BloodSample

Qiuhan Zhang

2023-02-14

My Repository (https://github.com/qiuhan1008/BIOL432_Assignment6.git)

```
#install.packages("BiocManager")
#install.packages("Biostrings")
#install.packages("annotater")
#install("genbankr")

#load library
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
```

```
library(BiocManager)
library(Biostrings)
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, rename
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges
```

```
##
## Attaching package: 'IRanges'
```

```
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
```

```
## Loading required package: XVector
```

```
## Loading required package: GenomeInfoDb
```

```
##
## Attaching package: 'Biostrings'
```

```
## The following object is masked from 'package:base':
##
##     strsplit
```

```
library(genbankr)
library(rentrez)
```

```
#load unknown sequence
unknseq <- "GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGAC
CTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCTCACGCCATCGGATGAAC
CCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGAT
CCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGT
CCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA"
```

```
#remove the 'carriage return' and 'newline' special character
unknseq <- gsub("[\r\n]", "", unknseq)
unknseq
```

```
## [1] "GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAG
GATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA"
```

```
library(annotate)
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'AnnotationDbi'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: XML
```

```
useqBLAST <- blastSequences(paste(unknseq),as = 'data.frame',
                            hitListSize = 20, timeout = 600 )
```

```
## estimated response time 21 seconds
```

```
## elapsed time 21 seconds
```

```
## elapsed time 32 seconds
```

```
## elapsed time 42 seconds
```

```
## elapsed time 53 seconds
```

```
## elapsed time 64 seconds
```

```
## elapsed time 74 seconds
```

```
## elapsed time 85 seconds
```

```
## elapsed time 96 seconds
```

```
## elapsed time 106 seconds
```

```
## elapsed time 117 seconds
```

```
## elapsed time 128 seconds
```

```
## elapsed time 138 seconds
```

```
## elapsed time 149 seconds
```

```
## elapsed time 160 seconds
```

```
## elapsed time 170 seconds
```

```
## elapsed time 181 seconds
```

```
## elapsed time 192 seconds
```

```
## elapsed time 202 seconds
```

```
## elapsed time 213 seconds
```

```
## elapsed time 224 seconds
```

```
## elapsed time 234 seconds
```

```
## elapsed time 245 seconds
```

```
## elapsed time 256 seconds
```

```
## elapsed time 266 seconds
```

```
## elapsed time 277 seconds
```

```
## elapsed time 288 seconds
```

```
## elapsed time 299 seconds
```

```
## elapsed time 310 seconds
```

```
## elapsed time 320 seconds
```

```
#Download ape library from CRAN repository
#install.packages("ape")

#Use library command to make ape functions accessible by this script
library(ape)
```

```
##
## Attaching package: 'ape'
```

```
## The following object is masked from 'package:Biostrings':
##
##     complement
```

```
## The following object is masked from 'package:dplyr':
##
##     where
```

```
# create a DNAbin object
useqHitsDF <- data.frame(ID = useqBLAST$Hit_accession, # specifying an ID column
                         Seq = useqBLAST$Hsp_hseq,
                         stringsAsFactors = FALSE)
```

```
 # length of each sequence
useqBLAST$Hit_len
```

```
##   [1] "4553685" "4553685" "4553685" "4553685" "4553685" "4553685" "4636015"
##   [8] "4636015" "4636015" "4636015" "4636015" "4647610" "4647610"
##  [15] "4647610" "4647610" "4647610" "4647610" "4647610" "4648824" "4648824"
##  [22] "4648824" "4648824" "4648824" "4648824" "4648824" "4731909" "4731909"
##  [29] "4731909" "4731909" "4731909" "4731909" "4731909" "4601712" "4601712"
##  [36] "4601712" "4601712" "4601712" "4601712" "4601712" "4636545" "4636545"
##  [43] "4636545" "4636545" "4636545" "4636545" "4636545" "4481542" "4481542"
##  [50] "4481542" "4481542" "4481542" "4481542" "4481542" "4612530" "4612530"
##  [57] "4612530" "4612530" "4612530" "4612530" "4612530" "4553687" "4553687"
##  [64] "4553687" "4553687" "4553687" "4553687" "4501826" "4501826" "4501826"
##  [71] "4501826" "4501826" "4501826" "4501826" "4644349" "4644349" "4644349"
##  [78] "4644349" "4644349" "4644349" "4644349" "4612708" "4612708" "4612708"
##  [85] "4612708" "4612708" "4612708" "4612708" "4616538" "4616538" "4616538"
##  [92] "4616538" "4616538" "4616538" "4616538" "4555347" "4555347" "4555347"
##  [99] "4555347" "4555347" "4555347" "4555347" "4499502" "4499502" "4499502"
## [106] "4499502" "4499502" "4499502" "4499502" "4617857" "4617857" "4617857"
## [113] "4617857" "4617857" "4617857" "4617857" "4604910" "4604910" "4604910"
## [120] "4604910" "4604910" "4604910" "4604910" "4610990" "4610990" "4610990"
## [127] "4610990" "4610990" "4610990" "4610990" "4639390" "4639390" "4639390"
## [134] "4639390" "4639390" "4639390" "4639390"
```

Those 137 sequences have similiar number of base pairs.

##To determine whether it's human or other organisms

```
results <- useqBLAST %>%
dplyr::select(Hit_def,
       Hit_accession,
       Hit_len,
       Hsp_score,
       Hsp_evalue,
       Hsp_gaps,
       Hsp_qseq)%>%
  arrange(desc(Hsp_score), .by_group = TRUE)
head(results)
```

```
##                                                  Hit_def Hit_accession Hit_len
## 1   Yersinia pestis EV76-CN chromosome, complete genome      CP096666 4553685
## 2   Yersinia pestis EV76-CN chromosome, complete genome      CP096666 4553685
## 3   Yersinia pestis EV76-CN chromosome, complete genome      CP096666 4553685
## 4   Yersinia pestis EV76-CN chromosome, complete genome      CP096666 4553685
## 5 Yersinia pestis strain 20 chromosome, complete genome      CP084343 4636015
## 6 Yersinia pestis strain 20 chromosome, complete genome      CP084343 4636015
##   Hsp_score   Hsp_evalue Hsp_gaps
## 1       500 1.70358e-122        0
## 2       500 1.70358e-122        0
## 3       500 1.70358e-122        0
## 4       500 1.70358e-122        0
## 5       500 1.70358e-122        0
## 6       500 1.70358e-122        0
##
Hsp_qseq
## 1 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCT
CACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGAT
GACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 2 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCT
CACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGAT
GACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 3 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCT
CACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGAT
GACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 4 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCT
CACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGAT
GACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 5 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCT
CACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGAT
GACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 6 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCT
CACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGAT
GACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
```

The isolated sequence suggests the unknown sequence is Yersinia pestis chromosome instead of a human genome.