—title: “BIOL432_A7” author: “Jennah Levac” date: “2025-02-24” output: html_document —

Jennah Levac (20169998), Grace Wolfe (20302888), Cameron DeBellefeuille (20324416), Harnaaz Sandhu (20258736)

Github link: https://github.com/Jennah2882/Rentrez/tree/main (https://github.com/Jennah2882/Rentrez/tree/main)

# Part 1

# Step 4

```
library(rentrez)
```

```
Data<- read.csv("Sequences.csv")
```

Creating a custom function to count the number of each nucleotide excluding headers, newlines, etc.

```
count_nucleotides <- function(seq) {
  seq <- gsub("[^ATCG]", "", seq)

  A_count <- nchar(gsub("[^A]", "", seq))
  T_count <- nchar(gsub("[^T]", "", seq))
  C_count <- nchar(gsub("[^C]", "", seq))
  G_count <- nchar(gsub("[^G]", "", seq))

  return(c(A = A_count, T = T_count, C = C_count, G = G_count))}
```

Creating an empty matrix to store nucleotide counts

```
nucleotide_counts <- matrix(0, nrow = nrow(Data), ncol = 4)
colnames(nucleotide_counts) <- c("A", "T", "C", "G")
```

Using a loop to apply counts to each sequence and converting to dataframe

```
for (i in 1:nrow(Data)) {
  nucleotide_counts[i, ] <- count_nucleotides(Data$Sequence[i])}

CountsDF <- data.frame(nucleotide_counts)
```

Printing each sequence

```
for (i in 1:nrow(Data)) {
  print(paste("Sequence ID:", Data$Name[i]))
  print(Data$Sequence[i])}
```

```
## [1] "Sequence ID: >HQ433692.1 Borrelia burgdorferi strain QLZP1 16S ribosomal RNA gene, parti
al sequence"
## [1] "AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTC
TTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGACACGGTCCA
GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAA
ATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGC
GGTAATACG"
## [1] "Sequence ID: >HQ433694.1 Borrelia burgdorferi strain CS4 16S ribosomal RNA gene, partial
sequence"
## [1] "AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTC
TTATTAGCTAGTTGGTAGGGTAAATGCCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGATACGGTCCA
GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAA
ATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACAAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCAGC
GGTAATACG"
## [1] "Sequence ID: >HQ433691.1 Borrelia burgdorferi strain GL18 16S ribosomal RNA gene, partia
l sequence"
## [1] "AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTC
TTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGACACGGTCCA
GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAA
ATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGC
GGTAATACG"
```

```
print(CountsDF)
```

```
##     A   T  C   G
## 1 154 114 82 131
## 2 155 114 81 131
## 3 154 115 81 131
```

Borrelia burgdorferi

Link: https://en.wikipedia.org/wiki/Borrelia_burgdorferi (https://en.wikipedia.org/wiki/Borrelia_burgdorferi)

Calculating GC content as percentages

```
calculate_gc_content <- function(seq) {
  seq <- gsub("[^ATCG]", "", seq)

  G_count <- nchar(gsub("[^G]", "", seq))
  C_count <- nchar(gsub("[^C]", "", seq))

  total_length <- nchar(seq)

  gc_content <- (G_count + C_count) / total_length * 100

  return(gc_content)}
```

Creating GC content table

```
Data$ID <- gsub("^>(.*?)(\\s|$).*", "\\1", Data$Name)

Data$GC_Content <- sapply(Data$Sequence, calculate_gc_content)

gc_content_table <- data.frame(
  Sequence_ID = Data$ID,
  GC_Content = paste(round(Data$GC_Content, 2), "%", sep = ""))

print(gc_content_table)
```

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

# Part 2

```
library(Biostrings)
library(annotate)
```

```
MySequence <- "GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGC
CTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAG
CCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA"
```

Performing blast to identify matching sequences

```
blast_results <- blastSequences(MySequence, as='data.frame',
                                timeout=240,
                                hitListSize=5)
```

Excluding duplicates

```
New <- data.frame(blast_results[!duplicated(blast_results$Hit_accession), ])
```

GC content of each hit

```
gc_content <- function(sequence) {
  g_count <- sum(grepl("G", strsplit(sequence, NULL)[[1]]))
  c_count <- sum(grepl("C", strsplit(sequence, NULL)[[1]]))

  g_c_count <- g_count + c_count
  return((g_c_count / nchar(sequence)) * 100)}

New$gc_content <- sapply(New$Hsp_hseq, gc_content)

gc_content_table2 <- data.frame(
  Sequence_ID = New$Hit_accession,
  GC_Content = paste(round(New$gc_content, 2), "%", sep = ""))

print(gc_content_table2)
```

```
##   Sequence_ID GC_Content
## 1    CP177420      56.4%
## 2    CP177416      56.4%
## 3    CP177432      56.4%
## 4    CP177412      56.4%
## 5    CP177464      56.4%
```

BLAST results

```
relevant_columns <- New[,c(10, 11, 14, 15)]
print(relevant_columns)
```

```
##    Hit_accession Hit_len Hsp_score   Hsp_evalue
## 1       CP177420 4528383       500 1.35431e-122
## 8       CP177416 4530751       500 1.35431e-122
## 15      CP177432 4598040       500 1.35431e-122
## 21      CP177412 4527029       500 1.35431e-122
## 28      CP177464 4612071       500 1.35431e-122
```

We can see from the results above in addition to similar GC content that the hits are very closely similar to our unknown sequence.

```
print(New$Hit_def)
```

```
## [1] "Yersinia pestis strain 41003 chromosome, complete genome"
## [2] "Yersinia pestis strain 41006 chromosome, complete genome"
## [3] "Yersinia pestis strain N010076 chromosome, complete genome"
## [4] "Yersinia pestis strain 38004 chromosome, complete genome"
## [5] "Yersinia pestis strain 101010 chromosome, complete genome"
```

Now we can confirm the sequence is not human DNA.

*Yersinia pestis* is a pathogenic bacteria capable of surviving at temperatures similar to the internal mammalian environment. It has a large number of plasmid-borne virulence factors and is responsible for plague, with the most virulent form being pneumonic. Other forms of plague associated with this bacteria are bubonic and septicemic.

The ability for this bacteria to avoid immune detection makes it infectious with great potential to progress to sepsis. Due to its extremely high mortality rate and virulence when left untreated, this is cause for concern. Any patient who is suspected of plague should undergo further testing until cleared. Additionally, patients exposed should be isolated and public health officials should be notified to investigate the source of infection.

The most common route of transmission is via fleas that have fed on infected rodents. The Pneumonic plague however is contagious and can be spread person-to-person through respiratory droplets. Due to the high virulence and potential for person-to-person spread Yersinia pestis poses public health concerns.

Several antibiotics have been shown to effectively treat *Yersinia pestis* infections. These antibiotics include tetracyclines, fluoroquinolones, aminoglycosides, sulfonamides, chloramphenicol, rifamycin, and β-lactams, however, there has been mixed results about the efficacy of each antibiotic depending on the strain.