Bioinformatics - Computer Lab 2

Group 7: Phillip Hölscher (phiho267), Lennart Schilling (lensc874), Thijs Quast (thiqu264), Mariano Maquieira Mariani (marma330)

20 November 2018

Question 1

At first, the dataset of the RAG1 gene sequences from 33 lizard species were downloaded from GenBank and saved in a fasta file using the provided R script *732A51 BioinformaticsHT2018 Lab02 GenBankGetCode.R.*. The code can be found in Appendix 1 (Data Import of original dataset).

Question 1.1

The saved fasta-file has to be read in R so that we can work with that. After analysing the sequences, it becomes clear that there can be found many whitespaces (“”). Since the artificial sequences should be simulated so that each nucleotide is to be independently and randomly drawn from the distribution given by the base composition in the true lizard sequences, the whitespaces have to be removed. Otherwise the artificial sequences are built on a probability distribution where the sum of all probabilities would not equal 1. The R code for the reading and preparation process can be found in Appendix 1.1 (Reading and preparing original data).

After preparing the data, the artificial dataset is built by considering that it contains 33 sequences (each length of the sequences is the same as in the lizard dataset) so that for each real sequence an artificial one is created. As mentioned, the simulation of the artificial sequences is based on the distribution given by the base composition of the original dataset.

The artificial dataset is submitted as the fasta file *artificial_dataset_1_1.fasta*. The written function for all these processes automatically prints the base composition in the simulated data compared to the base composition in the original data. An extract from the output can be seen here:

```
  get_artificial_sequence_dataset(lizards_sequences)
```

```
## [1] "comparison of base compositions between original and artificial datasets (values rounded):"
##   name_original name_artificial a_original a_artificial c_original
## 1 "JF806202"    "1"             "0.29"     "0.3"        "0.2"
## 2 "HM161150"    "2"             "0.31"     "0.32"       "0.21"
## 3 "FJ356743"    "3"             "0.31"     "0.31"       "0.21"
## 4 "JF806205"    "4"             "0.28"     "0.29"       "0.21"
## 5 "JQ073190"    "5"             "0.31"     "0.29"       "0.2"
##   c_artificial g_original g_artificial t_original t_artificial
## 1 "0.21"       "0.24"     "0.24"       "0.26"     "0.24"
## 2 "0.21"       "0.23"     "0.23"       "0.24"     "0.25"
## 3 "0.2"        "0.23"     "0.23"       "0.24"     "0.25"
## 4 "0.2"        "0.24"     "0.25"       "0.26"     "0.25"
## 5 "0.22"       "0.24"     "0.23"       "0.26"     "0.25"
```

It becomes clear that the base compositions are very similar. The entire code for the function can be seen in Appendix 1.1 (Function code).

Question 1.2

```
##
```

```
## Attaching package: 'ape'
```

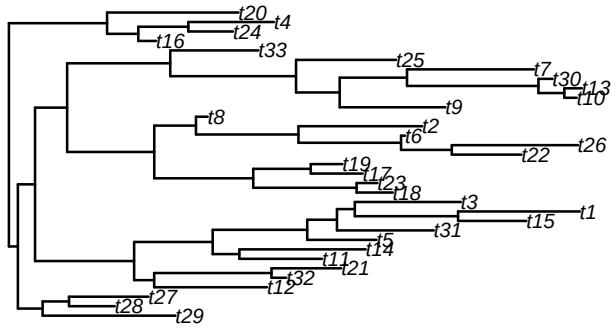
```
## The following objects are masked from 'package:seqinr':
##
##     as.alignment, consensus
```

In this part of the exercise do we use the prepared data from part 1, in Appendix 1 code can be found in (Data Import of original dataset).

We used the function *rtree* to create a tree object of the type phylo and the length of the original sequences.

```
tree <- rtree(n = length(lizards_sequences))
```

Here you can find the plot of the tree.



After the simulation of the phylogenetic tree, we had to simulate the sequence.

For this, we the had several things to do. 1. We simulated a transition rate matrix (Q-Matrix). In this case we choose one by yourself.

```
##      a    c    t    g
## a 0.25 0.25 0.25 0.25
## c 0.25 0.25 0.25 0.25
## t 0.25 0.25 0.25 0.25
## g 0.25 0.25 0.25 0.25
```

2. We had to choose the length of the sequence. To make it comparable with the original lizards dataset, we decided to create

```
lengths <- c()
for (i in 1:33){
lengths <- c(lengths, length(lizards_sequences[[i]]))
}
```

Now we can use the simulate the sequences by using the function *phangorn::simSeq()*.

```
sequences_artificial <- list()
for (j in 1:33){
sequences_artificial[j] <- simSeq(tree, l = lengths[j], Q=transition_matrix , type = "DNA")
}
```

Since in sequences are filled with integers from 1 to 4, do we have to replace the numbers by the letters a,b,c,d.

