An empirical example: *Rhynoclemmys* data

Daniel R. Miranda-Esquivel

2024 / 03 / 27

I am investigating whether branch lengths in the phylogenetic tree have any effect on the area(s) chosen for conservation. To do this, I will conduct a Phylogenetic Diversity (PD) analysis for the *Rhynoclemmys* genus. The topology I will use corresponds to a Total Evidence analysis from Romero-Alarcon (2020), and the distribution will be modified from Le and McCord (2008).

We read the distribution and the tree

```r
## For reproductibility purposes

set.seed(121)


options(warn=-1)
suppressMessages(library(ggtree))
#~ library(ape)
suppressMessages(library(blepd))
library(ggplot2)
#~
library(ggtree)

## Just in case, to install ggtree
## https://bioconductor.org/packages/release/bioc/html/ggtree.html
####
#  if (!require("BiocManager", quietly = TRUE))
#      install.packages("BiocManager")
#  BiocManager::install("ggtree")

## Version

cat("Analyses made with blepd version:",unlist(packageVersion("blepd")))
```

```
## Analyses made with blepd version: 0 1 98 2024 3 27
```

```r
## Create an object to place the distribution and the tree

RhinoclemmysData <- list()


## Read data

## distribution is a labeled csv file, areas by terminals
```

```r
#setwd("./csv/")

## getwd()


csvFile <- list.files(pattern="csv")

##csvFile


### the functions use a matrix object for the distributions

RhinoclemmysData$distribution <-  as.matrix(read.table(csvFile,
                                   stringsAsFactors=FALSE,
                                   header=TRUE,
                                   row.names=1,
                                   sep=",")
                                   )


print(t(RhinoclemmysData$distribution))
```

```
##                Al An-As Cho Nsa Nuh Ph Pl
## R_annulata      1     0   1   0   0  1  0
## R_areolata      1     0   0   0   0  0  0
## R_diademata     0     0   0   1   0  0  0
## R_funerea       1     0   0   0   1  0  0
## R_melanosterna  0     0   1   0   0  1  0
## R_nasuta        0     0   1   0   0  0  0
## R_pulcherrima   0     0   0   0   1  0  1
## R_punctularia   0     1   0   0   0  0  0
## R_rubida        0     0   0   0   0  0  1
```

```r
## tree(s) in nexus or newick format

##setwd("../tree/")

treeFiles <- list.files(pattern=".tre")

##treeFiles


RhinoclemmysData$tree  <-    read.tree(treeFiles)


## name of tree(s)

treeFiles <- gsub(".tre","",treeFiles)




