Assignment 5

Jieran Sun, Hui Jeong (HJ) Jung, Gudmundur Björgvin Magnusson

2023-03-28

Question 12

(1)

$$P(t+dt) = P(dt)P(t) = (I + Rdt)P(t) \tag{1}$$

Therefore we can show that

$$\frac{dP(t)}{dt} = RP(t) \tag{2}$$

(2)

As the Markov chain is homogeneous Markov chain and π is the ergodic stationary distribution,

$$\begin{align} t \to \infty \quad P(t)\pi &= \pi \tag{3} \\ \text{multiply } R, \quad RP(t)\pi &= R\pi \tag{4} \\ \frac{dP(t)}{dt}\pi &= R\pi \tag{5} \\ \tag{6} \end{align}$$

Given that when $t \to \infty$, $P(t)$ reach steady state, $\frac{dP(t)}{dt} = 0$. Hence $R\pi = 0$.

Question 13

(1) The joint probability of the tree is

$$\begin{align} P(X, Z|T) = P(Z_4) * P(X_5|Z_4) * P(Z_3|Z_4) * P(Z_2|Z_3) * P(Z_1|Z_3) \tag{7} \\ *P(X_4|Z_2) * P(X_3|Z_2) * P(X_2|Z_1) * P(X_1|Z_1) \tag{8} \end{align}$$

(2) To do the naive calculation of P(X|T) via brute-force marginalization over the hidden nodes Z,

for each node X, we have to marginalize out all the internal nodes Z that it is dependent on. Here is an example with X4.

$$\begin{aligned} P(X_4|T) = \sum_{Z} P(X_4, Z|T) = \sum_{Z_2,Z_3,Z_4} P(X_4|Z_2)P(Z_2|Z_3)P(Z_3|Z_4)P(Z_4) \tag{9} \\ \tag{10} \end{aligned}$$

Given each Z represent a nucleitode and can take 4 values(A,T,G,C), then the total number of cases for X_4 is $4^3 = 64$. Hence 64 summation operations are needed for X4.

Similarly, for X_1, X_2, X_3, they all have 64 cases and 64 summation operations, and for X_5 we have 4 summation cases as $P(X_5, Z) = \sum_{Z_4} P(X_5|Z_4)P(Z_4)$ and it only has 4 cases.

Hence in total the number of summation operation is $64 \times 4 + 4 = 260$.

(3)

$$\begin{aligned} P(X, Z|T) = P(Z_4) * P(X_5|Z_4) * P(Z_3|Z_4) * P(Z_2|Z_3) * P(Z_1|Z_3) \tag{11} \\ *P(X_4|Z_2) * P(X_3|Z_2) * P(X_2|Z_1) * P(X_1|Z_1) \tag{12} \end{aligned}$$

By rearranging the expression in such way we only need to do 4 summations.

Question 14

(1)

```
if(!require("phangorn")) {
  install.packages("phangorn")
}
```

```
## Loading required package: phangorn
```

```
## Loading required package: ape
```

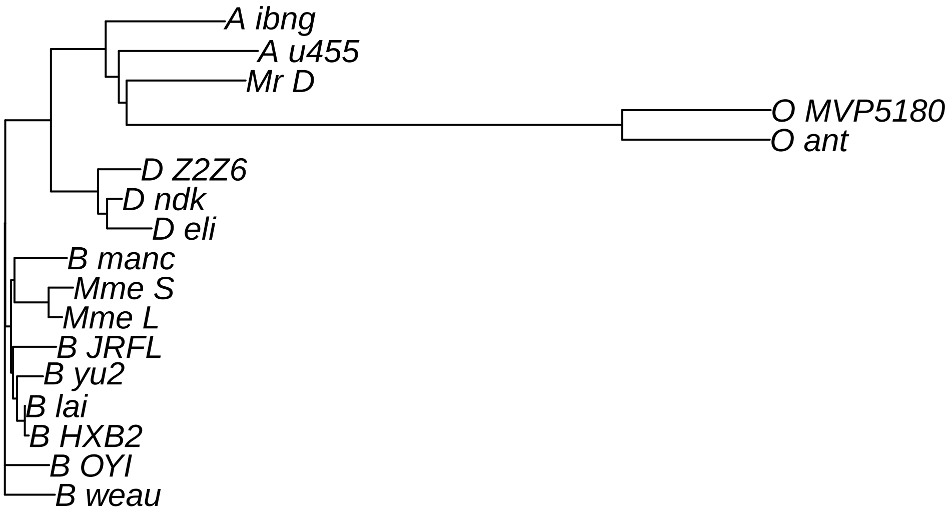
```
if(!require("ape")) {
  install.packages("ape")
}

library(phangorn)
library(ape)
```

```
ParisRT <- read.dna("ParisRT.txt")
```

(2)

```
distParis <- dist.dna(ParisRT)
initTree <- NJ(distParis)
plot(initTree)
```



(3)

```
kimura <- pml(tree= initTree, data= phyDat(ParisRT), model= "K80")
kimura$logLik
```

```
## [1] -3003.487
```

The log likelihood of the fitted model is -3003.487.

