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"     "0.27"       "0.2"
## 2 "HM161150"    "2"             "0.31"     "0.3"        "0.21"
## 3 "FJ356743"    "3"             "0.31"     "0.32"       "0.21"
## 4 "JF806205"    "4"             "0.28"     "0.26"       "0.21"
## 5 "JQ073190"    "5"             "0.31"     "0.31"       "0.2"
##   c_artificial g_original g_artificial t_original t_artificial
## 1 "0.24"       "0.24"     "0.22"       "0.26"     "0.27"
## 2 "0.22"       "0.23"     "0.22"       "0.24"     "0.25"
## 3 "0.2"        "0.23"     "0.24"       "0.24"     "0.23"
## 4 "0.21"       "0.24"     "0.27"       "0.26"     "0.26"
## 5 "0.19"       "0.24"     "0.24"       "0.26"     "0.26"
```

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).

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")
}
```