Assignment7

2025-10-21

```
knitr::opts_chunk$set(echo = TRUE)
options(repos = c(CRAN = ″https://cloud.r-project.org″))
install.packages(″rentrez″)
```

```
## Installing package into 'C:/Users/zyd11/AppData/Local/R/win-library/4.3’
## (as 'lib’ is unspecified)
```

```
##
##   There is a binary version available but the source version is later:
##         binary source needs_compilation
## rentrez  1.2.3  1.2.4             FALSE
```

```
## installing the source package 'rentrez’
```

```
library(rentrez)
library(ggplot2)
```

```
## Warning: package 'ggplot2’ was built under R version 4.3.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr’ was built under R version 4.3.3
```

```
##
## 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
```

```
NcbiIds <- c("HQ433692.1","HQ433694.1","HQ433691.1")
Bburg <- entrez_fetch(db = "nuccore", id = NcbiIds, rettype = "fasta")
cat(Bburg)
```

```
## >HQ433692.1 Borrelia burgdorferi strain QLZP1 16S ribosomal RNA gene, partial sequence
## AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTA
## CCTATGAGATGGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGA
## TGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTCTTATTAGTTAGTTGGTAGGGTAAATG
## CCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGACACGGTCCA
## GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTG
## AATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAG
## TGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG
##
## >HQ433694.1 Borrelia burgdorferi strain CS4 16S ribosomal RNA gene, partial sequence
## AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTA
## CCTATGAGATGGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGA
## TGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTCTTATTAGCTAGTTGGTAGGGTAAATG
## CCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGATACGGTCCA
## GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTG
## AATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACAAAG
## TGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCAGCGGTAATACG
##
## >HQ433691.1 Borrelia burgdorferi strain GL18 16S ribosomal RNA gene, partial sequence
## AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTA
## CCTATGAGATGGGGATAACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGA
## TGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCTGCGTCTTATTAGTTAGTTGGTAGGGTAAATG
## CCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGACACGGTCCA
## GACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTG
## AATGAAGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAG
## TGATGACGTTAATTTATGAATAAGCCCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG
```

```
TempList <- strsplit(Bburg, split = ">")
TempVector <- TempList[[1]]
NoEmpty <- c()
for (i in 1:length(TempVector)) {
   if (TempVector[i] != "") {
     NoEmpty <- c(NoEmpty, TempVector[i])
   }
}
SequencesRaw <- NoEmpty
print(SequencesRaw)
```

```
## [1] "HQ433692.1 Borrelia burgdorferi strain QLZP1 16S ribosomal RNA gene, partial sequenc
e\nAGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTA\nCCTATGAGATGGGGATA
ACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGA\nTGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGT
CTGCGTCTTATTAGTTAGTTGGTAGGGTAAATG\nCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACT
GAGACACGGTCCA\nGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTG\nAATGA
AGAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAG\nTGATGACGTTAATTTATGAATAAGC
CCCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG\n\n"
