Homework Assignment

Phylogenetic Trees

04-13-2023

Task A

```
seqs <- read.fasta(file = "usflu.fasta")
```

Task B

I just selected the first strain from each year.

```
annotations <- read.csv("usflu.annot.csv")
strains <- annotations %>%
    group_by(year) %>%
    slice_head(n = 1)
print(strains)
```

```
## # A tibble: 16 x 4
## # Groups:   year [16]
##        X accession  year misc
##    <int> <chr>     <int> <chr>
##  1     1 CY013200   1993 (A/New York/783/1993(H3N2))
##  2     6 CY012272   1994 (A/New York/729/1994(H3N2))
##  3    11 CY012480   1995 (A/New York/666/1995(H3N2))
##  4    16 CY009476   1996 (A/New York/565/1996(H3N2))
##  5    21 CY006259   1997 (A/New York/511/1997(H3N2))
##  6    26 CY006787   1998 (A/New York/506/1998(H3N2))
##  7    31 CY001453   1999 (A/New York/184/1999(H3N2))
##  8    36 CY000737   2000 (A/New York/180/2000(H3N2))
##  9    41 CY002816   2001 (A/New York/301/2001(H3N2))
## 10    46 CY000297   2002 (A/New York/96/2002(H3N2))
## 11    51 CY000105   2003 (A/New York/60A/2003(H3N2))
## 12    56 CY019245   2004 (A/New York/908/2004(H3N2))
## 13    61 CY003640   2005 (A/New York/463/2005(H3N2))
## 14    66 EF554795   2006 (A/Ohio/2006(H3N2))
## 15    71 EU199369   2007 (A/Minnesota/08/2007(H3N2))
## 16    76 FJ549055   2008 (A/Illinois/14/2008(H3N2))
```

```
filtered_seqs <- seqs[strains$accession]
write.fasta(sequences = filtered_seqs, names = getName(filtered_seqs),
    file.out = "filtered_seqs.fasta")
```

