Analysis

Qiuhan Zhang

2023-02-14

My Repository (https://github.com/qiuhan1008/BIOL432_Assignment6.git)

```
#load library
library(readr)
library(dplyr)
```

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

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

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

```
library(knitr)
```

```
#Import the Sequences.csv file.
Sequences <- read_csv("Sequences.csv")
```

```
## New names:
## Rows: 3 Columns: 3
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (2): Name, Sequence dbl (1): ...1
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...1`
```

##Count the number of each base pair (A, T, C and G), in each of the three sequences.

```
#convert to characters
seq_1 <- as.character(Sequences$Sequence[1])
seq_2 <- as.character(Sequences$Sequence[2])
seq_3 <- as.character(Sequences$Sequence[3])
```

```
#count the number of each base pair
##sequence 1
count1_A <- nchar(gsub("[^A]", "", seq_1))
count1_T <- nchar(gsub("[^T]", "", seq_1))
count1_C <- nchar(gsub("[^C]", "", seq_1))
count1_G <- nchar(gsub("[^G]", "", seq_1))

##sequence 2
count2_A <- nchar(gsub("[^A]", "", seq_2))
count2_T <- nchar(gsub("[^T]", "", seq_2))
count2_C <- nchar(gsub("[^C]", "", seq_2))
count2_G <- nchar(gsub("[^G]", "", seq_2))

##sequence 3
count3_A <- nchar(gsub("[^A]", "", seq_3))
count3_T <- nchar(gsub("[^T]", "", seq_3))
count3_C <- nchar(gsub("[^C]", "", seq_3))
count3_G <- nchar(gsub("[^G]", "", seq_3))
```

```
#Print out each sequence.
print(seq_1)
```

```
## [1] "AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGAT
GGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTG
TAGATGAGTCTGCGTCTTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTC
ACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACAC
TGCGTGAATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAA
TTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG"
```

```
print(seq_2)
```

```
## [1] "AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGAT
GGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTG
TAGATGAGTCTGCGTCTTATTAGCTAGTTGGTAGGGTAAATGCCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGGTGAACGGTC
ACACTGGAACTGAGATACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACAC
TGCGTGAATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACAAAGTGATGACGTTAA
TTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCAGCGGTAATACG"
```

```
print(seq_3)
```

```
## [1] "AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGAT
GGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTG
TAGATGAGTCTGCGTCTTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTC
ACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACAC
TGCGTGAATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAA
TTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG"
```

```
#Print out the number of each nucleotide as a table for each of the three sequences.
results <- data.frame(Sequence_Name = c("HQ433692.1", "HQ433694.", "HQ433691.1"),
                          A = c(count1_A, count2_A, count3_A),
                          T = c(count1_T, count2_T, count3_T),
                          C = c(count1_C, count2_C, count3_C),
                          G = c(count1_G, count2_G, count3_G))
print(results)
```

```
##   Sequence_Name   A   T  C   G
## 1    HQ433692.1 154 114 82 131
## 2     HQ433694. 155 114 81 131
## 3    HQ433691.1 154 115 81 131
```

##Include an image of a bacteria from the internet, and a link to the Wikipedia page about Borrelia burgdorferi

Photo of Borrelia burgdorferi (PIXNIO-38518-4252x2890.jpeg)

Borrelia burgdorferi wikipedia link (https://en.wikipedia.org/wiki/Borrelia_burgdorferi)

##Calculate GC Content (% of nucleotides that are G or C) and create a final table showing GC content for each sequence ID

```
GC_count <- results %>%
  group_by(Sequence_Name) %>%
  mutate(GC_count = ((C +G) / (A + T + C + G)) * 100) %>%
  select(Sequence_Name, GC_count)
GC_count
```

```
## # A tibble: 3 × 2
## # Groups:   Sequence_Name [3]
##   Sequence_Name GC_count
##   <chr>            <dbl>
## 1 HQ433692.1        44.3
## 2 HQ433694.         44.1
## 3 HQ433691.1        44.1
```

```
data1 <- read_csv("~/Desktop/data1.csv")
```

```
## Rows: 3 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): Sequence_Name, GC_count
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
table1 <- data1 %>%
  select(Sequence_Name, GC_count)
table1
```

```
## # A tibble: 3 × 2
##   Sequence_Name GC_count
##   <chr>         <chr>   
## 1 HQ433692.1    44.28%  
## 2 HQ433694.     44.07%  
## 3 HQ433694.     44.07%  
```