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 **1**

1.1 Question 1.1 1

1.2 Question 1.2 * 4

2 Question 2 **7**

2.1 Question 2.1 7

2.2 Question 2.2* 8

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

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)
```

Allign the sequences

```
sylvia.clus <- clustal(sylvia.seq)
library(phyloch)
```

```
## Loading required package: colorspace

## Loading required package: XML
```

```
sylvia.maff <- mafft(sylvia.seq, path = "/usr/bin/mafft")
identical(sylvia.clus[x, ], sylvia.maff[x, ]) #check that the result are equivalent
```

```
## [1] TRUE
```

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)
```

```
## 'data.frame':    26 obs. of  3 variables:
##  $ mig.dist : int  0 5000 7500 5900 5500 3400 2600 0 0 0 ...
##  $ mig.behav: chr  "resid" "short" "long" "long" ...
##  $ geo.range: chr  "trop" "temptrop" "temptrop" "temptrop" ...
```

```
rownames(sylvia.eco)
```

```
##  [1] "Sylvia_abyssinica"    "Sylvia_atricapilla"   "Sylvia_borin"
##  [4] "Sylvia_nisoria"       "Sylvia_curruca"       "Sylvia_hortensis"
##  [7] "Sylvia_crassirostris" "Sylvia_leucomelaena"  "Sylvia_buryi"
## [10] "Sylvia_lugens"        "Sylvia_layardi"       "Sylvia_subcaeruleum"
## [13] "Sylvia_boehmi"        "Sylvia_nana"          "Sylvia_deserti"
## [16] "Sylvia_communis"      "Sylvia_conspicillata" "Sylvia_deserticola"
## [19] "Sylvia_undata"        "Sylvia_sarda"         "Sylvia_balearica"
## [22] "Sylvia_cantillans"    "Sylvia_mystacea"      "Sylvia_melanocephala"
## [25] "Sylvia_rueppelli"     "Sylvia_melanothorax"
```

```
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.maff #or sylvia.clus or sylvia.eco
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)
```

```
##          syl.K80 syl.F84 syl.TN93 syl.GG95
## syl.K80    1.000   1.000    1.000    0.928
## syl.F84    1.000   1.000    1.000    0.927
## syl.TN93   1.000   1.000    1.000    0.925
## syl.GG95   0.928   0.927    0.925    1.000
```

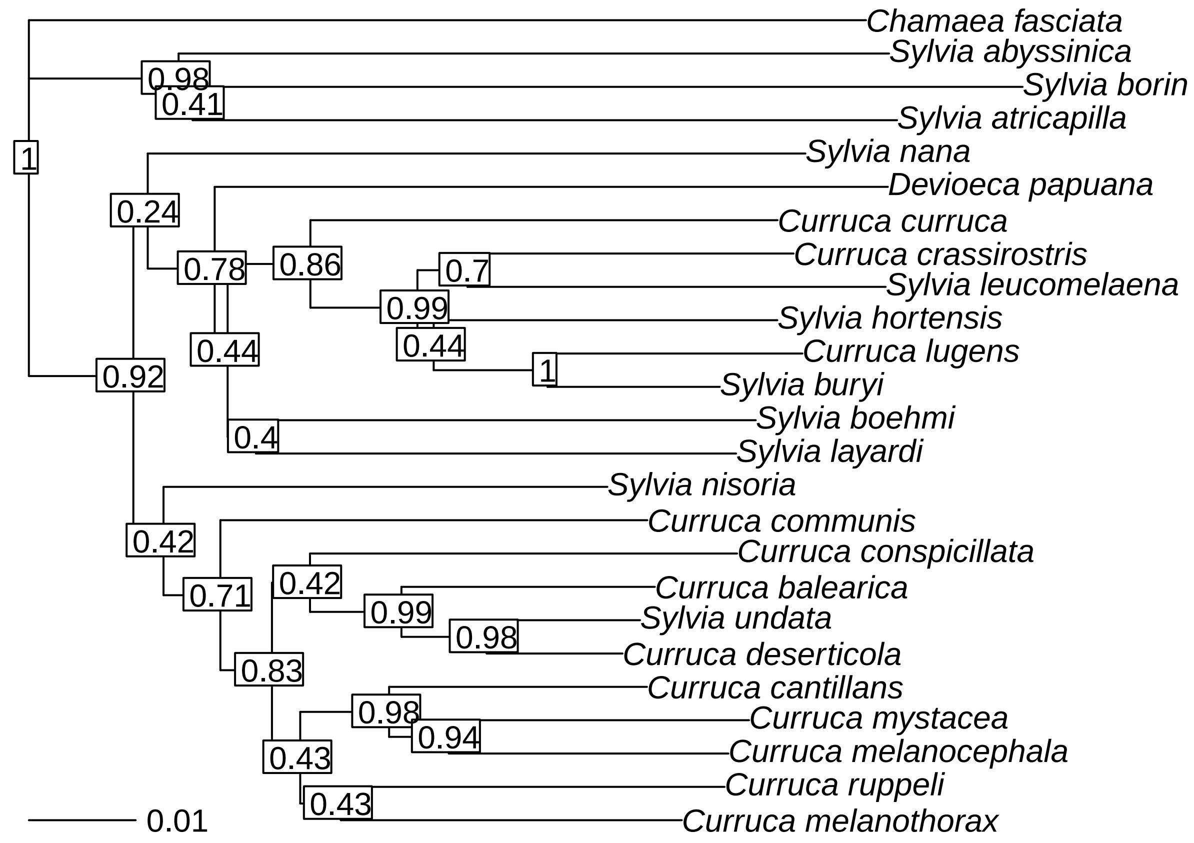
Bootstrap to Estimate significance of tree generating by Neighbor-Joining method.

