Project 5: Statistical phylogenetics

Team C - Minghang Li, Xiaocheng Yang, Xinyi Chen

March 22, 2024

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

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

Proof.

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

$\square$

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:

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

which leads us to the following equality:

$$\require{cancel}\cancel{\pi} + Rdt\pi = \cancel{\pi}$$
$$Rdt\pi = 0$$
$$R\pi = 0$$

$\square$

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 $P(X|T)$?

$$P(X|T) = \sum_{Z_4}\sum_{Z_3}\sum_{Z_2}\sum_{Z_1}\pi(Z_4)P(X_5|Z_4)P(Z_3|Z_4)P(Z_2|Z_3)P(Z_1|Z_3)P(X_4|Z_2)P(X_3|Z_2)P(X_2|Z_1)P(X_1|Z_1)$$

$$= \sum_{Z_4}\pi(Z_4)P(X_5|Z_4)\sum_{Z_3}P(Z_3|Z_4)\sum_{Z_2}P(Z_2|Z_3)P(X_4|Z_2)P(X_3|Z_2)\sum_{Z_1}P(Z_1|Z_3)P(X_2|Z_1)P(X_1|Z_1) \quad \text{(distribu}$$

$$= \underbrace{\sum_{Z_4}\pi(Z_4)P(X_5|Z_4)}_{4}\underbrace{\sum_{Z_3}P(Z_3|Z_4)}_{4\times4}\underbrace{\sum_{Z_2}P(Z_2|Z_3)P(X_4|Z_2)P(X_3|Z_2)}_{4\times4(Z_2\to X_4\text{ branch})+4\times4(Z_2\to X_3\text{ branch})}\underbrace{\sum_{Z_1}P(Z_1|Z_3)P(X_2|Z_1)P(X_1|Z_1)}_{4\times4(Z_1\to X_2\text{ branch})+4\times4(Z_1\to X_1\text{ branch})}$$

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
## ...
##
```
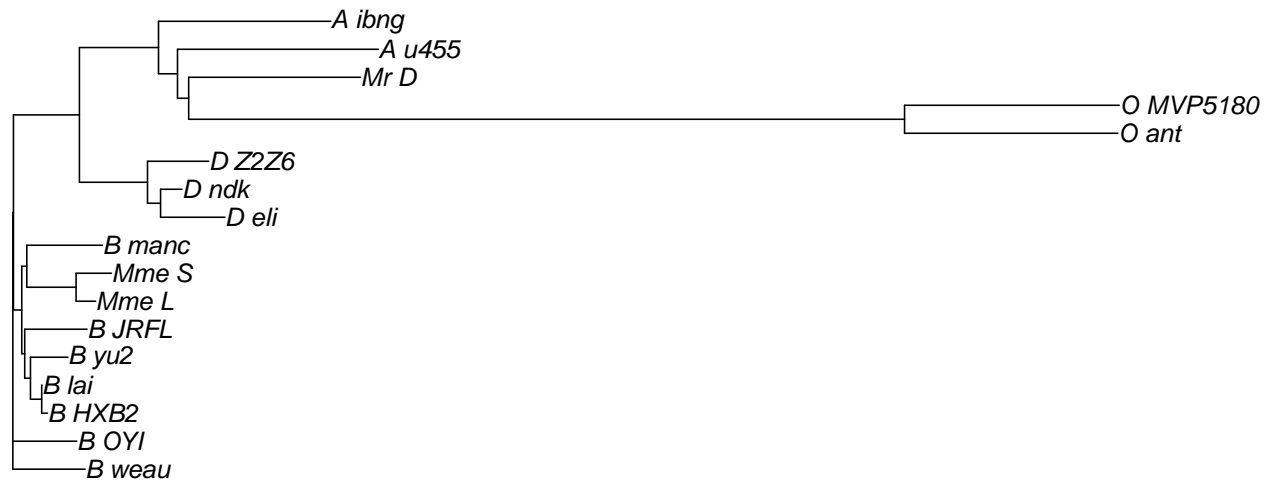
2

```
## 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.

```
set.seed(42)
tree.ini <- NJ(dist.dna(data, model = "K80"))
```

Plot the initial tree.

```
plot.phylo(tree.ini)
```



```
tree.ML <- pml(tree.ini, 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 5.262144 1.000000
## c 1.000000 0.000000 1.000000 5.262144
## g 5.262144 1.000000 0.000000 1.000000
## t 1.000000 5.262144 1.000000 0.000000
##
## Base frequencies:
##    a    c    g    t
## 0.25 0.25 0.25 0.25
```

Bootstrap on optimised model. The program resampled the nucleotides in the alignment (columns in the input matrix).

```
plotBS(tree.optim$tree, bootstrap, type = "phylogram")
```
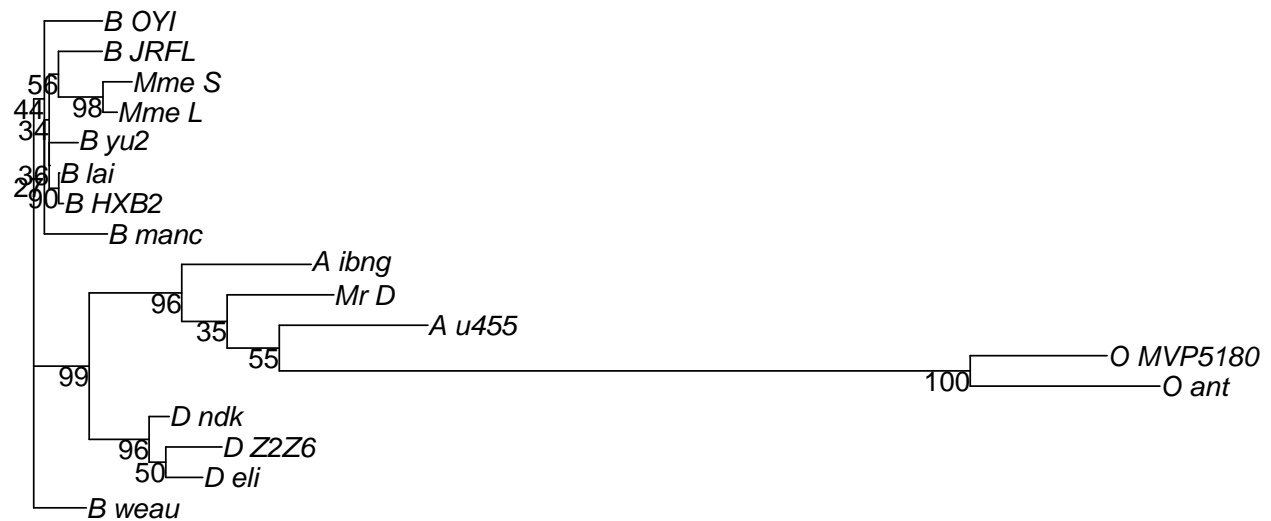

Mme_S is more likely to affect patient Mme_L. The bootstrap support is 98% (in my run with seed 42), suggesting Mme_S and Mme_L share a more recent ancestor (closer relationship) than Mr D