Lab 2 - Gr. 14 - Bioinformatics (732A93)

Andreas Stasinakis (andst745), Hector Plata (hecpl268), Julius Kittler (julki092), Mim Kemal Tekin (mimte666), Stefano Toffol (steto820)

Assignment 1

Question 1.1

Starting from 33 DNA sequence of various species of casque-headed lizard (Basiliscus basiliscus), other 33 sequences of nucleotides have been generated. The sampling probabilities are the same of the real proportions of the original dataset.
After the artificial DNA has been created, the base frequencies are compared in Table 1. As expected, the observed proportions of the generated data closely resamble the theoretical ones.

Question 1.2

- Created one phylogenetic tree with 33 tips
- For each original DNA sequence of the 33 available, used the function `simSeq(.)` from package `phangorn` to simulate the sequences.
- Result: 33 phylogenetic tree, one for DNA sequence, each with 33 tips.

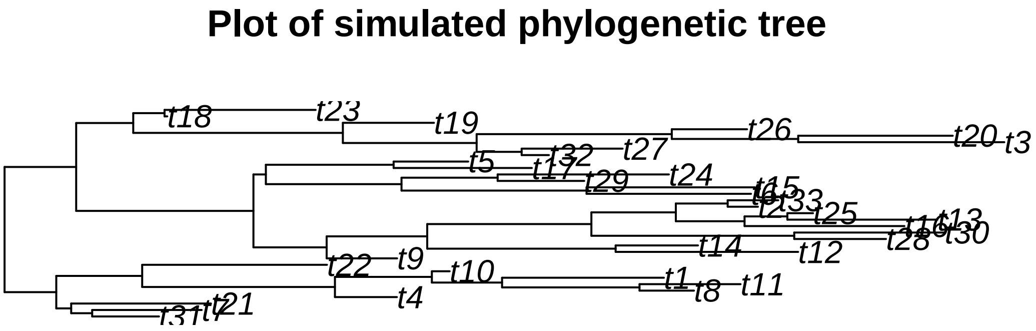

Table 1: Base frequencies of the 33 original and generated DNA sequences.

Base	Original frequency	Simulated frequency
a	0.3121	0.3120
c	0.2052	0.2045
g	0.2307	0.2329
t	0.2519	0.2505

Table 2: Base frequencies of the 33 original DNA sequences and of the 33 simulated phylogenetic trees.

Base	Original frequency	Simulated frequency
a	0.3121	0.3150
c	0.2052	0.2061
g	0.2307	0.2258
t	0.2519	0.2530

Assignment 2

2.1

2.2

2.3

Assignment 3

3.1

3.2

Appendix

```
knitr::opts_chunk$set(fig.width = 7, fig.height = 3, echo = FALSE)

library(dplyr)
library(tidyr)
library(magrittr)
library(ape)          # This is a general R-package for phylogenetics
                      # and comparative methods
library(seqinr)       # This is an specialized package for
                      # nucleotide sequence management
library(phangorn)
library(knitr)

source("732A51_BioinformaticsHT2018_Lab02_GenBankGetCode.R")

# -------------------------------------------------------------------------------
# Question 1.1
# -------------------------------------------------------------------------------

lizards_format_sequences = read.fasta(file = "lizard_seqs.fasta")
# Alternative version of the file. Useful in some ways?

n = length(lizards_accession_numbers)  # Number of sequences to reproduce
p = base.freq(lizards_sequences)  # Probability of the base sequences
simulated_lizards = list()  # Object that will contain our simulated data
```

```r
# The names of the simulated data are the original names + "_sim"
# NOTE: it does not follow the format from GenBank
simulated_names = paste(lizards_accession_numbers, "_sim", sep = "")

set.seed(1535)  # Set seed in order to reproduce the experiment
for(i in 1:n) {  # Cycle through every single object of the lizard_sequences
  len_seq = length(lizards_sequences[[i]])  # Lenght of each sequence
  simulated_lizards[[ simulated_names[i] ]] =
    sample(c("a", "c", "g", "t"), len_seq, replace = T, prob = p)
  # Creating the artificial sequence sampling with probabilities p
  # that are equal to the original ones.
  # NOTE: we use the general distribution for every single sequence
}

# Save as fasta file
write.dna(simulated_lizards, file = "simulated_lizards.fasta", format = "fasta",
          append = F, nbcol = 6, colsep = " ", colw = 10)

# Table with simulated base frequency
df_table = data.frame("Base" = c("a", "c", "g", "t"),
                       "Original\nfrequency" = p,
                       "Simulated\nfrequency" =
                        base.freq(as.DNAbin(simulated_lizards)),
                       row.names = NULL)
kableExtra::kable(df_table, booktabs = T, align = c("r", "l", "l"), digits = c(NA, 4, 4),
      col.names = c("Base", "Original\nfrequency", "Simulated\nfrequency"),
      format = "latex", caption = "Base frequencies of the 33 original and
      generated DNA sequences.")


# ------------------------------------------------------------------------------
# Question 1.2
# ------------------------------------------------------------------------------

# Simulate phylogenetic tree with 33 tips in phylo format (ape) ----------------
set.seed(1)
tree = ape::rtree(n = 33)

# Plot resulting tree ----------------------------------------------------------

plot(tree, edge.width = 1, main = "Plot of simulated phylogenetic tree")
# phytools::plotTree(tree) # Alternative

# Simulate sequences on this tree using phangorn::simSeq() ---------------------
Q = matrix(c(.1, .8, .05, .05,
             .35, .1, .1, .45,
             .3, .2, .2, .3,
             .6, .1, .25, .05), nrow = 4, byrow = TRUE)
rownames(Q) = c("a", "c", "g", "t")
colnames(Q) = c("a", "c", "g", "t")

Original = p
```

```r
tree_sequences_sim = phangorn::simSeq(tree, l = 2000, Q = Q, bf = Original)

# Explanation of parameters:
# l = 2000 because average sequence length in given data is ca. 2000
# bf = Original because this is the vector with the original base proportions
# Q = just chosen the matrix from Special Exercise 1 (Question 3)

# Convert to DNAbin
tree_sequences_sim = as.DNAbin(tree_sequences_sim)

# Save simulated sequences as fasta file ---------------------------------------

# Write simulated lizard sequences as fasta file
ape::write.dna(tree_sequences_sim, file ="lizard_seqs_tree_sim.fasta", format = "fasta",
               append = FALSE, nbcol = 6, colsep = " ", colw = 10)

# Report base composition ------------------------------------------------------

# Table with simulated base frequency
df_table = data.frame("Base" = c("a", "c", "g", "t"),
                       "Original\nfrequency" = Original,
                       "Simulated\nfrequency" = base.freq(tree_sequences_sim),
                       row.names = NULL)

kableExtra::kable(df_table, booktabs = T, align = c("r", "l", "l"), digits = c(NA, 4, 4),
      col.names = c("Base", "Original\nfrequency", "Simulated\nfrequency"),
      format = "latex", caption = "Base frequencies of the 33 original DNA
      sequences and of the 33 simulated phylogenetic trees.")

# ------------------------------------------------------------------------------
# Question 2.1
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 2.2
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 2.2
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 3.1
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Question 3.2
# ------------------------------------------------------------------------------
```