Solutions to The Exercises of Ziheng Yang's *Molecular Evolution: A Statistical Approach* (Yang 2014)

Version 1.0.5 (2024-03-15)

Chapter 1. Models of nucleotide substitution

1.1 Use the transition probabilities under the JC69 model (equation (1.4)) to confirm the Chapman–Kolmogorov theorem (equation (1.5)). It is sufficient to consider two cases: (**a**) $i = \mathrm{T}, j = \mathrm{T}$; and (**b**) $i = \mathrm{T}, j = \mathrm{C}$. For example, in case (**a**), confirm that $p_{TT}(t_1 + t_2) = p_{TT}(t_1)p_{TT}(t_2) + p_{TC}(t_1)p_{CT}(t_2) + p_{TA}(t_1)p_{AT}(t_2) + p_{TG}(t_1)p_{GT}(t_2)$.

1.2 Derive the transition probability matrix $P(t) = \mathrm{e}^{Qt}$ for the JC69 model. Set $\pi_T = \pi_C = \pi_A = \pi_G = 1/4$ and $\alpha_1 = \alpha_2 = \beta$ in the rate matrix (1.16) for the TN93 model to obtain the eigenvalues and eigenvectors of Q under JC69, using results of §1.2.3. Alternatively you can derive the eigenvalues and eigenvectors from equation (1.1) directly. Then apply equation (1.18).

1.3 Derive the transition probability matrix $P(t)$ for the Markov chain with two states 0 and 1 and rate matrix $Q = \begin{bmatrix} -u & u \\ v & -v \end{bmatrix}$. Confirm that the spectral decomposition of Q is given as

$$Q = U\Lambda U^{-1} = \begin{bmatrix} 1 & -u \\ 1 & v \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -u-v \end{bmatrix} \begin{bmatrix} \frac{v}{u+v} & \frac{u}{u+v} \\ -\frac{1}{u+v} & \frac{1}{u+v} \end{bmatrix}, \tag{1.77}$$

so that

$$P(t) = \mathrm{e}^{Qt} = U\mathrm{e}^{\Lambda t}U^{-1} = \frac{1}{u+v}\begin{bmatrix} v + u\mathrm{e}^{-(u+v)t} & u - u\mathrm{e}^{-(u+v)t} \\ v - v\mathrm{e}^{-(u+v)t} & u + v\mathrm{e}^{-(u+v)t} \end{bmatrix}. \tag{1.78}$$

Note that the stationary distribution of the chain is given by the first row of U^{-1}, as $(\frac{v}{u+v}, \frac{u}{u+v})$, which can also be obtained from $P(t)$ by letting $t \to \infty$. A special case is $u = v = 1$, when we have

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} \\ \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} \end{bmatrix}. \tag{1.79}$$

This is the binary equivalent of the JC69 model.

1.4 Confirm that the two likelihood functions for the JC69 model, equations (1.46) and (1.47), are proportional and the proportionality factor is a function of n and x but not of d. Confirm that the likelihood equation, $\frac{\mathrm{d}\ell}{\mathrm{d}d} = \frac{\mathrm{d}\log\{L(d)\}}{\mathrm{d}d} = 0$, is the same whichever of the two likelihood functions is used.

See Problems 1.1-1.4 in (Yang 2006).

1.5 Derive the equilibrium nucleotide frequencies for the K80 model. Solve the system of linear equations generated by equation (1.61) and the constraint $\sum_j \pi_j = 1$.

Solution.

The transition rate matrix for the model K80 is given as

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(2\beta+\alpha) & \alpha \\ \beta & \beta & \alpha & -(2\beta+\alpha) \end{bmatrix}.$$

According to Eq. (1.61) of (Yang 2014a), the equilibrium is achieved when the following holds

$$\pi Q = \mathbf{0}.$$

Thus, we have

$$\begin{cases} -(\alpha+2\beta)\pi_T + \alpha\pi_C + \beta\pi_A + \beta\pi_G = 0 \\ \alpha\pi_T - (\alpha+2\beta)\pi_C + \beta\pi_A + \beta\pi_G = 0 \\ \beta\pi_T + \beta\pi_C - (\alpha+2\beta)\pi_A + \alpha\pi_G = 0 \\ \beta\pi_T + \beta\pi_C + \alpha\pi_A - (\alpha+2\beta)\pi_G = 0 \end{cases}.$$

Because $\sum_j \pi_j = 1$, it is not difficult to see that due to symmetry, $\pi_T = \pi_C = \pi_A = \pi_G = 0.25$.

1.6 A large genomic region evolves neutrally according to the JC69 model, at the rate 2×10^{-8} substitutions/site/year (this is roughly the rate in the mitochondria in mammals). (**a**) Suppose initially the sequence consists of Ts only. What will be the proportions of T, C, A, and G in the sequence 10^6 and 10^8 years later? (**b**) Do the same calculation assuming that the sequence initially had Cs only. (**c**) Do the same calculation if the initial proportions of T, C, A, and G are $\pi_0 = (0.4,\ 0.3,\ 0.2,\ 0.1)$.

Solution.

According to the context of the problem, the rate, 3λ, is given in the problem as 2×10^{-8} substitutions/site/year.

a) The starting frequency $\pi^{(0)}$ is given by $\pi^{(0)} = (1,0,0,0)$. Under the JC69 model, according to Eq. (1.4) in (Yang 2014a), the transition probability matrix is defined as

$$P(t) = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}, \text{ with } \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \\ p_{1(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \end{cases}.$$

Thus, for evolution over 10^6 and 10^8 years, it follows that

$$\begin{aligned} \pi^{(10^6)} &= \pi^{(0)} P(t = 10^6) \\ &= \left[\frac{1}{4} + \frac{3}{4}e^{-\frac{0.02\times4}{3}} \quad \frac{1}{4} - \frac{1}{4}e^{-\frac{0.02\times4}{3}} \quad \frac{1}{4} - \frac{1}{4}e^{-\frac{0.02\times4}{3}} \quad \frac{1}{4} - \frac{1}{4}e^{-\frac{0.02\times4}{3}}\right] \\ &= [0.98 \quad 0.007 \quad 0.007 \quad 0.007], \end{aligned}$$

and

$$\pi^{(10^8)} = \pi^{(0)}P(t = 10^8)$$
$$= \left[\frac{1}{4}+\frac{3}{4}e^{-\frac{2\times4}{3}} \quad \frac{1}{4}-\frac{1}{4}e^{-\frac{2\times4}{3}} \quad \frac{1}{4}-\frac{1}{4}e^{-\frac{2\times4}{3}} \quad \frac{1}{4}-\frac{1}{4}e^{-\frac{2\times4}{3}}\right]$$
$$= [0.30 \quad 0.23 \quad 0.23 \quad 0.23].$$

b) The starting frequency $\pi^{(0)}$ is given by $\pi^{(0)} = (0,1,0,0)$. Hence,

$$\pi^{(10^6)} = [0.007 \quad 0.98 \quad 0.007 \quad 0.007],$$
$$\pi^{(10^8)} = [0.23 \quad 0.30 \quad 0.23 \quad 0.23].$$

c) The starting frequency $\pi^{(0)}$ is given by $\pi^{(0)} = (0.1,0.2,0.3,0.4)$. Hence,

$$\pi^{(10^6)} = \pi^{(0)}P(t = 10^6)$$
$$= [0.1 \quad 0.2 \quad 0.3 \quad 0.4]\begin{bmatrix} \frac{1}{4}+\frac{3}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} \\ \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}+\frac{3}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} \\ \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}+\frac{3}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} \\ \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}-\frac{1}{4}e^{-\frac{0.02\times4}{3}} & \frac{1}{4}+\frac{3}{4}e^{-\frac{0.02\times4}{3}} \end{bmatrix}$$
$$= [0.104 \quad 0.201 \quad 0.300 \quad 0.396].$$

Likewise, $\pi^{(10^8)} = [0.240 \quad 0.247 \quad 0.253 \quad 0.260]$.

R

```
f <- function(pi_0, t){ p0<-1/4+3/4*exp(-4*lambda*t); p1<-1/4-1/4*exp(-4*lambda*t);  P_t <-
matrix(c(p0,p1,p1,p1, p1,p0,p1,p1, p1,p1,p0,p1, p1,p1,p1,p0), ncol=4, byrow=T); pi_0 %*% P_t }
> pi_0s <- matrix(c(1,0,0,0, 0,1,0,0, 0.1,0.2,0.3,0.4), byrow=T, ncol=4)
> t(apply(pi_0s, 1, f, t=10^6))
> t(apply(pi_0s, 1, f, t=10^8))
```

1.7 Use the 12s rRNA data of Table 1.3 to conduct the LRT to compare K80 against HKY85, and HKY85 against GTR. The numbers of parameters under the models are listed in Table 1.1, and the log likelihood values are listed in Table 1.4, but you may prefer running a program (such as BASEML in the PAML package, Yang 2007b) to do the calculation yourself.

Solution.

If you were as lazy as I, you can find the data necessary to do the likelihood ratio test in Table 1.4 of (Yang 2014a) where the log-likelihoods under K80 and GTR are -1637.90 and -1610.36 respectively. Thus, the difference of log-likelihood is equal to $1637.90 - 1610.36 = 27.54$. The number of different parameters of these two nested models are $8 - 1 = 7$. The following R code

gives the *P*-value of the LRT test, which is 1.43747e-09, therefore the simpler model K80 is rejected at a significance level of 0.05. Note that it is $27.54 \times 2 = 55.08$ that should be the statistic to use.

R
> pchisq(27.54*2, 7, lower.tail=FALSE)

Chapter 2. Models of amino acid and codon substitution

2.1 Download the human and orangutan NADH6 gene sequences from GenBank (accession numbers X93334 for *Homo sapiens* and D38115 for *Pongo pygmaeus*), align them and apply the methods discussed in this chapter to estimate d_S and d_N. One way of aligning protein-coding DNA sequences is to use CLUSTAL (Thompson et al. 1994) to align the protein sequences first and then construct the DNA alignment based on the protein alignment, using, for example, MEGA (Tamura et al. 2011). Use CODEML to estimate $d_S, d_N, d_{1B}, d_{2B}, d_{3B}, d_S^*, d_N^*$, etc. Assess the impact of allowing for the transition–transversion rate difference on the estimation. Also examine the impact of the model assumption about codon frequencies, by using models such as Fequal, F1 × 4, F3 × 4, F61, and FMutSel.

See Problem 2.1 of (Yang 2006).

2.2* Are there really three nucleotide sites in a codon? How many synonymous and nonsynonymous sites are in the codon TAT (under the universal code)? Assume the transition/transversion rate ratio $\kappa = 1$.

See Problem 2.2 of (Yang 2006).

2.3 Verify $d_S = (d_{1B} + d_{2B} + d_{3B})/3$ (equation (2.27)).

Solution.

It is easy to verify using the result of Problem 2.1.

2.4 Conduct a computer simulation to examine the impact of sequence divergence on estimation of ω in comparison of two gene sequences. Assume the Fequal model ($\pi_j = 1/61$ for all j), with $\kappa = 2$, and set $\omega = 0.5$. Assume 500 codons in the gene. Use a few different sequence distances, such as $t = 0.1, 0.3, 0.5, 1.0, 1.5,$ 2, and 3 nucleotide substitutions per codon. Use EVOLVER to generate the data and CODEML to analyse them. Use 1,000 replicates to calculate the mean and variance of $\hat{\omega}$ for each sequence distance. What sequence divergence appears to be optimal for estimating ω? Take a guess before you conduct the simulation experiment.

