Project 5: Statistical phylogenetics

Team K - Minghang Li

March 29, 2023

Problem 12: Transition matrix, rate matrix, and stationary distribution

1. Show that $\frac{dP(t)}{dt} = R \cdot P(t)$

Proof.

$$\begin{aligned}
\frac{dP(t)}{dt} &= \frac{P(t+dt) - P(t)}{dt} && \text{(Definition of derivative)} \\
&= \frac{P(t)P(dt) - P(t)}{dt} && \text{(Chapman-Kolmogorov's equation)} \\
&= \frac{P(t)(I + Rdt) - P(t)}{dt} && \text{(Definition of } P(dt)\text{)} \\
&= \frac{\cancel{P(t)} + RP(t)dt - \cancel{P(t)}}{dt} \\
&= \frac{RP(t)dt}{dt} \\
&= RP(t)
\end{aligned}$$

□

2. Assume that the given Markov chain is ergodic with (unique) stationary distribution π, show that $R\pi = 0$

Proof. The stationary distribution π satisfies:

$$P(t)\pi = \pi$$

after a sufficiently long time t and any time point that follows. Hence, the following equation also holds:

$$P(t+dt)\pi = \pi$$

We can further re-write the left-hand side of the equation:

$$\begin{aligned}
P(t+dt)\pi &= P(t)P(dt)\pi && \text{(Chapman-Kologomorov's equation)} \\
&= P(t)(I + Rdt)\pi && \text{(Definition of } P(dt)\text{)} \\
&= P(t)\pi + Rdt \cdot P(t)\pi \\
&= \pi + Rdt\pi
\end{aligned}$$

which leads us to the following equality:

$$\begin{aligned} \cancel{\pi} + Rdt\pi &= \cancel{\pi} \\ Rdt\pi &= 0 \\ R\pi &= 0 \end{aligned}$$

□

Problem 13: Phylogenetic trees as Bayesian networks

1. What is the joint probability $P(X, Z|T)$ of the tree?

$$P(X, Z|T) = \pi(Z_4)P(X_5|Z_4)P(Z_3|Z_4)P(Z_2|Z_3)P(Z_1|Z_3)P(X_4|Z_2)P(X_3|Z_2)P(X_2|Z_1)P(X_1|Z_1)$$

2. How many summation steps would be required for the naive calculation of $P(X|T)$ via brute-force marginalization over the hidden nodes Z?

Since there are 4 possibilities (A, C, G, T) for each hidden node:

$$4 \times 4 \times 4 \times 4 = 256$$

3. Rearrange the expression $P(X|T)$ such that the number of operations is minimized. How many summation steps are required now for the calculation of $P(X|T)$?

$$\begin{aligned} P(X|T) &= \sum_{Z_4}\sum_{Z_3}\sum_{Z_2}\sum_{Z_1} \pi(Z_4)P(X_5|Z_4)P(Z_3|Z_4)P(Z_2|Z_3)P(Z_1|Z_3)P(X_4|Z_2)P(X_3|Z_2)P(X_2|Z_1)P(X_1|Z_1) \\ &= \sum_{Z_4} \pi(Z_4)P(X_5|Z_4) \sum_{Z_3} P(Z_3|Z_4) \sum_{Z_2} P(Z_2|Z_3)P(X_4|Z_2)P(X_3|Z_2) \sum_{Z_1} P(Z_1|Z_3)P(X_2|Z_1)P(X_1|Z_1) \qquad \text{(distributivity)} \\ &= \underbrace{\sum_{Z_4} \pi(Z_4)P(X_5|Z_4)}_{4} \underbrace{\sum_{Z_3} P(Z_3|Z_4)}_{4\times4} \underbrace{\sum_{Z_2} P(Z_2|Z_3)P(X_4|Z_2)P(X_3|Z_2)}_{4\times4(Z_2\to X_4 \text{ branch})+4\times4(Z_2\to X_3 \text{ branch})} \underbrace{\sum_{Z_1} P(Z_1|Z_3)P(X_2|Z_1)P(X_1|Z_1)}_{4\times4(Z_1\to X_2 \text{ branch})+4\times4(Z_1\to X_1 \text{ branch})} \end{aligned}$$

In total we have $(16 + 16) \times 2 + 16 + 4 = 52$ summations.

