An empirical example: *Rhynoclemmys* data

Daniel R. Miranda-Esquivel

2022/03/11

We want to test whether the branch lengths have any effect in the area(s) chosen for conservation, based on a Phylogenetic Diversity (PD) analysis for the *Rhynoclemmys* genus. The topology corresponds to a Total Evidence analysis from Romero-Alarcon (2020), and the distribution is modified from Le and McCord (2008).

We read the distribution and the tree

```
## For reproductibility purposes

set.seed(121)


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


## Version

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

```
## Analyses made with blepd version: 0 1 85 2022 1 7
```

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

RhinoclemmysData <- list()


## Read data

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


#setwd("./csv/")

##

getwd()
```

```
## [1] "/home/rafael/disco2/proyectosDRME/indices/blepd/docs"
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
## 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))
```

```r
## 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)


#~ Alternatively, we can plot the trees using APE
##
## plot.phylo(RhinoclemmysData$tree)
##
####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]

}
```

```r
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)
```
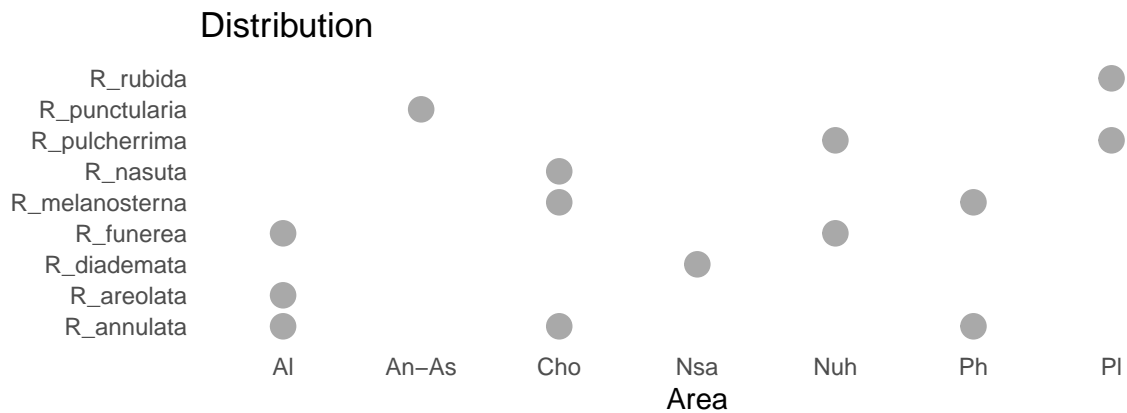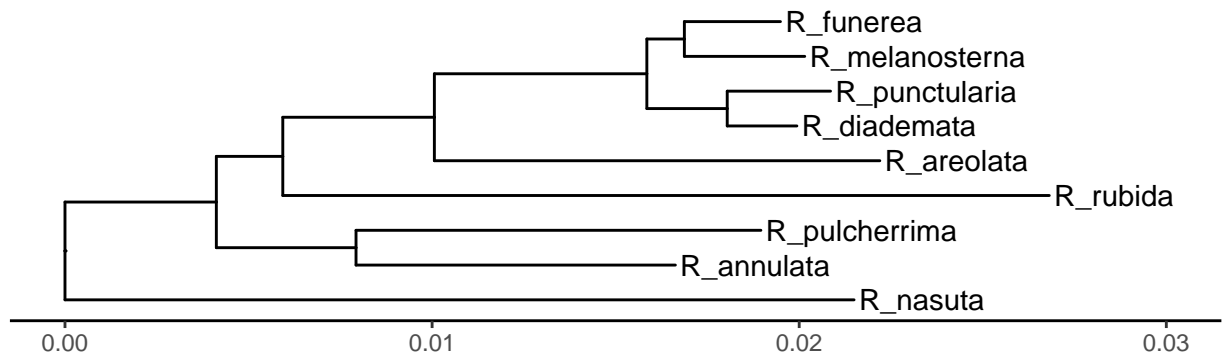
Rhinoclemmys_TotalEvidence_ML


```r
#~ plot(plotTree)

#~ plot(plotDistrib)

#~ print(t(RhinoclemmysData$distribution))


##dev.off()
```

