Package 'NicheBarcoding'

July 26, 2019

Title Species Identification using DNA Barcodes and Niche Models

Version 0.0.0.9000

Description To perform species identification using DNA barcodes and Niche Models.

Depends R (>= 3.3.0),BarcodingR,ape,class,dismo,maps,nnet,picante,randomForest,raster,sp,spider

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

Suggests testthat

NeedsCompilation no

Author Cai-qing YANG [aut, cre],
Xin-hai Li [aut],
Michael christopher ORR [aut],
Cong JIANG [aut],
Ai-bing ZHANG [aut]

Maintainer Cai-qing YANG <yangcq_ivy@163.com>

R topics documented:

BSI *Species Identification using Protein-coding Barcodes*

Description

Species identification using protein-coding barcodes with different methods,including a BP-based method (Zhang et al. 2008), a fuzzy-set based method (Zhang et al. 2012), and a Bayesian method (Jin et al. 2013).

Usage

```
BSI(ref, que, method = "bpNewTraining")
```

Arguments

ref	Object of class "DNAbin" used as a reference dataset, contains taxon information.
que	Object of class "DNAbin", whose identities (species names) need to be inferred.
method	A character string indicating which method will be used to train model and/or infer species membership. One of these methods ("fuzzyId", "bpNewTraining", "bpNewTrainingOnly", "bpUseTrained","Bayesian") should be specified.

Value

A list containing model parameters used, species identification success rates using references, query sequences, species inferred, and corresponding confidence levels (bp probability for BP-based method / FMF values for fuzzy set theory based method / posterior probability for Bayesian method) when available.

Note

Functions fasta2DNAbin() from package:adegenet and read.dna() from package:ape were used to obtain DNAbin objects in our package. The former is used to read large, aligned coding DNA barcodes, the latter unaligned ones. ref and que should be aligned with identical sequence length. We provided a pipeline to perform fast sequences alignment for reference and query sequences. Windows users could contact zhangab2008(at)mail.cnu.edu.cn for an exec version of the package. For very large DNA datasets, read.fas() package:phyloch is strongly suggested instead of fasta2DNAbin() since the latter is very slow.

Author(s)

Ai-bing ZHANG, PhD. CNU, Beijing, CHINA. zhangab2008(at)cnu.edu.cn

References

Zhang, A.B, Hao, M.D., Yang,C.Q., Shi, Z.Y. (2016). BarcodingR: an integrated R package for species identification using DNA barcodes. Methods in Ecology and Evolution. In press. Jin,Q., H.L. Han, X.M. Hu, X.H. Li,C.D. Zhu,S. Y. W. Ho, R. D. Ward, A.B. Zhang . (2013). Quantifying Species Diversity with a DNA Barcoding-Based Method: Tibetan Moth Species (Noctuidae) on the Qinghai-Tibetan Plateau. PloS One 8: e644. Zhang, A. B., C. Muster, H.B. Liang, C.D. Zhu, R. Crozier, P. Wan, J. Feng, R. D. Ward.(2012). A fuzzy-set-theory-based approach to analyse

species membership in DNA barcoding. Molecular Ecology, 21(8):1848-63. Zhang, A. B., D. S. Sikes, C. Muster, S. Q. Li. (2008). Inferring Species Membership using DNA sequences with Back-propagation Neural Networks. Systematic Biology, 57(2):202-215.

Examples

```
data(TibetanMoth)
ref<-as.DNAbin(as.character(TibetanMoth[1:5,]))
que<-as.DNAbin(as.character(TibetanMoth[50:55,]))
bsi<-barcoding.spe.identify(ref, que, method = "fuzzyId")
bsi
bsi<-barcoding.spe.identify(ref, que, method = "bpNewTraining")
bsi
bsi<-barcoding.spe.identify(ref, que, method = "Bayesian")
bsi
```

---

extractSpeInfo *Extraction of species taxonomic and distributional information*

Description

Splits comma-separated sample information into different columns of a dataframe.

Usage

```
extractSpeInfo(seqID.full)
```

Arguments

seqID.full	The sample ID, taxon information and longitude and latitude data, splited by commas in class character.

Value

A data frame of split sample ID, taxon information and longitude and latitude data for further analysis.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

Examples

```
data(Ref)
seqID.full<-rownames(Ref)

infor<-extractSpeInfo(seqID.full)
head(infor)
```

monophyly.prop *Analyses of phylogenetic monophyletic proportion*

Description

Calculate the proportion of monophyletic groups on a tree.