1 = a

2 = b

3 = c

4 = d

The code for this can be found in Appendix 1.2

The second simulate a artificial DNA sequence dataset do we save as *"artifical_dataset_1_2.fasta"*.

```
ape::write.dna(sequences_artificial, file ="artifical_dataset_1_2.fasta", format = "fasta", colsep = ""
```

Question 2

```
lizards_sequences = read.fasta("lizard_seqs.fasta")
original_dataset <- lizards_sequences
artificial_sequences_1 <- read.fasta("artificial_dataset_1_1.fasta")
artificial_sequences_2 <- read.fasta("artificial_dataset_1_2.fasta")
original_base_compositions <- list()
artificial_1_base_compositions <- list()
artificial_2_base_compositions <- list()
for (i in 1:length(original_dataset)) {
    # getting base compositions for each original sequence
    original_base_compositions[[i]] =
      seqinr::count(original_dataset[[i]],1)/length(original_dataset[[i]])
}

for (i in 1:length(artificial_sequences_1)) {
    # getting base compositions for each original sequence
    artificial_1_base_compositions[[i]] =
      seqinr::count(artificial_sequences_1[[i]],1)/length(artificial_sequences_1[[i]])
}

for (i in 1:length(artificial_sequences_2)) {
    # getting base compositions for each original sequence
    artificial_2_base_compositions[[i]] =
      seqinr::count(artificial_sequences_2[[i]],1)/length(artificial_sequences_2[[i]])
}
```

```
Reduce('+', original_base_compositions)
```

```
##
##        a        c        g        t
## 9.372287 6.239920 7.068938 7.764488
```

```
Reduce('+', artificial_1_base_compositions)
```

```
##
##         a         c         g         t
## 10.107689  6.864285  7.656544  8.371482
```

```
Reduce('+', artificial_2_base_compositions)
```

```
##
##        a        c        g        t
## 8.203887 8.339508 8.256393 8.200212
```

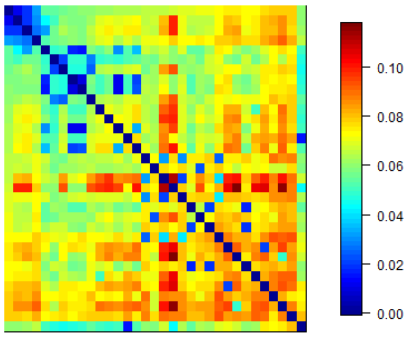
```
#install.packages("rDNAse")
GC_composition <- GC(lizards_sequences[[1]])
AT <- 1-GC_composition
```

```
library(rDNAse)
string1 <- paste(lizards_sequences[[1]], collapse = "")
string1 <- toupper(string1)
x <- kmer(string1)
```

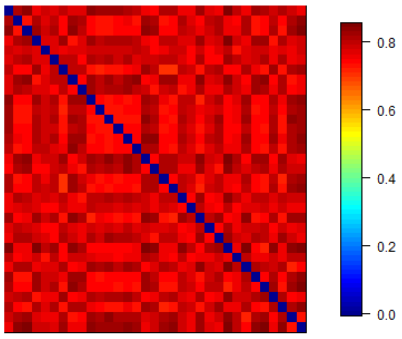
Question 2.3

To allign the sequences for each dataset (the original dataset *lizards_sequences*, the first artificial dataset *artificial_dataset_1_1* and the second artificial dataset *artificial_dataset_1_2*), the *plsgenomics* package was used. The *.fasta*-files for the datasets were transformed to a *DNAStringSet* - class within R. The uncorrected distance matrices created represent the hamming distance between each of the sequences in each dataset. The results of these distance matrices are plotted as heatmaps (using *plsgenomics* package) :

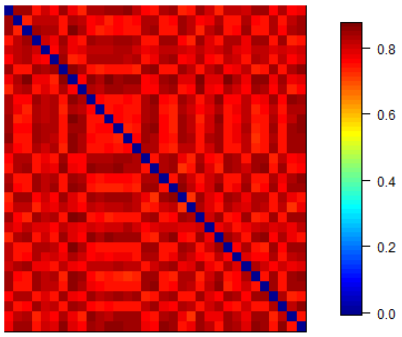
lizards_sequences



artificial_dataset_1_1



artificial_dataset_1_2



We see that for the original dataset, the allignment results are much better than for the artificial datasets. Based on the point that the artificial datasets were created by sampling randomly, the greater distances between the sequences compared to the distances within the original dataset make sense.

The R code for this Question 2.3 can be found in Appendix 2.3.