Solution.

To save time, I use only 100 replicates by setting ndata = 100 in *codeml.ctl*. Then, use the following R code to obtain the mean and variance of $\hat{\omega}$ for data sets simulated with each tree length. My output is shown as follows.

R
> d <- read.table("omega.tab", header=T);

```
> lapply(d, function(x){cat(mean(x), var(x), "\n", sep="\t")})
```

Tree length	Mean	Variance
0.1	0.568627	0.036107
0.3	0.500472	0.007161
0.5	0.515791	0.006978
1.0	0.508956	0.004473
1.5	0.501194	0.002129
2.0	0.508384	0.002945
3.0	0.499472	0.005358

Bash

```
for i in 0.1 0.3 0.5 1 1.5 2 3; do
        echo $i
        sed "/tree length/s/^[^ ]\+/$i/" -i MCcodon.dat
        evolver 6 MCcodon.dat >/dev/null
        codeml >/dev/null
        grep omega mlc | awk '{print $NF}' > omega-$i.list
        echo
done
```

2.5 How large a sample is large enough for the χ^2 approximation to the LRT statistic to be reliable? Conduct a computer simulation to examine the null distribution of the LRT statistic ($2\Delta\ell$) for testing the hypothesis $\omega = 1$, in comparison with χ^2_1 (the χ^2 distribution with df = 1) (see §1.4.3). Assume the Fequal model (with $\pi_j = 1/61$ for codon j), with $t = 0.5$, $\kappa = 2$, and $\omega = 0.5$. Use different numbers of codons in the gene, such as 50, 100, 200, 300, 500, or 1000. Use EVOLVER to generate the data and CODEML to analyse them. Use 1,000 or 10,000 replicates. Analyse each replicate dataset under H_0 (with $\omega = 1$ fixed) and H_1 (with ω estimated) to calculate the test statistic $2\Delta\ell = 2(\ell_1 - \ell_0)$. Plot the histogram of $2\Delta\ell$ (e.g. using the R command `hist`) and compare with χ^2_1. Also examine the estimated significance values and compare with the theoretical expectations (3.84 at 5% and 6.63 at 1%).

Solution.

I take the liberty to use AliSim instead of EVOLVER to generate the sequence.

Bash

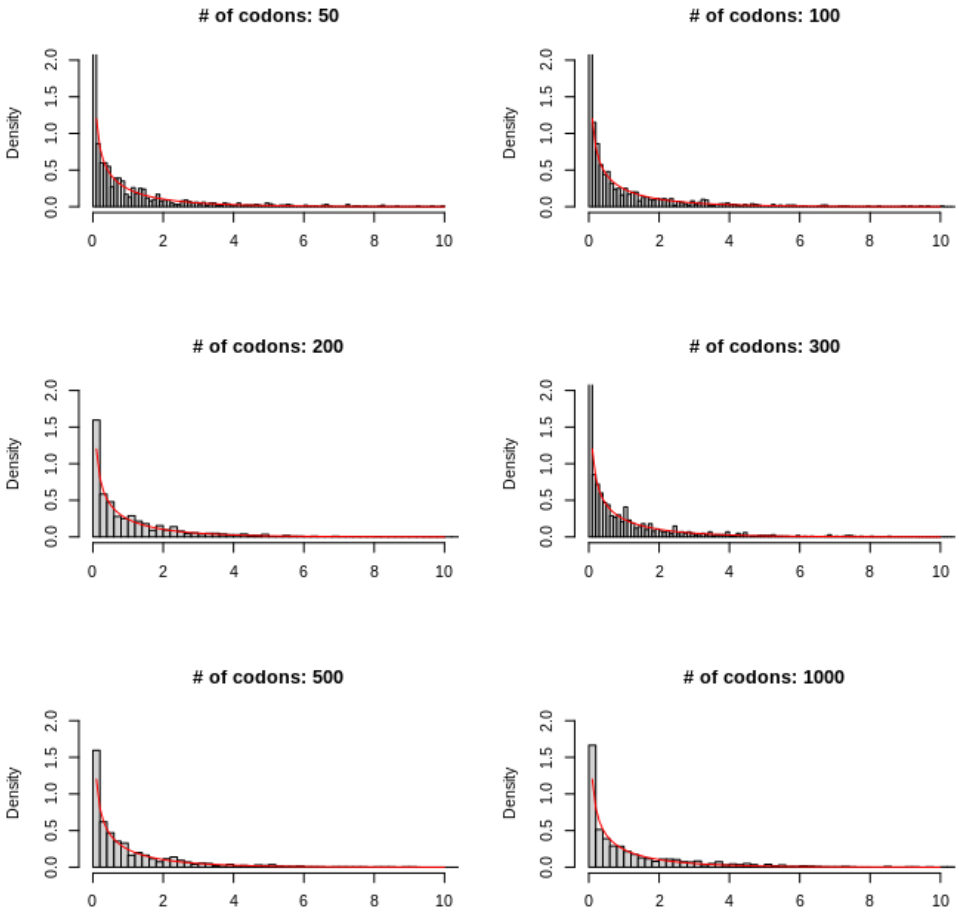
```
n=1000
```

```
n_codons=(50 100 200 300 500 1000)

for n_codon in ${n_codons[@]}; do
        echo -e "# of codons:\t$n_codon"
        [ -f deltaLL$n_codon.list ] && rm deltaLL$n_codon.list
        for i in `seq $n`; do
                mod=`echo "$i%100"|bc`
                [ $mod == 1 ] && echo $i >&2
                length=`expr 3 \* $n_codon`
                iqtree --alisim alignment_codon -m 'MG2K{1,2}+FQ' --length $length -t tree.nwk -af
phy -quiet
                # omega to be estimated
                sed -i 's/fix_omega.*/fix_omega = 0/' codeml.ctl
                codeml >/dev/null
                lnl1=`grep -P 'lnL' mlc | awk '{print $(NF-1)}'`
                # omega fixed as 1.0
                sed -i 's/fix_omega.*/fix_omega = 1/' codeml.ctl
                codeml >/dev/null
                lnl0=`grep -P 'lnL' mlc | awk '{print $(NF-1)}'`
                echo "scale=2; 2 * ($lnl1 - $lnl0)" | bc
        done > deltaLL$n_codon.list
        echo
done
```

The following result shows that with 100 codons the chi-square approximation is good enough but even with 50 codons it is not fine. This can be visualized by using the following R code.

R

```
> par(mfrow=c(2,3))
> for(i in c(50,100,200,300,500,1000)) {
    a<-unlist(read.table(paste0("deltaLL",i,".list")))
    hist(a, freq=F, breaks=80,ylim=c(0,2), xlim=c(0,10), main=paste0("# of codons: ", i), xlab="")
    curve(dchisq(x,df=1),add=T, col="RED") }
}
```


# of codons: 50
# of codons: 100
# of codons: 200
# of codons: 300
# of codons: 500
# of codons: 1000
Density
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8
10

Chapter 3. Phylogeny reconstruction: overview

3.1 Draw the tree

(((human: 0.040, chimpanzee: 0.052): 0.016, gorilla: 0.059): 0.047, orangutan: 0.090, gibbon: 0.125);

The branch lengths are the MLEs under JC69 obtained from the mitochondrial data of Brown et al. (1982). Identify the most distant pair of species and use midpoint rooting to root the tree. Draw the resulting rooted tree.

Solution.

Use the following R code to find the pair of taxa most distant to each other. The package *phytools* (Revell 2012) may need to be installed.

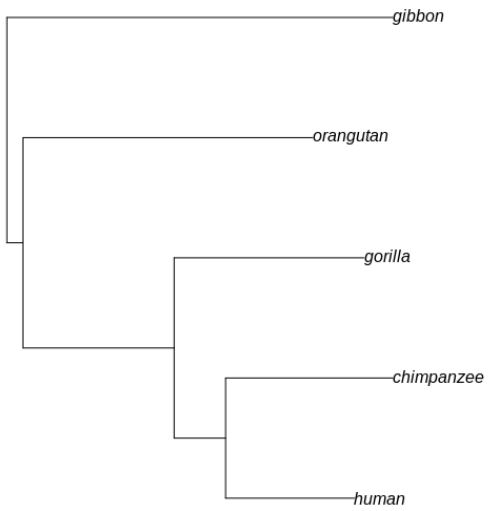
R

```
> library(phytools)
> tree <- read.tree( text="(((human: 0.040, chimpanzee: 0.052): 0.016, gorilla: 0.059): 0.047,
orangutan: 0.090, gibbon: 0.125);" )
> which(dist.nodes(tree) == max(dist.nodes(tree)), arr.ind = TRUE)
> tree$tip.label
```

This gives taxon 2 and taxon 5, which are chimpanzee and gibbon respectively. Then use the R code to root the tree and visualize it.

R

```
rooted_tree <- midpoint.root(tree)
plot(rooted_tree)
```



3.2 Write two equivalent Newick representations of the tree in Figure 3.9b.

(**b**) Strict consensus tree (**c**) Majority-rule consensus tree

Fig. 3.9 Three trees for eight species (**a**) and their strict consensus tree (**b**) and majority-rule consensus tree (**c**).

Solution.

Two possible ways are ((A,B,C),D,E,(F,G,H)), or (((A,B,C),D),E,(H,F,G)), among many others.

3.3 The following rooted tree is shown in Figure 3.26:

(a:0.05, (c: 0.07, ((b:0.015, f:0.12) :0.01, (d:0.01, e:0.4) :0.005) :0.03) :0.025);

Which of the following statements are incorrect?
(a) Species d and e are most closely related.
(b) Sequences b and d are most similar.
(c) Species b is more closely related to d than to e.
(d) Species d is more closely related to c than to f.

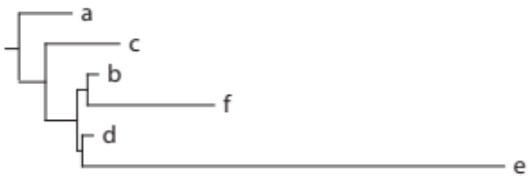


Fig. 3.26 A tree showing branch lengths for Problem 3.3.

Solution.

a) Correct. It is important to note that when we say that some taxa are more closely related it means that they a more recent common ancestor than other taxa [see Section 3.1.1.8 of (Yang 2014a)]. Because species d and e share the most recent common ancestor, they are more closely related to each other than to any others in the tree.

b) Correct. The sequence similarity may be measured by genetic distance. In view of the tree, this is calculated as the length of branches connecting any two tips. According to the Newick-formatted tree given in the problem, the distance between sequence b and sequence d is 0.01+0.005+0.01+0.015=0.04 which indeed is the smallest.

c) Incorrect. Because taxa b and e form a monophyly, b is more closely related to e than to any others in the tree.

d) Incorrect. Taxa d and f share a more recent common ancestor than with taxon c.

