flagged records removed is generated for each input fasta file. This output file is named the same as the input fasta with no_outlier.fas appended to the end.

Author(s)

Robert G. Young

References

<https://github.com/rgyoung6/MACER> Young RG, Gill R, Gillis D, Hanner RH (2021) Molecular Acquisition, Cleaning and Evaluation in R (MACER) - A tool to assemble molecular marker datasets from BOLD and GenBank. Biodiversity Data Journal 9: e71378. <https://doi.org/10.3897/BDJ.9.e71378>

See Also

auto_seq_download() create_fastas() align_to_ref()

Examples

```
## Not run:
barcode_clean(),
barcode_clean(AA_code = "vert", AGCT_only = TRUE),
barcode_clean(AA_code = "vert")

## End(Not run)
```

`create_fastas` *Table To FASTA*

Description

Using the output table from the download script and the user built genus-marker name parameter file to take the downloaded data and place them into fasta files.

Usage

```
create_fastas(
  data_file = NULL,
  input_file = NULL,
  output_folder = NULL,
  no_marker = FALSE,
  no_taxa = FALSE,
  no_seq = FALSE,
  name_issue = FALSE,
  taxa_digits = FALSE,
  taxa_punct = FALSE
)
```

Arguments

data_file	NULL prompts the user to indicate the location of the data file in the format of the auto_seq_download output, anything other than NULL then the string supplied will be used for the location; default NULL
input_file	NULL prompts the user to indicate the location of the input file used to select through point and click prompts, anything other than NULL then the string supplied will be used for the location; default NULL
output_folder	NULL prompts the user to indicate the location of the output file through point and click prompts, anything other than NULL then the string supplied will be used for the location; default NULL
no_marker	If set to TRUE then will include records filtered out due to no marker data. Default is FALSE to not include records with no marker data.
no_taxa	If set to TRUE then will include records filtered out due to no taxa data. Default is FALSE to not include records with no taxa data.
no_seq	If set to TRUE then will include records filtered out due to no sequence data. Default is FALSE to not include records with no sequence data.
name_issue	If set to TRUE then will include records filtered out due to genus and species names with more than two terms. Default is FALSE to not include records with taxonomic naming issues.
taxa_digits	If set to TRUE then will include records filtered out due to genus or species names containing digits. Default is FALSE to not include records with digits in the taxonomic naming.
taxa_punct	If set to TRUE then will include records filtered out due to the presence of punctuation in the genus or species names. Default is FALSE to not include records with punctuation in the taxonomic naming.

Details

Input: File with list of genera with the molecular markers names below the taxa. The information to create this parameters file can be obtained from A_Summary.txt file from the download script results. For further details please see the documentation.

Value

This script outputs a fasta file of sequences for each column in the submitted parameters file. These files are named with the genera of interest and the first marker name in the column of the parameters file. These files are located in the folder where the Total_tables.txt file is located.

Author(s)

Rekkab Singh Gill

References

<https://github.com/rgyoung6/MACER> Young RG, Gill R, Gillis D, Hanner RH (2021) Molecular Acquisition, Cleaning and Evaluation in R (MACER) - A tool to assemble molecular marker datasets from BOLD and GenBank. Biodiversity Data Journal 9: e71378. <https://doi.org/10.3897/BDJ.9.e71378>

See Also

create_fastas() align_to_ref() barcode_clean()

Examples

```
## Not run:
create_fastas()
create_fastas(no_marker = TRUE, no_taxa = TRUE)
create_fastas(no_seq  = TRUE, name_issue = TRUE)

## End(Not run)
```

Index

align_to_ref, 2
auto_seq_download, 4

barcode_clean, 5

create_fastas, 7