Now we plot the branch lengths

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

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


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


b  <- qplot(RhinoclemmysData$tree$edge.length[terminals],
```

```
        bins=12,
        #main=treeFiles,
        xlab="Branch length: terminals")  +
        xlim(c(0.001,0.025)) +
        ylim(c(0,5))


c  <- qplot(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)
```
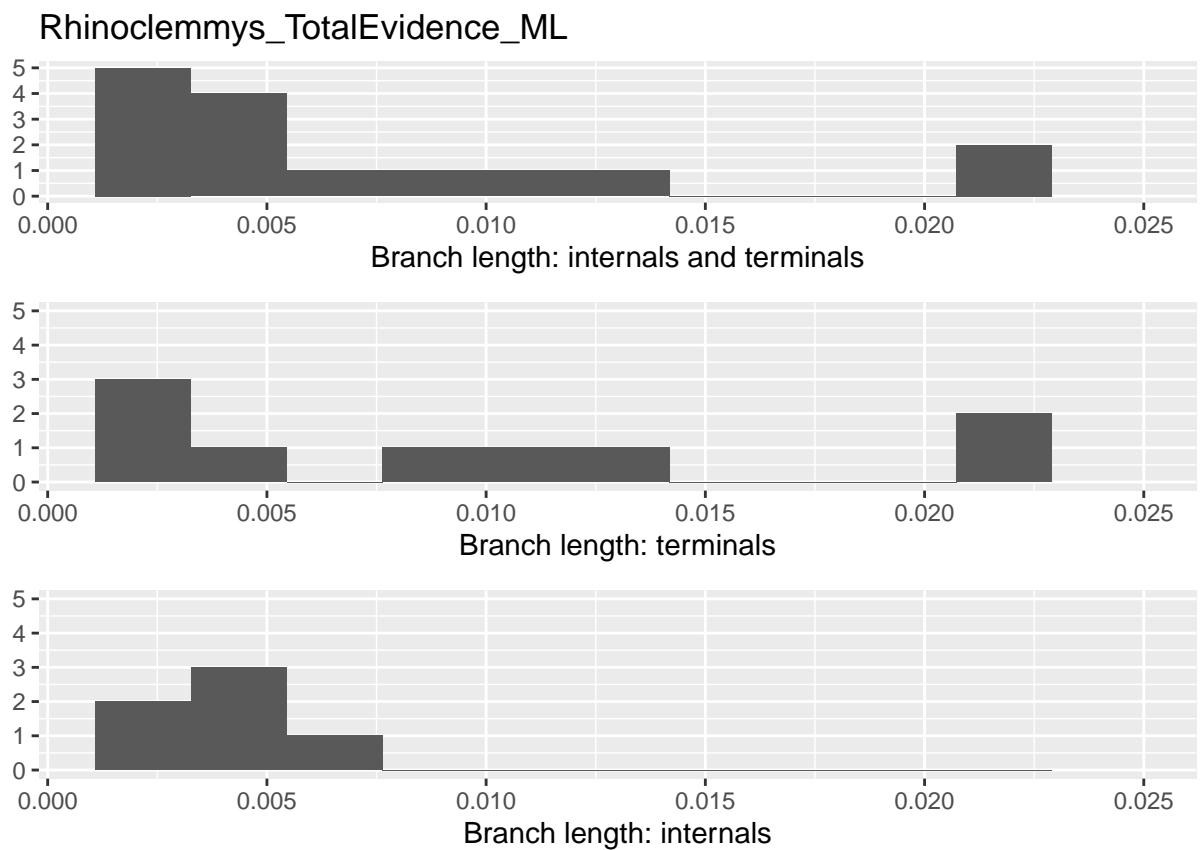
Rhinoclemmys_TotalEvidence_ML

Branch length: internals and terminals

Branch length: terminals

Branch length: internals

```
##
#
#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, we calculate the PD value for the areas, given this tree.