Task C

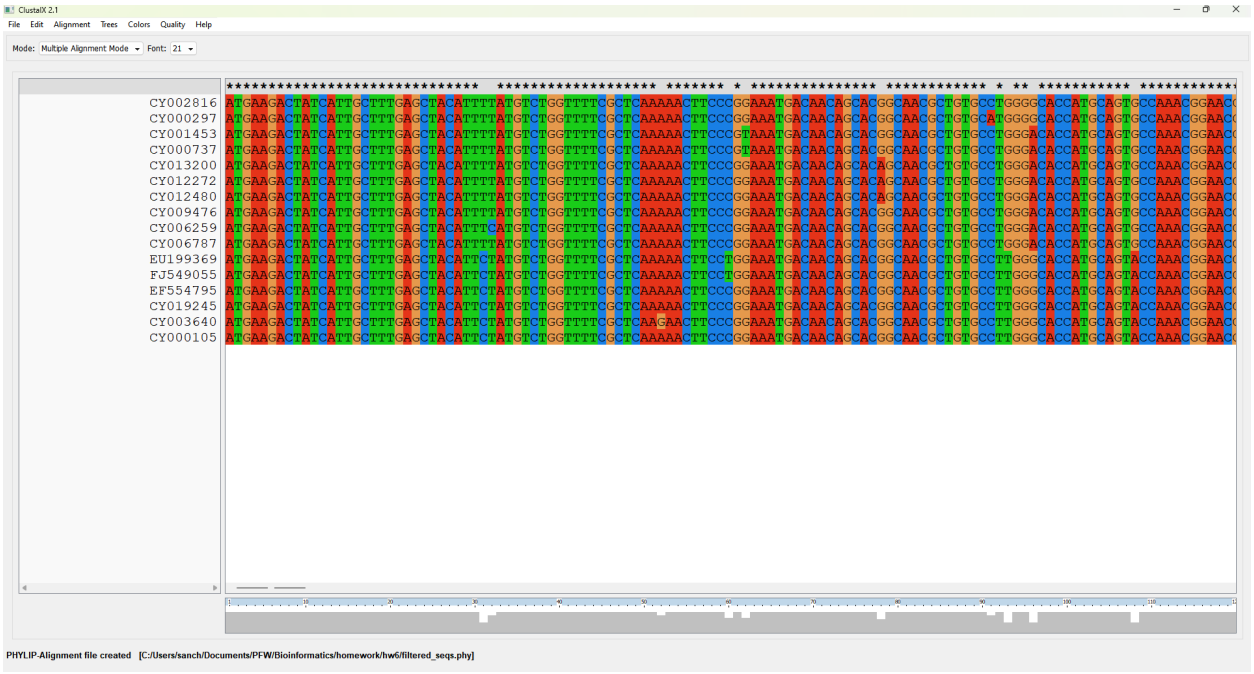
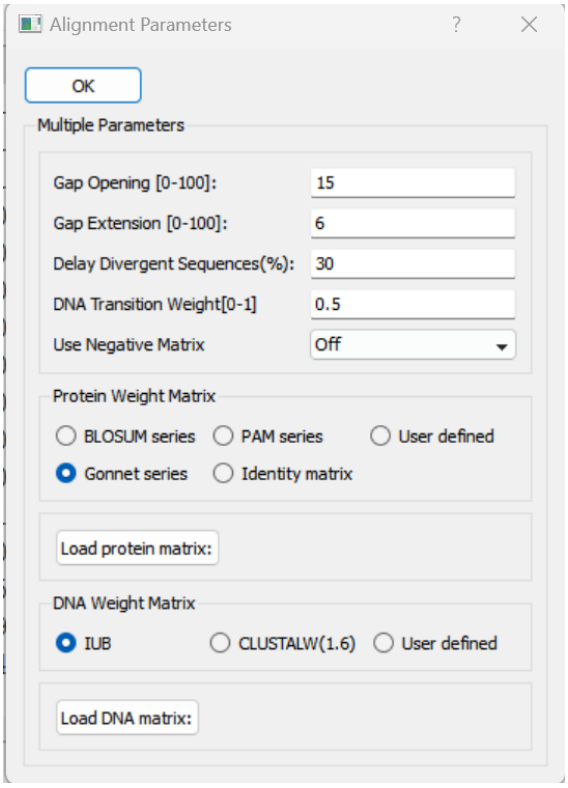