Usage

```
monophyly.prop(phy, sppVector, singletonsMono = TRUE)
```

Arguments

phy	A tree of class phylo.
sppVector	Species vector.
singletonsMono	Logical. Should singletons (i.e. only a single specimen representing that species) be treated as monophyletic? Default of TRUE. Additional possible values of FALSE and NA.

Value

A list containing proportion and number of monophyly group. a set monophyly and of non-monophyly group names.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

Examples

```
tree<-rtree(20)
tree$tip.label<-sample(tree$tip.label[1:10],size=20,replace = TRUE)
plot(tree)
sppVector<-tree$tip.label
MP<-monophyly.prop(tree,sppVector,singletonsMono = TRUE)
MP
```

niche.conserv *Niche conservatism of a multispecies dataset.*

Description

Calculate the K statistic of phylogenetic signals of ecological niche characters.

Usage

```
niche.conserv(spe.env, tree)
```

Arguments

spe.env	A data frame of ecological variables of species.
tree	The rooted phylogenetic tree of those species in class phylo.

Value

The K statistic of phylogenetic signal on top three principle components. The niche conservatism of the multispecies dataset.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)cnu.edu.cn.

References

Blomberg, S. P., and T. Garland, Jr. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. Journal of Evolutionary Biology 15:899-910.

Blomberg, S. P., T. Garland, Jr., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717-745.

Examples

```
data(Simulation)
spe.env<-Simulation$spe.env
tree<-Simulation$tree;plot(tree)

NC<-niche.conserv(spe.env,tree)
NC
```

---

niche.Model.Build *Ecological niche modele building using randomForest classifier.*

Description

Builds a niche model for a given species according to its distributional data.

Usage

```
niche.Model.Build(prese = NULL, absen = NULL, prese.env = NULL,
  absen.env = NULL, bak.vir = NULL, en.vir)
```

Arguments

prese	The longitude and latitude of the presence data of a species in class data.frame. (can be absent when providing prese.env parameter).
absen	The longitude and latitude of the absence data of a species in class data.frame. (can be absent when providing absen.env or back parameter).
prese.env	The bioclimatic variables of presence data in class data.frame (can be absent when providing prese parameter).

absen.env	The bioclimatic variables of absence data in class data.frame (can be absent when providing absen or back parameter).
bak.vir	Bioclimatic variables of random background points in class matrix (can be absent when providing absen or absen.env parameter).
en.vir	The globle bioclimate data from "raster::getData" function in class RasterBrick.

Value

A trained niche model in class randomForest.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)cnu.edu.cn.

References

Breiman, L. 2001. Random forests. Machine Learning 45(1): 5-32.

Liaw, A. and M. Wiener. 2002. Clasification and regression by randomForest. R News 2/3: 18-22.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15): 1965-1978.

Examples

```
## Not run:
data<-data.frame(species=rep("Acosmeryx anceus",3),
Lon=c(145.380,145.270,135.461),
Lat=c(-16.4800,-5.2500,-16.0810))
present.points<-pseudo.present.points(data,10,10,1)

# Note: set "download=FALSE", if raster::getData() has been run once!
envir<-raster::getData("worldclim",download=TRUE,var="bio",
res=2.5,lon=lon,lat=lat)
en.vir<-brick(envir)
back<-randomPoints(mask=en.vir,n=5000,ext=NULL,extf=1.1,
excludep=TRUE,prob=FALSE,cellnumbers=FALSE,tryf=3,
warn=2,lonlatCorrection=TRUE)
bak.vir<-extract(en.vir,back)

RF.out<-niche.Model.Build(prese=present.points,bak.vir=bak.vir,en.vir=en.vir)
RF.out
#prese.env<-extract(en.vir,present.points[,2:3])
### RF.out2<-niche.Model.Build(prese=NULL,absen=NULL,prese.env=prese.env,
absen.env=NULL,bak.vir=bak.vir,en.vir=en.vir)
### RF.out2

## End(Not run)
```

niche.overlap *Degree of niche overlap in a multispecies dataset.*

Description

Calculate the proportion of species pairs with overlapping niches across the total number of species pairs.

Usage

```
niche.overlap(ref, en.vir)
```

Arguments

ref	The reference dataset containing sample ID, taxon information, longitude and latitude and barcode sequences of each sample in class DNAbin.
en.vir	The global bioclimatic data from the "raster::getData" function in class RasterBrick.