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

# Estimate tree significance using bootstrap
nj.boot.sylvia <- boot.phylo(tr, sylvia.seq.ali, f, 200,
                             rooted = TRUE)
```

```
## Running bootstraps:       100 / 200Running bootstraps:       200 / 200
## Calculating bootstrap values... done.
```

```
nj.est <- tr #Neighbor Joining estimate
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)
```



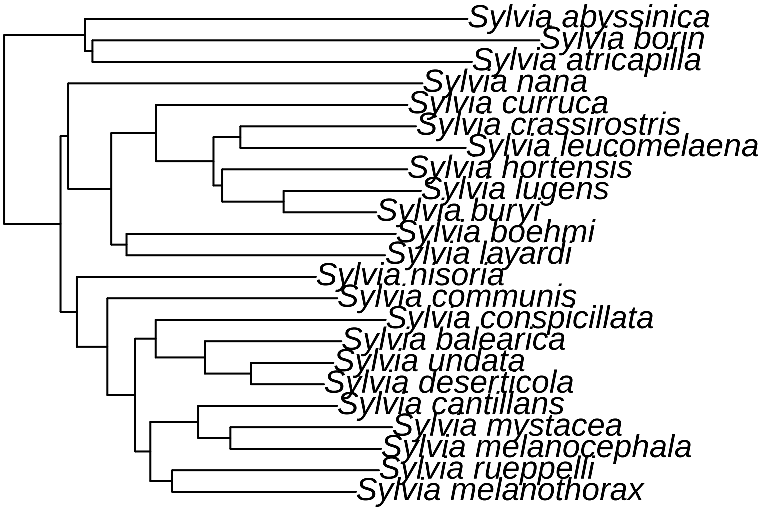
```
#Save the tree
write.tree(nj.est, "sylvia_nj_k80.tre")
```

1.2 Question 1.2 *

Analyze the discrete `(type=discrete)`geographical range variable (`DF$geo.range`) using `ape::ace`. Consider different models (parameter model). Report on the results and interpret the estimated rates and their standard errors.

```r
x <- factor(DF$geo.range)

rooted <- multi2di(tr)
plot(rooted)
```



```r
#Equal Rate (default)
ans <- ace(x, rooted, type = "discrete")
print(ans)
```

```
##
##     Ancestral Character Estimation
##
## Call: ace(x = x, phy = rooted, type = "discrete")
##
```

```
##     Log-likelihood: -20.89321
##
## Rate index matrix:
##          temp temptrop trop
## temp        .        1    1
## temptrop    1        .    1
## trop        1        1    .
##
## Parameter estimates:
##  rate index estimate std-err
##           1   5.1813  1.7578
##
## Scaled likelihoods at the root (type '...$lik.anc' to get them for all nodes):
##       temp   temptrop       trop
## 0.01474506 0.94014800 0.04510693
```

```r
#Symmetrical Model
ans <- ace(x, rooted, type = "discrete", model = "SYM")
```

```
## Warning in sqrt(diag(solve(h))): NaNs produced
```

```r
print(ans)
```

```
##
##     Ancestral Character Estimation
##
## Call: ace(x = x, phy = rooted, type = "discrete", model = "SYM")
##
##     Log-likelihood: -19.52549
##
## Rate index matrix:
##          temp temptrop trop
## temp        .        1    2
## temptrop    1        .    3
## trop        2        3    .
##
## Parameter estimates:
##  rate index estimate std-err
##           1   3.4054  1.8762
##           2   0.0000     NaN
##           3   9.3672  4.1510
##
## Scaled likelihoods at the root (type '...$lik.anc' to get them for all nodes):
##        temp    temptrop        trop
## 0.001748764 0.839713129 0.158538107
```