## [2] "HQ433694.1 Borrelia burgdorferi strain CS4 16S ribosomal RNA gene, partial sequence
\nAGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTA\nCCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGA\nTGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTC
TGCGTCTTATTAGCTAGTTGGTAGGGTAAATG\nCCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTG
AGATACGGTCCA\nGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTG\nAATGAA
GAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACAAAG\nTGATGACGTTAATTTATGAATAAGCC
CCGGCTAATTACGTGCCAGCAGCAGCGGTAATACG\n\n"
## [3] "HQ433691.1 Borrelia burgdorferi strain GL18 16S ribosomal RNA gene, partial sequence
\nAGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTA\nCCTATGAGATGGGGATAA
CTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGA\nTGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTC
TGCGTCTTATTAGTTAGTTGGTAGGGTAAATG\nCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTG
AGACACGGTCCA\nGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTG\nAATGAA
GAAGGTCGAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAG\nTGATGACGTTAATTTATGAATAAGCC
CCGGCTAATTACGTGCCAGCAGCCGCGGTAATACG\n\n"
```

```
Seq1Parts <- strsplit(SequencesRaw[1], "\n")[[1]]
Seq2Parts <- strsplit(SequencesRaw[2], "\n")[[1]]
Seq3Parts <- strsplit(SequencesRaw[3], "\n")[[1]]
header1 <- Seq1Parts[1]
header2 <- Seq2Parts[1]
header3 <- Seq3Parts[1]
sequence1 <- paste(Seq1Parts[-1], collapse = "")
sequence2 <- paste(Seq2Parts[-1], collapse = "")
sequence3 <- paste(Seq3Parts[-1], collapse = "")

Name <- c(header1, header2, header3)
Sequence <- c(sequence1, sequence2, sequence3)
Sequences <- data.frame(Name, Sequence)
Sequences
```

```
##                                                                                  Name
## 1 HQ433692.1 Borrelia burgdorferi strain QLZP1 16S ribosomal RNA gene, partial sequence
## 2   HQ433694.1 Borrelia burgdorferi strain CS4 16S ribosomal RNA gene, partial sequence
## 3  HQ433691.1 Borrelia burgdorferi strain GL18 16S ribosomal RNA gene, partial sequence
##
Sequence
## 1 AGCATGCAAGTCAAACGAGATGTAGCAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATA
ACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCT
GCGTCTTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGA
CACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTC
GAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAAT
TACGTGCCAGCAGCCGCGGTAATACG
## 2 AGCATGCAAGTCAAACGGGATGTAGCAATACATTCAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATA
ACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAGTTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCT
GCGTCTTATTAGCTAGTTGGTAGGGTAAATGCCTACCAAGGCAATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGA
TACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTC
GAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACAAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAAT
TACGTGCCAGCAGCAGCGGTAATACG
## 3 AGCATGCAAGTCAAACGAGATGTAGTAATACATCTAGTGGCGAACGGGTGAGTAACGCGTGGATGATCTACCTATGAGATGGGGATA
ACTATTAGAAATAGTAGCTAATACCGAATAAGGTCAATTAATTTGTTAATTGATGAAAGGAAGCCTTTAAAGCTTCGCTTGTAGATGAGTCT
GCGTCTTATTAGTTAGTTGGTAGGGTAAATGCCTACCAAGGCGATGATAAGTAACCGGCCTGAGAGGGTGAACGGTCACACTGGAACTGAGA
CACGGTCCAGACTCCTACGGGAGGCAGCAGCTAAGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCGACACTGCGTGAATGAAGAAGGTC
GAAAGATTGTAAAATTCTTTTATAAATGAGGAATAAGCTTTGTAGGAAATGACGAAGTGATGACGTTAATTTATGAATAAGCCCCGGCTAAT
TACGTGCCAGCAGCCGCGGTAATACG
```

```
UnknownSeq <- paste0(
  "GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGAC",
  "CTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCTCACGCCATCGGATGAAC",
  "CCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGAT",
  "CCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGT",
