Assignment 5, Part 1

Chenyang Wu

2022/2/16

Project Info

GitHub user name: Wuris

Date: 2022/2/16

GitHub Link: https://github.com/Wuris/Rentrez.git

Import the Sequences.csv file.

```
setwd("E:/Underguaduate/4th Fourth year/BIOL 432/Week 6/Assignment 5 - Sequence Data Assignment/Rentrez
A5Data <- read.csv("Sequences.csv")
```

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

```
# Separate each base pair for each sequence.
seq1 <- strsplit(A5Data$Sequence, "")[[1]]
seq2 <- strsplit(A5Data$Sequence, "")[[2]]
seq3 <- strsplit(A5Data$Sequence, "")[[3]]

# Calculate the total number of base pair.
tot1 <- sum(nchar(seq1))
tot2 <- sum(nchar(seq2))
tot3 <- sum(nchar(seq3))
```

```
# Create the for loop to calculate the number of each base pair for each sequence.
a1 = a2 = a3 = 0
t1 = t2 = t3 = 0
c1 = c2 = c3 = 0
g1 = g2 = g3 = 0

for(i in seq1){
    if(i == "A"){a1 <- a1 + 1}
    if(i == "T"){t1 <- t1 + 1}
    if(i == "C"){c1 <- c1 + 1}
    if(i == "G"){g1 <- g1 + 1}
}

for(i in seq2){
    if(i == "A"){a2 <- a2 + 1}
    if(i == "T"){t2 <- t2 + 1}
    if(i == "C"){c2 <- c2 + 1}
```

```
    if(i == "G"){g2 <- g2 + 1}
}

for(i in seq3){
    if(i == "A"){a3 <- a3 + 1}
    if(i == "T"){t3 <- t3 + 1}
    if(i == "C"){c3 <- c3 + 1}
    if(i == "G"){g3 <- g3 + 1}
}
```

Print out each sequence.

```
print(A5Data$Sequence)
```

```
## [1] "AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAACTATTAGA
## [2] "AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAACTATTAGA
## [3] "AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAACTATTAGA
```

Print out the number of each nucleotide as a table for each of the three sequences.

```
# Create the table
Seqs <- c("Sequence 1", "Sequence 2", "Sequence 3")
id <- c("HQ433692.1", "HQ433694.1", "HQ433691.1")
A <- c(a1, a2, a3)
T <- c(t1, t2, t3)
C <- c(c1, c2, c3)
G <- c(g1, g2, g3)
Total <- c(tot1, tot2, tot3)

Table <- data.frame(Sequences = Seqs, IDs = id, A = A, T = T, C = C, G = G, Total = Total)
Table
```

```
##     Sequences        IDs   A   T  C   G Total
## 1 Sequence 1 HQ433692.1 154 114 82 131   481
## 2 Sequence 2 HQ433694.1 155 114 81 131   481
## 3 Sequence 3 HQ433691.1 154 115 81 131   481
```

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

Wikipedia link for Borrelia burgdorferi

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

```
# Calculate the GC content
gc <- (C + G)/Total
GC <- format(round(gc*100, 2))
gc_content <- paste(GC,"%", sep = "")

# Make the table
GCTable <- data.frame(Sequence_ID = id, GC_content = gc_content)
GCTable
```
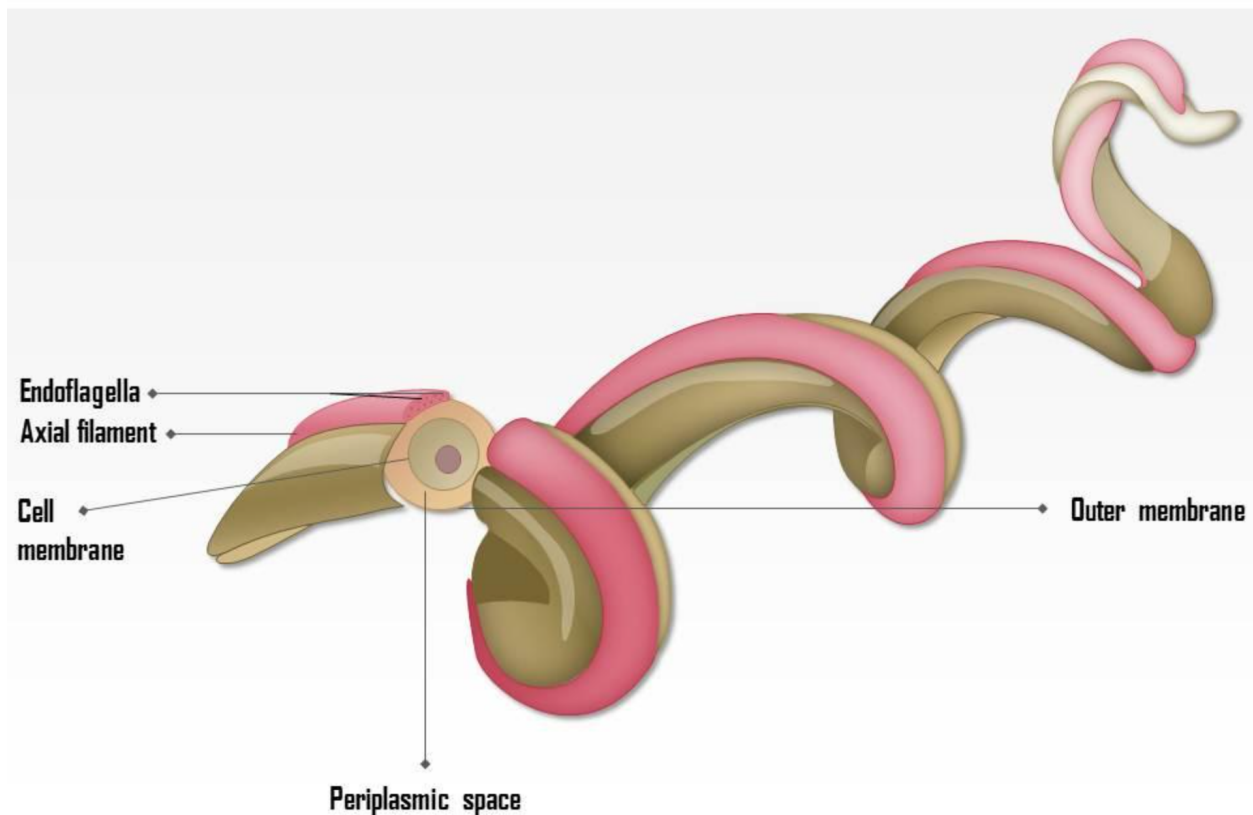
Borrelia Burgdorferi

Figure 1: Borrelia burgdorferi medical images from SlideTeam (https://www.slideteam.net/0714-borrelia-burgdorferi-medical-images-for-powerpoint.html). This medical image has been crafted with borrelia burgdorferi. Borrelia burgdorferi is a bacterial species of the spirochete class of the genus Borrelia. It is predominant causative agent of Lyme disease. This image contains structure of this bacteria and focuses on Endoflagella, Axial filament, Cell membrane and Periplasmic space in internal structure.

```
##   Sequence_ID GC_content
## 1  HQ433692.1     44.28%
## 2  HQ433694.1     44.07%
## 3  HQ433691.1     44.07%
```