Package 'adhoc'

March 17, 2017

Version 1.1

Type Package

Title Calculate Ad Hoc Distance Thresholds for DNA Barcoding Identification

Date 2017-03-17

Author Gontran Sonet

Maintainer Gontran Sonet <gosonet@gmail.com>

Description Two functions to calculate intra- and interspecific pairwise distances, evaluate DNA barcoding identification error and calculate an ad hoc distance threshold for each particular reference library of DNA barcodes. Specimen identification at this ad hoc distance threshold (using the best close match method) will produce identifications with an estimated relative error probability that can be fixed by the user (e.g. 5%).

URL http://jemu.myspecies.info/computer-programs

Depends R (>= 2.15), ape, pegas, polynom

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2017-03-17 18:56:42 UTC

R topics documented:

adhoc-package *Calculation of ad hoc distance thresholds for DNA barcoding identification.*

Description

The ad hoc distance threshold (see Virgilio et al. 2012) can be calculated for each particular reference library of DNA barcodes. Using this distance threshold in the "best close match method" (a method where a species name assignment is rejected when the distance between the query and its best match in a reference dataset is above the threshold) will provide identifications with an estimated relative error probability that is fixed by the user (e.g. 5%).

Details

Package:	adhoc
Type:	Package
Version:	1.0
Date:	2013-11-08
License:	GPL-2 | GPL-3

The procedure consists of two consecutive steps (Sonet et al. 2013). A first function (checkDNAbcd) evaluates a reference library in order to allow some quality control (sequence labelling, sequence lengths, etc.). The output of the first function is used by a second function (adhocTHR), which calculates an ad hoc distance threshold. For this, the second function takes each sequence of the reference library as a query and finds its best match(es) among all other DNA barcodes in the library. It then quantifies the relative identification errors (see Virgilio et al. 2012) obtained for a set of different arbitrary distance thresholds (30 values by default), performs a linear regression (or optionally a polynomial regression) and calculates the ad hoc distance threshold corresponding to an expected identification error probability (5% by default).

Author(s)

Gontran Sonet Maintainer: Gontran Sonet <gosonet@gmail.com>

References

Sonet G, Jordaens K, Nagy ZT, Breman FC, De Meyer M, Backeljau T & Virgilio M", "(2013) Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification, Zookeys, 365:329-336. http://zookeys.pensoft.net/articles.php?id=3057.

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M. (2012) Identifying insects with incomplete DNA barcode libraries, African Fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7(2) e31581. doi: 10.1371/journal.pone.0031581.

Examples

```
data(tephdata);
```

```
out1<-checkDNAbcd(tephdata);
out2<-adhocTHR(out1);
layout(matrix(1,1,1));
par(font.sub=8);
plot(RE~thres,out2$IDcheck,ylim=c(0,max(c(out2$IDcheck$RE,out2$ErrProb))),xlab=NA,ylab=NA);
title(main="Ad hoc threshold",xlab="Distance", ylab="Relative identification error (RE)")
title(sub=paste("For a RE of",round(out2$ErrProb,4), "use a threshold of", round(out2$THR,4)));
regcoef<-out2$reg$coefficient;
curve(regcoef[1] + regcoef[2]*x + regcoef[3]*x^2 + regcoef[4]*x^3, add=TRUE);
segments(-1,out2$ErrProb,out2$THR,out2$ErrProb);segments(out2$THR,-1,out2$THR,out2$ErrProb);
```

adhocTHR *Calculation of an ad hoc distance threshold for DNA barcoding identification*

Description

This function utilises the output of checkDNAbcd to quantify the relative identification errors obtained for a particular reference library at set of arbitrary distance thresholds. It then performs linear (or polynomial) regression and calculates the ad hoc distance threshold corresponding to an expected relative identification error. The method is described by Virgilio et al. (2012).

Usage

```
adhocTHR(a, NbrTh = 30, ErrProb = 0.05, Ambig = "incorrect", Reg = "linear")
```

Arguments

a	the output of the function checkDNAbcd.
NbrTh	the number of arbitrary distance thresholds used for the fitting (30 by default).
ErrProb	a value between 0 and 1 setting the relative identification error probability (5% by default).
Ambig	"incorrect", "correct" or "ignore" to treat ambiguous identifications as incorrect, correct, or to ignore them (default is "incorrect").
Reg	"linear" or " polynomial" to perform a linear or a polynomial fitting (default is linear).