## Plotting
```

```
#par(mfrow=c(2,1))


## The tree

## using ggtree

# RhinoclemmysData$tree <- reorder(RhinoclemmysData$tree, order = "cladewise")

plotTree <-  ggtree(RhinoclemmysData$tree, ladderize=TRUE,
                color="black", size=0.51, linetype="solid") +
          geom_tiplab(size=4, color="black") +
          xlim(0,0.030) +
          theme_tree2() +
          ggtitle(treeFiles[1])



##
print(plotTree)
```
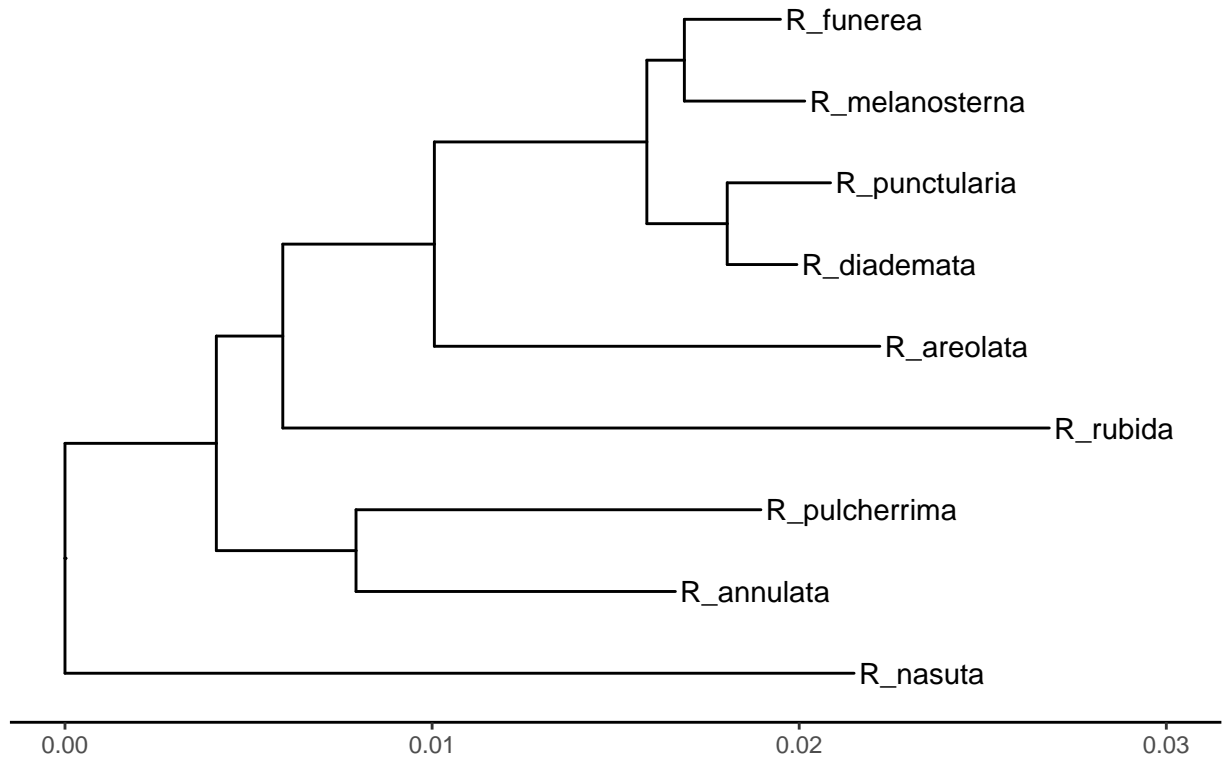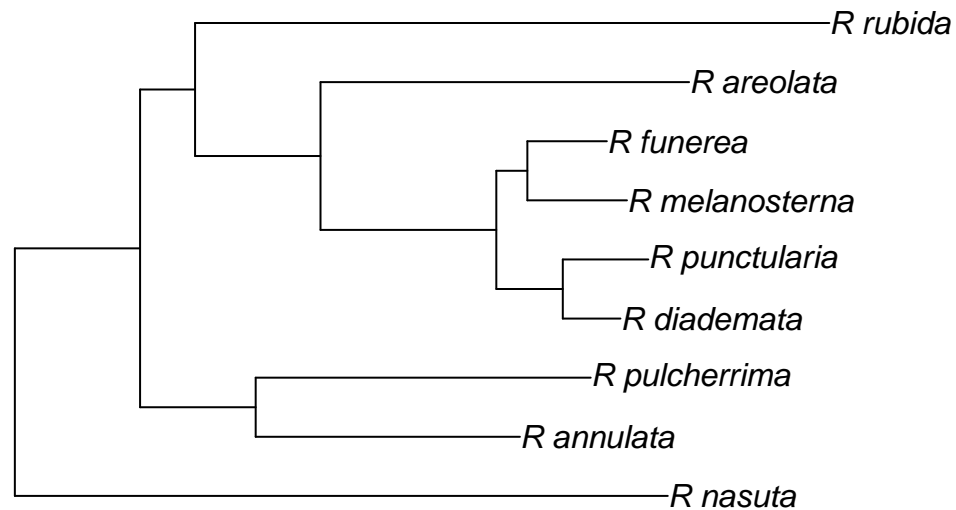
Rhinoclemmys_TotalEvidence_ML


```
#~ Alternatively, we can plot the trees using APE
##
##
```

```
plot.phylo(RhinoclemmysData$tree)
```

```
                                    ─R rubida
                              ─R areolata
                              ─R funerea
                              ─R melanosterna
                              ─R punctularia
                              ─R diademata
                              ─R pulcherrima
                              ─R annulata
                              ─R nasuta
```

```
##
####nodelabels()
####tiplabels()


