Bioinformatics Lab2

Milda Poceviciute, Fanny Karelius, Rab Nawaz Jan Sher and Saman Zahid

24 November 2018

Question 1: DNA sequence acquisition and simulation

In this question an analysis of 3 nucleotide sequences is performed. One sequence is downloaded from GenBank database, and saved into fatsa file. Here we load thev data set:

```
lizards_sequences<-ape::read.FASTA("lizard_seqs.fasta")
```

Q1.1

In this part we simulate DNA sequences with the same base composition and sequences length as the one from the GenBank:

```
len_list <- lengths(lizards_sequences)
new_seq_list <- list()
new_seq <- c()
for (i in 1:length(lizards_sequences)){
    # extract the base composition of each sequence
    freqx <- ape::base.freq(lizards_sequences[i])
    new_seq <- sample(c("a","c","g","t"),len_list[i],rep=TRUE,prob=freqx)
    new_seq_list[i] <- list(new_seq)
}

# convert result to the DNAbin class
simulated_seq1 <- as.DNAbin(new_seq_list)
names(simulated_seq1) <- names(lizards_sequences)
#ape::write.dna(simulated_seq1, file ="sim1_seq.fasta", format = "fasta", colsep = "")
```

```
## Base composition of the data from GenBank:
```

```
##         a         c         g         t
## 0.3121454 0.2052325 0.2307222 0.2518999
```

```
## Base composition of the simulated sequences:
```

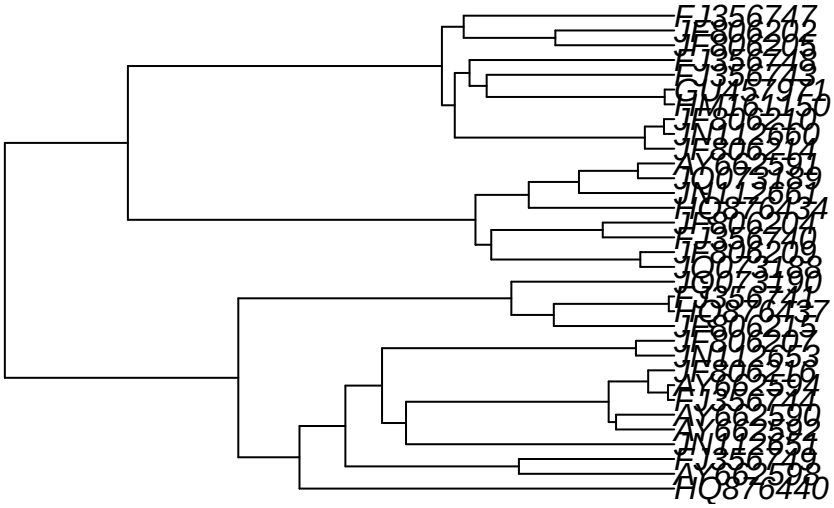
```
##         a         c         g         t
## 0.3088867 0.2068159 0.2320471 0.2522503
```

We can see that the original data, and the simulated data have quite similar base compositions.

Q1.2

In this part we simulate another set of DNA data with similar characteristics as the original one. Now we use a simulated tree with 33 nodes from the original data and package phangorn to do this:

```
# Simulate a tree with 33 nodes
tree<- TreeSim::sim.bd.taxa(n=33, numbsim=1, lambda=1, mu=0)[[1]]
tree$tip.label <- names(lizards_sequences)
plot(tree, yaxt = "n")
```



We need to create a transition matrix, so we use the full dataset to compute that:

```
data <- c()
simulated_seq_tree <- list()

DNA_list <- unlist(as.character(lizards_sequences),use.names = FALSE)
DNA_unique <- c("a","c","g","t")
matrix <-  matrix(0, ncol = length(DNA_unique),
                  nrow = length(DNA_unique))

for (i in 1:(length(DNA_list) - 1)) {
    index_of_i <- DNA_unique == DNA_list[i]
    index_of_i_plus_1 <- DNA_unique == DNA_list[i + 1]
    matrix[index_of_i, index_of_i_plus_1] = matrix[index_of_i, index_of_i_plus_1] + 1
}

# Transition matrix is finally computed
Q <- matrix / rowSums(matrix)
#Q_lower <- Q[lower.tri(Q)]
# Other parameters that are needed for phangorn package function
bf <-  ape::base.freq(lizards_sequences)
size <- round(mean(len_list))
```

```
# Generating the data based on the tree

simulated_seq_tree <-  phangorn::simSeq(
        tree,
        l = size,
        type = "DNA",
        bf = bf,
        Q = Q)

# change numbers into letters
simulated_seq_tree2 <- lapply(simulated_seq_tree,function(seq){
    #seq = a
    states = c("a","c","g","t")
    seq[which(seq==1)] <- states[1]
    seq[which(seq==2)] <- states[2]
```

```
    seq[which(seq==3)] <- states[3]
    seq[which(seq==4)] <- states[4]
    return(seq)
})

simulated_seq2 <- as.DNAbin(simulated_seq_tree2)
# Write the simulated sequence to a fasta file
ape::write.dna(simulated_seq2, file ="sim2_seqs.fasta", format = "fasta", colsep = "")
```

```
## Base composition of the data from GenBank:
```

```
##         a         c         g         t
## 0.3121454 0.2052325 0.2307222 0.2518999
```

```
## Base composition of the simulated sequences Q1.1:
```

```
##         a         c         g         t
## 0.3088867 0.2068159 0.2320471 0.2522503
```

```
## Base composition of the simulated sequences Q1.2:
```

```
##         a         c         g         t
## 0.3134981 0.2055349 0.2306881 0.2502789
```

The base compositions of all simulated sequences are quite similar to the original one.