Figure 1: Alignment



Task D

```
aln <- read.alignment(file = "filtered_seqs.phy", format = "phylip")
clean_aln <- cleanAlignment(aln, 75, 30)
dist_matrix <- dist.dna(as.DNAbin(clean_aln))
print(dist_matrix)
```

```
##             CY002816    CY000297    CY001453    CY000737    CY013200
## CY000297 0.004728428
## CY001453 0.014891427 0.019740370
## CY000737 0.014897424 0.019748781 0.003542187
## CY013200 0.037637595 0.042682008 0.033823346 0.033211603
## CY012272 0.038898630 0.043955083 0.035071787 0.034459264 0.003544020
## CY012480 0.043263526 0.048351787 0.040670136 0.040053408 0.011888264
## CY009476 0.031392047 0.035125027 0.025207677 0.024602128 0.019740370
## CY006259 0.014897424 0.019748781 0.010078234 0.009482720 0.025226380
## CY006787 0.017914440 0.022788741 0.013070497 0.012472841 0.027672253
## EU199369 0.046450280 0.051569301 0.045103350 0.045119950 0.068947847
## FJ549055 0.050924140 0.056084107 0.049557335 0.049576224 0.072305989
## EF554795 0.041994490 0.047070556 0.040670136 0.040683926 0.064372810
## CY019245 0.035097297 0.040112591 0.033801617 0.033812204 0.057224592
## CY003640 0.038882116 0.043935911 0.037564567 0.037578057 0.059853753
## CY000105 0.031392047 0.036379733 0.030725582 0.030735899 0.056596413
##             CY012272    CY012480    CY009476    CY006259    CY006787
## CY000297
## CY001453
## CY000737
## CY013200
## CY012272
## CY012480 0.012489949
## CY009476 0.020961372 0.027682852
## CY006259 0.026458666 0.031989776 0.019130986
## CY006787 0.028909112 0.034459264 0.021555549 0.004726415
## EU199369 0.070276752 0.076179174 0.063082156 0.046394648 0.049557335
## FJ549055 0.073644678 0.079580267 0.067675332 0.050861571 0.054046452
## EF554795 0.065691203 0.070887626 0.058503969 0.041947350 0.045087318
## CY019245 0.058525829 0.064324300 0.051428408 0.035059865 0.038166614
## CY003640 0.059853753 0.066986790 0.053379810 0.037578057 0.040698278
## CY000105 0.057896804 0.063690799 0.048916309 0.032600413 0.035697177
##             EU199369    FJ549055    EF554795    CY019245    CY003640
## CY000297
## CY001453
## CY000737
## CY013200
## CY012272
## CY012480
## CY009476
## CY006259
## CY006787
## EU199369
## FJ549055 0.004134956
## EF554795 0.011287303 0.013098328
## CY019245 0.015484899 0.018522343 0.010080806
## CY003640 0.017914440 0.020970529 0.013689694 0.007103052
```

```
## CY000105 0.024602128 0.028909112 0.021540269 0.014876640 0.017300881
```

```
print(max(dist_matrix))
```

```
## [1] 0.07958027
```

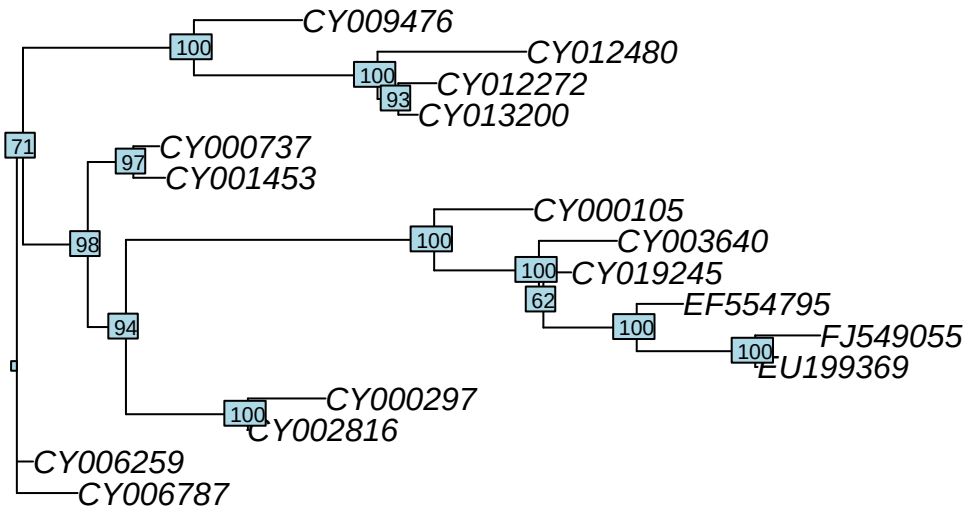
```
print(min(dist_matrix))
```

```
## [1] 0.003542187
```

The maximum distance is between FJ549055 and CY012480.
The minimum distance is between CY000737 and CY001453.

Task E

```
## Running bootstraps:       100 / 100
## Calculating bootstrap values... done.
```




```
##
## Phylogenetic tree with 16 tips and 14 internal nodes.
##
## Tip labels:
##   CY002816, CY000297, CY001453, CY000737, CY013200, CY012272, ...
## Node labels:
##   NA, 71, 98, 94, 97, 100, ...
##
## Unrooted; includes branch lengths.
```

Task F

The most closely related sequences based on the tree are CY000737 and CY001453.
Yes, this is consistent with my answer from part d.

