Analysis

Tianyi Zuo

2022/2/16

GitHub username: Lydia12138

Date: 2022-01-26 Repo:https://github.com/Lydia12138/Rentrez (https://github.com/Lydia12138/Rentrez)

load the packages

```
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

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

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

Import the Sequences.csv file

```
SeqData<-read.csv("output/Sequences.csv")#loading the data
str(SeqData) #check the structure of the Data
```

```
## 'data.frame':    3 obs. of  2 variables:
##  $ Name    : chr  ">HQ433692.1 Borrelia burgdorferi strain QLZP1 16S ribosomal RNA
gene, partial sequence" ">HQ433694.1 Borrelia burgdorferi strain CS4 16S ribosomal RN
A gene, partial sequence" ">HQ433691.1 Borrelia burgdorferi strain GL18 16S ribosomal
RNA gene, partial sequence"
##  $ Sequence: chr  "AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGA
TGATCTACCTATGAGATGGGGATAACTATTAGAAATAGTAGCTAATAC"| __truncated__ "AGCATGCAAGTCAAACGGG
ATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATAACTATTAGAAATAGTAG
CTAATAC"| __truncated__ "AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGT
GGATGATCTACCTATGAGATGGGGATAACTATTAGAAATAGTAGCTAATAC"| __truncated__
```

```
class(SeqData)
```

```
## [1] "data.frame"
```

```
dim(SeqData)
```

```
## [1] 3 2
```

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

```
# extract the certain sequence from table
seq1 <- strsplit(SeqData$Sequence,"")[[1]]
seq2 <- strsplit(SeqData$Sequence,"")[[2]]
seq3 <- strsplit(SeqData$Sequence,"")[[3]]



# Each base pair content in sequence 1
A1_number <- length(grep("A",seq1))
T1_number <- length(grep("T",seq1))
C1_number <- length(grep("C",seq1))
G1_number <- length(grep("G",seq1))

# Each base pair content in sequence 2
A2_number <- length(grep("A",seq2))
T2_number <- length(grep("T",seq2))
C2_number <- length(grep("C",seq2))
G2_number <- length(grep("G",seq2))

# Each base pair content in sequence 3
A3_number <- length(grep("A",seq3))
T3_number <- length(grep("T",seq3))
C3_number <- length(grep("C",seq3))
G3_number <- length(grep("G",seq3))
```

Print out each sequence

```
print (unlist(SeqData$Sequence))
```

```
## [1] "AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGA
GATGGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTT
CGCTTGTAGATGAGTCTGCGTCTTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGG
TGAACGGTCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTG
ACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACG
AAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG"
## [2] "AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGA
GATGGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTT
CGCTTGTAGATGAGTCTGCGTCTTATTAGCTAGTTGGTAGGGTAAATGCCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGG
TGAACGGTCACACTGGAACTGAGATACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTG
ACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACA
AAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCAGCGGTAATACG"
## [3] "AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGA
GATGGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTT
CGCTTGTAGATGAGTCTGCGTCTTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGG
TGAACGGTCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTG
ACGGAGCGACACTGCGTGAATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACG
AAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG"
```

Create a table with number of each nucleotide for each of the three sequences

```
Sequences_Id <- c("HQ433692.1","HQ433694.1","HQ433691.1")
A_content <- c(A1_number, A2_number, A3_number)
T_content <- c(T1_number, T2_number, T3_number)
C_content <- c(C1_number, C2_number, C3_number)
G_content <- c(G1_number, G2_number, G3_number)

SumTable<-data.frame(Sequences_Id, A_content, T_content, C_content, G_content, Total
 = nchar(SeqData$Sequence))
print(SumTable)
```

```
##   Sequences_Id A_content T_content C_content G_content Total
## 1   HQ433692.1       154       114        82       131   481
## 2   HQ433694.1       155       114        81       131   481
## 3   HQ433691.1       154       115        81       131   481
```

Upload Image of a bacteria from the internet, and a link to the Wikipedia page about Borrelia burgdorferi

Lyme Disease Bacteria : Borrelia burgdorferi, Image courtesy of Emily M. Eng

link to the Wikipedia page about Borrelia burgdorferi (https://en.wikipedia.org/wiki/Borrelia_burgdorferi)

Create a final table showing GC content for each sequence ID

```
# Calculate GC Content
FinalTable <- transmute(SumTable, Sequences_Id, GC_Content =  paste(round((C_content
 +G_content)/Total *100, 2), "%"))
print (FinalTable)
```

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