732A51 Bioinformatics

Lab 2

Johannes Hedström, Mikael Montén

STIMA
Institutionen för datavetenskap
Linköpings universitet

2024-11-20

Contents

1 Question 1: DNA sequence acquisition and simulation

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 file. 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.

```
library(ape)
library(seqinr)

## 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)
```

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 artificial 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.

```
#library(help = ape)
#library(help = seqinr)
set.seed(123456790)
#ape::as.character.DNAbin(lizards_sequences[[1]])

artif_dna_seq <- function(org_seq){
  new_seq <- list()

  for(i in 1:length(org_seq)){
    comp_dist <- base.freq(lizards_sequences[i][1]) # extract base compositions
    seq_len <- length(lizards_sequences[[i]]) # extract length of sequence
    # sample according to individual base comp distribution and length
    new_seq[[i]] <- sample(c("a", "c", "g", "t"), size = seq_len, replace = TRUE, prob = comp_dist)
  }
  # set unique names according to original sequence
  names(new_seq) <- paste0("ARTIFICAL_",names(org_seq))
  return(new_seq)
}

artif_liz_seq <- artif_dna_seq(lizards_sequences)
# save dataset as fasta
#seqinr::write.fasta(sequences =artif_liz_seq,
#                    names = names(artif_liz_seq),
#                    file.out = "artifical_lizard_seqs.fasta")

# report on base composition
artif_liz_seq_DNAbin <- as.DNAbin(artif_liz_seq)

seq_compare <- rbind(base.freq(lizards_sequences),base.freq(artif_liz_seq_DNAbin))
rownames(seq_compare) <- c("Lizards sequences", "Artif. lizards sequences")
knitr::kable(seq_compare, caption="Base compositions")
```

Table 1: Base compositions

	a	c	g	t
Lizards sequences	0.3121454	0.2052325	0.2307222	0.2518999
Artif. lizards sequences	0.3074960	0.2068006	0.2329029	0.2528005

Comparing the base compositions for the original lizards_sequences FASTA file and the artificially created lizards_sequences, we can see that the base compositions are nearly identical eachother with minimal differences between the sequences.