732A51 Bioinformatics

LAB 3 Bioinformatics

Hugo Morvan
William Wiik

STIMA
Department of Computer and Information Science
Linköpings universitet

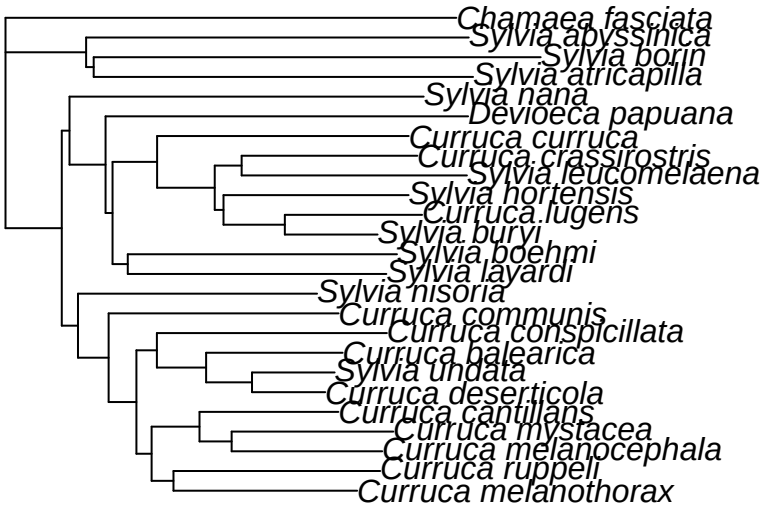
2024-11-29

Contents

1 Question 1

Using the script http://ape-package.ird.fr/APER/APER2/SylviaWarblers.R obtain the *Sylvia* warblers phylogeny (the script saves in in the file sylvia_nj_k80.tre). The geographical range data can be found in http://ape-package.ird.fr/APER/APER2/sylvia_data.txt and in the script is referenced as `DF$geo.range`. Notice that one tip is removed due to missing data

```
sw_phyl <- read.tree("sylvia_nj_k80.tre")
plot(sw_phyl)
```




```
DF$geo.range
```

```
##  [1] "temp"     "temptrop" "temptrop" "temptrop" "temptrop" "temp"
##  [7] "temp"     "temp"     "temptrop" "temptrop" "temptrop" "trop"
## [13] "trop"     "trop"     "trop"     "temptrop" "trop"     "temptrop"
## [19] "temptrop" "temptrop" "temptrop" "temptrop" "trop"
```

1.1 Question 1.1

Explain all the steps in the script required to obtain the phylogeny and trait data.

Answer:

Load the libraries:

```
library(ape)
library(phyloch)
```

Read the nucleotid sequence:

```
x <- paste("AJ5345", 26:49, sep = "")
x <- c("Z73494", x)
sylvia.seq <- read.GenBank(x)
```

Obtain species names and get rid of the rest:

```
taxa.sylvia <- attr(sylvia.seq, "species")
names(taxa.sylvia) <- names(sylvia.seq)
rm(sylvia.seq)
taxa.sylvia[1] <- "Sylvia_atricapilla"
taxa.sylvia[24] <- "Sylvia_abyssinica"
```

Read data from text file, then save the data:

```
sylvia.eco <- read.table("sylvia_data.txt")
str(sylvia.eco)
rownames(sylvia.eco)
save(sylvia.clus, taxa.sylvia, sylvia.eco,
     file = "sylvia.RData")
```

Load the DNA sequences, and calculate pairwise distance matrices from the DNA sequences using various DNA evolutionary models (K80, F84, TN93, GG95):

```
sylvia.seq.ali<-sylvia.seq
syl.K80 <- dist.dna(sylvia.seq.ali, pairwise.deletion = TRUE)
syl.F84 <- dist.dna(sylvia.seq.ali, model = "F84", p = TRUE)
syl.TN93 <- dist.dna(sylvia.seq.ali, model = "TN93", p = TRUE)
syl.GG95 <- dist.dna(sylvia.seq.ali, model = "GG95", p = TRUE)
```

Print correlation between the different model estimations:

```
round(cor(cbind(syl.K80, syl.F84, syl.TN93, syl.GG95)), 3)
```

More correlation analysis ?:

```
syl.JC69 <- dist.dna(sylvia.seq.ali, model = "JC69", p = TRUE)
syl.raw <- dist.dna(sylvia.seq.ali, model = "raw", p = TRUE)
layout(matrix(1:2, 1))
plot(syl.JC69, syl.raw)
abline(b = 1, a = 0) # draw x = y line
plot(syl.K80, syl.JC69)
abline(b = 1, a = 0)
```

Clustering analysis ?

```
layout(matrix(1:3, 1))
for (i in 1:3) {
    s <- logical(3); s[i] <- TRUE
    x <- sylvia.seq.ali[, s]
    d <- dist.dna(x, p = TRUE)
    ts <- dist.dna(x, "Ts", p = TRUE)
    tv <- dist.dna(x, "Tv", p = TRUE)
    plot(ts, d, xlab = "Number of Ts or Tv", col = "blue",
         ylab = "K80 distance", xlim = range(c(ts, tv)),
         main = paste("Position", i))
    points(tv, d, col = "red")
}
```

Some plotting:

```
y <- numeric()
for (i in 1:3) {
    s <- logical(3); s[i] <- TRUE
    y <- c(y, dist.dna(sylvia.seq.ali[, s], p = TRUE))
}
g <- gl(3, length(y) / 3)
library(lattice)
histogram(~ y | g, breaks = 20)
```