3.4 Calculate the partition distance between the two trees of Figure 3.11.

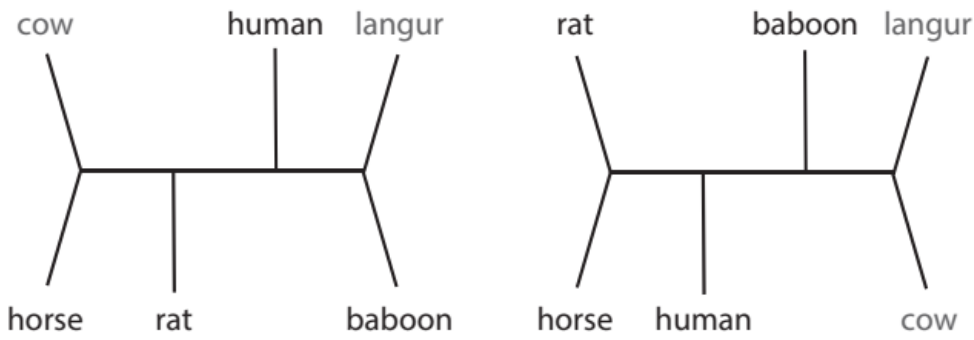


Fig. 3.11 Convergent evolution in the stomach lysozyme of

Solution.

In Section 3.1.1.6 of (Yang 2014a), it is clearly indicated that the partition distance between two trees is the number of bipartitions that are in one but not the other. Because none of the 3 bipartitions in the left tree is included in the right tree, and the same holds for the right tree, the partition distance equals $3 + 3 = 6$.

You can verify this by the following R code. Note that the package ape needs to be installed.

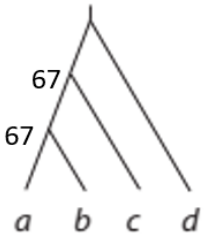
R

```
> library(ape)
> tree1 <- read.tree(text = "(((c,horse),r), human,(l,b));")
> tree2 <- read.tree(text = "(((r,horse),human), b,(l,c));")
> dist.topo(tree1,tree2)
```

3.5 Use the three trees of Figure 3.27 to construct the majority-rule consensus tree, and show the support values for the nodes on it.

Fig. 3.27 Three rooted trees for constructing the majority-rule consensus tree in Problem 3.5.

Solution.


67
67
a
b
c
d

Chapter 4. Maximum likelihood methods

4.1 Calculate the probabilities of site patterns *xxx*, *xxy*, *yxx*, and *xyx* as a function of the branch lengths t_{10} and t_{11} in the tree τ_1 of Figure 4.14. Assume the symmetrical substitution model for binary characters (equation (1.79)).

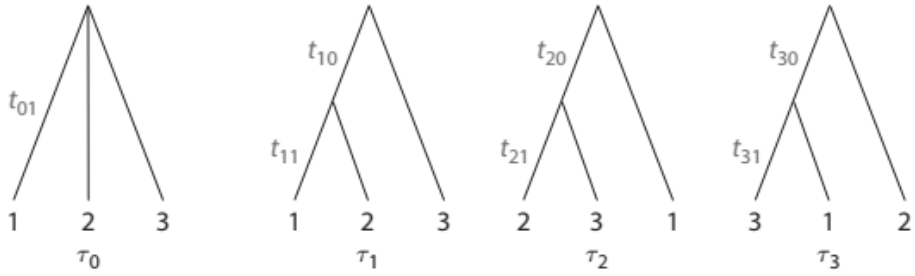


Fig. 4.14 The three rooted trees for three species: $\tau_1, \tau_2,$ and τ_3. Branch lengths t_{i0} and t_{i1} in each tree τ_i ($i = 1, 2, 3$) are measured by the expected number of character changes per site. The star tree τ_0 is also shown with its branch length t_{01}.

Solution.

In Eq. (1.79) of (Yang 2014a), the transition probability matrix is given as

$$P(t) = \begin{bmatrix} p_0 & p_1 \\ p_1 & p_0 \end{bmatrix},$$

where $p_0 = \frac{1}{2} + \frac{1}{2}e^{-2t}$ and $p_1 = \frac{1}{2} - \frac{1}{2}e^{-2t}$.

Denote the state at node i (either tip or ancestral node) as S_i. According to the problem statement, S_i are restricted to only two values, which we assign as A or B. The following can be easily calculated based on Section 4.2 of (Yang 2014a). Further, define $a = e^{-2t_{10}}$ and $b = e^{-2t_{11}}$, thus $e^{-2(t_{10}+t_{11})} = ab$.

Hence,

$$\begin{aligned} P(xxx) &= P(S_1 = A, S_2 = A, S_3 = A) + P(S_1 = B, S_2 = B, S_3 = B) \\ &= 2 \times P(S_1 = A, S_2 = A, S_3 = A) \\ &= 2 \times 0.5 \times \left(P(S_1 = A, S_2 = A, S_3 = A | S_0 = A) + P(S_1 = A, S_2 = A, S_3 = A | S_0 = B)\right) \\ &= P(S_1 = A, S_2 = A, S_3 = A, S_4 = A | S_0 = A) + P(S_1 = A, S_2 = A, S_3 = A, S_4 = B | S_0 = A) \\ &\quad + P(S_1 = A, S_2 = A, S_3 = A, S_4 = A | S_0 = B) \\ &\quad + P(S_1 = A, S_2 = A, S_3 = A, S_4 = B | S_0 = B) \end{aligned}$$

$$= \left(\frac{1}{2}+\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$+\left(\frac{1}{2}+\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$+\left(\frac{1}{2}-\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$+\left(\frac{1}{2}-\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$= \frac{1}{16}\big((1+ab)(1+a)(1+b)(1+b)+(1+ab)(1+a)(1-b)(1-b)$$
$$+(1-ab)(1-a)(1+b)(1+b)+(1-ab)(1+a)(1-b)(1-b)\big)$$
$$= \frac{1}{16}(4+2a-4ab+2a^2b+4b^2+2ab^2+a^2b^2+2a^2b^3),$$

$$P(xxy) = P(S_1=A,S_2=A,S_3=B)+P(S_1=B,S_2=B,S_3=A)$$
$$= 2\times P(S_1=A,S_2=A,S_3=B)$$
$$= P(S_1=A,S_2=A,S_3=B,S_4=A|S_0=A)+P(S_1=A,S_2=A,S_3=B,S_4=B|S_0=A)$$
$$+P(S_1=A,S_2=A,S_3=B,S_4=A|S_0=B)$$
$$+P(S_1=A,S_2=A,S_3=B,S_4=B|S_0=B)$$
$$= \left(\frac{1}{2}-\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$+\left(\frac{1}{2}-\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$+\left(\frac{1}{2}+\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$+\left(\frac{1}{2}+\frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2}+\frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2}-\frac{1}{2}e^{-2t_{11}}\right)^2$$
$$= \frac{1}{16}[(1-ab)(1+a)(1+b)(1+b)+(1-ab)(1-a)(1-b)(1-b)$$
$$+(1+ab)(1-a)(1+b)(1+b)+(1+ab)(1+a)(1-b)(1-b)]$$
$$= -\frac{1}{2}a^2b^2+\frac{1}{4}(b^2+1),$$

$$P(yxx) = P(S_1=A,S_2=B,S_3=B)+P(S_1=B,S_2=A,S_3=A)$$
$$= 2\times P(S_1=A,S_2=B,S_3=B)$$
$$= P(S_1=A,S_2=B,S_3=B,S_4=A|S_0=A)+P(S_1=A,S_2=B,S_3=B,S_4=B|S_0=A)$$
$$+P(S_1=A,S_2=B,S_3=B,S_4=A|S_0=B)$$
$$+P(S_1=A,S_2=B,S_3=B,S_4=B|S_0=B)$$

$$= \left(\frac{1}{2} - \frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2} + \frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2} + \frac{1}{2}e^{-2t_{11}}\right)\left(\frac{1}{2} - \frac{1}{2}e^{-2t_{11}}\right)$$
$$+ \left(\frac{1}{2} - \frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2} - \frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2} + \frac{1}{2}e^{-2t_{11}}\right)\left(\frac{1}{2} - \frac{1}{2}e^{-2t_{11}}\right)$$
$$+ \left(\frac{1}{2} + \frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2} - \frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2} + \frac{1}{2}e^{-2t_{11}}\right)\left(\frac{1}{2} - \frac{1}{2}e^{-2t_{11}}\right)$$
$$+ \left(\frac{1}{2} + \frac{1}{2}e^{-2(t_{10}+t_{11})}\right)\left(\frac{1}{2} + \frac{1}{2}e^{-2t_{10}}\right)\left(\frac{1}{2} - \frac{1}{2}e^{-2t_{11}}\right)^2$$
$$= \frac{1}{16}\big((1-ab)(1+a)(1+b)(1-b) + (1-ab)(1-a)(1+b)(1-b)$$
$$+ (1+ab)(1-a)(1+b)(1-b) + (1+ab)(1+a)(1-b)(1-b)\big)$$
$$= \frac{1}{8}\Big(a^2b^3 - a^2b^2 + (ab^3 - ab) + \big(2 - (b+b^2)\big)\Big).$$

***4.2** Try to estimate the single branch length under the JC69 model for the star tree of three sequences under the molecular clock (see Saitou (1988) and Yang (1994c, 2000a), for discussions of likelihood tree reconstruction under this model). The tree is shown in Fig. 4.8, where t is the only parameter to be estimated. Note that there are only three site patterns, with one, two, or three distinct nucleotides, respectively. The data are the observed numbers of sites with such patterns: n_0, n_1, and n_2, with the sum to be n. Let the proportions be $f_i = n_i/n$. The log likelihood is $\ell = n\sum_{i=0}^{2} f_i \log(p_i)$, with p_i to be the probability of observing site pattern i. Derive p_i by using the transition probabilities under the JC69 model, given in equation (1.3). You can calculate $p_0 = \Pr(\text{TTT})$, $p_1 = \Pr(\text{TTC})$, and $p_2 = \Pr(\text{TCA})$. Then set $\mathrm{d}\ell/\mathrm{d}t = 0$. Show that the transformed parameter $z = \mathrm{e}^{-4/3t}$ is a solution to the following quintic equation:

$$36z^5 + 12(6 - 3f_0 - f_1)z^4 + (45 - 54f_0 - 42f_1)z^3 + (33 - 60f_0 - 36f_1)z^2$$
$$+ (3 - 30f_0 - 2f_1)z + (3 - 12f_0 - 4f_1) \equiv 0.$$

See Problem 4.2 in (Yang 2006).

4.3 *Long-branch attraction for parsimony.* Calculate the probabilities of sites with data *xxyy*, *xyyx*, and *xyxy* in four species for the unrooted tree of Figure 4.18, using two branch lengths p and q under a symmetrical substitution model for binary characters (equation (1.79)). Here it is more convenient to define the branch length as the proportion of different sites at the two ends of the branch. Show that $\Pr(xxyy) < \Pr(xyxy)$ if and only if $q(1-q) < p^2$. With such branch lengths, parsimony for tree reconstruction is inconsistent (Felsenstein 1978a).

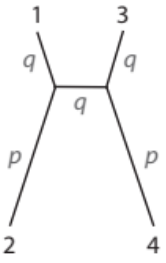


Fig. 4.18 A tree of four species with two branch lengths p and q, defined as the probability that any site is different at the two ends of the branch. For a binary character, this probability is $p = (1 - e^{-2t})/2$, where t is the expected number of character changes per site (see equation (1.79)).

See Problem 4.3 in (Yang 2006).