Question 2: Sequence analysis

Q2.1

```
## Individual base composition of sequence 1
## The data from the GenBank:
##         a         c         g         t
## 0.2898696 0.2026078 0.2437312 0.2637914
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2935872 0.2014028 0.2484970 0.2565130
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3076147 0.2072617 0.2279375 0.2571861
##
## Individual base composition of sequence 2
## The data from the GenBank:
##         a         c         g         t
## 0.3115541 0.2122554 0.2314507 0.2447398
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3093392 0.2056109 0.2296050 0.2554448
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3146747 0.1956631 0.2360061 0.2536561
##
## Individual base composition of sequence 3
## The data from the GenBank:
##         a         c         g         t
```

```
## 0.3130405 0.2099620 0.2348668 0.2421308
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3092356 0.2276029 0.2248357 0.2383258
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3227433 0.1991931 0.2264246 0.2516389
##
## Individual base composition of sequence 4
## The data from the GenBank:
##         a         c         g         t
## 0.2842942 0.2097416 0.2445328 0.2614314
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2666006 0.2150644 0.2646184 0.2537166
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3192133 0.2012103 0.2380232 0.2415532
##
## Individual base composition of sequence 5
## The data from the GenBank:
##         a         c         g         t
## 0.3059701 0.1987788 0.2360923 0.2591588
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3120760 0.1987788 0.2360923 0.2530529
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3292990 0.2032274 0.2284418 0.2390318
##
## Individual base composition of sequence 6
## The data from the GenBank:
##         a         c         g         t
## 0.2988084 0.2071494 0.2364803 0.2575619
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2997250 0.1979835 0.2355637 0.2667278
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3166919 0.1971760 0.2334846 0.2526475
##
## Individual base composition of sequence 7
## The data from the GenBank:
##         a         c         g         t
## 0.3137323 0.2068488 0.2338291 0.2455898
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3355240 0.2096160 0.2272570 0.2276029
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3015633 0.2153303 0.2254160 0.2576904
##
## Individual base composition of sequence 8
## The data from the GenBank:
```

```
##         a         c         g         t
## 0.2968127 0.2001992 0.2340637 0.2689243
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2888446 0.1982072 0.2390438 0.2739044
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3237519 0.2092789 0.2324760 0.2344932
##
## Individual base composition of sequence 9
## The data from the GenBank:
##         a         c         g         t
## 0.2998084 0.1992337 0.2346743 0.2662835
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2777778 0.2097701 0.2356322 0.2768199
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3025719 0.2027231 0.2360061 0.2586989
##
## Individual base composition of sequence 10
## The data from the GenBank:
##         a         c         g         t
## 0.3157534 0.2071918 0.2304795 0.2465753
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3102740 0.2123288 0.2253425 0.2520548
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3192133 0.2188603 0.2183560 0.2435703
##
## Individual base composition of sequence 11
## The data from the GenBank:
##         a         c         g         t
## 0.3121449 0.2041903 0.2325994 0.2510653
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3117898 0.1946023 0.2418324 0.2517756
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3192133 0.2017146 0.2400403 0.2390318
##
## Individual base composition of sequence 12
## The data from the GenBank:
##         a         c         g         t
## 0.3177408 0.2061677 0.2293832 0.2467082
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3240069 0.2169257 0.2148532 0.2442142
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3166919 0.1991931 0.2395361 0.2445789
##
## Individual base composition of sequence 13
```

```
## The data from the GenBank:
##         a         c         g         t
## 0.3187773 0.2034207 0.2241630 0.2536390
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3227802 0.2056041 0.2168850 0.2547307
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3091276 0.2012103 0.2289460 0.2607161
##
## Individual base composition of sequence 14
## The data from the GenBank:
##         a         c         g         t
## 0.3146384 0.2035273 0.2320988 0.2497354
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3006697 0.2040888 0.2368699 0.2583715
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3171962 0.2022189 0.2193646 0.2612204
##
## Individual base composition of sequence 15
## The data from the GenBank:
##         a         c         g         t
## 0.3196064 0.2033528 0.2255831 0.2514577
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3147744 0.1990539 0.2267103 0.2594614
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3106404 0.2037317 0.2501261 0.2355018
##
## Individual base composition of sequence 16
## The data from the GenBank:
##         a         c         g         t
## 0.2921236 0.2123629 0.2382851 0.2572283
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2908367 0.2171315 0.2300797 0.2619522
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3005547 0.2158346 0.2213817 0.2622289
##
## Individual base composition of sequence 17
## The data from the GenBank:
##         a         c         g         t
## 0.2902903 0.2052052 0.2412412 0.2632633
## The simulated data in Q1.1:
##     a     c     g     t
## 0.285 0.185 0.257 0.273
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3055976 0.2042360 0.2344932 0.2556732
##
```

```
## Individual base composition of sequence 18
## The data from the GenBank:
##         a         c         g         t
## 0.3203540 0.2042478 0.2269027 0.2484956
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3090265 0.2084956 0.2424779 0.2400000
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3171962 0.2143217 0.2148260 0.2536561
##
## Individual base composition of sequence 19
## The data from the GenBank:
##         a         c         g         t
## 0.3170475 0.2003515 0.2274165 0.2551845
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3151983 0.2060372 0.2316602 0.2471042
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3086233 0.2057489 0.2304589 0.2551689
##
## Individual base composition of sequence 20
## The data from the GenBank:
##         a         c         g         t
## 0.2946429 0.2043651 0.2341270 0.2668651
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3025794 0.1805556 0.2321429 0.2847222
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3081190 0.2062532 0.2198689 0.2657590
##
## Individual base composition of sequence 21
## The data from the GenBank:
##         a         c         g         t
## 0.3172554 0.1956522 0.2248641 0.2622283
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3023098 0.2044837 0.2228261 0.2703804
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3207262 0.2062532 0.2163389 0.2566818
##
## Individual base composition of sequence 22
## The data from the GenBank:
##         a         c         g         t
## 0.3219225 0.2074689 0.2285615 0.2420470
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3193916 0.2174214 0.2305565 0.2326305
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3040847 0.2047403 0.2435703 0.2476046
```

```
##
## Individual base composition of sequence 23
## The data from the GenBank:
##         a         c         g         t
## 0.3144044 0.1966759 0.2250693 0.2638504
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2978576 0.1955771 0.2349689 0.2715964
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3212305 0.2012103 0.2445789 0.2329803
##
## Individual base composition of sequence 24
## The data from the GenBank:
##     a     c     g     t
## 0.296 0.214 0.237 0.253
## The simulated data in Q1.1:
##     a     c     g     t
## 0.292 0.221 0.246 0.241
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3156833 0.1991931 0.2254160 0.2597075
##
## Individual base composition of sequence 25
## The data from the GenBank:
##         a         c         g         t
## 0.3175890 0.2054988 0.2284103 0.2485019
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3225238 0.2125485 0.2305252 0.2344025
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3292990 0.2087746 0.2178517 0.2440746
##
## Individual base composition of sequence 26
## The data from the GenBank:
##         a         c         g         t
## 0.3144453 0.2110934 0.2306464 0.2438148
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3168396 0.2210694 0.2238627 0.2382283
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3282905 0.2092789 0.2294503 0.2329803
##
## Individual base composition of sequence 27
## The data from the GenBank:
##         a         c         g         t
## 0.3094527 0.2059701 0.2218905 0.2626866
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3124378 0.2029851 0.2179104 0.2666667
## The simulated data in Q1.2:
##         a         c         g         t
```

```
## 0.3202219 0.2087746 0.2324760 0.2385275
##
## Individual base composition of sequence 28
## The data from the GenBank:
##         a         c         g         t
## 0.3160083 0.2065142 0.2286902 0.2487872
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3037011 0.1975095 0.2476652 0.2511242
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3111447 0.2002017 0.2390318 0.2496218
##
## Individual base composition of sequence 29
## The data from the GenBank:
##         a         c         g         t
## 0.3207612 0.2020761 0.2280277 0.2491349
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3030093 0.2030439 0.2348668 0.2590799
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3141704 0.1981846 0.2239032 0.2637418
##
## Individual base composition of sequence 30
## The data from the GenBank:
##         a         c         g         t
## 0.2943396 0.2037736 0.2377358 0.2641509
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2886792 0.1981132 0.2481132 0.2650943
## The simulated data in Q1.2:
##         a         c         g         t
## 0.2889561 0.2072617 0.2435703 0.2602118
##
## Individual base composition of sequence 31
## The data from the GenBank:
##         a         c         g         t
## 0.3200585 0.2071611 0.2232371 0.2495433
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3075237 0.2096421 0.2209642 0.2618700
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3086233 0.2022189 0.2410489 0.2481089
##
## Individual base composition of sequence 32
## The data from the GenBank:
##         a         c         g         t
## 0.2961117 0.2033898 0.2352941 0.2652044
## The simulated data in Q1.1:
##         a         c         g         t
## 0.3080758 0.2003988 0.2233300 0.2681954
## The simulated data in Q1.2:
```

```
##         a         c         g         t
## 0.2985376 0.2269289 0.2269289 0.2476046
##
## Individual base composition of sequence 33
## The data from the GenBank:
##         a         c         g         t
## 0.2932331 0.2030075 0.2406015 0.2631579
## The simulated data in Q1.1:
##         a         c         g         t
## 0.2953813 0.1944146 0.2448980 0.2653061
## The simulated data in Q1.2:
##         a         c         g         t
## 0.3141704 0.2052446 0.2269289 0.2536561
```