Problem 14: Learning phylogenetic trees from sequence alignment data

Load the dataset `ParisRT.txt`.

```
data <- read.dna("ParisRT.txt", format="sequential")
data
```

```
## 17 DNA sequences in binary format stored in a matrix.
##
## All sequences of same length: 618
##
## Labels:
```

```
## B_OYI
## B_HXB2
## D_eli
## Mr_D
## Mme_L
## A_ibng
## ...
##
## Base composition:
##    a    c    g    t
## 0.404 0.162 0.195 0.239
## (Total: 10.51 kb)
```

Create initial tree topology for the alignment using neighbor joining and K80 model.

```
ini_tree <- NJ(dist.dna(data, model = "K80"))
```

Plot the initial tree.

```
plot.phylo(ini_tree)
```


```
tree_ML <- pml(ini_tree, phyDat(data), model = "K80")
tree_ML
```

```
## model: K80
```

```
## loglikelihood: -3003.487
## unconstrained loglikelihood: -2098.897
##
## Rate matrix:
##  a c g t
## a 0 1 1 1
## c 1 0 1 1
## g 1 1 0 1
## t 1 1 1 0
##
## Base frequencies:
##   a   c   g   t
## 0.25 0.25 0.25 0.25
```

Find the optimal parameters of K80 model for rate matrix.

```
tree_optim_Q <- optim.pml(tree_ML,
          optQ = TRUE,
          optNni = FALSE,
          optBf = FALSE,
          optInv = FALSE,
          optGamma = FALSE,
          optEdge = FALSE,
          optRate = FALSE,
          optRooted = FALSE,
          model = "K80")
```

```
## optimize rate matrix:  -3003.487 --> -2884.408
## optimize rate matrix:  -2884.408 --> -2884.408
```

```
tree_optim_Q
```

```
## model: K80
## loglikelihood: -2884.408
## unconstrained loglikelihood: -2098.897
##
## Rate matrix:
##       a     c     g     t
## a 0.000000 1.000000 4.976955 1.000000
## c 1.000000 0.000000 1.000000 4.976955
## g 4.976955 1.000000 0.000000 1.000000
## t 1.000000 4.976955 1.000000 0.000000
##
## Base frequencies:
##   a   c   g   t
## 0.25 0.25 0.25 0.25
```

Optimize for branch lengths, nucelotide substitution rates and tree topology simultaneously.

```
tree_optim <- optim.pml(tree_ML,
          optQ = TRUE,     # rate matrix
          optNni = TRUE,    # tree topology
```

```
            optBf = FALSE,    # base frequencies
            optInv = FALSE,   # proportion of var size
            optGamma = FALSE, # gamma rate param
            optEdge = TRUE,   # edge lengths
            optRate = FALSE,  # overall rate
            optRooted = FALSE, # edge lengths of a rooted tree
            model = "K80")
```

```
## optimize edge weights: -3003.487 --> -2992.981
## optimize rate matrix: -2992.981 --> -2873.703
## optimize edge weights: -2873.703 --> -2872.892
## optimize topology: -2872.892 --> -2859.775 NNI moves: 5
## optimize rate matrix: -2859.775 --> -2859.682
## optimize edge weights: -2859.682 --> -2859.681
## optimize topology: -2859.681 --> -2859.681 NNI moves: 0
## optimize rate matrix: -2859.681 --> -2859.681
## optimize edge weights: -2859.681 --> -2859.681
```

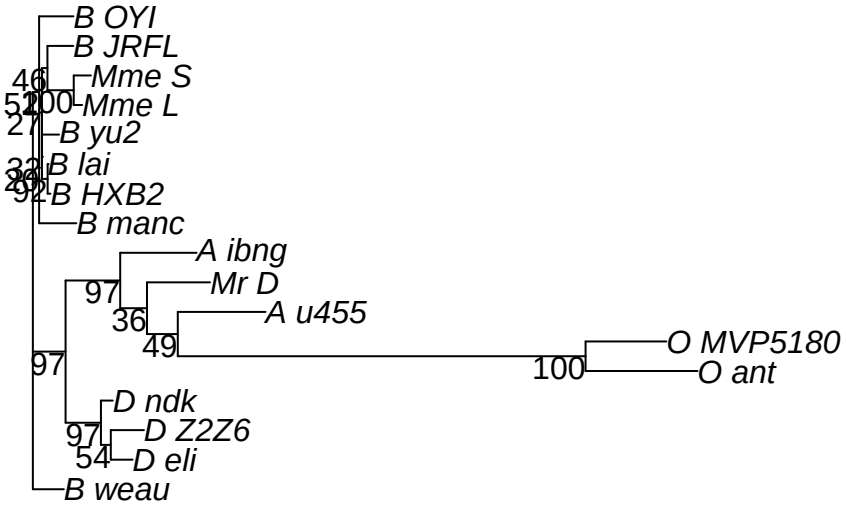
```
tree_optim
```

```
## model: K80
## loglikelihood: -2859.681
## unconstrained loglikelihood: -2098.897
##
## Rate matrix:
##       a     c     g     t
## a 0.000000 1.000000 5.262145 1.000000
## c 1.000000 0.000000 1.000000 5.262145
## g 5.262145 1.000000 0.000000 1.000000
## t 1.000000 5.262145 1.000000 0.000000
##
## Base frequencies:
##   a   c   g   t
## 0.25 0.25 0.25 0.25
```

Bootstrap on optimised model.

TODO: answer what is exactly being resampled.

```
plotBS(tree_optim$tree, bootstrap_trees, type = "phylogram")
```



Mme_S is more likely to affect patient Mme_L.