## to plot the distribution, we must transform it into a data.frame
## object


## distXY <- matrix2XY(RhinoclemmysData$distribution)


## We could reorder the data.frame following the names on the trees

## terminals <- colnames(RhinoclemmysData$distribution)

## realOrder <- match(terminals,RhinoclemmysData$tree$tip.label)

## equivalencias <- data.frame(terminals,realOrder)

## dXY2 <-  distXY

## for(cambiar in terminals){
```

```
##    dXY2$Terminal[distXY$Terminal == cambiar]    <-
##    equivalencias$realOrder[equivalencias$terminals==cambiar]
##}

## distGraficar <- distXY

#~ distGraficar <- dXY2


## plot using ggplot

# plotDistrib <- ggplot(distGraficar,
#                    aes(x= Area, y=Terminal), size =30) +
#             geom_point(shape=19, fill="white",
#                        color="darkgrey",
#                        size=4) +
#             labs(title = "Distribution",
#                   y = "",
#                   x = "Area") +
#             theme(axis.line=element_blank(),
#                   # axis.text.y=element_blank(),
#                 axis.ticks=element_blank(),
#                   # axis.title.y=element_blank(),
#                 legend.position="none",
#                 panel.background=element_blank(),
#                 panel.border=element_blank(),
#                 panel.grid.major=element_blank(),
#                 panel.grid.minor=element_blank(),
#                 plot.background=element_blank()
#                 )
#



#~