4.4 Bias in ancestral state reconstruction. Calculate the posterior probabilities for T, C, A, and G at the root of the tree of Figure 4.10 when the observed data at the site is $x_1x_2x_3$ = AAG. Assume the F81 substitution model, with base frequency parameters π_T = 0.2263, π_C = 0.3282, π_A = 0.3393, and π_G = 0.1062. Suppose that each branch length is 0.2, and the transition probability matrix is given in equation (4.26). Hint: Use equation (4.23).

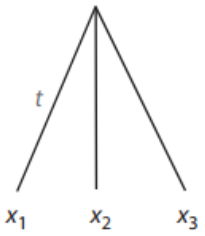


Fig. 4.10 A tree of three species for demonstrating the bias in ancestral reconstruction. The three branch lengths are equal, at t = 0.2 substitutions per site.

Solution.

Recall Eq. (4.26) in (Yang 2014a), where the transition matrix is given as

$$P(0.2) = \begin{bmatrix} 0.811138 & 0.080114 & 0.082824 & 0.025924 \\ 0.055240 & 0.836012 & 0.082824 & 0.025924 \\ 0.055240 & 0.080114 & 0.838722 & 0.025924 \\ 0.055240 & 0.080114 & 0.082824 & 0.781821 \end{bmatrix}.$$

Denote the ancestral node as node 0, such that its ancestral state is expressed as X_0 which can take any of the four nucleotides, denoted as x_0. The posterior probability that the ancestral node (node 0) takes character x_0 can be expressed as follows, according to Eq. (4.23) of (Yang 2014a)

$$f(x_0|x_h;\theta) = \frac{\pi_{x_0} L_0(x_0)}{\sum_{x_0 \in \{T,C,A,G\}} \pi_{x_0} L_0(x_0)}.$$

According to the problem statement, the equilibrium frequencies of each nucleotide are given as $\pi_T = 0.2263, \pi_C = 0.3282, \pi_A = 0.3393, \pi_G = 0.1062$.

For the first site $x_1 = A$, it can be calculated that

$$\pi_A L_0(T) = 0.2263 \times 0.055240 = 0.01250081,$$
$$\pi_C L_0(C) = 0.3282 \times 0.080114 = 0.02629341,$$
$$\pi_A L_0(A) = 0.3393 \times 0.838722 = 0.2845784,$$
$$\pi_G L_0(G) = 0.1062 \times 0.025924 = 0.002753129.$$

Accordingly, we obtain

$$\sum_{x_0 \in \{T,C,A,G\}} \pi_{x_0} L_0(x_0) = 0.01250081 + 0.02629341 + 0.2845784 + 0.002753129 = 0.3261257.$$

Hence, it is easy to calculate that the posterior probabilities at the ancestral node at the first site are $0.038331257, 0.080623533, 0.872603285, 0.008441925$ for T, C, A, G, respectively.

The result is the same for the second site which is also A.

The third site is G. We have

$$\pi_A L_0(G) = 0.2263 \times 0.055240 = 0.01250081,$$
$$\pi_C L_0(C) = 0.3282 \times 0.080114 = 0.02629341,$$
$$\pi_A L_0(A) = 0.3393 \times 0.082824 = 0.02810218,$$
$$\pi_G L_0(G) = 0.1062 \times 0.781821 = 0.08302939.$$

Hence, it is easy to calculate that the posterior probabilities at the ancestral node at the first site are 0.08337998, 0.17537616, 0.18744060, 0.55380325 for T, C, A, G, respectively.

4.5 Use the plastid *rbcL* genes from 12 plant species to test the goodness of fit of the JC69 model. Follow the example of §4.7.2. Use BASEML in the PAML package to analyse the original data to generate branch lengths and calculate $\Delta\ell = \ell_{\max} - \ell_{\mathrm{JC}}$. Use those branch lengths to simulate 1,000 datasets using the program SEQ-GEN or EVOLVER. Then use a likelihood program (such as BASEML) to analyse the 1,000 replicate datasets to calculate Δ_i to construct a histogram. Your results should be similar to Figure 4.17a.

Solution.

Data preparation.

Download the alignment and the tree from Yang's website (Yang 2014b) (see C2). Convert the alignment into FASTA format. Then, remove all sites that have any ambiguous sites or gaps. See pp. 146 of (Yang 2014a):

| *We do not have a model for alignment gaps, so we remove sites with gaps, with 1,312 sites left.*

This should generate an alignment of 1312 bp, available in the file "**rbcL.nogaps_amb.fas**".

Check the R script

I write the following R script to calculate log-likelihood under the multinomial distribution model where ℓ_{max} is calculated as

$$\ell_{max} = \sum_{i=1}^{4^s} n_i \log\left(\frac{n_i}{n}\right),$$

as given by Eq. (4.39) of (Yang 2014a). Of course, an alternative way is to use the value calculated by BASEML.

R: *4.5.R*

```
library(seqinr)
args = commandArgs(trailingOnly=TRUE)
seq_file <- args[1]
s <- read.fasta(seq_file)
v <- list()
for(i in 1:getLength(s)[1]){
    v[[i]] <- sapply(s, function(x) x[i])
    names(v[[i]]) <- names(s)
}
m <- t(sapply(v, unlist, list(use.names=F)))
dimnames(m) <- NULL
h <- apply(m, 1, function(x){paste(x,collapse="")})
t <- table(h)
n <- sum(t)
lnl <- 0
for(n_i in table(h)){
    lnl <- lnl + n_i * log(n_i/n)
}
cat(lnl,"\n")
```

Examine that the R script (*4.5.R*) yields $\ell_{max} = -4025.03$, exactly the same as that shown in Section 4.7.2 of (Yang 2014a) under the multinomial model. Also verify that the log-likelihood under the model HKY+G5 is generated by using IQ-Tree is -5703.967.

Bash

```
$ iqtree -s rbcL.nogaps_amb.fas -m HKY+G5 -te vegetables.nwk -redo
$ Rscript 4.5.R
```

Run analysis

Now, do. The following is my setting of the control file of EVOLVER. Note that I indicate “2” in the first row which indicates that the sequence output is *mc.nex* in Nexus format.

```
2          * 0,1:seqs or patters in paml format (mc.paml); 2:paup fo
rmat (mc.nex)
-1   * random number seed (odd number)

12 1312 1  * <# seqs>  <# nucleotide sites>  <# replicates>
-1         * <tree length, use -1 if tree below has absolute branch
lengths>

(kiwi_fruit:0.0298095265,((((agave:0.0293638477,garlic:0.0191949024)
:0.0158379569,rice:0.0828491683):0.0203804464,black_pepper:0.0585448
367):0.0265371030,((cabbage:0.0501439829,cotton:0.0334874188):0.0144
799379,(cucumber:0.0372660217,walnut:0.0214057950):0.0136341536):0.0
045273131):0.0047901507,(sunflower:0.0515783574,(tomato:0.0089322806
,tobacco:0.0041241907):0.0336170934):0.0096308085);

0          * model: 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85, 5:T92, 6:T
N93, 7:REV
1          * kappa or rate parameters in model
0.  0      * <alpha>  <#categories for discrete gamma>

//

0.25  0.25  0.25  0.25    * base frequencies
  T   C   A   G
```

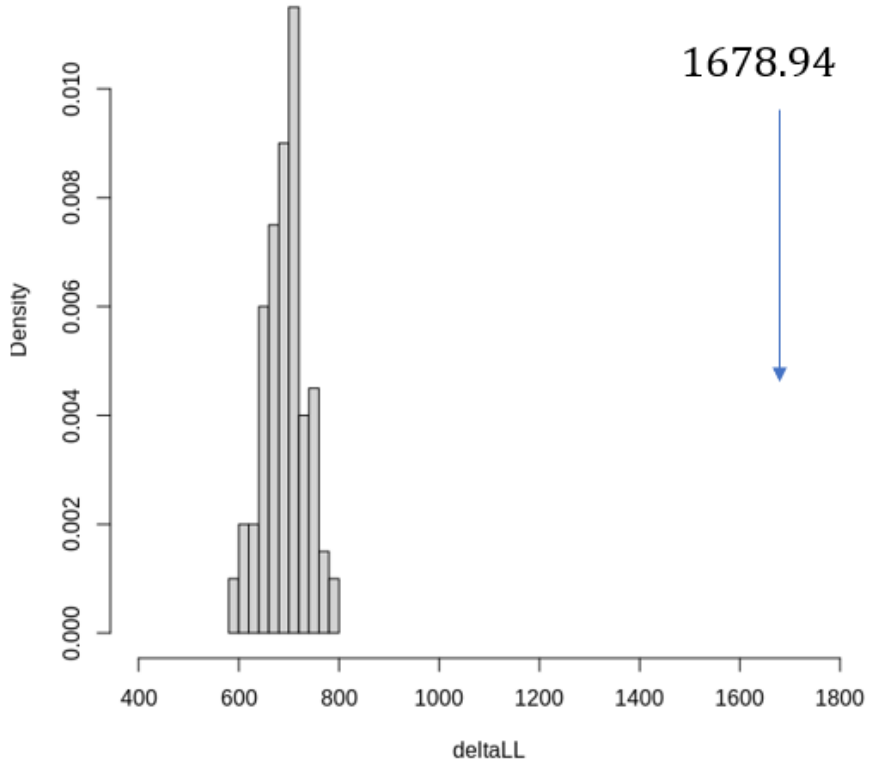
Bash (*4.5.sh*)

```
model=JC69
evolver 5 JC69.dat >/dev/null
python nexus2fasta.py mc.nex sample.fasta >/dev/null
iqtree -s sample.fasta -m $model -te vegetables.nwk -redo -quiet
lnl=`grep "Optimal log-likelihood" sample.fasta.log | awk '{print $NF}'`
lnl_max=`Rscript 4.5.R sample.fasta 2>/dev/null`
echo "scale=2; $lnl_max - $lnl" | bc
```

Run the following to generate the graph below.

Bash

```
$ for i in `seq 1000`; do bash 4.5.sh ; done > deltaLL.list
$ R -e 'v=unlist(read.table("deltaLL.list")); hist(v, freq=F, xlab="deltaLL")'
```



4.6 *Phylogenetic reconstruction using ML*. Use your own data or find a small dataset of 10–50 species from the literature to infer the ML phylogeny under various substitution models, such as JC69, K80, HKY85, GTR, and the gamma variants JC69+Γ_5, K80+Γ_5, HKY85+Γ_5, and GTR +Γ_5. You can use PHYML to run tree search under those models.

Solution.

You are welcome to follow what is told in the problem but I am lazy enough to just take the nucleotide alignment from Problem 4.5, i.e., the file "**rbcL.nogaps_amb.fas**". Here, I choose to use IQ-Tree and raxml-ng for tree construction. The command is as follows. Note that JC69 is specified as "JC", and HKY85 is specified as "HKY" in raxml-ng.

Bash

```
#! /bin/bash
# IQ-Tree
mkdir iqtree
cd iqtree
for m in JC69 K80 HKY85 GTR JC69+G5 K80+G5 HKY85+G5 GTR+G5; do
        echo $m
        iqtree -redo -m $m -pre $m -s ../sample.fasta -quiet
done
cd ..
```

```
# raxml-ng
mkdir raxml-ng
cd raxml-ng
for m in JC K80 HKY GTR JC+G5 K80+G5 HKY+G5 GTR+G5; do
        echo $m
        raxml-ng --redo --msa ../sample.fasta --model $m --prefix $m
done
cd ..
```

Then, use the following Bash command to summarize the result shown at the end.

