732A51 Bioinformatics

LAB 2 Bioinformatics

Hugo Morvan
William Wiik

STIMA
Department of Computer and Information Science
Linköpings universitet

2024-11-22

Contents

1 Question 1

In this exercise you will perform statistical analysis of three nucleotide data sets. First download the sequences from GenBank and save them in a fasta le. For this use the provided R script, `732A51_BioinformaticsHT2023_Lab02_GenBankGetCode.R`. This is a dataset of the RAG1 gene sequences from 33 lizard species. You are encouraged to read in detail the references in the script as they indicate many useful tools. Explore the dataset using the tools provided by the **ape** and **seqinr** packages. Take note of the lengths of all the sequences and the base composition.

```
## Gene bank accession numbers taken from http://www.jcsantosresearch.org/Class_2014_Spring_Comparative/
lizards_accession_numbers <- c("JF806202", "HM161150", "FJ356743", "JF806205",
                               "JQ073190", "GU457971", "FJ356741", "JF806207",
                               "JF806210", "AY662592", "AY662591", "FJ356748",
                               "JN112660", "AY662594", "JN112661", "HQ876437",
                               "HQ876434", "AY662590", "FJ356740", "JF806214",
                               "JQ073188", "FJ356749", "JQ073189", "JF806216",
                               "AY662598", "JN112653", "JF806204", "FJ356747",
                               "FJ356744", "HQ876440", "JN112651", "JF806215",
                               "JF806209")
lizards_sequences<-ape::read.GenBank(lizards_accession_numbers)
print(lizards_sequences)
```

```
## 33 DNA sequences in binary format stored in a list.
##
## Mean sequence length: 1982.879
##    Shortest sequence: 931
##     Longest sequence: 2920
##
## Labels:
## JF806202
## HM161150
## FJ356743
## JF806205
## JQ073190
## GU457971
## ...
##
## Base composition:
##     a     c     g     t
## 0.312 0.205 0.231 0.252
## (Total: 65.44 kb)
```

```
ape::write.dna(lizards_sequences, file ="lizard_seqs.fasta", format = "fasta",
               append =FALSE, nbcol = 6, colsep = " ", colw = 10)
```

1.1 Question 1.1

Simulate an artificial DNA sequence dataset. It should contain 33 sequence. The lengths of the sequences should be the same as in the lizard dataset, i.e. for each real sequence simulate an artffcial one. The simulation rule is as follows, each nucleotide is to be independently and randomly drawn from the distribution given by the base composition (frequencies) in the true lizard sequences. Save your dataset in a fasta format file. Remember to give unique names to your sequences. Report on the base composition in your simulated data.

Answer:

```
true_base_composition <- c(0.312, 0.205, 0.231, 0.252) # a c g t

# Length of each real sequence
sequence_lengths <- sapply(lizards_sequences, length)

# Simulating DNA
sim_sequences <- lapply(sequence_lengths, function(length) {
   rDNAbin(n = length, base.freq = true_base_composition)[[1]]
})

names(sim_sequences) <- paste0("sim_",names(lizards_sequences))

sim_sequences <- structure(sim_sequences, class = "DNAbin")

ape::write.dna(sim_sequences, file = "sim_lizard_seqs.fasta",
               format = "fasta", append = FALSE, nbcol = 6, colsep = " ", colw = 10)

sim_lizard_seqs <- read.FASTA("sim_lizard_seqs.fasta")
print(sim_lizard_seqs)
```

```
## 33 DNA sequences in binary format stored in a list.
##
## Mean sequence length: 1982.879
##    Shortest sequence: 931
##     Longest sequence: 2920
##
## Labels:
## sim_JF806202
## sim_HM161150
## sim_FJ356743
## sim_JF806205
## sim_JQ073190
## sim_GU457971
## ...
##
## Base composition:
##     a     c     g     t
## 0.312 0.203 0.230 0.255
## (Total: 65.44 kb)
```

a: 31.2%

c: 20.3%

g: 23.0%

t: 25.5%

We obtain almost the same base composition as the true base composition for lizard sequences.