The CG contents are:

```
## In the data from the GenBank:
```

```
## [1] 0.4359547
```

```
## In the simulated data in Q1.1:
```

```
## [1] 0.438863
```

```
## In the simulated data in Q1.2:
```

```
## [1] 0.436223
```

A,C,G,T contents are:

```
# A,C,G,T content
cat("In the data from the GenBank:")
```

```
## In the data from the GenBank:
```

```
cat("\n")
```

```
print(ape::base.freq(lizards_sequences))
```

```
##         a         c         g         t
## 0.3121454 0.2052325 0.2307222 0.2518999
```

```
cat("In the simulated data in Q1.1:")
```

```
## In the simulated data in Q1.1:
```

```
cat("\n")
```

```
print(ape::base.freq(simulated_seq1))
```

```
##         a         c         g         t
## 0.3088867 0.2068159 0.2320471 0.2522503
```

```
cat("In the simulated data in Q1.2:")
```

```
## In the simulated data in Q1.2:
```

```
cat("\n")
```

```
print(ape::base.freq(simulated_seq2))
```

```
##         a         c         g         t
## 0.3134981 0.2055349 0.2306881 0.2502789
```

The nucleotide sequences were transformed into amino acid sequences using the EMBOSS Transeq function, and saved to a fasta file:

```
amino_lizard_seq <- read.fasta("lizard_amino_seqs.fasta")
amino_sim1_seq <- read.fasta("sim1_amino_seq.fasta")
amino_sim2_seq <- read.fasta("sim2_amino_seq.fasta")


amino_rev_com_lizard_seq <- read.fasta("rev_com_lizard_amino_acid.fasta")
amino_rev_com_sim1_seq <- read.fasta("rev_com_sim1_amino_acid.fasta")
amino_rev_com_sim2_seq <- read.fasta("rev_com_sim2_amino_acid.fasta")
```

```
stop_code_count <- function(amino_acid_seq){

  stop_count = c()
  for (i in 1:length(amino_acid_seq)) {

    sequence = amino_acid_seq[[i]]
    stop <- sequence[which(sequence == "*")]
    stop_count[i] = length(stop)
  }
  return(stop_count)
}
df = data.frame(lizard_seq = stop_code_count(amino_lizard_seq),
           "sim1" = stop_code_count(amino_sim1_seq),
           "sim2" = stop_code_count(amino_sim2_seq),
           "rev_comp_lizard_seq" = stop_code_count(amino_rev_com_lizard_seq),
           "rev_comp_sim1_seq" = stop_code_count(amino_rev_com_sim1_seq),
           "rev_comp_sim2_seq" = stop_code_count(amino_rev_com_sim2_seq)
          )

knitr::kable(df)
```

lizard_seq	sim1	sim2	rev_comp_lizard_seq	rev_comp_sim1_seq	rev_comp_sim2_seq
16	15	39	9	11	38
0	38	36	27	62	33
78	55	44	45	43	29
33	25	43	8	21	36
0	30	48	28	26	30
38	28	42	19	22	32
79	54	41	48	52	33
38	26	38	17	19	36
41	25	52	10	7	34
78	69	36	31	48	43
0	55	37	33	57	36
81	53	49	34	51	33
0	61	37	29	38	26
82	47	38	51	56	41
0	64	46	30	47	35
0	21	35	11	16	23
19	23	38	13	17	37
85	60	44	49	54	33
87	45	42	51	45	43
37	19	33	12	19	40

lizard_seq	sim1	sim2	rev_comp_lizard_seq	rev_comp_sim1_seq	rev_comp_sim2_seq
0	41	32	18	36	37
75	58	49	38	46	38
0	29	36	29	23	33
18	19	32	10	11	39
75	50	43	46	38	44
0	66	37	37	44	29
0	19	38	10	18	33
81	68	42	49	57	32
76	56	43	45	51	30
38	21	54	10	20	43
61	48	36	30	44	27
17	15	46	13	19	33
36	16	32	7	15	39

For some sequences the stop codons are more in true in sequence than in simulated sequences but there are some sequences for which the stop codon does not appear at all in the original sequence but it appears many times in the simulated sequences. It could be possible because the nucleotides forming the stop codon might have been drawn more in simulated sequences.

Q2.2

In this part we do markov chain analysis of the data sets. We combine all the sequences from each dataset to seperate lists, and use the *markovchainFit* function with bootstrapping to estimate the Markov models. This function fits the first order markov model to all three datasets.

```
lizard <- as.character(lizards_sequences)
sim1 <- as.character(simulated_seq1)
sim2 <- as.character(simulated_seq2)


flat_lizard <- unlist(lizard,use.names = FALSE)
mc_lizard <-  markovchainFit(flat_lizard,method = "bootstrap", nboot = 3,
                             name = "Bootstrap Mc",
                             parallel = TRUE)

# the original data contains some "weird nucleotides" - what to do with them?
which(flat_lizard == "y")
```

```
##  [1]   269  6787 21562 21706 21859 22675 23419 28171 30281 31547 33732
## [12] 37937 45594 56259 58504
```

```
which(flat_lizard == "m")
```

```
## [1] 21470 38532 45235
```

```
which(flat_lizard == "r")
```

```
##  [1]  6653  6836 21395 21582 23621 26779 30533 32015 32784 36968 37225
## [12] 44900 46386 54146 54560 56322 61914
```

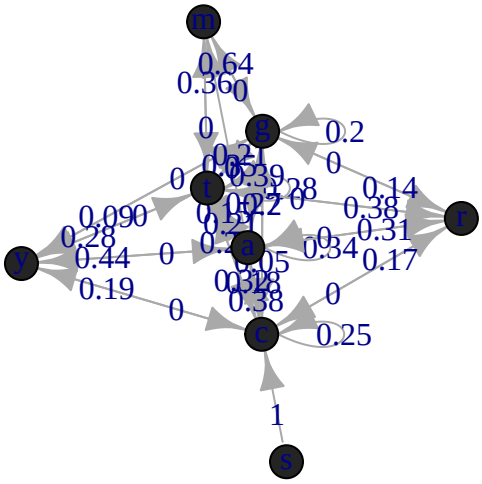
```
which(flat_lizard == "s")
```

```
## [1] 54451
```

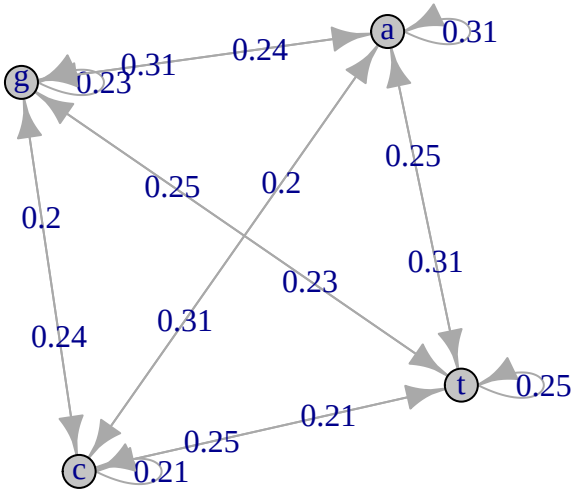
```
flat_sim1 <- unlist(sim1,use.names = FALSE)
mc_sim1 <-  markovchainFit(flat_sim1,method = "bootstrap", nboot = 3,
                           name = "Bootstrap Mc",
                           parallel = TRUE
                           )

flat_sim2 <- unlist(sim2,use.names = FALSE)
mc_sim2 <-  markovchainFit(flat_sim2,method = "bootstrap", nboot = 3,
                           name = "Bootstrap Mc",
                           parallel = TRUE
                           )
```

We see that the original data set from GenBank contains a small amount of additional "nucleotides": they are called "y", "m", and "r". These are the nucleotides which we are not sure about where Y indicates unsurity in A or G while R shows ambiguity in C or T and M appears when there is no surity at all.