Bash

```
for m in JC69 K80 HKY85 GTR JC69+G5 K80+G5 HKY85+G5 GTR+G5; do grep "BEST SCORE FOUND" iqtree/$m.log | awk '{print $NF}'; done
for m in JC K80 HKY GTR JC+G5 K80+G5 HKY+G5 GTR+G5; do grep "Final LogLikelihood" raxml-ng/$m.raxml.log | awk '{print $NF}'; done
```

model	IQ-Tree	raxml-ng
JC69	-5703.967	-5703.966563
K80	-5565.204	-5565.20372
HKY85	-5551.769	-5549.292507
GTR	-5531.846	-5531.350899
JC69+G5	-5405.809	-5405.808531
K80+G5	-5254.135	-5254.135172
HKY85+G5	-5242.351	-5240.447588
GTR+G5	-5228.581	-5228.470939

Chapter 5. Comparison of phylogenetic methods and tests on trees

To make the result more "interesting", I change in all exercises of Chapter 5 the number of sites from 1000 to 100.

5.1 Conduct a computer simulation to examine the efficiency of tree reconstruction methods. Use SEQ-GEN or EVOLVER to generate 1,000 datasets, each of 1,000 sites, under the JC69 model, and then construct the maximum parsimony and maximum likelihood trees from each to calculate the probability of recovering the correct tree. You can use PHYLIP or some other programs for tree reconstruction. Use three different shapes of four-species trees:

i. $((a: 0.1,\ b: 0.1): 0.05,\ c: 0.1,\ d: 0.1)$;
ii. $((a: 0.01,\ b: 0.01): 0.01,\ c: 0.05,\ d: 0.05)$;
iii. $((a: 0.01,\ b: 0.05): 0.01,\ c: 0.01,\ d: 0.05)$.

5.2 Conduct a computer simulation to examine the robustness of tree reconstruction methods to transition–transversion rate difference. Redo the simulation of Problem 5.1, but simulate the data under the K80 model with transition/transversion rate ratio $\kappa = 5$. Analyse the data using three methods: (i) parsimony, (ii) ML assuming JC69, and (iii) ML assuming K80 (with κ estimated).

5.3 Conduct a computer simulation to examine the robustness of tree reconstruction methods to rate variation among sites. Redo the simulation of Problem 5.1, but simulate the data under the JC69 + Γ_5 model with the gamma shape parameter $\alpha = 0.5$. Analyse the data using three methods: (i) parsimony, (ii) ML assuming JC69, and (iii) ML assuming JC69 + Γ_5 (with α fixed at 0.5).

5.4 Conduct a computer simulation to examine the relative performance of parsimony and likelihood methods of tree reconstruction when the sequences are highly divergent. Redo the simulation of Problem 5.1 using trees (ii) and (iii), with all branch lengths in the trees multiplied by 5.

Solution.

	5.1		***5.2***			***5.3***			***5.4***	
	MP	ML (JC69)	MP	ML (K80)	ML (JC69)	MP	ML (JC69)	ML (JC69+G5{0.5})	MP	ML (JC69)
Tree1	0.89	0.91	0.90	0.90	0.91	0.80	0.88	0.87	-	-
Tree2	0.71	0.80	0.82	0.88	0.87	0.80	0.82	0.82	0.86	0.80
Tree3	0.59	0.55	0.66	0.66	0.64	0.61	0.60	0.60	0.76	0.88

Chapter 6. Bayesian theory

6.1 In Example 6.1 of testing for infection, suppose that a person was tested twice and found to be positive both times. What is the probability that he has the infection?

See Problem 5.1 in (Yang 2006).

6.2* The sequence distance θ under JC69 and the probability of difference (p) are related by $p = \frac{3}{4} - \frac{3}{4}e^{-4\theta/3}$, with $0 \le \theta < \infty$ and $0 \le p < \frac{3}{4}$. (a) Given $p \sim U(0, \frac{3}{4})$, derive the density for θ. (b) Given that p has the truncated beta distribution of equation (6.58), derive the density for θ. (Hint. Use Theorem 1 in Appendix A.)

Solution.

a)

$$\frac{dp}{d\theta} = -\frac{3}{4}\left(-\frac{4}{3}\right)e^{\frac{-4}{3}\theta} = e^{-\frac{4}{3}\theta}.$$

Hence,

$$f_\Theta(\theta) = \left(\frac{3}{4}\right)^{-1} e^{-\frac{4}{3}\theta} = \frac{4}{3}e^{-\frac{4}{3}\theta}.$$

Thus, this corresponds to an exponential distribution with the parameter $\theta = \frac{4}{3}$.

b)

Refer back to Eq. (6.58) of (Yang 2014a), which is expressed as

$$f(p) \propto p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}, 0 \le p < \frac{3}{4}.$$

Note that there might be some confusion by using *P* to denote the variable the probability of difference but as that is what is stated in the problem, we follow it in the solution. Derive the PDF of *P* as

$$f(p) = \frac{p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}}{\int_0^{\frac{3}{4}} p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}dp}.$$

The denominator can be calculated as follows

$$\begin{aligned}\int_0^{\frac{3}{4}} p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}dp &= \int_0^{\frac{3}{4}} \frac{1}{\sqrt{(1-p)p}}dp \\ &= 2\int_0^{\frac{\sqrt{3}}{2}} \frac{1}{\sqrt{1-u^2}}du \qquad (u = \sqrt{p}) \\ &= 2\arcsin\left(\frac{\sqrt{3}}{2}\right).\end{aligned}$$

Therefore, we have

$$f_\Theta(\theta) = \frac{\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4\theta}{3}}\right)^{-\frac{1}{2}}\left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4\theta}{3}}\right)^{-\frac{1}{2}} e^{-\frac{4}{3}\theta}}{2\arcsin\left(\frac{\sqrt{3}}{2}\right)}.$$

Note.

This problem likely originates from the study (dos Reis and Yang 2011).

6.3* Use the example of normal distributions to study the sensitivity of Bayesian model selection to the prior on parameters in the models, in contrast to the sensitivity of Bayesian parameter estimation to the prior. We use an i.i.d. sample from the normal distribution $N(\mu, 1)$ to estimate the population mean μ and to compare the null hypothesis H_0: $\mu = 0$ against the alternative H_1: $\mu \neq 0$. Let the sample size be $n = 100$ and the sample mean be $\bar{x} = 0.3$. Calculate the p-value for the LRT. Assign the prior probability ½ for each of the two models, and $\mu \sim N(\mu_0, \sigma_0^2)$ under H_1 to calculate the posterior probability for H_0. Use different priors, for example, (i) $\mu_0 = 0, \sigma_0^2 = 0.3$; (ii) $\mu_0 = 0, \sigma_0^2 = 9$; and (iii) $\mu_0 = -2.95, \sigma_0^2 = 1$. Review the theory of §6.2.3.3 and use equation (6.32). Also calculate the posterior of μ under H_1 (equation (6.44)).

Solution.

a)

$$\begin{aligned} 2\Delta lnL &= -2\ell(0) + 2\ell(\hat{\mu}) \\ &= -2\log\left(\frac{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_i^2}{2}}}{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{\frac{-(x_i-\bar{x})^2}{2}}}\right) \\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= n\bar{x}^2 \end{aligned}$$

Given the context, $n = 100$ and $\bar{x} = 0.3$, hence $2\Delta lnL = 9$. This is enough to reject the null hypothesis because $\chi_{0.05,1} = 3.84 < 9$ (the P-value is 0.0027).

b)

Referring back to Eq. (6.32) of (Yang 2014a), P_0 is defined as follows

$$P_0 = \frac{1}{1 + \frac{1}{\sqrt{1+n\sigma_0^2}}\exp\left\{\frac{n^2\sigma_0^2\bar{x} + 2n\mu_0\bar{x} - n\mu_0^2}{2(1+n\sigma_0^2)}\right\}}.$$

Hence, using the following R code, it can be easily calculated that i) $P_0 = 0.067$, ii) $P_0 = 0.25$, iii) $P_0 = 0.95$.

R

```
calculate_p0 <- function(n,x,mu,var0){ denominator=1+1/sqrt(1+n*var0)*exp( (n^2*var0*x^2 +
2*n*mu*x - n*mu^2)/(2+2*n*var0) );   1/denominator}
> calculate_p0(n=100,x=0.3,mu=0.3,var0=0.3)
> calculate_p0(n=100,x=0.3,mu=0.3,var0=9)
> calculate_p0(n=100,x=0.3,mu=-2.95,var0=1)
```

Note.

If you use Eq. (6.33) of (Yang 2014a) to calculate i) and ii) where $\mu_0 = 0$, you may want to be aware of a typo in the original equation, as also pointed out in Errata I of the book (Yang 2022).

6.4 We use a prior in our analysis but take great effort to ensure that it does not have any undue influence on our results. Why don't we avoid the prior in the first place?

Solution.

This is an open question. Based on my limited experience, people would like to make a balance between leveraging prior knowledge and maintaining the integrity of the empirical evidence. Using a properly chosen prior could lead to more reasonable estimates, particularly when there is only limited data. But on the other side, we also hope to perform sensitivity analyses to ensure that the results are not too heavily affected by our prior beliefs and the extent to which the results are affected by different priors. By evaluating the posteriors under different priors, we can have a better idea of how robust the posterior is robust to the prior.

Chapter 7. Bayesian computation (MCMC)

7.1 Write a program to implement the MCMC algorithm of Example 7.1 to estimate the distance between the human and orangutan 12S rRNA genes under the JC69 model. Use any programming language of your choice, such as C/C++, Java, or R. Investigate how the acceptance proportion changes with the window size w. (Note: Calculate the logarithms of the prior, the likelihood and the acceptance ratio to avoid overflows and underflows. If the logarithm of the acceptance ratio is ≥ 0, the proposal is accepted. Otherwise take the exponential to calculate the acceptance ratio to decide whether the proposal is accepted.)

See Problem 5.4 of (Yang 2006).

7.2 Modify the program of Problem 7.1 to estimate the sequence distance θ under the K80 model. Use the exponential prior $f(\theta) = \frac{1}{\mu}e^{-\theta/\mu}$ with mean $\mu = 0.2$ for distance θ and exponential prior with mean 5 for the transition/transversion rate ratio κ. Implement two proposal steps, one for updating θ and another for updating κ. Compare the posterior estimates with the MLEs of §1.4.2.

See Problem 5.6 of (Yang 2006).

7.3 Study the tail behaviour of the Markov chain for estimation of the sequence distance under JC69 of Example 7.1. Suppose the current state is $\theta = 10$. Calculate the acceptance ratios $\alpha_{\text{left}} = \pi(9.9)/\pi(10)$ and $\alpha_{\text{right}} = \pi(10.1)/\pi(10)$ for moves of size $\Delta\theta = 0.1$. Use different starting values such as $\theta = 10$ and 100 to run the MCMC program of Problem 7.1 to examine convergence of the chain. Then do the same calculation using the uniform prior $\theta \sim U(0, 200)$, and modify the MCMC program to do the same test. Use a sliding window with window size $w = 0.2$.

Solution.

It is straightforward to use the following R code to get $\alpha_{left} = \frac{\pi(9.9)}{\pi(10)} = 1.000574$, and $\alpha_{right} = \frac{\pi(10.1)}{\pi(10)} = 0.9994979$.