1.2 Question 1.2 *

2 Question 2

2.1 Question 2.1

Report some basic statistics on each sequence dataset: individual base composition, GC content, CG, AT content. Also translate your sequences into protein sequences (see Lab 1) and report on the amino acid composition. In your simulated sequences, how many times did you observe a stop codon inside your sequence? Does this occur in your true sequences? Comment.

```
true_sequences <- read.FASTA("lizard_seqs.fasta")
sim_sequences <- read.FASTA("sim_lizard_seqs.fasta")

calc_seq_stats <- function(all_seq, index_seq) {


 seq_DNAbin  <- structure(all_seq[[index_seq]], class = "DNAbin")
 seq <- as.character(seq_DNAbin) # from raw to character


 # How many g and c that are in the sequence
 GC_content <- ((sum(seq == "g") + sum(seq == "c")) / length(seq))

 # Base composition
 base <- base.freq(seq_DNAbin) %>% round(3)

 # CpG (How many c is followed by a g)
 # seq[-length(seq)] removes the last element
 # seq[-1] removes the first element
 CpG <- sum(seq[-length(seq)] == "c" & seq[-1] == "g") / length(seq)


 # ApT (How many a is followed by a t)
 ApT <- sum(seq[-length(seq)] == "a" & seq[-1] == "t") / length(seq)



 print(paste0("Base composition for ", names(all_seq)[index_seq],
              ": a=",base["a"],", c=",base["c"],", g=",base["g"],", t=",base["t"]))

 print(paste0("GC content for ", names(all_seq)[index_seq],": ",round(GC_content,3)))

 print(paste0("CpG content for ", names(all_seq)[index_seq],": ",round(CpG,3)))

 print(paste0("ApT content for ", names(all_seq)[index_seq],": ",round(ApT,3)))

}
```

```r
# True lizard sequence 1 (JF806202)
calc_seq_stats(all_seq = true_sequences, index_seq = 1)
```

```
## [1] "Base composition for JF806202: a=0.29, c=0.203, g=0.244, t=0.264"
## [1] "GC content for JF806202: 0.446"
## [1] "CpG content for JF806202: 0.013"
## [1] "ApT content for JF806202: 0.077"
```

```r
# Simulate lizard sequence 1 (sim_JF806202)
calc_seq_stats(all_seq = sim_sequences, index_seq = 1)
```

```
## [1] "Base composition for sim_JF806202: a=0.286, c=0.207, g=0.257, t=0.251"
## [1] "GC content for sim_JF806202: 0.464"
## [1] "CpG content for sim_JF806202: 0.045"
## [1] "ApT content for sim_JF806202: 0.07"
```

```r
# ============================================================================
# From chatGPT, how to get frequency every charchater per sequence for an out file
# ============================================================================


process_protein_file <- function(file_path) {
  # Read the file
  lines <- readLines(file_path)

  # Extract headers (lines starting with ">")
  headers <- grep("^>", lines, value = TRUE)

  # Extract sequence lines (lines not starting with ">")
  header_indices <- grep("^>", lines)

  # Create a list of sequences
  sequences <- mapply(function(start, end) {
    paste(lines[(start + 1):(end - 1)], collapse = "")
  }, header_indices, c(header_indices[-1] - 1, length(lines)), SIMPLIFY = TRUE)

  # Remove ">" from headers
  headers <- gsub("^>", "", headers)

  # Create the data frame
  protein_data <- data.frame(
    Header = headers,
    Sequence = sequences,
    stringsAsFactors = FALSE
  )

}
```

```r
# File from Ying Luo
codon_and_ORF <- read.csv("codon_and_ORF.csv")

# https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20241120-223345-0023
# Codon start = Frame = 3
transeq_true <- process_protein_file("transeq_true.out")
chars <- strsplit(transeq_true[1,2], NULL)[[1]] # JF806202
char_table <- prop.table(table(chars))
round(char_table[order(table(chars))],2)
```

```
## chars
##    X    Y    D    H    F    G    E    N    C    I    W    P    A    V    *    R
## 0.00 0.01 0.02 0.02 0.02 0.02 0.03 0.03 0.03 0.04 0.04 0.05 0.05 0.05 0.05 0.05
##    T    Q    K    M    S    L
## 0.05 0.06 0.06 0.07 0.09 0.15
```

```r
# https://www.ebi.ac.uk/jdispatcher/st/emboss_transeq/summary?jobId=emboss_transeq-I20241120-221152-0336
transeq_sim  <- process_protein_file("transeq_sim.out")
chars <- strsplit(transeq_sim[1,2], NULL)[[1]] # seq_1
char_table <- prop.table(table(chars))
round(char_table[order(table(chars))],2)
```

```
## chars
##    F    W    C    Y    M    H    D    N    P    *    E    K    T    I    V    A
## 0.02 0.02 0.02 0.02 0.03 0.03 0.04 0.04 0.04 0.04 0.04 0.05 0.05 0.06 0.06 0.06
##    Q    G    S    R    L
## 0.06 0.07 0.08 0.08 0.09
```

2.2 Question 2.2*

2.3 Question 2.3

Align your sequences using software of your choice (a starter for R: https://stackoverflow.com/questions/4497747/how-to-perform-basic-multiple-sequence-alignments-in-r, you can also look what `Biopython`, `BioPerl` offer, use the `Clustal` family of programs or something else of your choice). Choose a distance measure between sequences, calculate for each alignment the distances between all pairs of sequences. Then, plot heatmaps visualizing the distances. Comment on what you can observe.

```r
if (!requireNamespace("msa", quietly = TRUE)) {
  install.packages("BiocManager")
  BiocManager::install("msa")
}
```

```r
library(seqinr)
library(msa)
```

```
## Loading required package: Biostrings
```

```
## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
##     tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'
```

```
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:seqinr':
##
##     translate

## The following object is masked from 'package:ape':
##
##     complement

## The following object is masked from 'package:base':
##
##     strsplit
```

```
library(Biostrings)

dna_seqs <- readDNAStringSet("lizard_seqs.fasta")
```

```
## Warning in .Call2("fasta_index", filexp_list, nrec, skip, seek.first.rec, :
## reading FASTA file lizard_seqs.fasta: ignored 5451 invalid one-letter sequence
## codes
```

```
cleaned_seqs <- DNAStringSet(gsub("[^ATCG]", "", dna_seqs))
clust_lizard_seqs <- msaClustalW(cleaned_seqs, type = "dna")
```

```
## use default substitution matrix
```

```
dna_seqs_sim <- readDNAStringSet("sim_lizard_seqs.fasta")
```