Value

A matrix of whether species pairs have overlapping niches (1 means this pair of species have overlapped in the range of 95 percentage confidence intervals with each other on at least one ecological variables, while 0 means not), and a numeric, the proportion representing the degree of niche overlap of the multispecies dataset.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

Examples

```
## Not run:
data(Ref)
ref<-Ref
# Note: set "download=FALSE", if raster::getData() has been run once!
envir<-raster::getData("worldclim",download=TRUE,var="bio",
res=2.5,lon=lon,lat=lat)
en.vir<-brick(envir)

n.overlap<-niche.overlap(ref,en.vir)
n.overlap$niche.overlap

## End(Not run)
```

niche.PCA *Principal component analysis of ecological niche among unknown species and the potential species to which they may belong*

Description

Determine whether unknown species belong to a known species through principal component analysis of their ecological niches according to their distributional information.

Usage

```
niche.PCA(ref.lonlat, que.lonlat, en.vir)
```

Arguments

ref.lonlat	A data frame of coordinates of a known species.
que.lonlat	A data frame of coordinates of unknown species.
en.vir	The globle bioclimate data from "raster::getData" function in class RasterBrick.

Value

A list containing importance and loadings of the components.

A figure showing whether the query points (blue solid circles) are located in the 95 percentage confidence intervals of the niche space of reference species.

Author(s)

Cai-qing YANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

Examples

```
## Not run:
data<-data.frame(species=rep("Acosmeryx anceus",3),
Lon=c(145.380,145.270,135.461),
Lat=c(-16.4800,-5.2500,-16.0810))
simuSites<-pseudo.present.points(data,10,500,30)
ref.lonlat<-simuSites[1:10,]
que.lonlat<-simuSites[481:500,]
# Note: set "download=FALSE", if raster::getData() has been run once!
envir<-raster::getData("worldclim",download=TRUE,var="bio",res=2.5,lon=lon,lat=lat)
en.vir<-brick(envir)

PCA.summary<-niche.PCA(ref.lonlat,que.lonlat,en.vir)
PCA.summary

## End(Not run)
```

NicoB *Niche-models corrected barcoding identification.*

Description

Species identification using both the barcoding and niche models.

Usage

```
NicoB(ref.seq, que.seq, barcode.method, en.vir, bak.vir,
  bio.variables = c(1:19))
```

Arguments

ref.seq	The reference dataset containing sample ID, taxon information, longitude and latitude and barcode sequences of each sample in class DNAbin.
que.seq	The query file containing sample ID, longitude and latitude and barcode sequences of each sample in class DNAbin.
barcode.method	A character string indicating which method will be used to train the model and/or infer species membership. One of these methods ("fuzzyId", "bpNewTraining", "Bayesian") should be specified.
en.vir	The global bioclimatic data from the "raster::getData" function in class RasterBrick.
bak.vir	Bioclimatic variables of random background points in class matrix.
bio.variables	The identifier of selected bioclimate variables in class integer. Default of c(1:19), representing the 19 commonly used bioclimate variables.

Value

A data frame of barcoding identification results for each query sample and their niche-based reliabilities.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

References

Breiman, L. 2001. Random forests. Machine Learning 45(1): 5-32.

Liaw, A. and M. Wiener. 2002. Clasification and regression by randomForest. R News 2/3: 18-22.

Zhang, A.B, Hao, M.D., Yang,C.Q., Shi, Z.Y. (2016). BarcodingR: an integrated R package for species identification using DNA barcodes. Methods in Ecology and Evolution. https://doi.org/10.1111/2041-210X.12682.

Jin,Q., H.L. Han, X.M. Hu, X.H. Li,C.D. Zhu,S. Y. W. Ho, R. D. Ward, A.B. Zhang. (2013). Quantifying Species Diversity with a DNA Barcoding-Based Method: Tibetan Moth Species (Noctuidae) on the Qinghai-Tibetan Plateau. PloS One 8: e644.

Zhang, A. B., C. Muster, H.B. Liang, C.D. Zhu, R. Crozier, P. Wan, J. Feng, R. D. Ward.(2012). A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. Molecular Ecology, 21(8):1848-63.

Zhang, A. B., D. S. Sikes, C. Muster, S. Q. Li. (2008). Inferring Species Membership using DNA sequences with Back-propagation Neural Networks. Systematic Biology, 57(2):202-215.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15): 1965-1978.

Examples

```
## Not run:
data(Ref)
data<-Ref
ref.seq=data[1:25,]
ref.seq
que.seq=data[26:30,];rownames(que.seq)<-gsub("\\,[A-Za-z0-9\\.\\-\_]*\\,","\\,",rownames(que.seq))
que.seq
# Note: set "download=FALSE", if raster::getData() has been run once!
envir<-raster::getData("worldclim",download=TRUE,var="bio",
res=2.5,lon=lon,lat=lat) ## first run, download=true, second run, download=FALSE!
en.vir<-brick(envir)
back<-randomPoints(mask=en.vir,n=5000,ext=NULL,extf=1.1,
excludep=TRUE,prob=FALSE,cellnumbers=FALSE,tryf=3,
warn=2,lonlatCorrection=TRUE)
bak.vir<-extract(en.vir,back)

NMCB<-NicoB(ref.seq,que.seq,barcode.method="Bayesian",en.vir,bak.vir,bio.variables=c(1:19))
NMCB

## End(Not run)
```

`NicoB2` *Niche-model corrected for a given barcoding identification.*

Description

Correct the barcoding identification result according to a given niche model.