Appendix 1

Data Import of original dataset

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

Appendix 1.1

Reading and preparing original data

```
library(seqinr)
# reading original_dataset from fasta file
lizards_sequences = read.fasta("lizard_seqs.fasta")

# preparing data in fasta file (dna sequences include emtpy spaces which will be removed)
for (i in 1:length(lizards_sequences)) {
  lizards_sequences[[i]] = lizards_sequences[[i]][lizards_sequences[[i]] != " "]
}
```

Function code

```
library(seqinr)
get_artificial_sequence_dataset = function(original_dataset) {
  # creating empty varibales which will be filled in following for-loop
  original_base_compositions = list()
  artificial_dataset = list()
  artificial_base_compositions = list()
  a_original = c(); c_original = c(); g_original = c(); t_original = c()
  a_artificial = c(); c_artificial = c(); g_artificial = c(); t_artificial = c()
  for (i in 1:length(original_dataset)) {
    # getting base compositions for each original sequence
    original_base_compositions[[i]] =
      seqinr::count(original_dataset[[i]],1)/length(original_dataset[[i]])
    # creating artificial sequences randomly drawn from the distribution
    # given by the base composition
    artificial_dataset[[as.character(i)]] = sample(x = c("a","c","g","t"),
                                                   size = length(original_dataset[[i]]),
                                                   rep = TRUE,
                                                   prob = original_base_compositions[[i]])
    # creating dataframe to compare base compositions
    # between original and artificial sequences
    artificial_base_compositions[[i]] =
      seqinr::count(artificial_dataset[[i]],1)/length(artificial_dataset[[i]])
    a_original = c(a_original, round(original_base_compositions[[i]][1],2))
    a_artificial = c(a_artificial, round(artificial_base_compositions[[i]][1],2))
    c_original = c(c_original, round(original_base_compositions[[i]][2],2))
    c_artificial = c(c_artificial, round(artificial_base_compositions[[i]][2],2))
```

```
    g_original = c(g_original, round(original_base_compositions[[i]][3],2))
    g_artificial = c(g_artificial, round(artificial_base_compositions[[i]][3],2))
    t_original = c(t_original, round(original_base_compositions[[i]][4],2))
    t_artificial = c(t_artificial, round(artificial_base_compositions[[i]][4],2))
  }
  comparison_base_compositions = cbind(
    name_original = names(original_dataset), name_artificial = names(artificial_dataset),
    a_original, a_artificial, c_original, c_artificial,
    g_original, g_artificial, t_original, t_artificial
  )
  rownames(comparison_base_compositions) = 1:nrow(comparison_base_compositions)
  print("comparison of base compositions
        between original and artificial datasets (values rounded): ")
  print(comparison_base_compositions)
  # saving fasta file
  ape::write.dna(artificial_dataset, file ="artificial_dataset_1_1.fasta", format = "fasta", colsep = "
}
```

Appendix 1.2

Replace the integers by letters

```
for (k in 1:33){
sequences_artificial[[k]][sequences_artificial[[k]] == 1] = "a"
sequences_artificial[[k]][sequences_artificial[[k]] == "2"] = "c"
sequences_artificial[[k]][sequences_artificial[[k]] == "3"] = "g"
sequences_artificial[[k]][sequences_artificial[[k]] == "4"] = "t"
}
```

Appendix 2

Appendix 2.1

Appendix 2.2

Appendix 2.3

```
library(seqinr)
library(DECIPHER)
library(plsgenomics)
library(ape)

# getting all datasets in DNAStringSet format

  # original dataset
    # readAAStringSet-function needs path of fasta file as input. The original
    # dataset needs to be prepared and saved so that the fasta file does not
    # inlcude whitespaces anymore.
      # reading original_dataset from fasta file
      lizards_sequences = read.fasta("lizard_seqs.fasta")
      # preparing data in fasta file (dna sequences include emtpy spaces which will be removed)
      for (i in 1:length(lizards_sequences)) {
        lizards_sequences[[i]] = lizards_sequences[[i]][lizards_sequences[[i]] != " "]
      }
```

```
      # saving prepared fasta file
      ape::write.dna(lizards_sequences, file ="lizards_sequences_no_whitespaces.fasta",
                     format = "fasta", colsep = "")
    # reading prepared fasta file as biostrings-object
    lizards_sequences = readDNAStringSet("lizards_sequences_no_whitespaces.fasta")

  # artificial_dataset_1_1
  artificial_dataset_1_1 = readDNAStringSet("artificial_dataset_1_1.fasta")

  # artificial_dataset_1_2
  artificial_dataset_1_2 = readDNAStringSet("artificial_dataset_1_2.fasta")

# aligning sequences for each dataset
sequence_aligning = function(dataset, name) {
  # aligning process
  sequences_aligned = AlignSeqs(dataset)
  # creating distance matrix
  dm_sequences_aligned = DistanceMatrix(sequences_aligned)
  # creating matrix heatmap
  heatmap_dm_sequences_aligned = matrix.heatmap(dm_sequences_aligned)
  dev.copy(png,paste("heatmap_", name, ".png", sep=""))
  dev.off()
  return(sequences_aligned)
}

lizards_sequences_aligned = sequence_aligning(dataset = lizards_sequences,
                                              name = "lizards_sequences")
artificial_dataset_1_1_aligned = sequence_aligning(artificial_dataset_1_1,
                                                   name = "artificial_dataset_1_1")
artificial_dataset_1_2_aligned = sequence_aligning(artificial_dataset_1_2,
                                                   name = "artificial_dataset_1_2")
```