  "CCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA"
)
cat("UnknownSeq Length: ", nchar(UnknownSeq), "bp\n")
```

```
## UnknownSeq Length:  250 bp
```

```
SearchTerm <- "Borrelia burgdorferi 16S[Title] OR 16S ribosomal RNA"
Results <- entrez_search(db = "nuccore", term = SearchTerm, retmax = 3)
Results
```

```
## Entrez search result with 8546129 hits (object contains 3 IDs and no web_history object)
##  Search term (as translated):  Borrelia burgdorferi 16S[Title] OR 16s ribosomal r ...
```

```
TopIds <- Results$ids
fetched <- entrez_fetch(db = "nuccore", id = TopIds, rettype = "fasta")
Parts <- strsplit(fetched, ">")[[1]]
Parts <- Parts[Parts != ""]
ref1 <- strsplit(Parts[1], "\n")[[1]]
RefHeader <- ref1[1]
RefSeq <- paste(ref1[-1], collapse = "")
cat("Reference Seq Name: ", RefHeader, "\n")
```

```
## Reference Seq Name:  PX457416.1 Paenibacillus polymyxa strain S437 16S ribosomal RNA gen
e, partial sequence
```

```
cat("Reference Seq Length: ", nchar(RefSeq), "bp\n")
```

```
## Reference Seq Length:  1490 bp
```

```
Length <- min(nchar(RefSeq), nchar(UnknownSeq))
RefBases <- strsplit(RefSeq, "")[[1]][1:Length]
UnkBases <- strsplit(UnknownSeq, "")[[1]][1:Length]

AAA <- 0
for (i in 1:Length) {
  if (RefBases[i] == UnkBases[i]) {
    AAA <- AAA + 1
  }
}
SeqLength <- nchar(UnknownSeq)
SeqSimilarity <- round(AAA / SeqLength * 100, 2)
cat(" Seq Similarity: ", SeqSimilarity, "%\n")
```

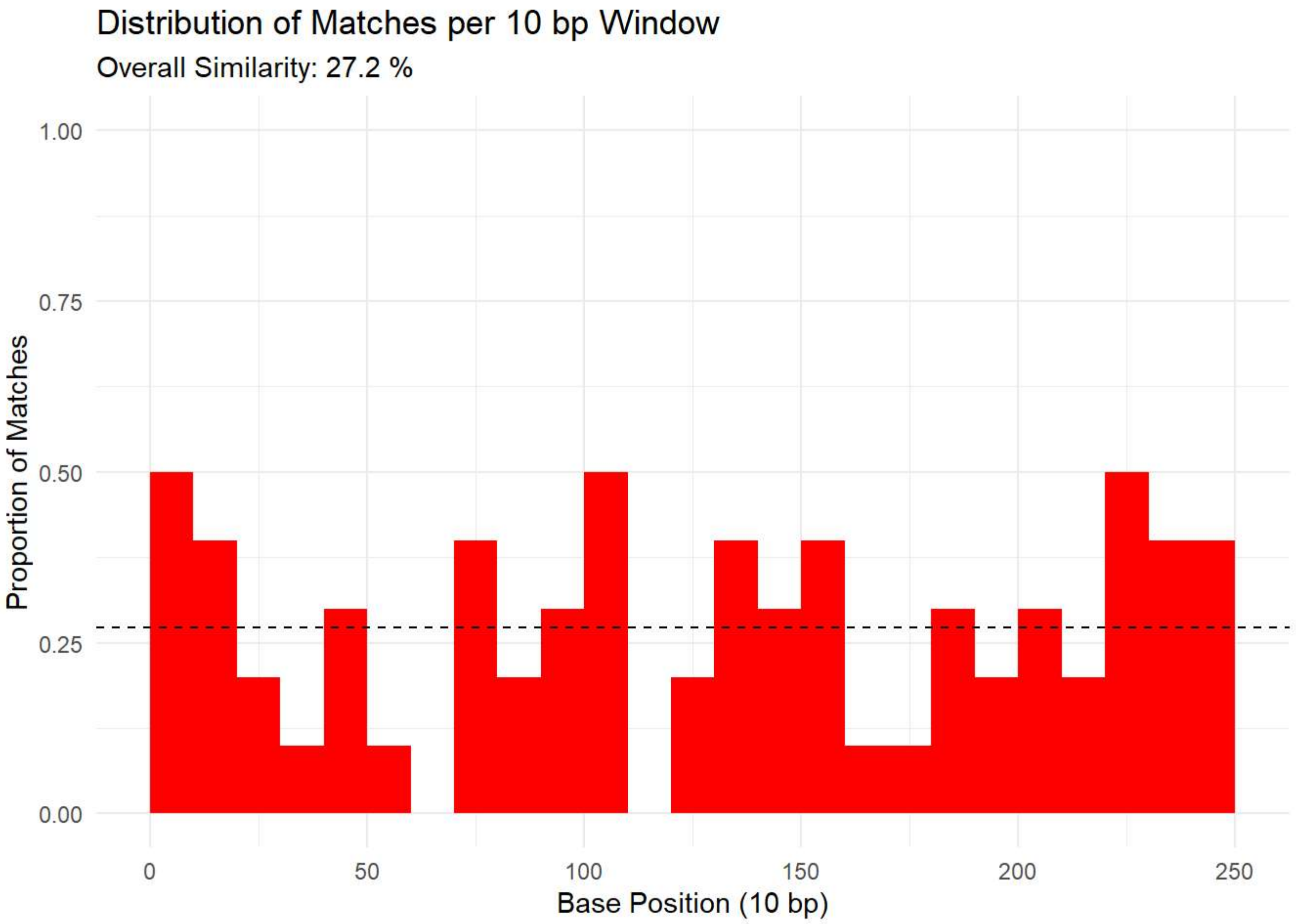
```
##  Seq Similarity:  27.2 %
```

##Unknown_Length: Unknown sequence length (250 bp) ##Reference_Length: Total length of reference 16S sequence (1490 bp) ##Similarity_Percent: The similarity calculated after base-by-base alignment (27.2%) ##Reference_Name: Species name and sequence ID from GenBank

```
MatchStatus <- ifelse(RefBases == UnkBases, 1, 0)
df <- data.frame(Position = 1:Length, Match = MatchStatus)

df_summary <- df %>%
  mutate(group = ceiling(Position / 10)) %>%
  group_by(group) %>%
  summarise(Match_Rate = mean(Match), .groups = "drop")

ggplot(df_summary, aes(x = (group - 0.5) * 10, y = Match_Rate)) +
  geom_col(width = 10, fill = "red") +
  geom_hline(yintercept = mean(df$Match), linetype = "dashed", color = "black") +
  scale_x_continuous(limits = c(0, Length), breaks = seq(0, Length, by = 50)) +
  scale_y_continuous(limits = c(0, 1)) +
  labs(title = "Distribution of Matches per 10 bp Window",
       subtitle = paste("Overall Similarity:", SeqSimilarity, "%"),
       x = "Base Position (10 bp)",
       y = "Proportion of Matches") +
  theme_minimal()
```



```
if (SeqSimilarity > 80) {
  message("The sequence is very similar(high similarity)(SeqSimilarity: ",  SeqSimilarity,
"%) to reference bacterial sequences and is likely from the genus Borrelia.")
} else {
  message("This sequence is quite different(low similarity)(SeqSimilarity: ",  SeqSimilarit
y, "%) from the reference bacterial sequence and may be of human or other origin.")
}
```

```
## This sequence is quite different(low similarity)(SeqSimilarity: 27.2%) from the reference
bacterial sequence and may be of human or other origin.
```

```
Result1 <- data.frame(Unknown_Length = nchar(UnknownSeq),
                     Reference_Length = nchar(RefSeq),
                     Similarity_Percent = SeqSimilarity,
                     Reference_Name = RefHeader)
Result1
```

```
##   Unknown_Length Reference_Length Similarity_Percent
## 1            250             1490               27.2
##                                                                            Reference_Name
## 1 PX457416.1 Paenibacillus polymyxa strain S437 16S ribosomal RNA gene, partial sequence
```