

BIOL432_Assignment 6

SiruiZHAO

2023-02-14

Part II

First, we need to load library

```
library(annotater)
library(BiocManager)
library(annotate)
```

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

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

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

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

```
## 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: Biobase
```

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

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

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

```
## Warning: package 'S4Vectors' was built under R version 4.2.2
```

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

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

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

```
library(ape)  
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'
```

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

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

```
## The following objects are masked from 'package:IRanges':  
##  
## collapse, desc, intersect, setdiff, slice, union
```

```
## The following objects are masked from 'package:S4Vectors':  
##  
## first, intersect, rename, setdiff, setequal, union
```

```
## The following object is masked from 'package:Biobase':  
##  
## combine
```

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

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

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

Then, copy and paste the sequence we found

```
Sequences <- "GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGAC  
CTCGCAAGAGCAAAGTGGGGGACCTTAGGGCCTCACGCCATCGGATGAAC  
CCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGAT  
CCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAAGTGGAGACACGGT  
CCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA"
```

```
#remove special character  
Sequences <- gsub("[\r\n]", "", Sequences)
```

There will be 33 sequences matched when we choose 5 hits.

```
seqBLAST <- blastSequences(paste(Sequences), as = 'data.frame',  
                           hitListSize = 5, timeout = 600)
```

```
## estimated response time 29 seconds
```

```
## elapsed time 29 seconds
```

```
## elapsed time 40 seconds
```

```
## elapsed time 51 seconds
```

```
## elapsed time 62 seconds
```

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

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

elapsed time 96 seconds

elapsed time 107 seconds

elapsed time 118 seconds

elapsed time 129 seconds

elapsed time 140 seconds

elapsed time 151 seconds

elapsed time 163 seconds

elapsed time 173 seconds

elapsed time 184 seconds

elapsed time 195 seconds

elapsed time 205 seconds

elapsed time 216 seconds

elapsed time 227 seconds

elapsed time 238 seconds

elapsed time 249 seconds

elapsed time 260 seconds

elapsed time 272 seconds

elapsed time 282 seconds

elapsed time 294 seconds

elapsed time 305 seconds

elapsed time 315 seconds

elapsed time 326 seconds

elapsed time 337 seconds

elapsed time 348 seconds

elapsed time 359 seconds

elapsed time 370 seconds

elapsed time 381 seconds

elapsed time 391 seconds

elapsed time 402 seconds

elapsed time 413 seconds

elapsed time 424 seconds

elapsed time 435 seconds

elapsed time 446 seconds

elapsed time 457 seconds

elapsed time 468 seconds

elapsed time 478 seconds

elapsed time 489 seconds

elapsed time 500 seconds

elapsed time 511 seconds

elapsed time 522 seconds

elapsed time 533 seconds

elapsed time 543 seconds

elapsed time 555 seconds

elapsed time 565 seconds

elapsed time 576 seconds

elapsed time 587 seconds

elapsed time 597 seconds

elapsed time 608 seconds

elapsed time 619 seconds

elapsed time 630 seconds

elapsed time 640 seconds

elapsed time 651 seconds

elapsed time 662 seconds

elapsed time 673 seconds

elapsed time 683 seconds

elapsed time 694 seconds

elapsed time 705 seconds

elapsed time 716 seconds

elapsed time 727 seconds

elapsed time 738 seconds

elapsed time 748 seconds

elapsed time 760 seconds

elapsed time 771 seconds

elapsed time 782 seconds

elapsed time 792 seconds

elapsed time 803 seconds

elapsed time 814 seconds

elapsed time 824 seconds

elapsed time 835 seconds

elapsed time 846 seconds

elapsed time 857 seconds

elapsed time 868 seconds

elapsed time 878 seconds

elapsed time 889 seconds

elapsed time 900 seconds

elapsed time 911 seconds

elapsed time 922 seconds

elapsed time 933 seconds

elapsed time 943 seconds

elapsed time 954 seconds

elapsed time 965 seconds

elapsed time 976 seconds

elapsed time 987 seconds

elapsed time 997 seconds

elapsed time 1008 seconds

elapsed time 1019 seconds

elapsed time 1030 seconds

elapsed time 1041 seconds

elapsed time 1052 seconds

elapsed time 1063 seconds

elapsed time 1074 seconds

elapsed time 1085 seconds

elapsed time 1096 seconds

elapsed time 1107 seconds

elapsed time 1118 seconds

elapsed time 1129 seconds

elapsed time 1140 seconds

elapsed time 1151 seconds

elapsed time 1162 seconds

elapsed time 1172 seconds

elapsed time 1183 seconds

elapsed time 1194 seconds

elapsed time 1204 seconds

elapsed time 1215 seconds

elapsed time 1225 seconds

elapsed time 1237 seconds

elapsed time 1247 seconds

elapsed time 1258 seconds

elapsed time 1269 seconds

elapsed time 1280 seconds

elapsed time 1290 seconds

elapsed time 1301 seconds

elapsed time 1311 seconds

elapsed time 1322 seconds

elapsed time 1333 seconds

elapsed time 1345 seconds

elapsed time 1357 seconds

elapsed time 1368 seconds

elapsed time 1379 seconds

elapsed time 1390 seconds

elapsed time 1401 seconds

elapsed time 1411 seconds

elapsed time 1422 seconds

elapsed time 1433 seconds

elapsed time 1444 seconds

elapsed time 1455 seconds

```
## elapsed time 1465 seconds
```

```
## elapsed time 1476 seconds
```

```
## elapsed time 1487 seconds
```

```
## elapsed time 1497 seconds
```

```
## elapsed time 1508 seconds
```

```
## elapsed time 1519 seconds
```

```
## elapsed time 1529 seconds
```

```
## elapsed time 1540 seconds
```

```
## elapsed time 1550 seconds
```

```
## elapsed time 1561 seconds
```

```
## elapsed time 1571 seconds
```

```
## elapsed time 1582 seconds
```

```
nrow(seqBLAST)
```

```
## [1] 33
```

Now, we can select some columns and make them a new data frame. To compare these sequences, I sorted the data in decreasing order on Hsp_score. Because it represents the strength of the match, and higher scores mean stronger matches. We can also compare them based on Hsp_evalue, and smaller values generally indicate stronger matches. In hit_def, it shows most of them come from Yersinia but are not human. The Hsp_gaps shows the number of gaps between the query sequence and its match, in which we all got 0 gaps.

```

results <- seqBLAST %>%
  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 1.69196e-122         0
## 2         500 1.69196e-122         0
## 3         500 1.69196e-122         0
## 4         500 1.69196e-122         0
## 5         500 1.69196e-122         0
## 6         500 1.69196e-122         0
##
Hsp_qseq
## 1 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGA
GAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 2 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGA
GAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 3 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGA
GAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 4 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGA
GAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 5 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGA
GAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA
## 6 GCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGCAAGAGCAAAGTGGGGGACCTTAGGG
CCTCACGCCATCGGATGAACCCAGATGGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGA
GAGGATGACCAGCCACACTGGAAGTGGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAA

```

Thanks!