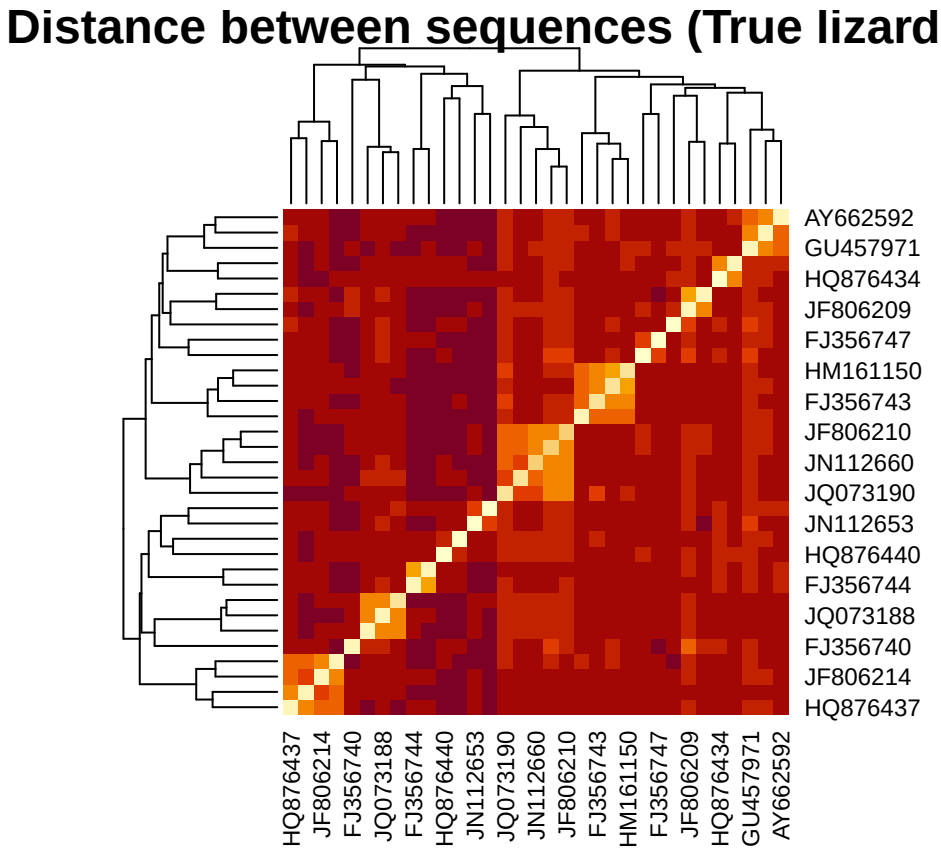
```
## Warning in .Call2("fasta_index", filexp_list, nrec, skip, seek.first.rec, :
## reading FASTA file sim_lizard_seqs.fasta: ignored 5451 invalid one-letter
## sequence codes
```

```
cleaned_seqs_sim <- DNAStringSet(gsub("[^ATCG]", "", dna_seqs_sim))
clust_sim_lizard_seqs <- msaClustalW(cleaned_seqs_sim, type = "dna")
```

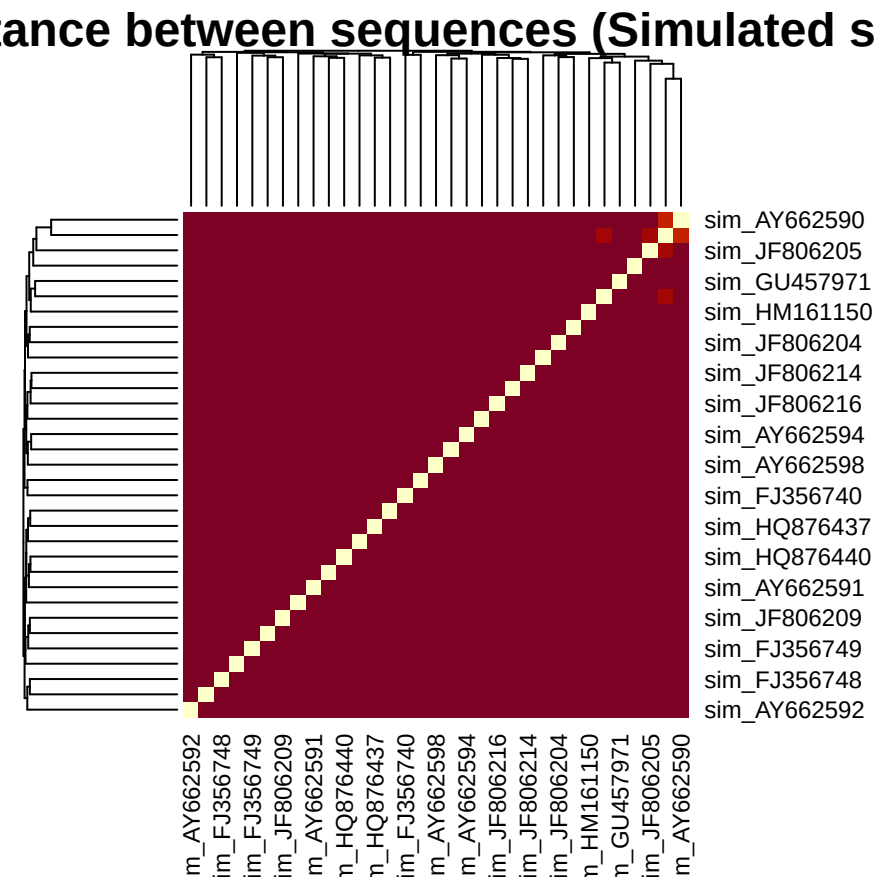
```
## use default substitution matrix
```

```r
# Convert MSA object to formats used in other sequence
# type = "seqinr::alignment" because we want to use dist.alignment() later
converted_lizard_seqs <- msaConvert(clust_lizard_seqs, type = "seqinr::alignment")
converted_sim_lizard_seqs <- msaConvert(clust_sim_lizard_seqs, type = "seqinr::alignment")


# Calculate pairwise distances using dist.alignment() from seqinr package
dist_matrix_lizard_seqs <- as.matrix(dist.alignment(converted_lizard_seqs), matrix = "identity")
dist_matrix_sim_lizard_seqs <- as.matrix(dist.alignment(converted_sim_lizard_seqs), matrix = "identity")
```

```r
heatmap(dist_matrix_lizard_seqs, main = "Distance between sequences (True lizard seq.)")
```




```r
heatmap(dist_matrix_sim_lizard_seqs, main = "Distance between sequences (Simulated seq.)")
```