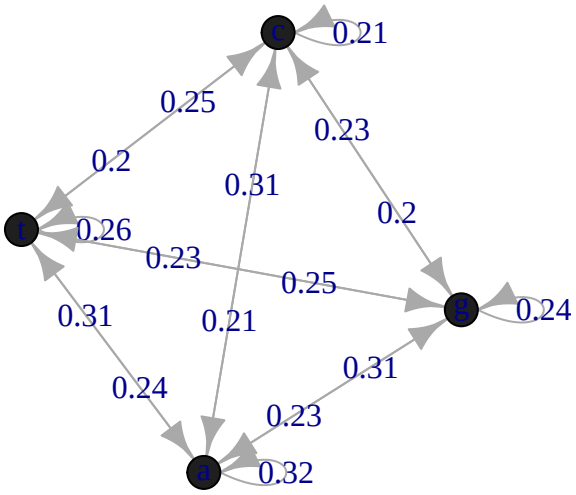
```
plot(mc_lizard$estimate)
```




```
plot(mc_sim1$estimate)
```



```
plot(mc_sim2$estimate)
```



In the first order Markov chains we assume that the nucleotide X_n only depends on a previous nucleotide, namely X_{n-1}. This is not a reasonable assumption, as DNA (or RNA) sequences contain the codons: the triplets of the nulceotides that determine what amino acids will be produced from it. Hence, it may be more reasonable that the most important dependences are between nucleotides in these triplets. In the lecture it was mentioned that there is some research supporting that a sixth order Markov Chain models perform better. That implies that the nth nucleotide depends on what were the previous 5 nucleotides.

Q2.3

In this part the three DNA collections of the sequences are aligned using the *msa* package. Furthermore the distances between the alignments are calculated. They are used to produce the heatmaps

```
aligment_lizard <- msa("lizard_seqs.fasta", type="dna")
```

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

```
aligment_sim1 <- msa("sim1_seq.fasta", type="dna")
```

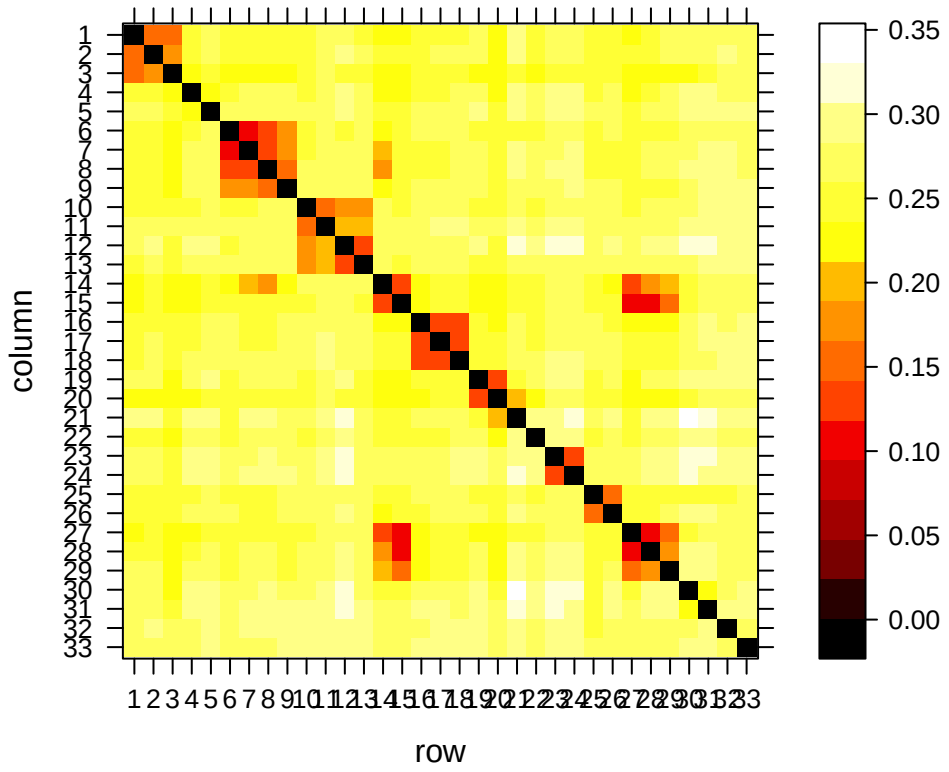
```
## use default substitution matrix
```

```
aligment_sim2 <- msa("sim2_seqs.fasta", type="dna")
```

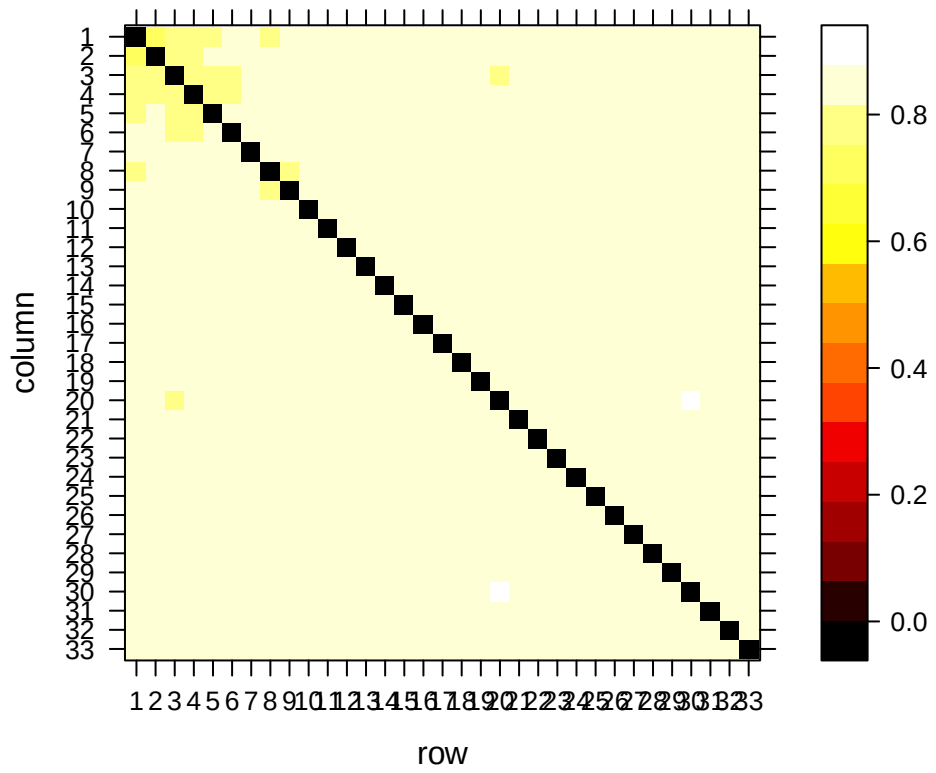
```
## use default substitution matrix
```

```
aligment_lizard2 <- msaConvert(aligment_lizard, type="seqinr::alignment")
dist_lizard <- seqinr::dist.alignment(aligment_lizard2,matrix="identity")

aligment_sim1_2 <- msaConvert(aligment_sim1, type="seqinr::alignment")
dist_sim1 <- seqinr::dist.alignment(aligment_sim1_2,matrix="identity")

aligment_sim2_2 <- msaConvert(aligment_sim2, type="seqinr::alignment")
dist_sim2 <- seqinr::dist.alignment(aligment_sim2_2,matrix="identity")

# the seqinr::dist.alignment computes the square root distance
dm_lizard <- as.matrix(dist_lizard)
dm_sim1 <- as.matrix(dist_sim1)
dm_sim2 <- as.matrix(dist_sim2)
```