(4)

The values of the optimised rate matrix can be found below.

```
optimParam <- optim.pml(kimura, optQ= TRUE)
```

```
## optimize edge weights:  -3003.487 --> -2992.981
## optimize rate matrix:  -2992.981 --> -2863.264
## optimize edge weights:  -2863.264 --> -2862.477
## optimize rate matrix:  -2862.477 --> -2862.456
## optimize edge weights:  -2862.456 --> -2862.455
## optimize rate matrix:  -2862.455 --> -2862.455
## optimize edge weights:  -2862.455 --> -2862.455
## optimize rate matrix:  -2862.455 --> -2862.455
## optimize edge weights:  -2862.455 --> -2862.455
```

```
optimParam
```

```
## model: K80
## loglikelihood: -2862.455
## unconstrained loglikelihood: -2098.897
##
## Rate matrix:
##          a         c         g        t
## a 0.000000 2.4318480 6.8571651 1.118323
## c 2.431848 0.0000000 0.6119506 7.262319
## g 6.857165 0.6119506 0.0000000 1.000000
## t 1.118323 7.2623187 1.0000000 0.000000
##
## Base frequencies:
##    a    c    g    t
## 0.25 0.25 0.25 0.25
```

(5)

After optimizing with respect to branch lengths, nucleotide substitution rates, and tree topology, the results are as below.

```
optimParam2 <- optim.pml(kimura, optQ = TRUE, optNni = TRUE, optEdge = TRUE)
```

```
## optimize edge weights:  -3003.487 --> -2992.981
## optimize rate matrix:  -2992.981 --> -2863.264
## optimize edge weights:  -2863.264 --> -2862.477
## optimize topology:  -2862.477 --> -2849.886  NNI moves:  5
## optimize rate matrix:  -2849.886 --> -2849.791
## optimize edge weights:  -2849.791 --> -2849.789
## optimize topology:  -2849.789 --> -2849.789  NNI moves:  0
## optimize rate matrix:  -2849.789 --> -2849.789
## optimize edge weights:  -2849.789 --> -2849.789
## optimize rate matrix:  -2849.789 --> -2849.789
## optimize edge weights:  -2849.789 --> -2849.789
```

```
optimParam2$logLik
```

```
## [1] -2849.789
```

The log likelihood is -2849.789.

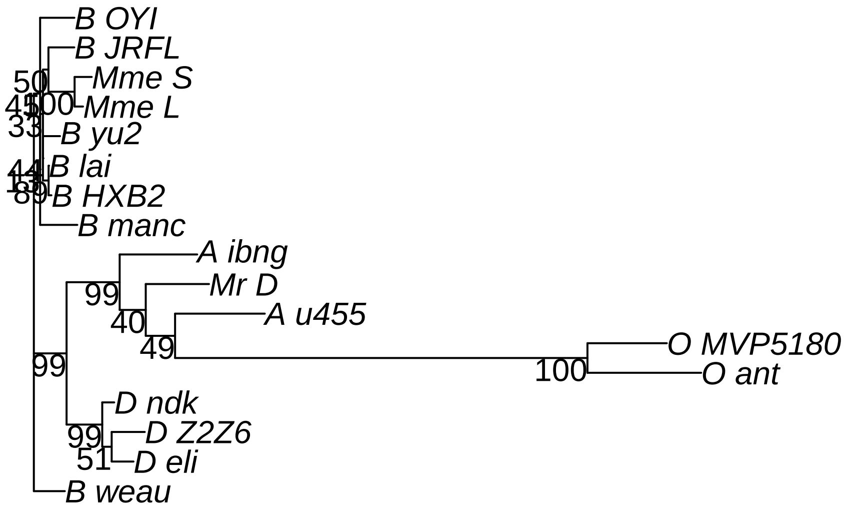
(6)

```
bootPML <- bootstrap.pml(optimParam2, optNni= TRUE)
```

In this bootstrapping function, we are resampling the sequences that we are using to construct a phylogenetic tree, and seeing if the same branch is observed even when generating a new tree based on bootstrapped data, which would indicate confidence in the observed branch.

(7)

```
plotBS(tree= optimParam2$tree, BStrees = bootPML, type= "phylogram")
```



Judging from this plot, it is more likely that Mme_S was more likely to have infected the patient Mme_L.