```r
#All rates Different Model
ans <- ace(x, rooted, type = "discrete", model = "ARD")
```

```
## Warning in sqrt(diag(solve(h))): NaNs produced
```

```r
print(ans)
```

```
##
##     Ancestral Character Estimation
##
## Call: ace(x = x, phy = rooted, type = "discrete", model = "ARD")
##
##     Log-likelihood: -20.07269
##
## Rate index matrix:
##          temp temptrop trop
## temp        .        3    5
## temptrop    1        .    6
## trop        2        4    .
##
## Parameter estimates:
##  rate index estimate std-err
##           1   2.7162  2.0550
##           2   0.0000 15.5967
##           3   0.0000 31.8116
##           4   0.0000     NaN
##           5   0.0000 26.6413
##           6   6.5245  3.9666
##
## Scaled likelihoods at the root (type '...$lik.anc' to get them for all nodes):
##     temp temptrop     trop
##        0        1        0
```

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.

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

data(carni70)

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

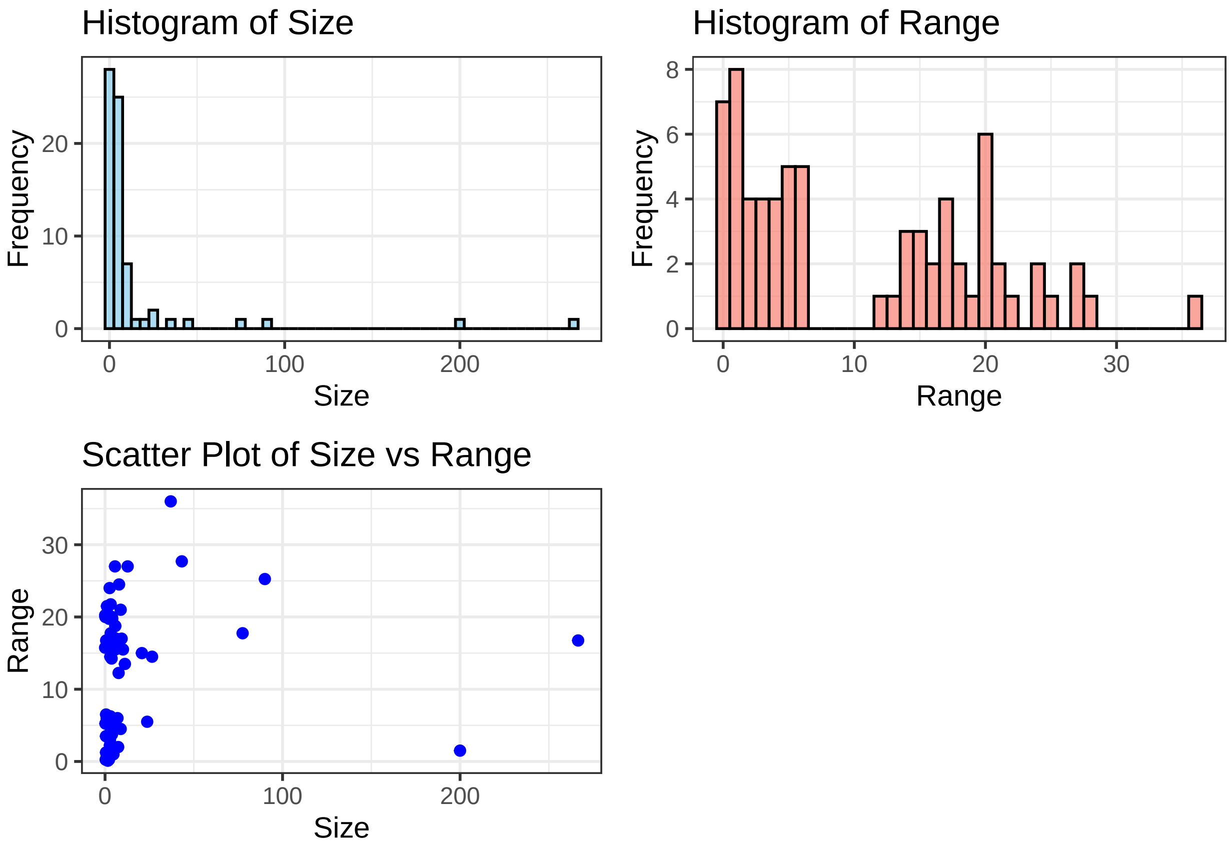
```
##       size             range       
##  Min.   :  0.040   Min.   : 0.120  
##  1st Qu.:  1.282   1st Qu.: 2.062  
##  Median :  3.200   Median : 6.125  
##  Mean   : 14.288   Mean   :10.721  
##  3rd Qu.:  7.293   3rd Qu.:17.750  
##  Max.   :266.500   Max.   :36.000  
```

```r
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*