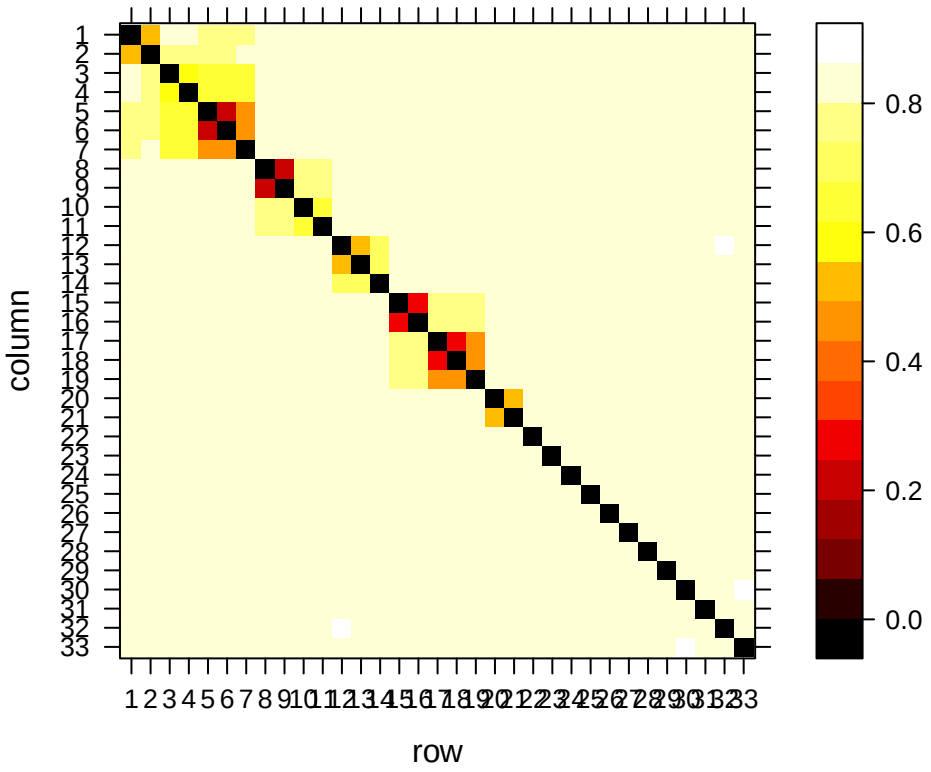
```
# heatmaps
new.palette=colorRampPalette(c("black","red","yellow","white"),space="rgb")
levelplot(dm_lizard[1:ncol(dm_lizard),ncol(dm_lizard):1],col.regions=new.palette(20))
```




```
levelplot(dm_sim1[1:ncol(dm_sim1),ncol(dm_sim1):1],col.regions=new.palette(20))
```



```
levelplot(dm_sim2[1:ncol(dm_sim2),ncol(dm_sim2):1],col.regions=new.palette(20))
```



The plots illustrate how much overlaying the sequences have in each dataset. The smaller the value (the darker the colour), the more nucleotides are matching between the corresponding sequences of the DNA. Therefore, the sequences from the data from GenBank has much more overlaying parts then the two randomly generated sequences. This result is expected as sequencing of a DNA produces "fragments" of the real DNA sequence, and those "fragments" do contain the same bits of the original DNA. The artificially simulated DNA sequences did not actually go through a process of the sequencing, hence they are way less likely to randomly contain long matching parts in their sequences.