## cowplot::plot_grid(plotTree,plotDistrib, ncol=1)

#~
plot(plotTree)
```
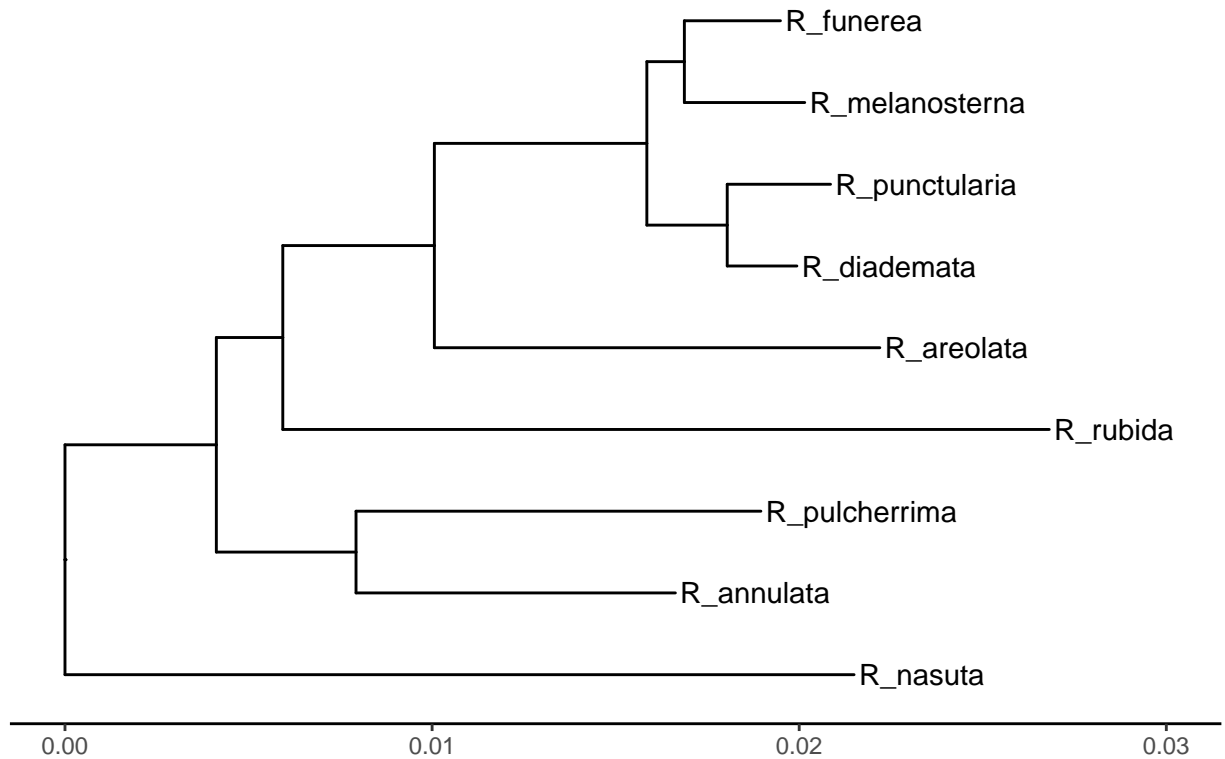
Rhinoclemmys_TotalEvidence_ML


```
#~ plot(plotDistrib)

#~
print(t(RhinoclemmysData$distribution))
```

```
##                  Al An-As Cho Nsa Nuh Ph Pl
## R_annulata        1     0   1   0   0  1  0
## R_areolata        1     0   0   0   0  0  0
## R_diademata       0     0   0   1   0  0  0
## R_funerea         1     0   0   0   1  0  0
## R_melanosterna    0     0   1   0   0  1  0
## R_nasuta          0     0   1   0   0  0  0
## R_pulcherrima     0     0   0   0   1  0  1
## R_punctularia     0     1   0   0   0  0  0
## R_rubida          0     0   0   0   0  0  1
```

```
##dev.off()
```
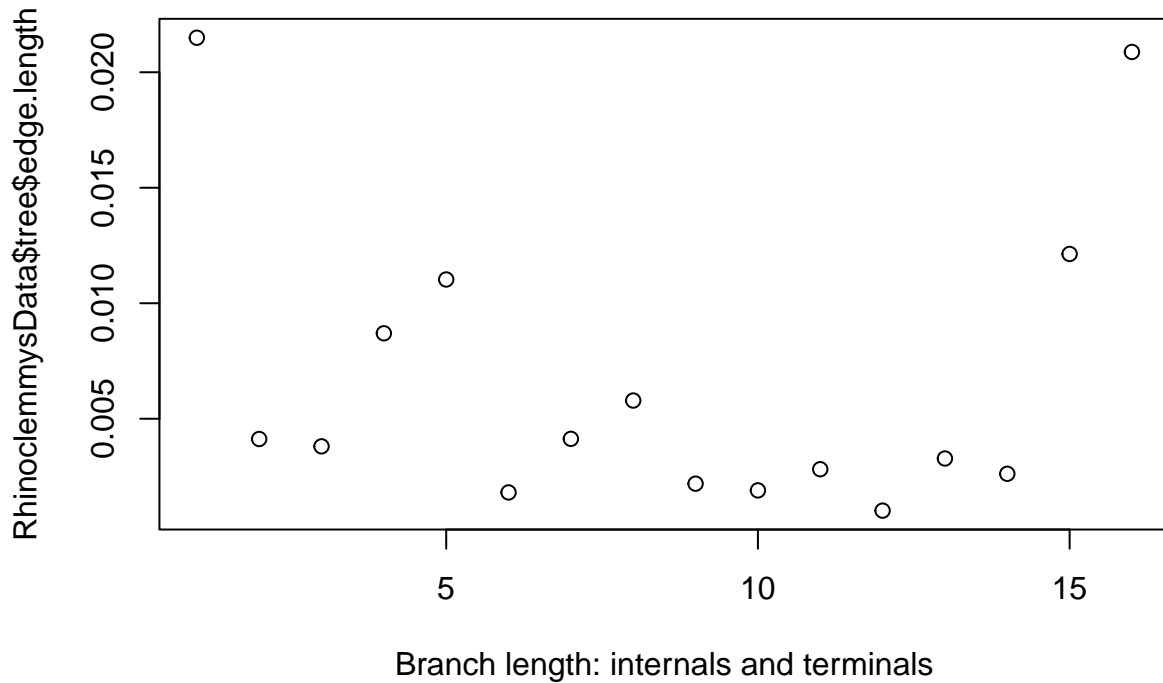
Now we plot the branch lengths

```
#par(mfrow=c(2,1))

a  <- plot(RhinoclemmysData$tree$edge.length,
           bins=12,
           main=treeFiles,
           xlab="Branch length: internals and terminals") +
```
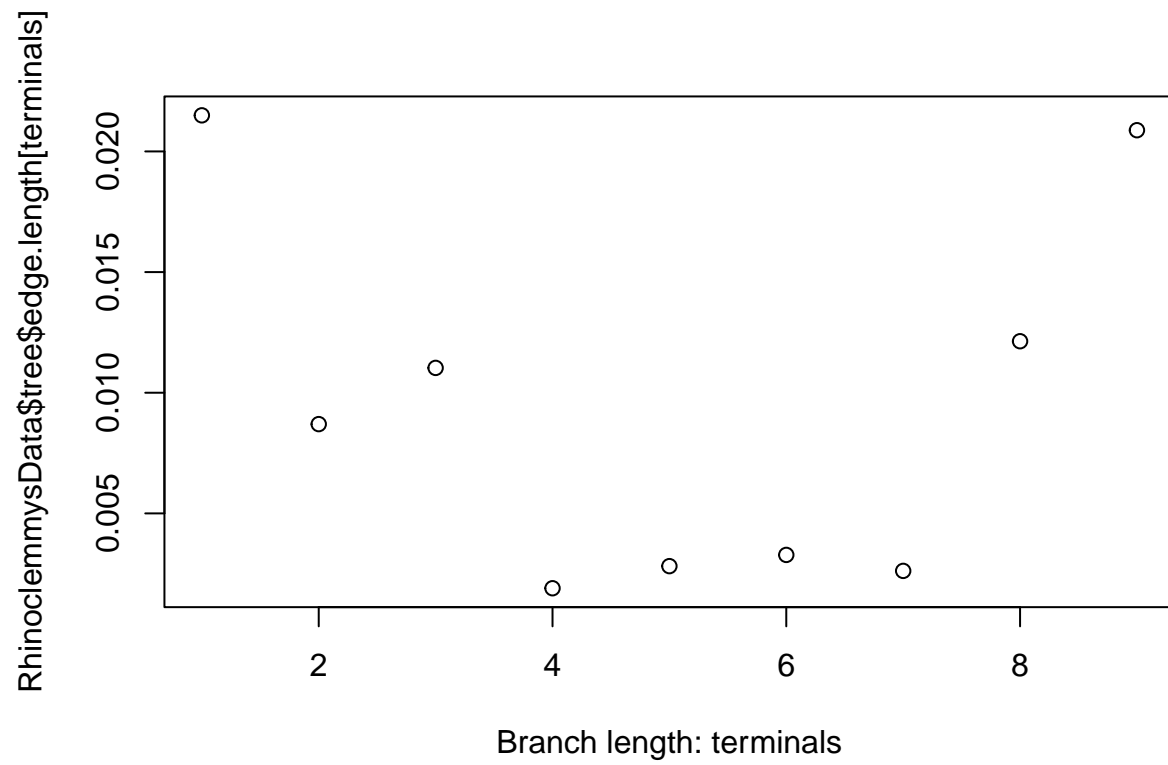
```
        xlim(c(0.001,0.025)) +
        ylim(c(0,5))
```