Task G

```
# Using the fasta file directly
sars_seq <- read.fasta(file = "sars.fasta")
name <- getName(sars_seq)
sars_seq <- getSequence(sars_seq)
print(DNAString(c2s(sars_seq[[1]])))
```

```
## 29632-letter DNAString object
## seq: GATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA...TAGTGCTATCCCCATGTGATTTTAATAGCTTCTTAG
```

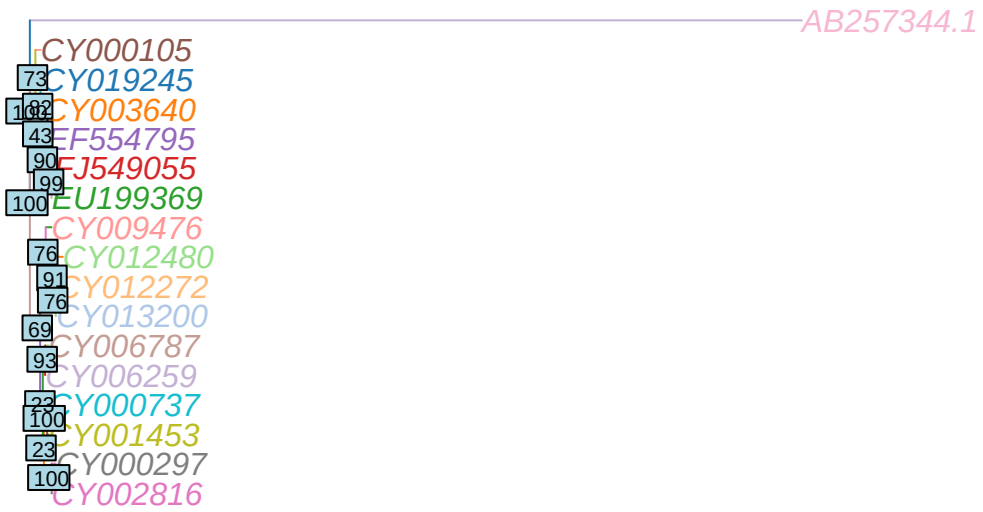
```
write.fasta(sars_seq, name, file.out = "filtered_seqs.fasta",
    open = "a")
new_aln <- read.alignment(file = "filtered_sequences_1.phy",
    format = "phylip")
dist_matrix_1 <- dist.dna(as.DNAbin(new_aln))
print(dist_matrix_1)
```

```
##               CY019245    CY003640    EU199369    FJ549055    EF554795
## CY003640   0.007214078
## EU199369   0.015728467 0.018196989
## FJ549055   0.018196989 0.020681307 0.003597406
## EF554795   0.010238667 0.013904888 0.011464386 0.012691917
## CY000105   0.015110436 0.017573488 0.024992101 0.028740181 0.021880703
## CY002816   0.035022782 0.038865400 0.046551033 0.050440528 0.042026116
## CY000297   0.040115068 0.043997382 0.051750287 0.055675538 0.047181064
## CY001453   0.033708181 0.037528255 0.045183598 0.049056002 0.040681960
## CY000737   0.033718713 0.037541781 0.045200327 0.049074668 0.040695792
## CY013200   0.056857101 0.059526403 0.068759505 0.071491676 0.064114208
## CY012272   0.058178105 0.059526403 0.070109033 0.072849451 0.065452897
## CY012480   0.064064635 0.066768295 0.076102743 0.078870836 0.070728997
## CY009476   0.050973563 0.052954393 0.062803896 0.066795180 0.058155788
## CY006259   0.034985338 0.037541781 0.046494826 0.050378450 0.041978667
## CY006787   0.038139876 0.040710205 0.049707044 0.053610131 0.045167453
## AB257344.1 0.950842344 0.955497898 0.966958137 0.962222106 0.952229310
##               CY000105    CY002816    CY000297    CY001453    CY000737
## CY003640
## EU199369
## FJ549055
## EF554795
## CY000105
## CY002816   0.031260665
## CY000297   0.036324493 0.004802226
## CY001453   0.030584920 0.015125694 0.020052279
## CY000737   0.030595175 0.015131883 0.020060959 0.003597406
## CY013200   0.056219512 0.037618030 0.042743066 0.033741496 0.033120433
## CY012272   0.057539645 0.038899261 0.044036728 0.035009727 0.034387861
## CY012480   0.063421562 0.043332583 0.048502890 0.040695792 0.040069590
## CY009476   0.048424050 0.031273591 0.035065376 0.024992101 0.024377425
```

```