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



RhinoclemmysData$matrixPD <- as.data.frame(
                                    matrix(
                                        unlist(RhinoclemmysData$tablePD),
                                        nrow= length(treeFiles),
                                        byrow=TRUE))



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

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

print(t(sort(RhinoclemmysData$matrixPD,decreasing = TRUE)))
```

```
##          PD value
## Cho         21.86
## Al          17.82
## Pl          16.81
## Nuh         13.86
## Ph          13.18
## An-As        8.42
## Nsa          8.05
```

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" )

        print.blepd(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   86      86
## 3           Pl    1       1
##
##
##
## 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.

```
for(repetir in 1:5){

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

        print.blepd(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
## 5           Pl    1       1
## model to test allswap reps 1000
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al  394    39.4
## 2        An-As   12     1.2
## 3          Cho  415    41.5
## 4          Nsa    7     0.7
## 5          Nuh  144    14.4
## 6           Pl   28     2.8
## model to test allswap reps 10000
##
## BestInitial:Cho
##   AreaSelected Freq Percent
## 1           Al 3892   38.92
## 2        An-As   85    0.85
## 3          Cho 4427   44.27
## 4       ChoNuh    2    0.02
## 5          Nsa   88    0.88
## 6          Nuh 1212   12.12
## 7           Pl  294    2.94
## model to test allswap reps 1e+05
##
## BestInitial:Cho
##   AreaSelected  Freq Percent
## 1           Al 39724   39.72
## 2        An-As   762    0.76
## 3          Cho 43652   43.65
## 4       ChoNuh    14    0.01
## 5          Nsa   768    0.77
## 6          Nuh 12263   12.26
## 7           Pl  2817    2.82
```

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 possible difference in results could be assigned to the longest branches.

```
## Testing the effect if we change the branch length from the actual
## value to 0.0. We use evalTerminal(approach = "lower")

testEvalTerminalALL <- evalTerminal( RhinoclemmysData$tree,
                                     RhinoclemmysData$distribution,
                                     tipToEval = "all",
                                     approach = "lower" )
```

```r
print.multiBlepd( testEvalTerminalALL )
```

```
##
## Delta (%)    Terminal     Initial Selected    Evaluating lower
## -47.18    R_nasuta    Cho Al
## 0      R_annulata      Cho Cho
## 0      R_pulcherrima  Cho Cho
## 0      R_diademata     Cho Cho
## 0      R_punctularia  Cho Cho
## 0      R_melanosterna     Cho Cho
## 0      R_funerea  Cho Cho
## 0      R_areolata      Cho Cho
## 0      R_rubida    Cho Cho
```

```r
## Testing the effect if we change the branch length from the actual
## value to the sum of all branch lengths. We use evalTerminal(approach = "upper")


testEvalTerminalALL <- evalTerminal( RhinoclemmysData$tree,
                                     RhinoclemmysData$distribution,
                                     tipToEval = "all",
                                     approach = "upper" )

print.multiBlepd( testEvalTerminalALL )
```

```
##
## Delta (%)    Terminal     Initial Selected    Evaluating upper
## 0      R_nasuta    Cho Cho
## 0      R_annulata      Cho Cho
## 116.08    R_pulcherrima  Cho Pl
## 1819.58    R_diademata     Cho Nsa
## 1202.58    R_punctularia  Cho An-As
## 0      R_melanosterna     Cho Cho
## 391.02    R_funerea  Cho Al
## 84.95      R_areolata      Cho Al
## 61.32      R_rubida    Cho Pl
```

The area selected depends on the branch length. 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.