Rhinoclemmys_TotalEvidence_ML
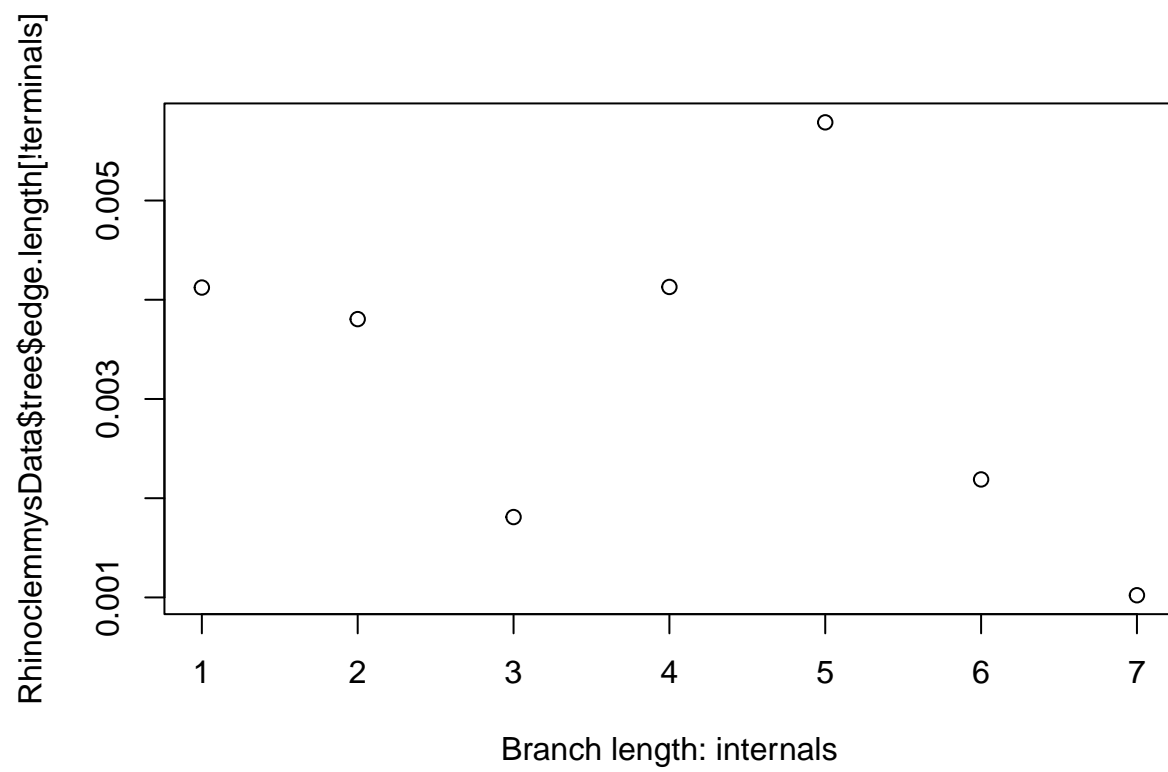

Branch length: internals and terminals

```
terminals <- RhinoclemmysData$tree$edge[,2] < 1 + length(RhinoclemmysData$tree$tip.label)


b  <- plot(RhinoclemmysData$tree$edge.length[terminals],
        bins=12,
        #main=treeFiles,
        xlab="Branch length: terminals")  +
        xlim(c(0.001,0.025)) +
        ylim(c(0,5))
```

```
c  <- plot(RhinoclemmysData$tree$edge.length[!terminals],
          bins=12,
          #main=treeFiles,
          xlab="Branch length: internals")  +
          xlim(c(0.001,0.025)) +
          ylim(c(0,5))
```

```
##
```

```
cowplot::plot_grid(a, b, c, nrow=3)
```

```
##
#
#cat("Terminals, with BL larger than 0.02 = ",findTerminalgivenLength(RhinoclemmysData$tree,0.02))
#
##


###dev.off()
```

The branch length histograms and the tree plot show that the internal length branches are similar among each other, and different to terminals'; there are two longer branches (larger than 0.02), *R. nasuta* (inhabiting Cho) and *R. rubida* (Pl), while the areas Al and Cho are the richest.