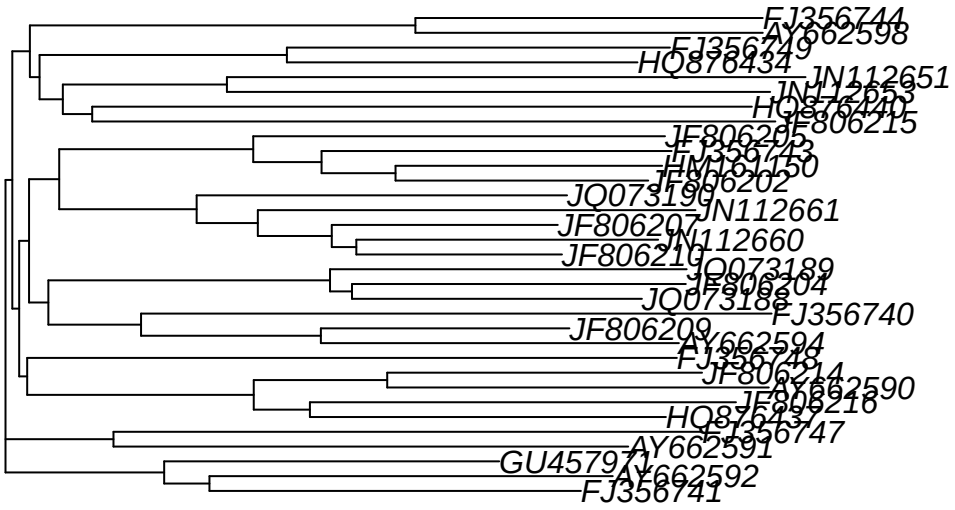


3 Question 3

3.1 Question 3.1

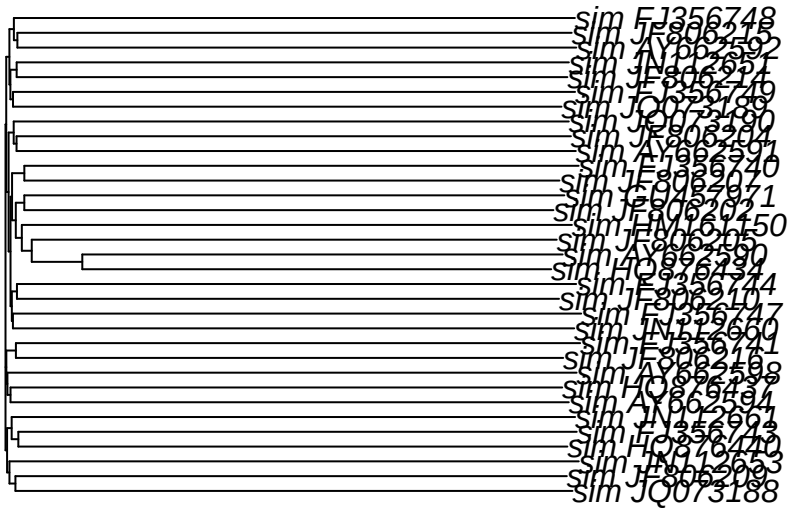
Construct (using algorithm and software of your choice) phylogenetic trees from the three multiple alignments (or distance matrices) done in Question 2.3. You might want to look at the functions offered by `ape`, `phangorn` (https://cran.r-project.org/web/packages/phangorn/vignettes/Trees.pdf) or go for some completely different software. Plot the inferred trees. Are the two based on the simulated data similar to expected? Perform a phylogenetic bootstrap analysis and report the bootstrap support for the individual clades, you can look at `ape::boot.phylo()`.

```
# FASTME Tree Estimation Based on the Minimum Evolution Algorithm
phyl1 <- nj(dist_matrix_lizard_seqs)
plot(phyl1)
```




```
#Other options to create the tree : fastme.bal bionj, mvr(X, V), nj
#print.phylo(phyl1)
```