Details

NbrTh: By default, the function samples 30 distance thresholds equidistantly distributed between 0.0 and the largest distance between all pairs of query - best match observed in the reference library. Using more arbitrary distance thresholds may improve the fitting but will also require more computing time.

ErrProb: By default, the function sets an estimated error probability of 5%. In some cases, the selected relative identification error (e.g. 5%) cannot be reached, even when using the most restrictive distance threshold (viz. when setting distance threshold = 0.00). See below.

Ambig: We recommend to use this option with caution. If "correct", it will treat all red-flagged species involved in the same ambiguous identification as a single species. Setting this option to "ignore" will allow the user to evaluate the consequences of the ambiguous identifications on the relative identification errors and hence, on the estimated ad hoc distance threshold.

Reg: The user has the possibility to estimate the ad hoc distance threshold on the basis of polynomial regression (rather than linear). For this, check that the package polynom (Venables et al. 2013) is installed.

Value

`"adhocTHR"` returns a list of 7 components:

`BM` a data.frame describing the best matches obtained for each query: distance (distBM), identification (idBM), evaluation of the identification (IDeval) according to Sonet et al. (2013).

`IDcheck` a data.frame describing the identification success at each arbitrary threshold: numbers of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), the relative identification error (RE), the overall identification error (OE), the accuracy and the precision (Sonet et al. 2013).

`reg` an object of class `"lm"` describing the regression performed by `"adhocTHR"`.

`ErrProb` the value of the relative identification error probability fixed as an argument by the user (default is 5%).

`THR` the value of the ad hoc distance threshold producing the desired probability (ErrProb) of relative identification error.

`redflagged` lists all ambiguous matches obtained for each ambiguous identification: the number of species involved in the ambiguous identification (Nb_species), the number of reference sequences with the same species name as the query (Nb_conspecific_seq), the number of reference sequences with a different species name than the query (Nb_heteroallospecific_seq) and the labels of all best matches involved in the ambiguous identification (List_of_best-matches).

`redflaggedSP` lists all species names that are involved in the same ambiguous identification.

Note

In particular cases, (e.g. reference libraries with low taxon coverage) all best matches might result in correct identifications, with RE = 0.0 at all distance thresholds and the regression line being parallel to the x axis. In this situation, adhocTHR will give the following message: "All identifications are correct when using the best match method (no distance threshold considered). An ad hoc threshold for best close match identification cannot be calculated".

In other cases, reaching an estimated RE of 5% might not be possible, even at the most restrictive distance threshold (distance threshold = 0.00) and regression fitting will intercept the y axis above the RE value. In this case, adhocTHR will give the following message: "The estimated RE cannot be reached using this reference library". The user should then either increase the relative error probability (RE) or try and increase the taxon coverage of the library (see Virgilio et al. 2012).

Author(s)

Gontran Sonet

References

Sonet G, Jordaens K, Nagy ZT, Breman FC, De Meyer M, Backeljau T & Virgilio M", "(2013) Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification, Zookeys, 365:329-336. http://zookeys.pensoft.net/articles.php?id=3057.

Venables B, Hornik K, Maechler M (2013) polynom: a collection of functions to implement a class for univariate polynomial manipulations. R package version 1.3-7. http://cran.r-project.org/web/packages/polynom/index.htm

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M. (2012) Identifying insects with incomplete DNA barcode libraries, African Fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7(2) e31581. doi: 10.1371/journal.pone.0031581.