## CY006259   0.032487922 0.015131883 0.020060959 0.010236013 0.009631133
## CY006787   0.035632167 0.018196989 0.023149754 0.013275572 0.012668482
## AB257344.1 0.958287327 0.963519460 0.972351036 0.963463521 0.970971723
##               CY013200    CY012272    CY012480    CY009476    CY006259
## CY003640
## EU199369
## FJ549055
## EF554795
## CY000105
## CY002816
## CY000297
## CY001453
## CY000737
## CY013200
## CY012272   0.003599296
## CY012480   0.012074886 0.012686133
## CY009476   0.020052279 0.021292974 0.028122982
## CY006259   0.025011750 0.026263315 0.031880019 0.018822569
## CY006787   0.027495235 0.028751517 0.034387861 0.021284080 0.004800150
## AB257344.1 0.955831137 0.955831137 0.970552273 0.961728038 0.961581741
##               CY006787
## CY003640
## EU199369
## FJ549055
## EF554795
## CY000105
## CY002816
## CY000297
## CY001453
## CY000737
## CY013200
## CY012272
## CY012480
## CY009476
## CY006259
## CY006787
## AB257344.1 0.969668171
```

```
rootedNJtree(new_aln, "AB257344.1", "DNA")
```

```
## Running bootstraps:       100 / 100
## Calculating bootstrap values... done.
```




```
##
## Phylogenetic tree with 17 tips and 16 internal nodes.
##
## Tip labels:
##   CY019245, CY003640, EU199369, FJ549055, EF554795, CY000105, ...
## Node labels:
##   100, 23, 23, 93, 69, 100, ...
##
## Rooted; includes branch lengths.
```

Task H

For the tree in part (e), most of the groups have a confidence value of more than 90%. A couple of nodes have values less than 80 (71 and 62). I would say that the groups have been formed with moderately high accuracy.
For the tree in part (g), the bootstrap values seem to be on the lower side, and this suggests that the groups haven't been formed with great accuracy.

Task I

```
long_substr <- function(data) {
    is_substr <- function(find, data) {
        if (length(data) < 1 && nchar(find) < 1) {
            return(FALSE)
```

```
        }

        for (i in 1:length(data)) {
            if (!stri_detect_fixed(data[[i]], find)) {
                return(FALSE)
            }
        }

        return(TRUE)
    }

    data <- data$seq
    substr <- " "
    if (length(data) > 1 && nchar(data[[1]]) > 0) {
        for (i in 1:nchar(data[[1]])) {
            for (j in 1:(nchar(data[[1]]) - i + 1)) {
                if (j > nchar(substr) && is_substr(substring(data[[1]],
                  i, i + j), data)) {
                  substr <- substring(data[[1]], i,
                    i + j)
                }
            }
        }
    }

    return(substr)
}

result <- long_substr(clean_aln)
print(result)
```

```
## [1] "atgtgggcctgccaaaaaggcaacattaggtgcaacatttgcatttga"
```