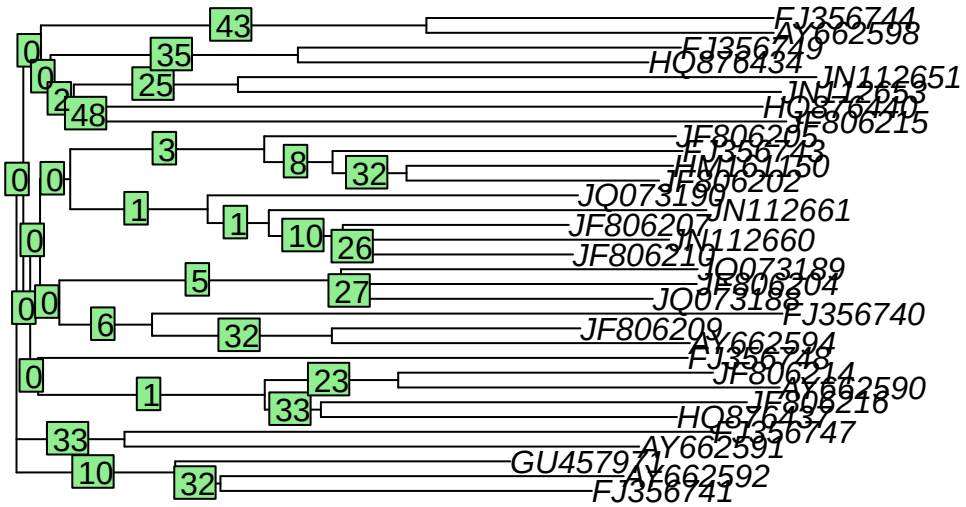
```
# FASTME Tree Estimation Based on the Minimum Evolution Algorithm
phyl2 <-nj(dist_matrix_sim_lizard_seqs)
plot(phyl2)
```




```
#Boostrapping
## from ?boot.phylo:
f <- function(x) nj(x)
tw <- f(dist_matrix_lizard_seqs)
set.seed(1)
## bootstrap with 1000 replications:
(bp <- boot.phylo(tw, dist_matrix_lizard_seqs, f, B=1000, quiet=TRUE))
```

```
##  [1] NA  0  0  0  0  0  0  0  0  2 48 33  6  1 10 32 25  1  3  1 10 26 35  8 33
## [26] 32  5 27 32 23 43
```

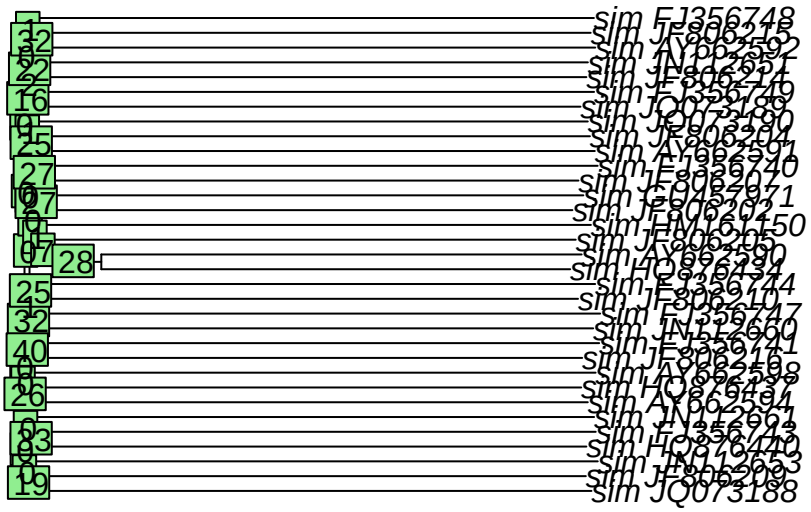
```
## more readable but the tree is really unrooted:
plot(tw)
drawSupportOnEdges(bp)
```




```
#Boostrapping
## from ?boot.phylo:
f <- function(x) nj(x)
tw <- f(dist_matrix_sim_lizard_seqs)
set.seed(1)
## bootstrap with 100 replications:
(bp <- boot.phylo(tw, dist_matrix_sim_lizard_seqs, f, B=1000, quiet=TRUE))
```

```
##  [1] NA  0  0  0  0  0  0  2  0  0  1  0 32 26 16  1  0  1  0 25 25 22 19 32 40
## [26]  1 33 27 27  7 28
```

```
plot(tw)
drawSupportOnEdges(bp)
```



Interpretation Scale:

70-95%: Moderately supported branch 95-100%: Highly supported branch <70%: Weak support, less confidence in that branching pattern

For both true and simulated trees, none of the branches have significant suppport (>70%, i.e. >700) Tree building functions tested : fastme.bal, bionj, nj.