Question 3: Phylogeny reconstruction

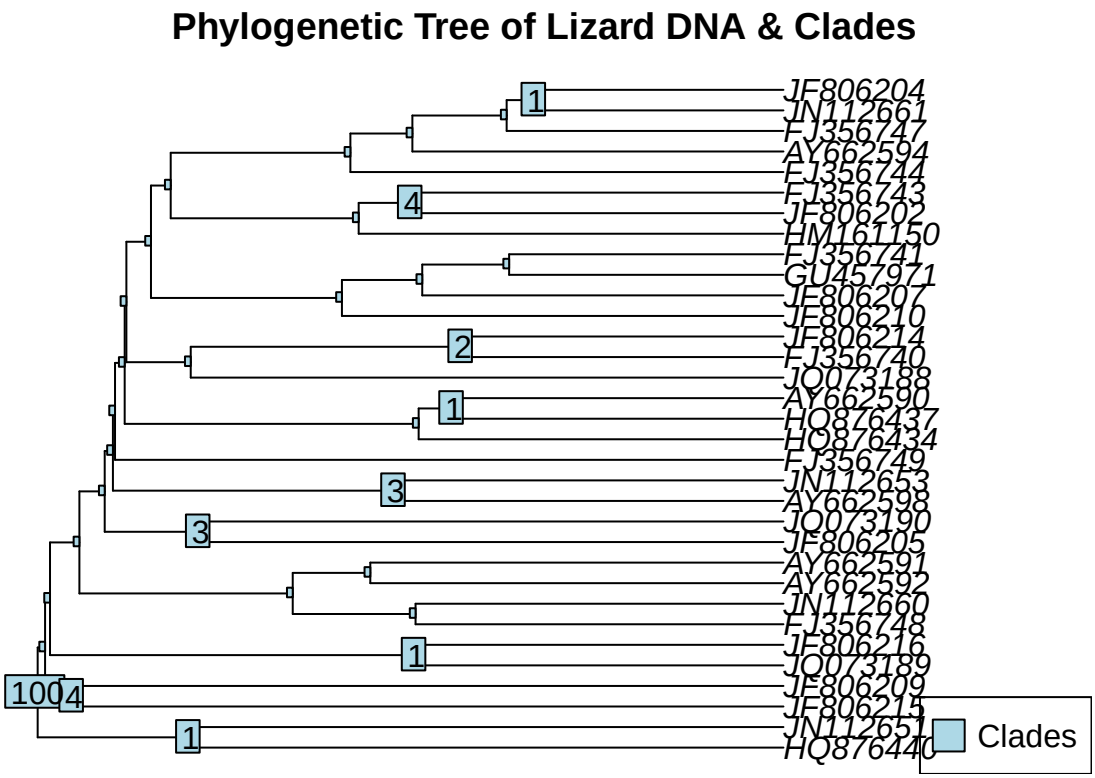
Q3.1

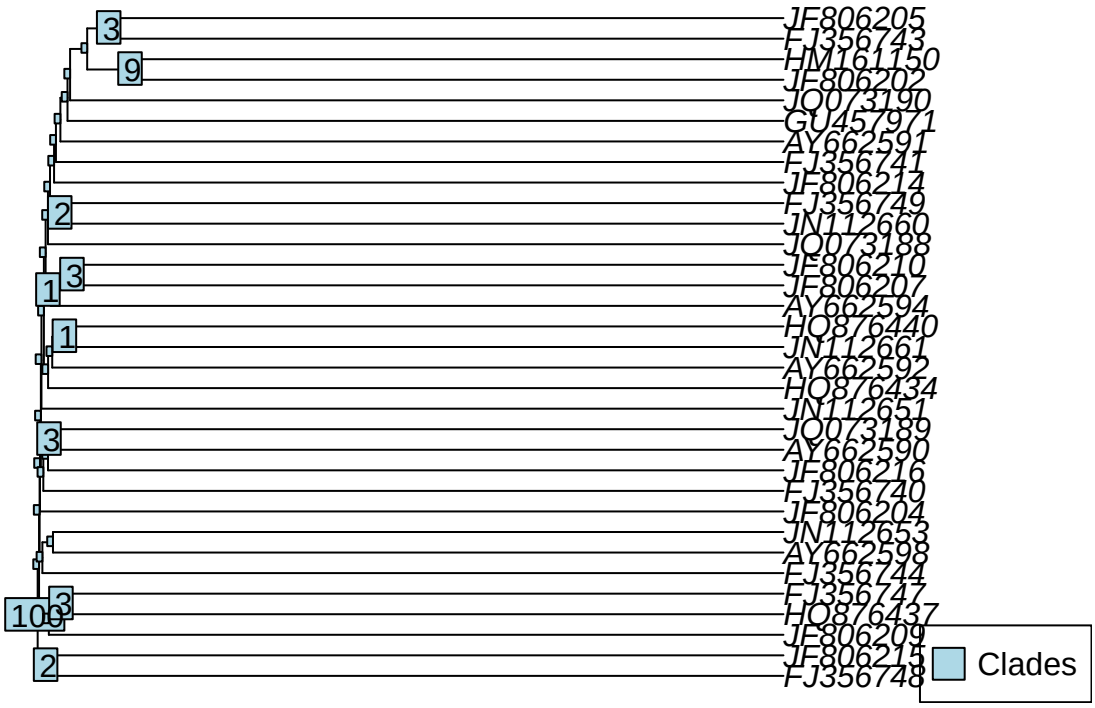
```
library(phangorn)


upgma_tree <- function(x){
  tree <- upgma(x)
  tree$tip.label <- names(lizards_sequences)
  return(tree)
}


lizard_seq_tree <- upgma_tree(dm_lizard)
#plot(lizard_seq_tree, main="Phylogenetic Tree of Lizard DNA")
```

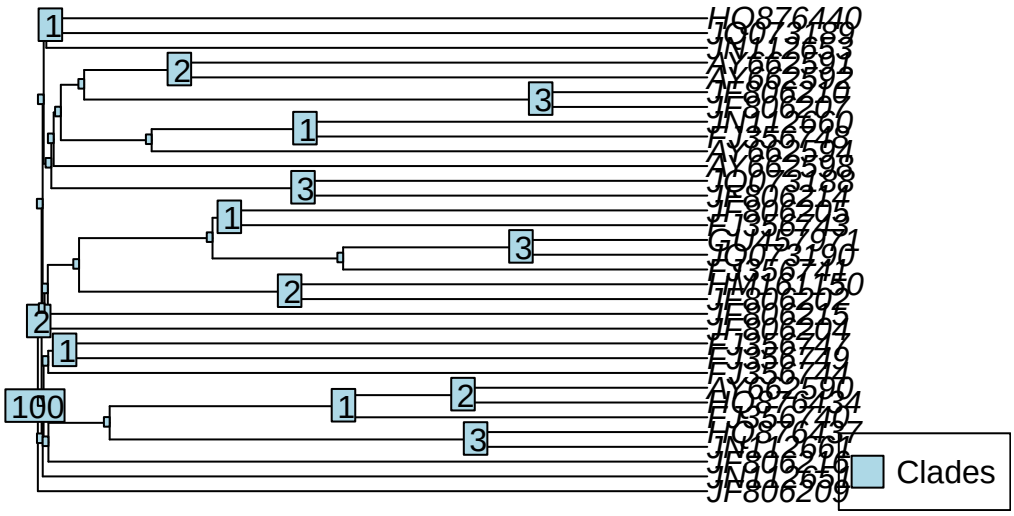
```
sim1_seq_tree <- upgma_tree(dm_sim1)
#plot(sim1_seq_tree, main="Phylogenetic Tree of Lizard DNA sim1")


sim2_seq_tree <- upgma_tree(dm_sim2)
#plot(sim2_seq_tree, main="Phylogenetic Tree of Lizard DNAs sim2")


boot_lizard <-  boot.phylo(lizard_seq_tree, dm_lizard, FUN=upgma_tree, trees  = TRUE,quiet = TRUE)
sim1_boot     <- boot.phylo(sim1_seq_tree, dm_sim1, FUN=upgma_tree, trees  = TRUE,quiet = TRUE)
sim2_boot     <- boot.phylo(sim2_seq_tree, dm_sim2, FUN=upgma_tree, trees  = TRUE,quiet = TRUE)
```

Extracting Clades

```
lizard_clade = prop.clades(lizard_seq_tree,boot_lizard$trees,rooted = TRUE)
sim1_clade = prop.clades(sim1_seq_tree,sim1_boot$trees,rooted = TRUE)
sim2_clade = prop.clades(sim2_seq_tree,sim2_boot$trees,rooted = TRUE)
```







Q3.2

```
library(phangorn)
path_diff_sim1_lizard <- path.dist(sim1_seq_tree,lizard_seq_tree)
path_diff_sim2_lizard <- path.dist(sim2_seq_tree,lizard_seq_tree)
path_diff_sim1_sim2 <- path.dist(sim1_seq_tree,sim2_seq_tree)
```

```
felenstein_diff_sim1_lizard <- KF.dist(sim1_seq_tree,lizard_seq_tree)
felenstein_diff_sim2_lizard <- KF.dist(sim2_seq_tree,lizard_seq_tree)
felenstein_diff_sim1_sim2 <- KF.dist(sim1_seq_tree,sim2_seq_tree)

result = data.frame(path_difference=c(path_diff_sim1_lizard,
                                      path_diff_sim2_lizard,
                                      path_diff_sim1_sim2),
                    felenstein_distance=c(felenstein_diff_sim1_lizard,
                                          felenstein_diff_sim2_lizard,
                                          felenstein_diff_sim1_sim2)
                    )
rownames(result) = c("sim1 vs lizard","sim2 vs lizard","sim1 vs sim2")
knitr::kable(result)
```

	path_difference	felenstein_distance
sim1 vs lizard	118.4314	1.909449
sim2 vs lizard	106.5270	1.468367
sim1 vs sim2	104.3264	1.050960

Geodesic Distance (Distance between edeges)

```
library(distory)
#distance between edeges using geodesic
geodesic_dist = dist.multiPhylo(list(lizard_seq_tree,sim1_seq_tree,sim2_seq_tree), method = "geodesic")
geodesic_dist = as.matrix(geodesic_dist)
geodesic_dist
```

```
##          1        2        3
## 1 0.000000 1.912268 1.517070
## 2 1.912268 0.000000 1.059108
## 3 1.517070 1.059108 0.000000
```

