Bioinformatics Lab 2

Simon Jorstedt

2024-11-22

```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,
               colsep = " ",
               colw = 10)
```

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.

```
set.seed(8182746)

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, convert to DNAbin, and save the sequences to a fasta file
names(sim_sequences) <- lizards_accession_numbers
```

```
sim_sequences <- as.DNAbin(sapply(sim_sequences, strsplit, split = ""))

ape::write.dna(sim_sequences,
               file ="simulated_sequences.fasta",
               format = "fasta")
```

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
base.freq(sim_sequences[1])
```

```
##         a         c         g         t
## 0.2855711 0.2084168 0.2294589 0.2765531
```

Question 2

Q 2.1

Report some basic statistics on each sequence dataset: individual base composition, GC content, CpG, ApT content. Also translate your sequences into protein sequences (see Lab 1) and report on the amino acid composition. In your simulated sequences, how many times did you observe a stop codon inside your sequence? Does this occur in your true sequences? Comment.

```
# Create empty dataframe
df_base_stats <- data.frame(Accession=character(),
                            GC_cont=numeric(),
                            CpG_cont=numeric(),
                            ApT_cont=numeric())


for(i in 1:33){
  # GC content
  gc_cont <- base.freq(lizards_sequences[1])[c(2,3)] %>% sum()

  # CpG and ApT contents
  dna_string <- as.character(lizards_sequences)[[i]] %>% paste(collapse = "")
  CpG_cont <- gregexpr(pattern="cg", text=dna_string)[[1]] %>% length() %>% {2*./nchar(dna_string)}
  ApT_cont <- gregexpr(pattern="at", text=dna_string)[[1]] %>% length() %>% {2*./nchar(dna_string)}

  # Store values
  new_row <- data.frame(Accession=lizards_accession_numbers[i],
                        GC_cont=gc_cont,
                        CpG_cont=CpG_cont,
```

```
                        ApT_cont=ApT_cont)

  df_base_stats <- rbind(df_base_stats, new_row)
}

df_base_stats
```

```
##    Accession  GC_cont   CpG_cont  ApT_cont
## 1   JF806202 0.446339 0.02605210 0.1543086
## 2   HM161150 0.446339 0.02288667 0.1247693
## 3   FJ356743 0.446339 0.02421308 0.1189900
## 4   JF806205 0.446339 0.02973241 0.1427156
## 5   JQ073190 0.446339 0.02713704 0.1492537
## 6   GU457971 0.446339 0.02199817 0.1429881
## 7   FJ356741 0.446339 0.02144587 0.1224490
## 8   JF806207 0.446339 0.02191235 0.1593625
## 9   JF806210 0.446339 0.02107280 0.1551724
## 10  AY662592 0.446339 0.02054795 0.1273973
## 11  AY662591 0.446339 0.02201705 0.1264205
## 12  FJ356748 0.446339 0.02210708 0.1291883
## 13  JN112660 0.446339 0.01819505 0.1375546
## 14  AY662594 0.446339 0.02044413 0.1290095
## 15  JN112661 0.446339 0.02037846 0.1317322
## 16  HQ876437 0.446339 0.02191235 0.1314741
## 17  HQ876434 0.446339 0.03000000 0.1440000
## 18  AY662590 0.446339 0.01345133 0.1267257
## 19  FJ356740 0.446339 0.01684802 0.1375921
## 20  JF806214 0.446339 0.02182540 0.1468254
## 21  JQ073188 0.446339 0.01630435 0.1521739
## 22  FJ356749 0.446339 0.01935707 0.1251296
## 23  JQ073189 0.446339 0.01520387 0.1478922
## 24  JF806216 0.446339 0.03200000 0.1360000
## 25  AY662598 0.446339 0.02044413 0.1254847
## 26  JN112653 0.446339 0.01516361 0.1332801
## 27  JF806204 0.446339 0.01393035 0.1512438
## 28  FJ356747 0.446339 0.02075406 0.1272916
## 29  FJ356744 0.446339 0.02006226 0.1286752
## 30  HQ876440 0.446339 0.02641509 0.1415094
## 31  JN112651 0.446339 0.01533966 0.1322133
## 32  JF806215 0.446339 0.02193420 0.1395813
## 33  JF806209 0.446339 0.02363050 0.1460795
```

We will now translate the sequences to proteins.

```
# Translate to proteins
# Get codon table and Open Reading Frame (ORF)
cod_orf_info <- read.delim("codon_and_orf.txt", header = TRUE, sep = ",")

lizard_proteins <- list()
sim_proteins <- list()
for(i in 1:33){
  # Translate true lizard sequences
  true_protein <- ape::trans(lizards_sequences[[i]],
```

```
                              codonstart = cod_orf_info["codon_start"][[1]][i],
                              code = cod_orf_info["transl_table"][[1]][i])
  lizard_proteins[[i]] <- true_protein

  # Translate simulated sequences
  sim_prot <- ape::trans(sim_sequences[[i]],
                         codonstart = cod_orf_info["codon_start"][[1]][i],
                         code = cod_orf_info["transl_table"][[1]][i])
  sim_proteins[[i]] <- sim_prot
}
```

Now we can determine the composition of amino acids in the proteins.