3.2 Question 3.2*

Compare your inferred trees and also your simulated one. Apart from visualizing the trees one may calculate various indices related to them and distances between the trees. Explore what indices and metrics the `ape`, `distory`, `phangorn`, `phyloTop`, `TotalCopheneticIndex` or `treespace` R packages offer, choose some and report the results in a meaningful way. You might have to save your tree to drive and then read in it using e.g. ape's tree reading functionality.

```
ape::comparePhylo(phyl1, phyl2)
```

```
## => Comparing phyl1 with phyl2.
## Both trees have the same number of tips: 33.
## Tips in phyl1 not in phyl2 : JF806210, JN112660, JN112661, JF806202, HM161150, FJ356743, JF806205, JQ
```

```
## Tips in phyl2 not in phyl1 : sim_HQ876434, sim_AY662590, sim_JF806205, sim_JF806202, sim_GU457971, si
## Both trees have the same number of nodes: 31.
## Both trees are unrooted.
## Both trees are not ultrametric.
```

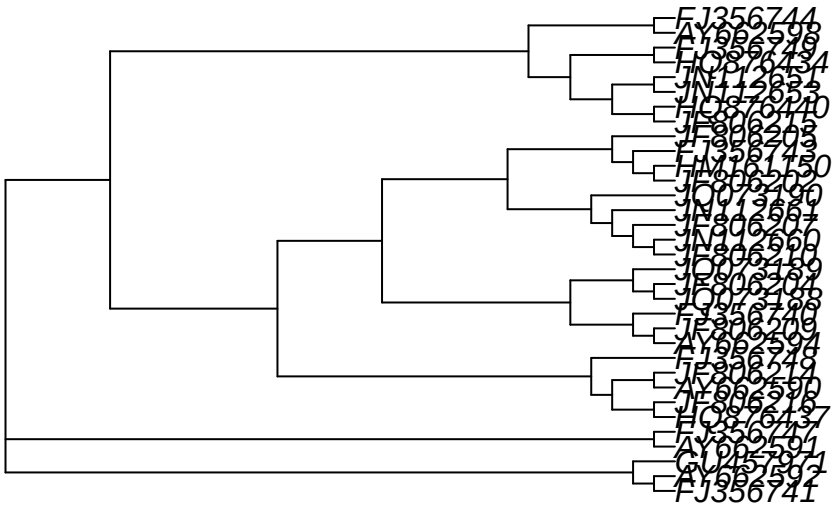
```
#Metrics :
branching.times(phyl1)
```

```
##         34         35         36         37         38         39         40
## 0.15316931 0.15179162 0.15040138 0.14878227 0.14840381 0.14823928 0.14627806
##         41         42         43         44         45         46         47
## 0.14449828 0.14230621 0.14157215 0.13562936 0.13132725 0.12575191 0.11449990
##         48         49         50         51         52         53         54
## 0.12107332 0.11191276 0.10837965 0.10212980 0.10305594 0.10295326 0.08716938
##         55         56         57         58         59         60         61
## 0.08220290 0.09624948 0.08923640 0.09159566 0.08937464 0.08752518 0.08308446
##         62         63         64
## 0.07427249 0.07595380 0.07025427
```

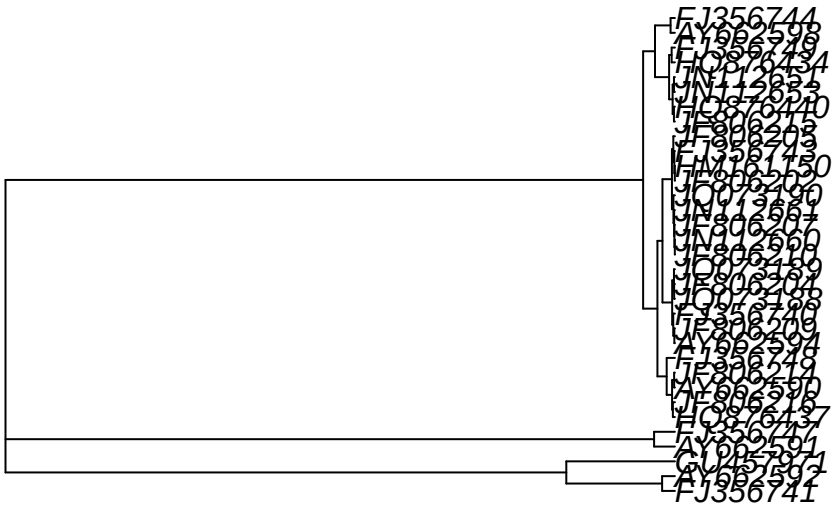
```
branching.times(phyl2)
```

```
##        34        35        36        37        38        39        40        41
## 0.4275893 0.4270439 0.4269598 0.4261432 0.4263924 0.4252625 0.4246524 0.4239261
##        42        43        44        45        46        47        48        49
## 0.4246124 0.4238175 0.4228314 0.4226558 0.4220246 0.4238257 0.4219514 0.4214440
##        50        51        52        53        54        55        56        57
## 0.4198596 0.4212946 0.4229515 0.4189981 0.4192971 0.4192182 0.4201420 0.4186683
##        58        59        60        61        62        63        64
## 0.4197654 0.4153275 0.4179461 0.4134634 0.4134462 0.4078837 0.3699204
```