It can be seen from the path distance, Kuhner Felensteins distance and edge distance that tree from simulation 1 differs more from expected lizard tree than simulation 2. It can also be observed that both simulations are closer to each other that is in terms of distance.

```
comp1 <- comparePhylo(lizard_seq_tree,sim1_seq_tree)
comp2 <- comparePhylo(lizard_seq_tree,sim2_seq_tree)
comp12 <- comparePhylo(sim1_seq_tree,sim2_seq_tree)
```

```
comp1$messages
```

```
## [1] "=> Comparing lizard_seq_tree with sim1_seq_tree."
## [2] "Both trees have the same number of tips: 33."
## [3] "Both trees have the same tip labels."
## [4] "Both trees have the same number of nodes: 32."
## [5] "Both trees are rooted."
## [6] "Both trees are ultrametric."
## [7] "30 clades in lizard_seq_tree not in sim1_seq_tree."
## [8] "30 clades in sim1_seq_tree not in lizard_seq_tree."
## [9] "Branching times of clades in common between both trees: see ..$BT\n(node number in parentheses)
```

```
comp2$messages
```

```
## [1] "=> Comparing lizard_seq_tree with sim2_seq_tree."
## [2] "Both trees have the same number of tips: 33."
## [3] "Both trees have the same tip labels."
## [4] "Both trees have the same number of nodes: 32."
## [5] "Both trees are rooted."
## [6] "Both trees are ultrametric."
## [7] "29 clades in lizard_seq_tree not in sim2_seq_tree."
## [8] "29 clades in sim2_seq_tree not in lizard_seq_tree."
## [9] "Branching times of clades in common between both trees: see ..$BT\n(node number in parentheses)
```

```
comp12$messages
```

```
## [1] "=> Comparing sim1_seq_tree with sim2_seq_tree."
## [2] "Both trees have the same number of tips: 33."
## [3] "Both trees have the same tip labels."
## [4] "Both trees have the same number of nodes: 32."
## [5] "Both trees are rooted."
## [6] "Both trees are ultrametric."
## [7] "28 clades in sim1_seq_tree not in sim2_seq_tree."
## [8] "28 clades in sim2_seq_tree not in sim1_seq_tree."
## [9] "Branching times of clades in common between both trees: see ..$BT\n(node number in parentheses)
```