R

```
> lnl_JC69 <- function(d, n, x) x*log(3/4-3/4*exp(-4/3*d)) + (n-x)*log(1/4+3/4*exp(-4/3*d))
> n <- 948; x <- 90;
> exp(lnl_JC69(9.9,n=n,x=x) - lnl_JC69(10,n=n,x=x))
> exp(lnl_JC69(10.1,n=n,x=x) - lnl_JC69(10,n=n,x=x))
```

Using a starting value of $\theta_{start} = 10$ and sliding window width $w = 0.2$ would lead to the correct estimate. However, $\theta_{start} = 100$ will lead to poor mixing given $w = 0.2$, as there is little difference between $\pi(99.8)$, $\pi(100)$ and $\pi(100.2)$. Changing w to a bigger value can overcome this problem.

Note

See also Section 7.3.1 of (Yang 2014a).

7.4 Write an MCMC program to sample from the 2-D posterior density of equation (7.30). Use a 2-D sliding window to propose moves: $x' \sim U(x - \frac{\varepsilon}{2}, x + \frac{\varepsilon}{2})$ and $y' \sim U(y - \frac{\varepsilon}{2}, y + \frac{\varepsilon}{2})$. Start the chain at $(10000, 0)$. Observe the fraction of the proposals that are accepted. Then use two 1-D sliding windows to propose moves, changing x and y, respectively.

Solution.

Refer to Eq. (7.30) in (Yang 2014a):

$$\pi(x, y) \propto e^{-(x^2+x^2y^2+y^2)}, -\infty < x, y < \infty.$$

7.5 Modify the program of Problem 7.1 to estimate two parameters under the JC69 model: the time of species divergence t and the substation rate $r = 3\lambda$, instead of the distance $\theta = 2t \times r$. Consider one time unit to be 100 million years. Assign the gamma prior $t \sim G(2, 12)$ with mean 0.167 (meaning 16.7 million years for the human–orangutan divergence) and another gamma prior for the rate $r \sim G(2, 2)$ with mean 1 (meaning a prior mean rate of 10^{-8} substitutions/site/year). Implement one or more of the following proposals:

i. A 2-D sliding window updating t and r in one step.
ii. Two separate 1-D sliding windows updating t and r, respectively.
iii. In both the above cases add an extra step of multiplying t by a random variable c that is close to 1 and dividing r by the same c (see §7.2.5).

Study the convergence of the chain at the tail by using different initial states, such as $(100, 100)$, $(5, 0.01)$, etc., and by comparing different proposals. Study the sensitivity of the posterior to the prior by changing the parameters in the priors.

See Problem 5.5 of (Yang 2006).

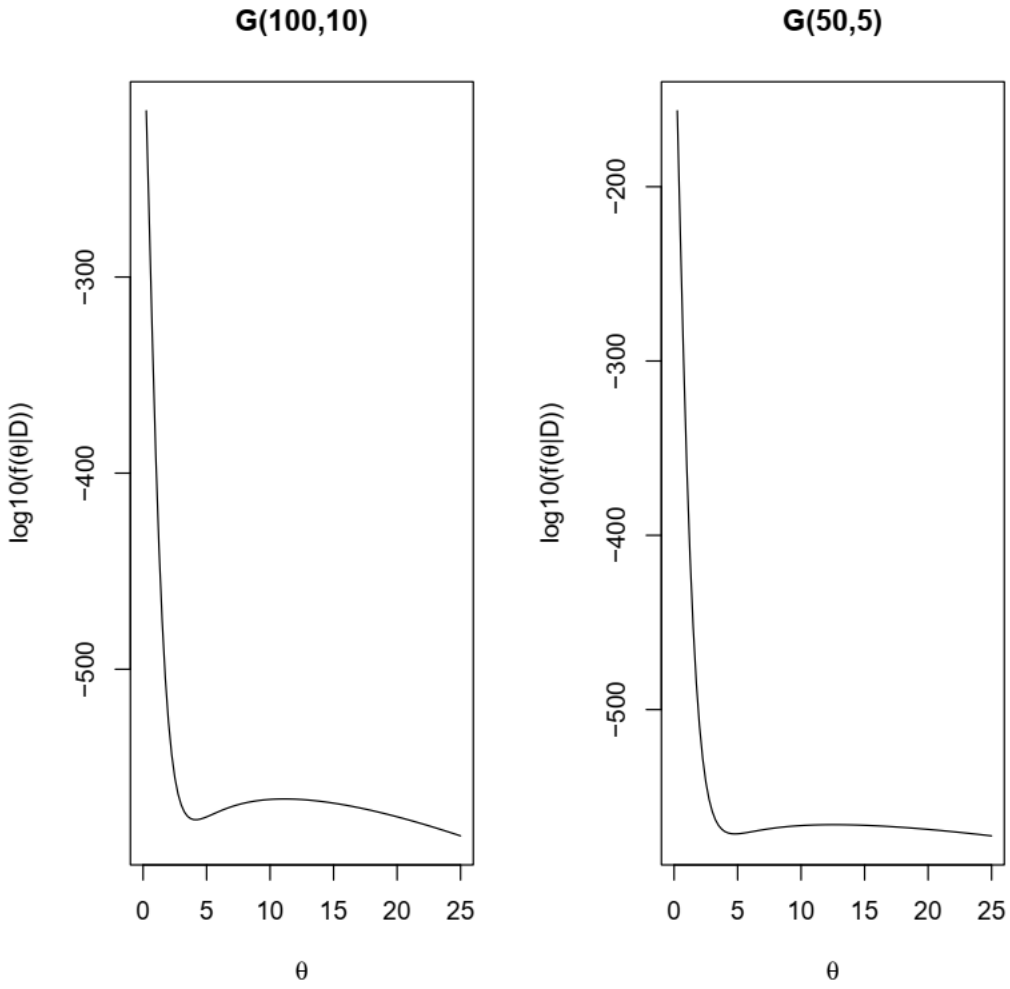
7.6 Modify the program of Problem 7.1 to estimate the JC69 distance with a gamma prior and confirm the two modes in the posterior of Example 7.2 by using different initial values $(0.01, 1, 10, 100)$ and window sizes $(0.01,\ 0.1,$ or $1)$. Note that the constant in the gamma prior cancels so that you need not calculate $\Gamma(100) = 99!$. As in Problem 7.1, calculate the logarithms of the prior and likelihood to avoid numerical problems.

Solution.

The code and the figure are given as follows. Two Gamma priors are used Gamma(100,10) and Gamma(50,5). I use "curve()" instead of MCMC. Note that $\log_{10}(f(\theta|D))$ is displayed as y-axis and θ denotes the JC69 distance to estimate.

R *7.6.R*

```
log10_lik_JC69 <- function(d, n, x, a, b){
   log(dgamma(d, a, b)) + x*log(3/4-3/4*exp(-4/3*d)) + (n-x)*log(1/4+3/4*exp(-4/3*d)) / log(10)
}
##########################################
x <- 90; n<-948
pdf("7.6.pdf"); par(mfrow=c(1,2))
posterior <- expression('log10(f('*theta*'|D))'); theta <- expression(italic(theta))
df <- data.frame(c(100,10), c(50,5))
for(i in 1:length(df)){
   v <- df[,i]
   curve(log10_lik_JC69(d=x,n=n,x=x,a=v[1],b=v[2]), from=0, to=25, ylab=posterior, xlab=theta,
main=paste0("G(",v[1],',',v[2],")"))
}
dev.off()
```



7.7 Write down the transition matrix P for the Markov chain generated in the MCMC algorithm for the robot-on-box example of §7.1. Consider both the symmetrical move of §7.1.1 and the asymmetrical move of §7.1.2. Assume $\pi_1 \geq \pi_2 \geq \pi_3$.

Solution.

According to Eq. (7.9) in (Yang 2014a), the transition probability is defined as

$$p(x,y) = \begin{cases} q(y|x) \cdot \alpha(x,y), y \neq x, \\ 1 - \sum_{y \neq x} q(y|x)\alpha(x,y), y = x, \end{cases}$$

where $\alpha(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)} \times \frac{q(x|y)}{q(y|x)}\right)$.

a)

Consider the symmetrical move, where by definition $q(\cdot \mid \cdot) = \frac{1}{2}$. As given in the problem, $\pi_1 \geq \pi_2 \geq \pi_3$. Hence, the transition matrix P is given as

$$\alpha(1,2) = \frac{\pi_2}{\pi_1},$$
$$\alpha(1,3) = \frac{\pi_3}{\pi_1},$$
$$\alpha(2,1) = 1,$$
$$\alpha(2,3) = \frac{\pi_3}{\pi_2},$$
$$\alpha(3,1) = \alpha(3,2) = 1.$$

Accordingly,

$$P = \begin{bmatrix} 1 - \frac{\pi_2}{2\pi_1} - \frac{\pi_3}{2\pi_1} & \frac{\pi_2}{2\pi_1} & \frac{\pi_3}{2\pi_1} \\ \frac{1}{2} & 1 - \frac{1}{2} - \frac{\pi_3}{2\pi_2} & \frac{\pi_3}{2\pi_2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{3}{2} - \frac{1}{2\pi_1} & \frac{\pi_2}{2\pi_1} & \frac{\pi_3}{2\pi_1} \\ \frac{1}{2} & \frac{1}{2} - \frac{\pi_3}{2\pi_2} & \frac{\pi_3}{2\pi_2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

b)

Consider the asymmetrical move where $q(3|1) = q(1|2) = q(2|3) = \frac{2}{3}$ and $q(2|1) = q(3|2) = q(1|3) = \frac{1}{3}$.

$$\alpha(1,2) = \min\left(1, \frac{2\pi_2}{\pi_1}\right),$$

$$\alpha(1,3) = \min\left(1, \frac{\pi_3}{2\pi_1}\right) = \frac{\pi_3}{2\pi_1},$$

$$\alpha(2,1) = \min\left(1, \frac{\pi_1}{2\pi_2}\right),$$

$$\alpha(2,3) = \min\left(1, \frac{2\pi_3}{\pi_2}\right),$$

$$\alpha(3,1) = \min\left(1, \frac{2\pi_1}{\pi_3}\right) = 1,$$

$$\alpha(3,2) = \min\left(1, \frac{\pi_2}{2\pi_3}\right).$$

Consequently, we can derive the transition matrix P as

$$P = \begin{bmatrix} 1 - \frac{1}{3}\min\left(1, \frac{2\pi_2}{\pi_1}\right) - \frac{\pi_3}{3\pi_1} & \frac{1}{3}\min\left(1, \frac{2\pi_2}{\pi_1}\right) & \frac{\pi_3}{3\pi_1} \\ \frac{2}{3}\min\left(1, \frac{\pi_1}{2\pi_2}\right) & 1 - \frac{2}{3}\min\left(1, \frac{\pi_1}{2\pi_2}\right) - \frac{1}{3}\min\left(1, \frac{2\pi_3}{\pi_2}\right) & \frac{1}{3}\min\left(1, \frac{2\pi_3}{\pi_2}\right) \\ \frac{1}{3} & \frac{2}{3}\min\left(1, \frac{\pi_2}{2\pi_3}\right) & \frac{2}{3} - \frac{2}{3}\min\left(1, \frac{\pi_2}{2\pi_3}\right) \end{bmatrix}.$$

7.8* Show or confirm that the eigenvalues and eigenvectors of the P matrix of equation (7.42), as defined in equation (7.36), are given as

$$\Lambda = \begin{bmatrix} 1 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & \ddots & & \\ & & & & 0 & \\ & & & & & 1-\frac{1}{\pi_1} \end{bmatrix},$$

$$E = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & \left(1-\frac{1}{\pi_1}\right)a_K \\ 1 & a_2 & a_3 & \cdots & a_{K-1} & a_K \\ 1 & -\frac{\pi_2}{\pi_3}a_2 & 0 & \cdots & 0 & a_K \\ 1 & 0 & -\frac{\pi_2}{\pi_4}a_3 & \cdots & 0 & a_K \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & -\frac{\pi_2}{\pi_K}a_{K-1} & a_K \end{bmatrix}, \qquad (7.105)$$

where the factors $a_k = \left[\frac{\pi_{k+1}}{\pi_2(\pi_2+\pi_{k+1})}\right]^{\frac{1}{2}}, k = 2, \ldots, K-1$, and $a_K = \left[\frac{\pi_1}{1-\pi_1}\right]^{\frac{1}{2}}$ are for normalizing the eigenvectors so that $E^{\mathrm{T}}BE = I$ or $\sum_{i=1}^{K} \pi_i e_{ik}^2 = 1$ for each k. Note that the kth column in E is the eigenvector corresponding to λ_k.