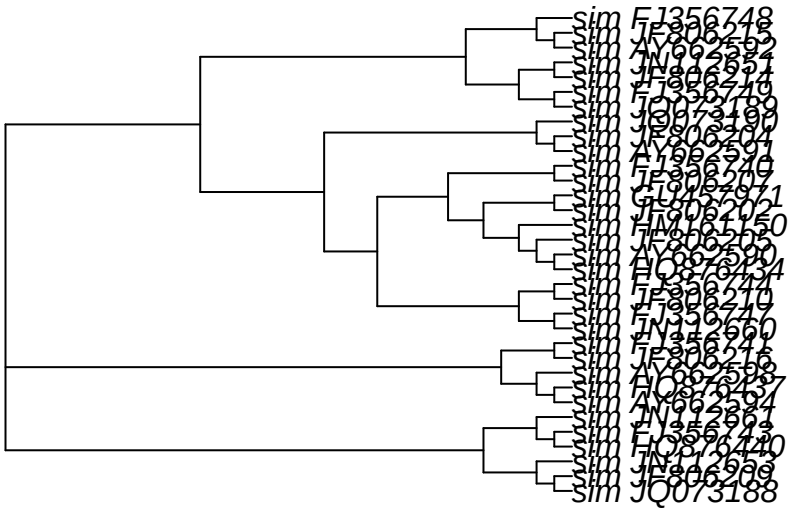
```
#compar.cheverud          Cheverud's Comparative Method
#compar.gee               Comparative Analysis with GEEs
#compar.lynch             Lynch's Comparative Method
#compar.ou                Ornstein-Uhlenbeck Model for Continuous Characters

a = compute.brlen(phyl1)           #Branch Lengths Computation
plot(a)
```



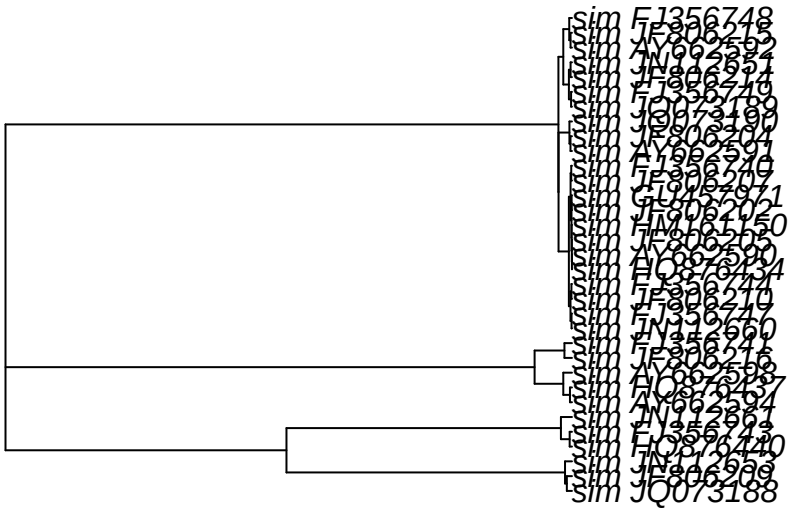

```
b = compute.brtime(phyl1)          #Compute and Set Branching Times
plot(b)
```



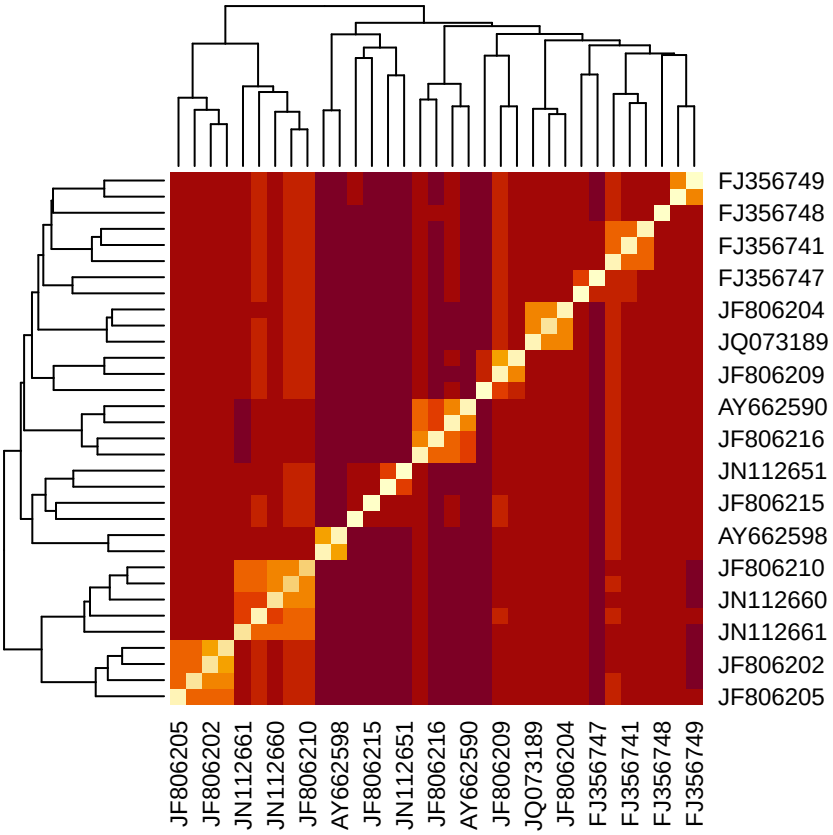

```
a = compute.brlen(phyl2)           #Branch Lengths Computation
plot(a)
```