PD values

First, I calculate the PD value for the areas, given this tree.

```
RhinoclemmysData$tablePD <- PDindex(RhinoclemmysData$tree,
                           distribution=RhinoclemmysData$distribution,
                                   root=TRUE,
                                   percentual = TRUE)



RhinoclemmysData$matrixPD <- as.data.frame(
                           matrix(
```

```r
                                         unlist(RhinoclemmysData$tablePD),
                                         nrow= length(treeFiles),
                                         byrow=TRUE))
```

```r
colnames(RhinoclemmysData$matrixPD) <- row.names(RhinoclemmysData$distribution)

row.names(RhinoclemmysData$matrixPD) <- "PD value"

new <- (t(RhinoclemmysData$matrixPD))

## officially I am an Idiot

values <-  new[order(new[,1])]
names   <-  rownames(new)[order(new[,1])]

print(as.data.frame(paste(names,values)))
```

```
##   paste(names, values)
## 1            Nsa 8.05
## 2           An-As 8.42
## 3             Ph 13.18
## 4            Nuh 13.86
## 5             Pl 16.81
## 6             Al 17.82
## 7            Cho 21.86
```
cat("Max value is:",max(t(RhinoclemmysData$matrixPD)),"Area:",colnames(RhinoclemmysData$matrixPD)[wh.

The highest PD is for area Cho, followed by Al, the two richest areas, and the difference in PD value is not given by the richness but the species inhabiting each area.

Effect of branch lengths in the PD

We test the effect of the branch length on the PD values by swapping terminal and/or internal branch lengths, using the three available models.

```r
for( modelo in c("simpleswap","allswap","uniform") ){

        for( rama in c("terminals","internals","all") ){

            val <- swapBL(RhinoclemmysData$tree,
                        RhinoclemmysData$distribution,
                        model = modelo,
                        branch = rama
            )

        cat( "\n\t Tree=",treeFiles,"\n\t Model=",modelo,
            "\n\t Branchs swapped=",rama,"\n" )

        printBlepd(val)

        cat( "\n\n\n" )

    }
```

```
  }
```

```
## model to test simpleswap reps 100
##
##    Tree= Rhinoclemmys_TotalEvidence_ML
##    Model= simpleswap
##    Branchs swapped= terminals
##
## BestInitial:Cho
##    AreaSelected Freq Percent
## 1            Al   13        13
## 2           Cho   85        85
## 3            Pl    2         2
##
##
##
## model to test simpleswap reps 100
##
##    Tree= Rhinoclemmys_TotalEvidence_ML
##    Model= simpleswap
##    Branchs swapped= internals
##
## BestInitial:Cho
##    AreaSelected Freq Percent
## 1           Cho  100       100
##
##
##
## model to test simpleswap reps 100
##
##    Tree= Rhinoclemmys_TotalEvidence_ML
##    Model= simpleswap
##    Branchs swapped= all
##
## BestInitial:Cho
##    AreaSelected Freq Percent
## 1            Al    8         8
## 2           Cho   91        91
## 3            Pl    1         1
##
##
##
## model to test allswap reps 100
##
##    Tree= Rhinoclemmys_TotalEvidence_ML
##    Model= allswap
##    Branchs swapped= terminals
##
## BestInitial:Cho
##    AreaSelected Freq Percent
## 1            Al   35        35
## 2           Cho   53        53
## 3           Nuh   10        10
```

```
## 4            Pl    2        2
##
##
##
## model to test allswap reps 100
##
##   Tree= Rhinoclemmys_TotalEvidence_ML
##   Model= allswap
##   Branchs swapped= internals
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1          Cho  100      100
##
##
##
## model to test allswap reps 100
##
##   Tree= Rhinoclemmys_TotalEvidence_ML
##   Model= allswap
##   Branchs swapped= all
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al   37       37
## 2        An-As    4        4
## 3          Cho   33       33
## 4          Nsa    4        4
## 5          Nuh   19       19
## 6           Pl    3        3
##
##
##
## model to test uniform reps 100
##
##   Tree= Rhinoclemmys_TotalEvidence_ML
##   Model= uniform
##   Branchs swapped= terminals
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al   46       46
## 2          Cho   50       50
## 3          Nuh    4        4
##
##
##
## model to test uniform reps 100
##
##   Tree= Rhinoclemmys_TotalEvidence_ML
##   Model= uniform
##   Branchs swapped= internals
##
## BestInitial:Cho
```

```
##   AreaSelected Freq Percent
## 1          Cho  100     100
##
##
##
## model to test uniform reps 100
##
##   Tree= Rhinoclemmys_TotalEvidence_ML
##   Model= uniform
##   Branchs swapped= all
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al   38      38
## 2        An-As    1       1
## 3          Cho   53      53
## 4          Nsa    1       1
## 5          Nuh    7       7
```

The terminal branch length is critical in the decision taken, if we swap all terminal branch lengths (model= "allswap"), or if we replace them with a uniform distribution (model= "uniform"), the area selected might change from Cho to Al.