Usage

```
NicoB2(ref.infor, que.infor, barcode.identi.result, en.vir, bak.vir,
  bio.variables = c(1:19))
```

Arguments

`ref.infor`	The reference dataset containing sample ID, taxon information and longitude and latitude of each sample in class data.frame.
`que.infor`	The query file containing sample ID and longitude and latitude of each sample in class data.frame.
`barcode.identi.result`	The previous results of species identification containing query ID, target species and credibility in class data.frame.

en.vir	The global bioclimatic data from the "raster::getData" function in class RasterBrick.
bak.vir	Bioclimatic variables of random background points in class matrix.
bio.variables	The identifier of selected bioclimatic variables in class integer. Default of c(1:19), representing the 19 commonly-used bioclimate variables.

Value

A data frame of barcoding identification results for each query sample and their niche-based reliabilities.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

References

Breiman, L. 2001. Random forests. Machine Learning 45(1): 5-32.

Liaw, A. and M. Wiener. 2002. Clasification and regression by randomForest. R News 2/3: 18-22.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15): 1965-1978.

Examples

```
## Not run:
data(Identified_results)
# Note: set "download=FALSE", if raster::getData() has been run once!
 envir<-raster::getData("worldclim",download=TRUE,var="bio",
 res=2.5,lon=lon,lat=lat)
 en.vir<-brick(envir)
back<-randomPoints(mask=en.vir,n=5000,ext=NULL,extf=1.1,
excludep=TRUE,prob=FALSE,cellnumbers=FALSE,tryf=3,
warn=2,lonlatCorrection=TRUE)
bak.vir<-extract(en.vir,back)

NMCB2<-NicoB2(ref.infor,que.infor,barcode.identi.result,en.vir,
bak.vir,bio.variables=c(1:19))
NMCB2

## End(Not run)
```

---

pseudo.absent.points *Generation of pseudo absence points for niche model building*

Description

Randomly generates pseudo points outside the 95 percentage confidence intervals (CI) of the ecological space of the presence data when there is no absence data for building a niche model.

Usage

```
pseudo.absent.points(data, en.vir, outputNum = 500)
```

Arguments

data	A data frame containing longitudes and latitudes of a single species in class data.frame.
en.vir	The global bioclimatic data from the "raster::getData" function in class RasterBrick.
outputNum	The expected number of points.

Value

A data frame of simulated pseudo points. A data frame of bioclimatic variables for these pseudo points.

Author(s)

Cai-qing YANG, Cong JIANG, and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

Examples

```
## Not run:
data<-data.frame(species=rep("Acosmeryx anceus",3),
Lon=c(145.380,145.270,135.461),
Lat=c(-16.4800,-5.2500,-16.0810))
# Note: set "download=FALSE", if raster::getData() has been run once!
envir<-raster::getData("worldclim",download=TRUE,var="bio",
res=2.5,lon=lon,lat=lat)
en.vir<-brick(envir)

absent.points<-pseudo.absent.points(data,en.vir,outputNum=100)
head(absent.points$lonlat)
head(absent.points$envir)

## End(Not run)
```

---

pseudo.present.points *Generation of pseudo present points for niche model building*

Description

Randomly generates pseudo points around actual presence distribution sites when the number of present points is inadequate for building a niche model.

Usage

```
pseudo.present.points(data, minNum = 20, outputNum = 50,
  squareRange = 1)
```

Arguments

data	The longitude and latitude of a single species in class data.frame.
minNum	The allowed minimum number of points.
outputNum	The expected number of points.
squareRange	Range of points generated (How many degrees around the actual present points).

Value

A data frame containing actual present points and simulated pseudo points.

Author(s)

Cai-qing YANG and Ai-bing ZHANG, CNU, Beijing, CHINA. Emails: yangcq_ivy(at)163.com; zhangab2008(at)mail.cnu.edu.cn.

Examples

```
data<-data.frame(species=rep("Acosmeryx anceus",3),
Lon=c(145.380,145.270,135.461),
Lat=c(-16.4800,-5.2500,-16.0810))

present.points<-pseudo.present.points(data,10,10,1)
present.points
```

Index