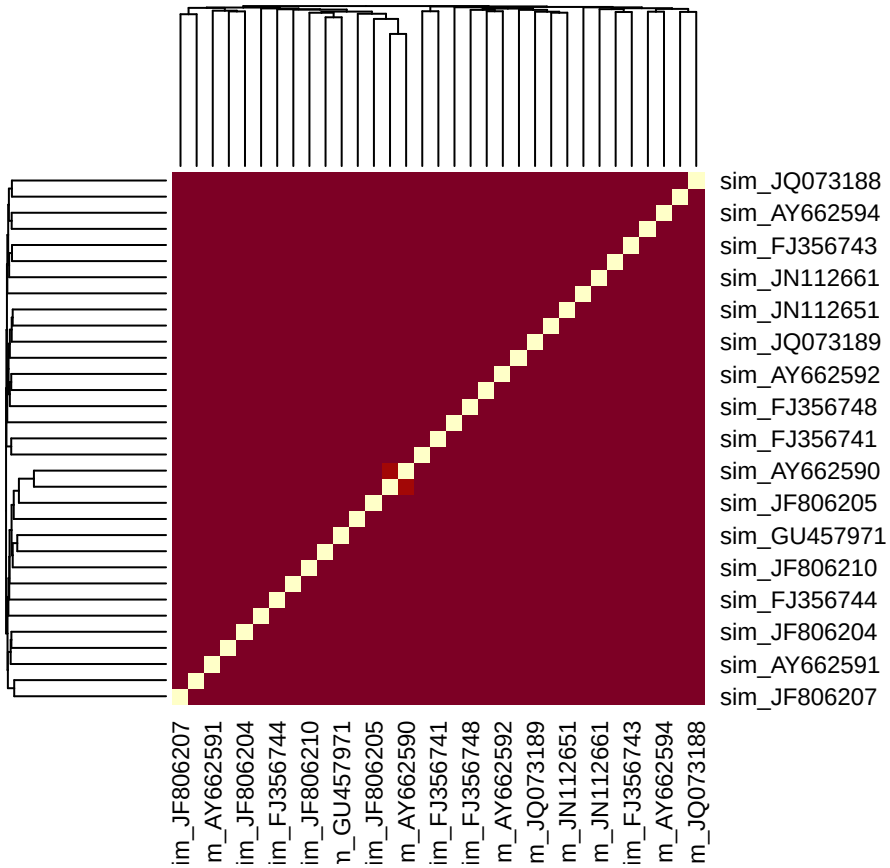

```
b = compute.brtime(phyl2)          #Compute and Set Branching Times
plot(b)
```




```
heatmap(cophenetic.phylo(phyl1))
```




```
heatmap(cophenetic.phylo(phyl2))
```




```
# corBlomberg              #Blomberg et al.'s Correlation Structure
# corBrownian              #Brownian Correlation Structure
# corClasses               #Phylogenetic Correlation Structures
# corGrafen                #Grafen's (1989) Correlation Structure
# corMartins               #Martins's (1997) Correlation Structure
# corPagel                 #Pagel's "lambda" Correlation Structure
# corphylo                 #Correlations among Multiple Traits with Phylogenetic Signal
# correlogram.formula      #Phylogenetic Correlogram
#
# dist.topo(phyl1, phyl2) #needs similar leaf names
#
# gammaStat                #Gamma-Statistic of Pybus and Harvey, requires ultrametric trees
#
# root                     #Roots Phylogenetic Trees

summary.phylo(phyl1)
```

```
##
## Phylogenetic tree: phyl1
##
##   Number of tips: 33
```

```
##   Number of nodes: 31
##   Branch lengths:
##     mean: 0.05191795
##     variance: 0.001511728
##     distribution summary:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## 0.001377696 0.014389981 0.051113051 0.075027992 0.138188147
##   No root edge.
##   First ten tip labels: JF806210
##                         JN112660
##                         JN112661
##                         JF806202
##                         HM161150
##                         FJ356743
##                         JF806205
##                         JQ073190
##                         JF806207
##                         JQ073188
##   No node labels.
```

```
summary.phylo(phyl2)
```

```
##
## Phylogenetic tree: phyl2
##
##   Number of tips: 33
##   Number of nodes: 31
##   Branch lengths:
##     mean: 0.2178854
##     variance: 0.04231989
##     distribution summary:
##         Min.      1st Qu.       Median      3rd Qu.         Max.
## 0.0005453872 0.0027112159 0.3607140035 0.4162849315 0.4275080112
##   No root edge.
##   First ten tip labels: sim_HQ876434
##                         sim_AY662590
##                         sim_JF806205
##                         sim_JF806202
##                         sim_GU457971
##                         sim_HM161150
##                         sim_JQ073189
##                         sim_JQ073190
##                         sim_JF806207
##                         sim_JN112660
##   No node labels.
```

True data phylogenetic tree Branch lengths stats:

mean: 0.05191795 , variance: 0.001511728

Min.: 0.001377696

1st Qu.: 0.014389981

Median: 0.051113051

3rd Qu.: 0.075027992

Max.: 0.138188147

Simulated data phylogenetic tree Branch lengths stats:

mean: 0.2178854 , variance: 0.04231989

Min.: 0.0005453872

1st Qu.: 0.0027112159

Median: 0.3607140035

3rd Qu. : 0.4162849315

Max.: 0.4275080112

The simulated data phylogeneteic tree branch length has on average much longer branch lengths (0.05 vs 0.22), but with also much more variance (0.0015 vs 0.0423). This these results correlate with the obeservation from the heatmaps which shows that in the simulated data, no clade is close to any other, which makes sense since the data was randomly generated and should be i.i.d.