See Also

`checkDNAbcd`

Examples

```
data(tephdata);
out1<-checkDNAbcd(tephdata);
out2<-adhocTHR(out1);
layout(matrix(1,1,1));
par(font.sub=8);
plot(RE~thres,out2$IDcheck,ylim=c(0,max(c(out2$IDcheck$RE,out2$ErrProb))),xlab=NA,ylab=NA);
title(main="Ad hoc threshold",xlab="Distance", ylab="Relative identification error (RE)");
title(sub=paste("For a RE of",round(out2$ErrProb,4), "use a threshold of", round(out2$THR,4)));
regcoef<-out2$reg$coefficient;
curve(regcoef[1] + regcoef[2]*x + regcoef[3]*x^2 + regcoef[4]*x^3, add=TRUE);
segments(-1,out2$ErrProb,out2$THR,out2$ErrProb);segments(out2$THR,-1,out2$THR,out2$ErrProb);
```

`checkDNAbcd` *Evaluation of a reference library of aligned DNA barcodes*

Description

This function provides an overview of the content of a reference library of aligned DNA barcodes (Sonet et al. 2013). It calculates all pairwise distances and delivers an output that can be used by the function adhocTHR.

Usage

```
checkDNAbcd(seq, DistModel = "K80")
```

Arguments

`seq` an object of class `"DNAbin"`.

`DistModel` "K80" (for Kimura two-parameter) or "raw" (for p-distances) or any other nucleotide substitution model available in the function `"dist.dna"` (Paradis et al. 2004).

Details

Sequence labels of "seq" should have the following structure: ">species_name_any_additional_information" as in the following example (note that character strings have to be separated by underscores): ">Bactrocera_amplexa_Kenya_voucher1052_JEMU".

Value

checkDNAbcd returns a list of 6 components:

`mylabels` a data.frame providing both parts of the species names and the complete label of each sequence (as extracted from the first argument).

`listsp` a data.frame listing the number of sequences (Nseq) and haplotypes (Nhap) for each species of the reference library.

`DNAlength` a numeric vector of the sequence lengths of each DNA sequence.

`dist` a matrix of all distances obtained by pairwise comparison.

`spdist` a list of all pairwise interspecific distances (inter) and all pairwise intraspecific distances (intra).

`seq` an object of class "DNAbin" with all sequences in the reference library (= first argument).

Author(s)

Gontran Sonet

References

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

Sonet G, Jordaens K, Nagy ZT, Breman FC, De Meyer M, Backeljau T & Virgilio M", "(2013) Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification, Zookeys, 365:329-336. http://zookeys.pensoft.net/articles.php?id=3057.

See Also

adhocTHR

Examples

```
data(tephdata);
out1<-checkDNAbcd(tephdata);

#Plot distribution of sequence lengths
hist(out1$DNAlength,main="Seq. lengths",xlab="Seq. length (bp)");

#Plot distribution of pairwise interspecific distances
hist(out1$spdist$inter,main="Intersp. dist",xlab="Distance",col="#0000ff99");

#Plot distribution of pairwise intraspecific distances
```

```
hist(out1$spdist$intra, main="Intrasp. dist.",xlab="Distance",col="#0000ff22");

#Plot distribution of both pairwise intra- and interspecific distances
hist(out1$spdist$inter,main="Intra- & intersp. dist",xlab="Distance",col="#0000ff99");
hist(out1$spdist$intra, add=TRUE,col="#0000ff22");

#Idem as previous example with zoom on intraspecific values
hist(out1$spdist$intra,main="Zoom intra- & intersp. dist",xlab="Distance",col="#0000ff99");
hist(out1$spdist$inter, add=TRUE,col="#0000ff22");
```

tephdata *DNA sequences of Tephritids to use as an example dataset for package adhoc.*

Description

DNA sequences (326 sequences of 656 base pairs) of the cytochrome c oxidase subunit I (COI) gene of Tephritids (Diptera) (see Virgilio et al. 2012) to use as an example dataset for package adhoc (Sonet et al. 2013).

Usage

```
data(tephdata)
```

Format

326 DNA sequences in binary format stored in a matrix.

References

Sonet G, Jordaens K, Nagy ZT, Breman FC, De Meyer M, Backeljau T & Virgilio M", "(2013) Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification, Zookeys, 365:329-336. http://zookeys.pensoft.net/articles.php?id=3057.

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M. (2012) Identifying insects with incomplete DNA barcode libraries, African Fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7(2) e31581. doi: 10.1371/journal.pone.0031581.

Examples

```
data(tephdata);
labels(tephdata);
```

Index