[Hint. (**a**) To confirm the results, simply check that Λ and E above satisfy $Px = \lambda x$, with x to be a column in E. (**b**) To derive the eigenvalues, one way is to solve the characteristic equation $|P-\lambda I| = 0$. By Laplace's formula, the determinant $|P-\lambda I| = p_{11} \cdot |P_{11}| - p_{12} \cdot |P_{12}| + p_{13} \cdot |P_{13}| - \ldots$, where P_{1k} is the $(K-1) \times (K-1)$ matrix that results from removing the 1st row and kth column of $P-\lambda I$. Thus show that $|P-\lambda I| = \frac{1}{\pi_1}(-\lambda)^{K-2}(\lambda-1)(\pi_1\lambda+1-\pi_1)$. Note that the determinant of a triangular matrix is the product of the diagonal elements and that interchanging two rows of a matrix multiplies its determinant by –1.]

Solution.

We actually have reservations about part of this problem. The reason is because E does not meet the criterion $E^T BE = I$ as given in the problem. As detailed below, denote

$$E^T = [e_1 \quad \ldots \quad e_K],$$

where $e_1 = [e_{11} \quad \ldots \quad e_{1K}]^T, e_1 = [e_{21} \quad \ldots \quad e_{1K}]^T, \ldots, e_K = [e_{11} \quad \ldots \quad e_{1K}]^T$.

According to the problem statement, we have

$$E^T BE = [e_1 \quad \ldots \quad e_K] \times \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_K \end{bmatrix} \times \begin{bmatrix} e_1^T \\ \vdots \\ e_K^T \end{bmatrix}$$

$$= \pi_1 e_1 e_1^T + \cdots + \pi_K e_K e_K^T$$

$$= \begin{bmatrix} \sum_{i=1}^{K} \pi_i e_{i1}^2 & \sum_{i=1}^{K} \pi_i e_{i2} e_{i1} & \cdots & \sum_{i=1}^{K} \pi_i e_{iK} e_{i1} \\ \sum_{i=1}^{K} \pi_i e_{i1} e_{i2} & \sum_{i=1}^{K} \pi_i e_{i2}^2 & \cdots & \sum_{i=1}^{K} \pi_i e_{iK} e_{i2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{K} \pi_i e_{i1} e_{iK} & \sum_{i=1}^{K} \pi_i e_{i2} e_{iK} & \cdots & \sum_{i=1}^{K} \pi_K e_{iK}^2 \end{bmatrix}$$

$$= I.$$

Hence, we have

$$\sum_{i=1}^{K} \pi_i e_{ik}^2 = 1, \text{for } k = 1, \ldots, K,$$

$$\sum_{i=1}^{K} \pi_i e_{im} e_{in} = 0, \text{for } m \neq n, 1 \leq m, n \leq K.$$

Unfortunately, E as given in the problem does not seem to meet the second criterion above (readers can easily verify themselves).

Nevertheless, as follows, we still follow the "answer" given in the problem to present a solution but readers may want to be aware of the above. Interestingly, this does not affect the calculation of Problem 7.9.

a)

Refer to Eq. 7.42 in (Yang 2014a) as follows

$$P = \begin{bmatrix} 1 - \frac{1-\pi_1}{\pi_1} & \frac{\pi_2}{\pi_1} & \frac{\pi_3}{\pi_1} & \cdots & \frac{\pi_K}{\pi_1} \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

To find the eigenvalue λ, set $|P - \lambda I| = 0$. Assume $\lambda \neq 0$, it follows that

$$|P - \lambda I|$$

$$= \begin{vmatrix} 1 - \frac{1-\pi_1}{\pi_1} - \lambda & \frac{\pi_2}{\pi_1} & \frac{\pi_3}{\pi_1} & \cdots & \frac{\pi_K}{\pi_1} \\ 1 & -\lambda & 0 & \cdots & 0 \\ 1 & 0 & -\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$\underline{\underline{\pi_1 r_1}} \frac{1}{\pi_1} \begin{vmatrix} (2-\lambda)\pi_1 - 1 & \pi_2 & \pi_3 & \cdots & \pi_K \\ 1 & -\lambda & 0 & \cdots & 0 \\ 1 & 0 & -\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -\lambda \end{vmatrix}$$

$$\begin{matrix}\frac{1}{\lambda}c_2\\ \vdots\\ \frac{1}{\lambda}c_{K-1}\end{matrix}\frac{\lambda^{K-1}}{\pi_1}\begin{vmatrix}(2-\lambda)\pi_1-1 & \frac{\pi_2}{\lambda} & \frac{\pi_3}{\lambda} & \cdots & \frac{\pi_K}{\lambda}\\ 1 & -1 & 0 & \cdots & 0\\ 1 & 0 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & 0 & 0 & \cdots & -1\end{vmatrix}$$

$$\begin{matrix}c_1+c_2\\ \vdots\\ c_1+c_K\end{matrix}\frac{\lambda^{K-1}}{\pi_1}\begin{vmatrix}(2-\lambda)\pi_1-1+\sum_{i=2}^{K}\frac{\pi_i}{\lambda} & \frac{\pi_2}{\lambda} & \frac{\pi_3}{\lambda} & \cdots & \frac{\pi_K}{\lambda}\\ 0 & -1 & 0 & \cdots & 0\\ 0 & 0 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & -1\end{vmatrix}$$

$$=(-1)^{K-1}\frac{\lambda^{K-1}}{\pi_1}\left((2-\lambda)\pi_1-1+\sum_{i=2}^{K}\frac{\pi_i}{\lambda}\right)$$

$$=(-1)^{K-1}\times\frac{\lambda^{K-1}}{\pi_1}\times\left(\lambda(2-\lambda)\pi_1-\lambda+(1-\pi_1)\right)$$

$$=\frac{1}{\pi}(-1)^{\mathrm{K-1}}(\lambda-0)^{K-2}(\lambda-\lambda_1)(\lambda-\lambda_K),$$

where λ_1 and λ_K are the two roots of the unary quadratic equation $\lambda(2-\lambda)\pi_1-\lambda+(1-\pi_1)=0$. By solving it, we have

$$\lambda=\frac{-(1-2\pi_1)\pm\sqrt{(1-2\pi_1)^2-4\pi_1(\pi_1-1)}}{2\pi_1}$$

$$=\frac{-(1-2\pi_1)\pm 1}{2\pi_1}$$

$$=1 \text{ or } \frac{\pi_1-1}{\pi_1}.$$

Note that in Section 7.3.2.1 of (Yang 2014a) it is indicated that $1=\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_K\geq-1$, therefore we have $\lambda_1=1$, $\lambda_K=\frac{\pi_1-1}{\pi_1}$, $\lambda_2=\lambda_3=\cdots=\lambda_{K-1}=0$. Accordingly,

$$\Lambda=\begin{bmatrix}1 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & \ddots & & \\ & & & & 0 & \\ & & & & & \frac{\pi_1-1}{\pi_1}\end{bmatrix}.$$

b)

As given in the problem, the eigenvectors should be normalized so that the following holds for each *k*

$$\sum_{i=1}^{K}\pi_i e_{ik}^2=1.$$

Denote $e_k = (e_{1k}, e_{2k}, \dots, e_{Kk})^T$ as the k^{th} eigenvector of the matrix E. In other words, $E = (e_1, e_2, \dots, e_K)$. Solve the normalizing eigenvector(s) for each eigenvalue as follows.

i) For $k = 1$, we have $\lambda_k = \lambda_1 = 1$. The eigenvector can be found by solving $(P - \lambda_1)e_1 = \mathbf{0}$. Hence,

$$\begin{bmatrix} 1 - \frac{1-\pi_1}{\pi_1} - 1 & \frac{\pi_2}{\pi_1} & \frac{\pi_3}{\pi_1} & \cdots & \frac{\pi_K}{\pi_1} \\ 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} e_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

By solving the above, we get $e_1 = c(1,1,\dots,1)^T$ where c is a constant. Due to the constraint $\sum_{i=1}^{K} \pi_i e_{i1}^2 = 1$, it is easy to see $c^2(\pi_1 + \pi_2 + \cdots + \pi_K) = 1$, thus $c = 1$ and $e_1 = (1,1,\dots,1)^T$.

For $k = K$, we have $\lambda_k = \lambda_K = \frac{1-\pi_1}{\pi_1}$. It follows that

$$(P - \lambda_K)e_K = \begin{bmatrix} 1 - \frac{1-\pi_1}{\pi_1} - \frac{\pi_1 - 1}{\pi_1} & \frac{\pi_2}{\pi_1} & \frac{\pi_3}{\pi_1} & \cdots & \frac{\pi_K}{\pi_1} \\ 1 & \frac{1-\pi_1}{\pi_1} & 0 & \cdots & 0 \\ 1 & 0 & \frac{1-\pi_1}{\pi_1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & \frac{1-\pi_1}{\pi_1} \end{bmatrix} e_K = \mathbf{0}.$$

It can be further shown that

$$\begin{cases} \pi_1 e_{1K} + \sum_{i \geq 2}^{K} \pi_i e_{iK} = 0, \\ e_{1K} = \left(\frac{1-\pi_1}{\pi_1}\right) e_{2K}, \\ e_{1K} = \left(\frac{1-\pi_1}{\pi_1}\right) e_{3K}, \\ \vdots \\ e_{1K} = \left(\frac{1-\pi_1}{\pi_1}\right) e_{KK}. \end{cases}$$

This is equivalent to saying

$$\frac{e_{1K}}{1 - \frac{1}{\pi_1}} = e_{2K} = \cdots = e_{KK}.$$

Applying the constraint $\sum_{i=1}^{K} \pi_i e_{iK}^2 = 1$, it can be shown that

$$\pi_1 \left(\left(1 - \frac{1}{\pi_1}\right) e_{KK} \right)^2 + \pi_2 (e_{KK})^2 + \pi_3 (e_{KK})^2 + \cdots + \pi_K (e_{KK})^2 = \frac{e_{KK}^2 (1 - \pi_1)}{\pi_1} = 1.$$

Hence

$$e_{iK} = \begin{cases} -\dfrac{\sqrt{1-\pi_1}}{\pi_1}, i = 1 \\ \sqrt{\dfrac{\pi_1}{1-\pi_1}}, i \neq 1 \end{cases}.$$

Further, define $a_K = \sqrt{\frac{\pi_1}{1-\pi_1}}$, such that e_K can be rewritten as

$$e_K = \left(\frac{\pi_1 - 1}{\pi_1} a_K, a_K, \dots, a_K\right)^T$$

where $a_K = \sqrt{\frac{\pi_1}{1-\pi_1}}$.

