Bioinformatics Lab 2

Simon Jorstedt

2024-11-21

```r
# Dependencies and data
library(ape)
```

```
## Warning: package 'ape' was built under R version 4.4.2
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
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
```

Question 1

Q 1.1

Simulate an artificial DNA sequence dataset. It should contain 33 sequences. The lengths of the sequences should be the same as in the lizard dataset, i.e. for each real sequence simulate an arti cial 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 le. Remember to give unique names to your sequences. Report on the base composition in your simulated data.

```
sim_sequences <- list()

# Simulate sequences
for (i in 1:33){
  # Concatenate a vector like c("a", ..., "g")
  sim_sequences[[i]] <- paste(
    # Sample
    sample(x = c("a", "c", "g", "t"),
           size = lizards_sequences[[i]] %>% length(),
           replace = TRUE,
           prob = base.freq(lizards_sequences[i])),
    collapse = "")
}

# Rename and save the sequences to a fasta file
#names(sim_sequences) <- lizards_accession_numbers
#ape::write.dna(x=sim_sequences, file="simulated_sequences.fasta", format = "fasta", nbcol = 6, colsep

ape::write.dna(sim_sequences,
               file ="simulated_sequences.fasta",
               format = "fasta",
               append =FALSE,
               nbcol = 10,
               colsep = " ",
```

```
                colw = 10000)
#due to ape bug, set colw to longer than the longest sequence
```

The base composition of the respective simulated sequences will be roughly the same as the true base compositions, due to the law of large numbers. Example:

```
# Base frequences of the first real sequence
base.freq(lizards_sequences[1])
```

```
##         a         c         g         t
## 0.2898696 0.2026078 0.2437312 0.2637914
```

```
# Base frequences of the first simulated sequence
(strsplit(sim_sequences[[1]], split = "")[[1]] %>% table())/nchar(sim_sequences[[1]])
```

```
## .
##         a         c         g         t
## 0.2945892 0.2214429 0.2294589 0.2545090
```

Question 2

Q 2.1