The comparision in terms of tips,nodes,label and clades also implies that simulation 2 is more similar to lizard tree as it has 27 different clades from expected lizard tree while simulation 1 has 28 different clades

```
knitr::opts_chunk$set(echo = TRUE)

library(ape)
library(phangorn)
library(markovchain)
library(seqinr)
library(msa)
library(lattice)
lizards_sequences<-ape::read.FASTA("lizard_seqs.fasta")
len_list <- lengths(lizards_sequences)
new_seq_list <- list()
new_seq <- c()
for (i in 1:length(lizards_sequences)){
    # extract the base composition of each sequence
    freqx <- ape::base.freq(lizards_sequences[i])
    new_seq <- sample(c("a","c","g","t"),len_list[i],rep=TRUE,prob=freqx)
    new_seq_list[i] <- list(new_seq)
}

# convert result to the DNAbin class
simulated_seq1 <- as.DNAbin(new_seq_list)
names(simulated_seq1) <- names(lizards_sequences)
#ape::write.dna(simulated_seq1, file ="sim1_seq.fasta", format = "fasta", colsep = "")

cat("Base composition of the data from GenBank:")
cat("\n")
ape::base.freq(lizards_sequences)
cat("\n")
```

```
cat("Base composition of the simulated sequences:")
cat("\n")
ape::base.freq(simulated_seq1)

# Simulate a tree with 33 nodes
tree<- TreeSim::sim.bd.taxa(n=33, numbsim=1, lambda=1, mu=0)[[1]]
tree$tip.label <- names(lizards_sequences)
plot(tree, yaxt = "n")

data <- c()
simulated_seq_tree <- list()

DNA_list <- unlist(as.character(lizards_sequences),use.names = FALSE)
DNA_unique <- c("a","c","g","t")
matrix <-  matrix(0, ncol = length(DNA_unique),
                  nrow = length(DNA_unique))

for (i in 1:(length(DNA_list) - 1)) {
    index_of_i <- DNA_unique == DNA_list[i]
    index_of_i_plus_1 <- DNA_unique == DNA_list[i + 1]
    matrix[index_of_i, index_of_i_plus_1] = matrix[index_of_i, index_of_i_plus_1] + 1
}

# Transition matrix is finally computed
Q <- matrix / rowSums(matrix)
#Q_lower <- Q[lower.tri(Q)]
# Other parameters that are needed for phangorn package function
bf <-  ape::base.freq(lizards_sequences)
size <- round(mean(len_list))
# Generating the data based on the tree

simulated_seq_tree <-  phangorn::simSeq(
        tree,
        l = size,
        type = "DNA",
        bf = bf,
        Q = Q)

# change numbers into letters
simulated_seq_tree2 <- lapply(simulated_seq_tree,function(seq){
    #seq = a
    states = c("a","c","g","t")
    seq[which(seq==1)] <- states[1]
    seq[which(seq==2)] <- states[2]
    seq[which(seq==3)] <- states[3]
    seq[which(seq==4)] <- states[4]
    return(seq)
})

simulated_seq2 <- as.DNAbin(simulated_seq_tree2)
# Write the simulated sequence to a fasta file
ape::write.dna(simulated_seq2, file ="sim2_seqs.fasta", format = "fasta", colsep = "")
```

```r
cat("Base composition of the data from GenBank:")
cat("\n")
ape::base.freq(lizards_sequences)
cat("\n")
cat("Base composition of the simulated sequences Q1.1:")
cat("\n")
ape::base.freq(simulated_seq1)
cat("\n")
cat("Base composition of the simulated sequences Q1.2:")
cat("\n")
ape::base.freq(simulated_seq2)

for (i in 1:length(lizards_sequences)){
    cat(paste0("Individual base composition of sequence ",i))
    cat("\n")
    cat("The data from the GenBank:")
    cat("\n")
    print(ape::base.freq(lizards_sequences[i]))
    cat("The simulated data in Q1.1:")
    cat("\n")
    print(ape::base.freq(simulated_seq1[i]))
    cat("The simulated data in Q1.2:")
    cat("\n")
    print(ape::base.freq(simulated_seq2[i]))
    cat("\n")
}
# CG content
comp_true <- ape::base.freq(lizards_sequences)
cg_true <- sum(comp_true[2:3])/sum(comp_true)
comp_s1 <- ape::base.freq(simulated_seq1)
cg_s1 <- sum(comp_s1[2:3])/sum(comp_s1)
comp_s2 <- ape::base.freq(simulated_seq2)
cg_s2 <- sum(comp_s2[2:3])/sum(comp_s2)

cat("In the data from the GenBank:")
cat("\n")
print(cg_true)
cat("In the simulated data in Q1.1:")
cat("\n")
print(cg_s1)
cat("In the simulated data in Q1.2:")
cat("\n")
print(cg_s2)
cat("\n")
# A,C,G,T content
cat("In the data from the GenBank:")
cat("\n")
print(ape::base.freq(lizards_sequences))
cat("In the simulated data in Q1.1:")
cat("\n")
print(ape::base.freq(simulated_seq1))
cat("In the simulated data in Q1.2:")
cat("\n")
```

```
print(ape::base.freq(simulated_seq2))

amino_lizard_seq <- read.fasta("lizard_amino_seqs.fasta")
amino_sim1_seq <- read.fasta("sim1_amino_seq.fasta")
amino_sim2_seq <- read.fasta("sim2_amino_seq.fasta")


amino_rev_com_lizard_seq <- read.fasta("rev_com_lizard_amino_acid.fasta")
amino_rev_com_sim1_seq <- read.fasta("rev_com_sim1_amino_acid.fasta")
amino_rev_com_sim2_seq <- read.fasta("rev_com_sim2_amino_acid.fasta")

stop_code_count <- function(amino_acid_seq){

  stop_count = c()
  for (i in 1:length(amino_acid_seq)) {

    sequence = amino_acid_seq[[i]]
    stop <- sequence[which(sequence == "*")]
    stop_count[i] = length(stop)
  }
  return(stop_count)
}
df = data.frame(lizard_seq = stop_code_count(amino_lizard_seq),
           "sim1" = stop_code_count(amino_sim1_seq),
           "sim2" = stop_code_count(amino_sim2_seq),
           "rev_comp_lizard_seq" = stop_code_count(amino_rev_com_lizard_seq),
           "rev_comp_sim1_seq" = stop_code_count(amino_rev_com_sim1_seq),
           "rev_comp_sim2_seq" = stop_code_count(amino_rev_com_sim2_seq)
          )

knitr::kable(df)
lizard <- as.character(lizards_sequences)
sim1 <- as.character(simulated_seq1)
sim2 <- as.character(simulated_seq2)


flat_lizard <- unlist(lizard,use.names = FALSE)
mc_lizard <-  markovchainFit(flat_lizard,method = "bootstrap", nboot = 3,
                             name = "Bootstrap Mc",
                             parallel = TRUE)

# the original data contains some "weird nucleotides" - what to do with them?
which(flat_lizard == "y")
which(flat_lizard == "m")
which(flat_lizard == "r")
which(flat_lizard == "s")

flat_sim1 <- unlist(sim1,use.names = FALSE)
mc_sim1 <-  markovchainFit(flat_sim1,method = "bootstrap", nboot = 3,
                           name = "Bootstrap Mc",
                           parallel = TRUE
                           )
```

```
flat_sim2 <- unlist(sim2,use.names = FALSE)
mc_sim2 <-  markovchainFit(flat_sim2,method = "bootstrap", nboot = 3,
                           name = "Bootstrap Mc",
                           parallel = TRUE
                           )


plot(mc_lizard$estimate)
plot(mc_sim1$estimate)
plot(mc_sim2$estimate)
aligment_lizard <- msa("lizard_seqs.fasta", type="dna")
aligment_sim1 <- msa("sim1_seq.fasta", type="dna")
aligment_sim2 <- msa("sim2_seqs.fasta", type="dna")

aligment_lizard2 <- msaConvert(aligment_lizard, type="seqinr::alignment")
dist_lizard <- seqinr::dist.alignment(aligment_lizard2,matrix="identity")

aligment_sim1_2 <- msaConvert(aligment_sim1, type="seqinr::alignment")
dist_sim1 <- seqinr::dist.alignment(aligment_sim1_2,matrix="identity")

aligment_sim2_2 <- msaConvert(aligment_sim2, type="seqinr::alignment")
dist_sim2 <- seqinr::dist.alignment(aligment_sim2_2,matrix="identity")

# the seqinr::dist.alignment computes the square root distance
dm_lizard <- as.matrix(dist_lizard)
dm_sim1 <- as.matrix(dist_sim1)
dm_sim2 <- as.matrix(dist_sim2)
# heatmaps
new.palette=colorRampPalette(c("black","red","yellow","white"),space="rgb")
levelplot(dm_lizard[1:ncol(dm_lizard),ncol(dm_lizard):1],col.regions=new.palette(20))
levelplot(dm_sim1[1:ncol(dm_sim1),ncol(dm_sim1):1],col.regions=new.palette(20))
levelplot(dm_sim2[1:ncol(dm_sim2),ncol(dm_sim2):1],col.regions=new.palette(20))

library(phangorn)


upgma_tree <- function(x){
  tree <- upgma(x)
  tree$tip.label <- names(lizards_sequences)
  return(tree)
}


lizard_seq_tree <- upgma_tree(dm_lizard)
#plot(lizard_seq_tree, main="Phylogenetic Tree of Lizard DNA")



sim1_seq_tree <- upgma_tree(dm_sim1)
#plot(sim1_seq_tree, main="Phylogenetic Tree of Lizard DNA sim1")


sim2_seq_tree <- upgma_tree(dm_sim2)
```

```
#plot(sim2_seq_tree, main="Phylogenetic Tree of Lizard DNAs sim2")


boot_lizard <-  boot.phylo(lizard_seq_tree, dm_lizard, FUN=upgma_tree, trees  = TRUE,quiet = TRUE)
sim1_boot     <- boot.phylo(sim1_seq_tree, dm_sim1, FUN=upgma_tree, trees  = TRUE,quiet = TRUE)
sim2_boot     <- boot.phylo(sim2_seq_tree, dm_sim2, FUN=upgma_tree, trees  = TRUE,quiet = TRUE)


lizard_clade = prop.clades(lizard_seq_tree,boot_lizard$trees,rooted = TRUE)
sim1_clade = prop.clades(sim1_seq_tree,sim1_boot$trees,rooted = TRUE)
sim2_clade = prop.clades(sim2_seq_tree,sim2_boot$trees,rooted = TRUE)
layout(1)
par(mar = rep(2, 4))
plot(lizard_seq_tree, main = "Phylogenetic Tree of Lizard DNA & Clades")
nodelabels(lizard_clade)
legend("bottomright", legend = c("Clades"), pch = 22,
       pt.bg = c("lightblue"), pt.cex = 2.5)


layout(1)
par(mar = rep(2, 4))
plot(sim1_seq_tree, main = "Sim 1 Phylogenetic Tree of Lizard DNA & Clades")
nodelabels(sim1_clade)
legend("bottomright", legend = c("Clades"), pch = 22,
       pt.bg = c("lightblue"), pt.cex = 2.5)



plot(sim2_seq_tree, main = "Sim 2 Phylogenetic Tree of Lizard DNA & Clades")
nodelabels(sim2_clade)
legend("bottomright", legend = c("Clades"), pch = 22,
       pt.bg = c( "lightblue"), pt.cex = 2.5)


library(phangorn)
path_diff_sim1_lizard <- path.dist(sim1_seq_tree,lizard_seq_tree)
path_diff_sim2_lizard <- path.dist(sim2_seq_tree,lizard_seq_tree)
path_diff_sim1_sim2 <- path.dist(sim1_seq_tree,sim2_seq_tree)


felenstein_diff_sim1_lizard <- KF.dist(sim1_seq_tree,lizard_seq_tree)
felenstein_diff_sim2_lizard <- KF.dist(sim2_seq_tree,lizard_seq_tree)
felenstein_diff_sim1_sim2 <- KF.dist(sim1_seq_tree,sim2_seq_tree)

result = data.frame(path_difference=c(path_diff_sim1_lizard,
                                      path_diff_sim2_lizard,
                                      path_diff_sim1_sim2),
                    felenstein_distance=c(felenstein_diff_sim1_lizard,
                                          felenstein_diff_sim2_lizard,
                                          felenstein_diff_sim1_sim2)
                    )
rownames(result) = c("sim1 vs lizard","sim2 vs lizard","sim1 vs sim2")
knitr::kable(result)
```

```r
library(distory)
#distance between edeges using geodesic
geodesic_dist = dist.multiPhylo(list(lizard_seq_tree,sim1_seq_tree,sim2_seq_tree), method = "geodesic")
geodesic_dist = as.matrix(geodesic_dist)
geodesic_dist

comp1 <- comparePhylo(lizard_seq_tree,sim1_seq_tree)
comp2 <- comparePhylo(lizard_seq_tree,sim2_seq_tree)
comp12 <- comparePhylo(sim1_seq_tree,sim2_seq_tree)

comp1$messages
comp2$messages
comp12$messages
```