As the terminal branch lengths are distributed unequally, we might suspect that the results could depend on the longest branches that inhabit the areas Cho/Al.

But first, we test if the number of replicates has any effect.

```r
for(repetir in 1:2){

        val <- swapBL( RhinoclemmysData$tree,
                       RhinoclemmysData$distribution,
                       model = "allswap",
                       branch = "terminals",
                       nTimes = 10**repetir
                     )

        printBlepd(val)

}
```

```
## model to test allswap reps 10
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al    1      10
## 2          Cho    6      60
## 3          Nuh    3      30
## model to test allswap reps 100
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al   39      39
## 2          Cho   40      40
## 3          Nsa    1       1
## 4          Nuh   19      19
```

14

```
## 5              Pl      1        1
```

Roughly speaking, from 1000 on the results are alike, and the largest difference is using 10 or 100 replicates. As a rule of thumb, we must use at least 100 replicates, but 1000 will be better.

Now, let us see if the difference in results could be assigned to the longest branches (or not).

```r
## we could use a single command (*evalBranch*)


AllTerminals <- evalBranch(tree=RhinoclemmysData$tree,
                           distribution=RhinoclemmysData$distribution,
                           branchToEval="terminals",
                           approach="all")
```

```
## Loading required package: maps

##
## Attaching package: 'phytools'

## The following object is masked from 'package:vegan':
##
##      scores
```

```r
print.evalBranchAll(AllTerminals)
```

```
##
## BranchNumber initArea    Mod-upper   Delta%  Mod-lower   Delta%
## NA     Cho Cho 0    Al  -47.18
## NA     Cho Cho 0    Cho 0
## NA     Cho Pl   116.08  Cho 0
## NA     Cho Nsa 1819.58 Cho 0
## NA     Cho An-As   1202.58 Cho 0
## NA     Cho Cho 0    Cho 0
## NA     Cho Al   391.02  Cho 0
## NA     Cho Al   84.95   Cho 0
## NA     Cho Pl   61.32   Cho 0
##
```

```r
AllInternals <- evalBranch(tree=RhinoclemmysData$tree,
                           distribution=RhinoclemmysData$distribution,
                           branchToEval="internals",
                           approach="all")
```

```r
print.evalBranchAll(AllInternals)
```

```
##
## BranchNumber initArea     Mod-upper   Delta%  Mod-lower   Delta%
## 11    Cho Cho 0    Cho 0
## 12    Cho Cho 0    Cho 0
## 13    Cho Cho 0    Cho 0
## 14    Cho Cho 0    Cho 0
## 15    Cho Cho 0    Cho 0
## 16    Cho An-As   1529.13 Cho 0
## 17    Cho Cho 0    Cho 0
##
```

The area selected depends on the branch length. While the internal terminals had almost no impact, if the terminal branch length of *R. nasuta* is -47.18% shorter, or if the terminal branch length is 84.95% larger for

R. areolata, the area selected will change from Cho to Al, and if the terminal branch length of *R. rubida* is 61.32% larger, the area selected will change from Cho to Pl.

Literature cited

Le, Minh, and William P. McCord. 2008. "Phylogenetic relationships and biogeographical history of the genus Rhinoclemmys Fitzinger, 1835 and the monophyly of the turtle family Geoemydidae (Testudines: Testudinoidea)." *Zoological Journal of the Linnean Society* 153 (4): 751–67. https://doi.org/10.1111/j. 1096-3642.2008.00413.x.

Romero-Alarcon, L. Viviana. 2020. "A total-evidence phylogeny of the crown and stem-groups of turtles (Pantestudines: Testudinata)." Master's thesis, Colombia: Escuela de Biología, UIS.