Analysis.Rmd

Roshael Chellappah (20103016)

16/02/2022

```
MyData <- read.csv("./Sequences.csv")
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
```

Properties of the Sequences

```
for (i in 1:nrow(MyData)){
  ID <- MyData[i, "Name"]
  Sequence <- MyData[i, "Sequence"]

  print(paste("Sequence ID:", substr(ID, 2, 44))) # identify the ID of the sequence
  print(Sequence) # prints the nucleotide sequence
  print(table(strsplit(Sequence, ""))) # prints table with nucleotides and number of occurrences
within the sequence
}
```

```
## [1] "Sequence ID: HQ433692.1 Borrelia burgdorferi strain QLZP"
## [1] "AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTC
TTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGACACGGTCCA
GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAA
ATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGC
GGTAATACG"
##
##   A   C   G   T
## 154  82 131 114
## [1] "Sequence ID: HQ433694.1 Borrelia burgdorferi strain CS4 "
## [1] "AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTC
TTATTAGCTAGTTGGTAGGGTAAATGCCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGATACGGTCCA
GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAA
ATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACAAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCAGC
GGTAATACG"
##
##   A   C   G   T
## 155  81 131 114
## [1] "Sequence ID: HQ433691.1 Borrelia burgdorferi strain GL18"
## [1] "AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTC
TTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGACACGGTCCA
GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAA
ATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGC
GGTAATACG"
##
##   A   C   G   T
## 154  81 131 115
```

Calculate GC Content

```
df1 <- data.frame(table(strsplit(MyData$Sequence, "")[[1]]))
gc_cont1 <- sum(df1$Freq[2:3]) # total number of G & C nucleotides
nclt_total1<- sum(df1$Freq) # sum of all nucleotides
gc_percent1 <- round((gc_cont1/nclt_total1) * 100, digits = 2)

df2 <- data.frame(table(strsplit(MyData$Sequence, "")[[2]]))
gc_cont2 <- sum(df2$Freq[2:3]) # total number of G & C nucleotides
nclt_total2 <- sum(df2$Freq) # sum of all nucleotides
gc_percent2 <- round((gc_cont2/nclt_total2) * 100, digits = 2)

df3 <- data.frame(table(strsplit(MyData$Sequence, "")[[3]]))
gc_cont3 <- sum(df3$Freq[2:3]) # total number of G & C nucleotides
nclt_total3 <- sum(df3$Freq) # sum of all nucleotides
gc_percent3 <- round((gc_cont3/nclt_total3) * 100, digits = 2)

GC_Content <- data.frame(Sequence_ID = c("HQ433692.1", "HQ433694.1", "HQ433691.1"),
                         GC_Content = c(gc_percent1, gc_percent2, gc_percent3))
GC_Content
```

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

Borrelia burgorferi

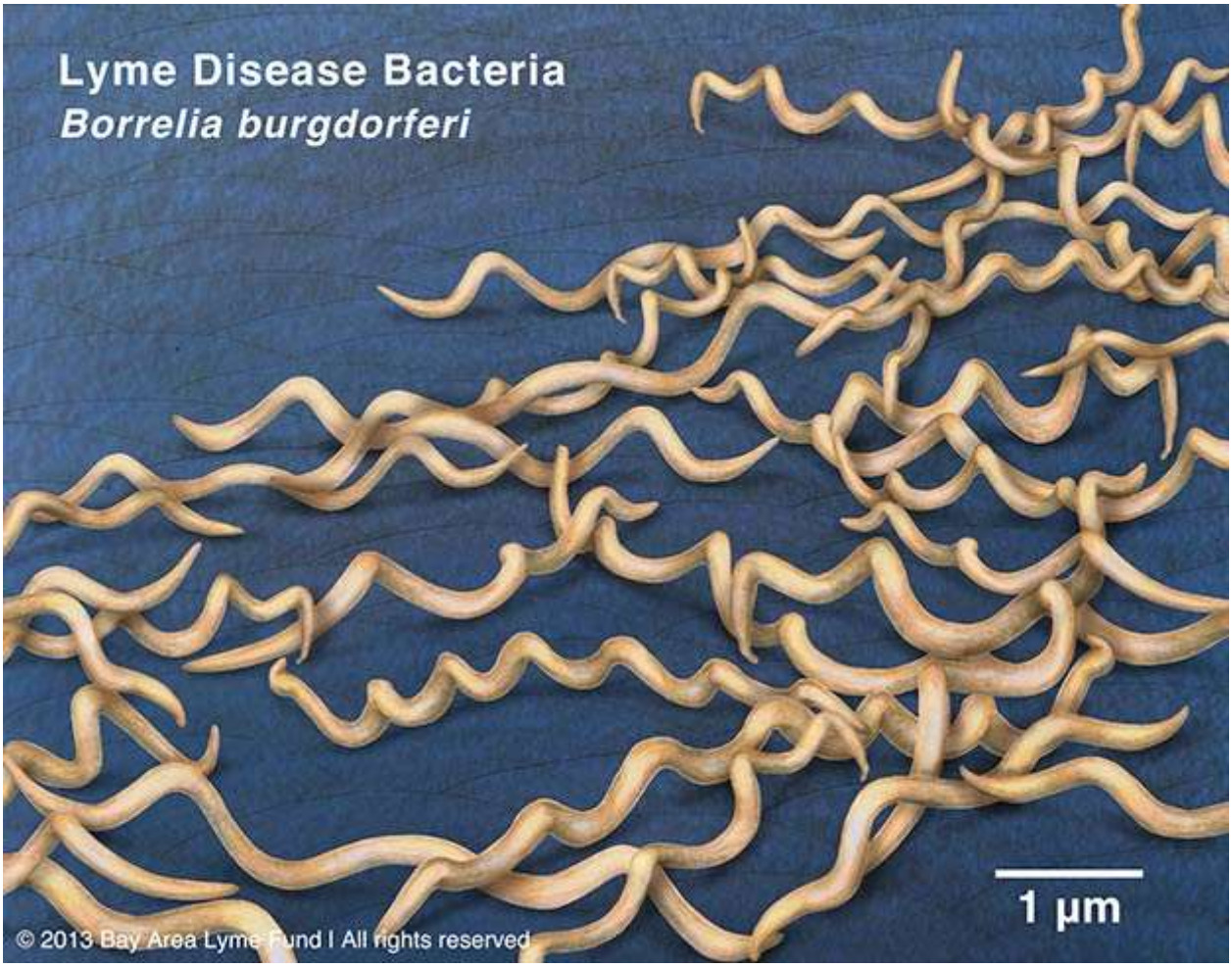

Image of *Borrelia burgdorferi* bacteria

Link to Wikipedia page on *Borrelia burgdorferi* (https://en.wikipedia.org/wiki/Borrelia_burgdorferi)