```
# Report amino acid compositions
for(i in 1:33){
  true_composition <- as.character(lizard_proteins[[1]]) %>%
    table(dnn = paste("Composition of Amino Acids in",
                      lizards_accession_numbers[i]))

  sim_compositions <- as.character(sim_proteins[[1]]) %>%
    table(dnn = paste("Composition of Amino Acids in",
                      lizards_accession_numbers[i], "(simulated)"))

  print(true_composition)
  print(sim_compositions)
  cat("\n")
}
```

```
## Composition of Amino Acids in JF806202
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806202 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in HM161150
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in HM161150 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356743
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356743 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806205
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806205 (simulated)
```

```
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JQ073190
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JQ073190 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in GU457971
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in GU457971 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356741
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356741 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806207
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806207 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806210
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806210 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in AY662592
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in AY662592 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in AY662591
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in AY662591 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356748
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
```

```
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356748 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JN112660
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JN112660 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in AY662594
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in AY662594 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JN112661
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JN112661 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in HQ876437
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in HQ876437 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in HQ876434
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in HQ876434 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in AY662590
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in AY662590 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356740
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356740 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
```

```
## Composition of Amino Acids in JF806214
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806214 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JQ073188
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JQ073188 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356749
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356749 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JQ073189
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JQ073189 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806216
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806216 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in AY662598
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in AY662598 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JN112653
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JN112653 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806204
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806204 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
```

```
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356747
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356747 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in FJ356744
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in FJ356744 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in HQ876440
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in HQ876440 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JN112651
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JN112651 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806215
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806215 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
##
## Composition of Amino Acids in JF806209
##  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 23  8 23 32 13 19  9 21 20 30  9 11 15  5 19 25 15 22  3 10
## Composition of Amino Acids in JF806209 (simulated)
##  *  A  C  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
## 17 13 11 14 12 11 16  4 20 13 28  6  9 19  7 26 36 25 21  3 21
```

Now we will count how many times the stop codons (taa, tag, tga) occur in the respective sequences. Because the simulated sequences are random, the stop codons will occur frequently.

```
# Positions of stop codons in simulated sequence (example)

sim_seq_n_stopcodons <- c()
liz_seq_n_stopcodons <- c()
for(i in 1:33){
  ### Get number of stop codons in *simulated* sequences
  n_taa <- as.character(sim_sequences[[i]] %>%
```

```
                          paste(collapse="")) %>%
    gregexpr(pattern="taa", text=.) %>%
    unlist() %>% length()

  n_tag <- as.character(sim_sequences[[i]] %>%
                          paste(collapse="")) %>%
    gregexpr(pattern="tag", text=.) %>%
    unlist() %>% length()

  n_tga <- as.character(sim_sequences[[i]] %>%
                          paste(collapse="")) %>%
    gregexpr(pattern="tga", text=.) %>%
    unlist() %>% length()

  sim_seq_n_stopcodons <- c(sim_seq_n_stopcodons, n_taa+n_tag+n_tga)



  ### Get number of stop codons in *true* sequences
  n_taa <- as.character(lizards_sequences)[[i]] %>%
                          paste(collapse="") %>%
    gregexpr(pattern="taa", text=.) %>%
    unlist() %>% length()

  n_tag <- as.character(lizards_sequences)[[i]] %>%
                          paste(collapse="") %>%
    gregexpr(pattern="tag", text=.) %>%
    unlist() %>% length()

  n_tga <- as.character(lizards_sequences)[[i]] %>%
                          paste(collapse="") %>%
    gregexpr(pattern="tga", text=.) %>%
    unlist() %>% length()

  liz_seq_n_stopcodons <- c(liz_seq_n_stopcodons, n_taa+n_tag+n_tga)
}
print(sim_seq_n_stopcodons)
```

```
##  [1]  59 149 174  57  87  65 165  62  78 189 177 184 185 173 160  62  66 179 188
## [20]  65  88 184  88  57 188 145  60 178 180  60 156  61  57
```

```
print(liz_seq_n_stopcodons)
```

```
##  [1]  52 131 140  50  78  61 145  57  60 143 138 144 143 144 132  54  55 151 144
## [20]  56  76 141  78  50 136 124  56 145 139  57 141  57  54
```

```
#cat("Number of stop codons in simulated sequences:\n", sim_seq_n_stopcodons, "\n")
#cat("Number of stop codons in true sequences:\n", liz_seq_n_stopcodons, "\n")

# Positions of stop codons in true sequenes
#as.character(lizards_sequences)[[1]] %>% paste(collapse = "") %>%
#  gregexpr(pattern="taa", text=.) %>% unlist()
```