Calculate distance between to topo:

```
nj.sylvia.K80 <- nj(syl.K80)
nj.sylvia.GG95 <- nj(syl.GG95)
dist.topo(nj.sylvia.K80, nj.sylvia.GG95)
```

Bootstrap something then use result to create the tree:

```
grep("Chamaea", taxa.sylvia, value = TRUE)
f <- function(xx) root(nj(dist.dna(xx, p=TRUE)), "AJ534526")
tr <- f(sylvia.seq.ali)
## same than: tr <- root(nj.sylvia.K80, "AJ534526")
nj.boot.sylvia <- boot.phylo(tr, sylvia.seq.ali, f, 200,
```

```
                              rooted = TRUE)
nj.boot.codon <- boot.phylo(tr, sylvia.seq.ali, f, 200, 3,
                            rooted = TRUE)
nj.est <- tr
nj.est$tip.label <- taxa.sylvia[tr$tip.label]
plot(nj.est, no.margin = TRUE)
nodelabels(round(nj.boot.sylvia / 200, 2), bg = "white")
add.scale.bar(length = 0.01)
write.tree(nj.est, "sylvia_nj_k80.tre")
```

1.2 Question 1.2 *

2 Question 2

Install the `ade4` package. Included with it you will find the carnivores dataset, `data(carni70)`

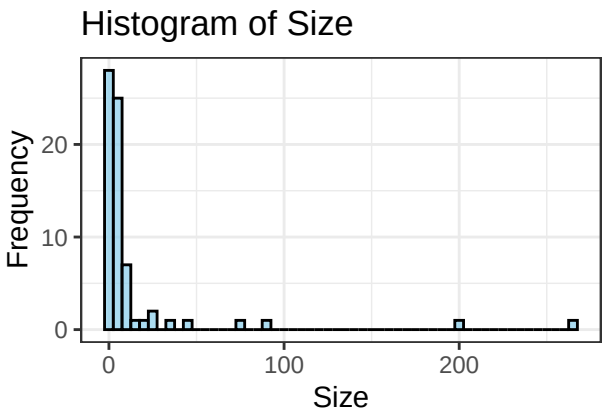
2.1 Question 2.1

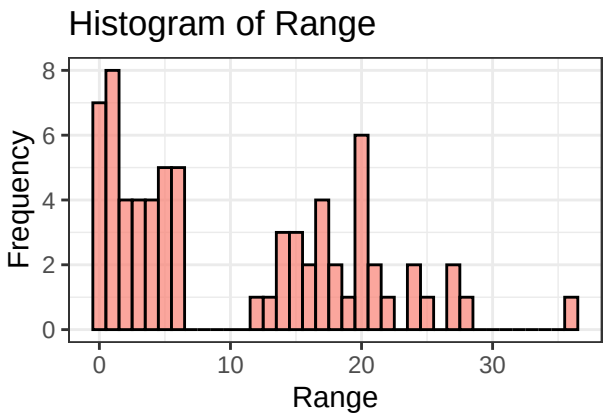
Explore the data set and report what can be found in it. Provide some plots.

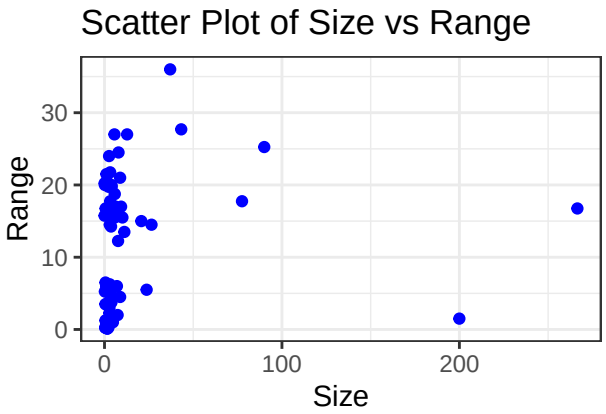
```
library(ade4)
library(ggplot2)
library(cowplot)

data(carni70)

tab_df <- as.data.frame(carni70$tab)
summary(tab_df)
```

```
##       size              range        
##  Min.   :  0.040   Min.   : 0.120  
##  1st Qu.:  1.282   1st Qu.: 2.062  
##  Median :  3.200   Median : 6.125  
##  Mean   : 14.288   Mean   :10.721  
##  3rd Qu.:  7.293   3rd Qu.:17.750  
##  Max.   :266.500   Max.   :36.000
```

```
p1 <- ggplot(tab_df, aes(x = size)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Size", x = "Size", y = "Frequency") + theme_bw()

p2 <- ggplot(tab_df, aes(x = range)) +
  geom_histogram(binwidth = 1, fill = "salmon", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Range", x = "Range", y = "Frequency") + theme_bw()

p3 <- ggplot(tab_df, aes(x = size, y = range)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot of Size vs Range", x = "Size", y = "Range") + theme_bw()


cowplot::plot_grid(p1, p2, p3, ncol = 2)
```








```
## [1] "Total number of carnivores: 70"
```

```
## [1] "The carnivore with biggest size: Ursus_arctos"
```

```
## [1] "The carnivore with smallest size: Mustela_nivalis"
```

```
## [1] "The carnivore with biggest range: Puma_concolor"
```

```
## [1] "The carnivore with smallest range: Bassariscus_pauli"
```

There are 70 carnivores, with a median size of 3.2 and a median range of 6.1. Two clear outliers in size are Ursus arctos (Brown bear) and Tremarctos ornatus (Spectacled bear).

2.2 Question 2.2*