For $2 \leq k \leq K-1$, $\lambda_k = 0$. It follows that

$$(P - \lambda_k)e_k = \begin{bmatrix} 1 - \frac{1-\pi_1}{\pi_1} & \frac{\pi_2}{\pi_1} & \frac{\pi_3}{\pi_1} & \cdots & \frac{\pi_K}{\pi_1} \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} e_k = \mathbf{0}.$$

It can be shown that

$$\begin{cases} e_{1k} = 0, \\ \pi_2 e_{2k} + \pi_3 e_{3k} + \cdots \pi_K e_{Kk} = 0. \end{cases}$$

By solving the above, we get

$$e_{2k} = \begin{bmatrix} 0 \\ 1 \\ -\pi_2/\pi_3 \\ 0 \\ \vdots \\ 0 \end{bmatrix} a_2, e_{3k} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -\pi_2/\pi_4 \\ 0 \\ \vdots \\ 0 \end{bmatrix} a_3, \dots, e_{K-1,k} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ -\pi_2/\pi_K \end{bmatrix} a_{K-1},$$

where a_k is a normalizing constant. Applying the constraint $\sum_{i=1}^{K} \pi_i e_{ik}^2 = 1$, it can be calculated that $\left(\pi_2 + \pi_{k+1} \times \frac{\pi_2^2}{\pi_{k+1}^2}\right) a_k^2 = 1$, thus $a_k = \sqrt{\frac{\pi_{k+1}}{\pi_2 \pi_{k+1} + \pi_2^2}}$ for $2 \leq k \leq K-1$.

By integrating the above results, finally we obtain

$$E = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & \left(1 - \frac{1}{\pi_1}\right) a_K \\ 1 & a_2 & a_3 & \cdots & a_{K-1} & a_K \\ 1 & -\frac{\pi_2}{\pi_3} a_2 & 0 & \cdots & 0 & a_K \\ 1 & 0 & -\frac{\pi_2}{\pi_4} a_3 & \cdots & 0 & a_K \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & -\frac{\pi_2}{\pi_K} a_{K-1} & a_K \end{bmatrix},$$

where $a_k = \sqrt{\frac{\pi_{k+1}}{\pi_2 \pi_{k+1} + \pi_2^2}}$ for $2 \le k \le K-1$ and $a_K = \sqrt{\frac{\pi_1}{1-\pi_1}}$.

7.9* Show that the asymptotic variance when a Markov chain sample from P of equation (7.42) is used to estimate π_1 is $v = \pi_1(1-\pi_1)(2\pi_1 - 1)$. [Hint: One way is to use equation (7.37) with the eigenvalues and eigenvectors given in Problem 7.8.]

Solution.

Referring to Eq. (7.37) in (Yang 2014a), the asymptotic variance v is defined as

$$v = \sum_{k \ge 2}^{K} \frac{1+\lambda_k}{1-\lambda_k} (E^T B h)_k^2.$$

According to Eqs. (7.35-7.36) of (Yang 2014a), we have

$$B = \mathrm{diag}\{\pi_1, \ldots, \pi_K\} = \begin{bmatrix} \pi_1 & 0 & 0 & \cdots & 0 \\ 0 & \pi_2 & 0 & \cdots & 0 \\ 0 & 0 & \pi_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \pi_K \end{bmatrix},$$

$$h = (1,0,0, \ldots, 0)^T.$$

Noting that multiplying a matrix by the column vector h on the right retrieves its first column, we have

$$Bh = \begin{bmatrix} \pi_1 & 0 & 0 & \cdots & 0 \\ 0 & \pi_2 & 0 & \cdots & 0 \\ 0 & 0 & \pi_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \pi_K \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \pi_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Likewise, multiplying E^T by Bh is equivalent to retrieving the first column of E^T and by π_1 times. Hence, it is not difficult to see the following

$$E^T B h = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & a_2 & -\frac{\pi_2}{\pi_3} a_2 & 0 & \cdots & 0 \\ 0 & a_3 & 0 & -\frac{\pi_2}{\pi_4} a_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{K-1} & 0 & 0 & 0 & -\frac{\pi_2}{\pi_K} a_{K-1} \\ \left(1 - \frac{1}{\pi_1}\right) a_K & a_K & a_K & a_K & \cdots & a_K \end{bmatrix} \times \begin{bmatrix} \pi_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \pi_1 \\ 0 \\ 0 \\ \vdots \\ \left(1 - \frac{1}{\pi_1}\right) a_K \cdot \pi_1 \end{bmatrix}.$$

Thus, we have

$$(E^T Bh)_k^2 = \begin{cases} \pi_1^2, k = 1, \\ 0, 2 \le k \le K-1, \\ \left(\left(1 - \frac{1}{\pi_1}\right) a_K \cdot \pi_1\right)^2, k = K. \end{cases}$$

Because $a_K = \left[\frac{\pi_1}{1-\pi_1}\right]^{\frac{1}{2}}$, by plugging the above into Eq. (7.37) of (Yang 2014a), the asymptotic variance v can be written as

$$\begin{aligned} v &= \sum_{k\ge 2}^{K} \frac{1+\lambda_k}{1-\lambda_k} (E^T Bh)_k^2 \\ &= \frac{1+\lambda_2}{1-\lambda_2}(E^T Bh)_2^2 + \frac{1+\lambda_3}{1-\lambda_3}(E^T Bh)_3^2 + \cdots + \frac{1+\lambda_K}{1-\lambda_K}(E^T Bh)_K^2 \\ &= \frac{1+0}{1-0} \times 0 + \cdots + \frac{1+0}{1-0} \times 0 + \frac{1+1-\frac{1}{\pi_1}}{1-\left(1-\frac{1}{\pi_1}\right)} \times \left(\left(1-\frac{1}{\pi_1}\right) a_K \cdot \pi_1\right)^2 \\ &= \frac{2-\frac{1}{\pi_1}}{\frac{1}{\pi_1}} \times \pi_1(1-\pi_1) \\ &= \pi_1(1-\pi_1)(2\pi_1 - 1). \end{aligned}$$

7.10* Calculate the efficiency of the following Markov chain sampler for estimating $\pi_1 = 1/K$:

$$P = \begin{bmatrix} 0 & \frac{1}{K-1} & \cdots & \frac{1}{K-1} \\ \frac{1}{K-1} & 0 & \cdots & \frac{1}{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{K-1} & \frac{1}{K-1} & \cdots & 0 \end{bmatrix}. \tag{7.106}$$

Note that efficiency is defined as $\pi_1(1-\pi_1)/v$, where v is the asymptotic variance of equation (7.35).

Solutions.

Two solutions are provided, the first solving the problem by Eq. (7.35), the second using Eq. (7.37) in (Yang 2014a).

Solution 1.

Eq. (7.35) in (Yang 2014a) is written as

$$v = h^T \cdot B(2Z - I - A) \cdot h.$$

Because $h = (1,0,0,\ldots,0)^T$, it is apparent that v equals the element located at the intersection of the first row and the first column of matrix $B(2Z - I - A)$.

Define

$$B(2Z - I - A) = C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1K} \\ c_{21} & c_{22} & \cdots & c_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{K1} & c_{K2} & \cdots & c_{KK} \end{bmatrix}.$$

In other words, $v = c_{11}$ is what to be calculated.

Due to symmetry, we have

$$\pi_1 = \pi_2 = \cdots = \pi_K = \frac{1}{K}.$$

Hence,

$$A = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_K \\ \pi_1 & \pi_2 & \cdots & \pi_K \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_K \end{bmatrix} = \frac{1}{K} 1_{K \times K},$$

and

$$B = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_K \end{bmatrix} = \frac{1}{K} I_K.$$

According to Section 7.3.2.1 in (Yang 2014a), the matrix Z is defined as

$$Z = (I - P + A)^{-1}.$$

Define $D = I - P + A$, hence $Z = D^{-1}$. It follows that

$$D = I_K - \begin{bmatrix} 0 & \frac{1}{K-1} & \cdots & \frac{1}{K-1} \\ \frac{1}{K-1} & 0 & \cdots & \frac{1}{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{K-1} & \frac{1}{K-1} & \cdots & 0 \end{bmatrix} + \frac{1}{K} 1_{K \times K}$$

$$= \begin{bmatrix} 1 & \frac{1}{1-K} & \cdots & \frac{1}{1-K} \\ \frac{1}{1-K} & 1 & \cdots & \frac{1}{1-K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-K} & \frac{1}{1-K} & \cdots & 1 \end{bmatrix} + \begin{bmatrix} \frac{1}{K} & \frac{1}{K} & \cdots & \frac{1}{K} \\ \frac{1}{K} & \frac{1}{K} & \cdots & \frac{1}{K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{K} & \frac{1}{K} & \cdots & \frac{1}{K} \end{bmatrix}$$

$$= \frac{1}{K} \begin{bmatrix} K+1 & \frac{1}{1-K} & \cdots & \frac{1}{1-K} \\ \frac{1}{1-K} & K+1 & \cdots & \frac{1}{1-K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-K} & \frac{1}{1-K} & \cdots & K+1 \end{bmatrix}.$$

Further define

$$a = K + 1,$$

$$b = \frac{1}{1-K},$$

$$E=\frac{1}{K}\begin{bmatrix} K+1 & \frac{1}{1-K} & \cdots & \frac{1}{1-K} \\ \frac{1}{1-K} & K+1 & \cdots & \frac{1}{1-K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-K} & \frac{1}{1-K} & \cdots & K+1 \end{bmatrix}=\begin{bmatrix} a & b & \cdots & b \\ b & a & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & a \end{bmatrix}.$$

D may be rewritten as follows $D=\frac{1}{K}E$. Hence, $D^{-1}=K\cdot E^{-1}$.

To get the inverse of matrix D, we apply the following trick in linear algebra. Define

$$E^{-1}=\begin{bmatrix} x & y & \cdots & y \\ y & x & \cdots & y \\ \vdots & \vdots & \ddots & \vdots \\ y & y & \cdots & x \end{bmatrix}.$$

Because $EE^{-1}=I$, we have

$$\begin{bmatrix} ax+(K-1)by & ay+bx+(K-2)by & \cdots & ay+bx+(K-2)by \\ ay+bx+(K-2)by & ax+(K-1)by & \cdots & ay+bx+(K-2)by \\ \vdots & \vdots & \ddots & \vdots \\ ay+bx+(K-2)by & ay+bx+(K-2)by & \cdots & ax+(K-1)by \end{bmatrix}=\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Thus, finding E^{-1} is equivalent to solving the following system of linear equations with two variables

$$\begin{cases} ax+(K-1)by=1 \\ ay+bx+(K-2)by=0 \end{cases}.$$

By solving it, we get

$$x=\frac{2b-a-Kb}{(K-1)b^2-\left(a^2+ab(K-2)\right)},$$

$$y=\frac{b}{(K-1)b^2-\left(a^2+ab(K-2)\right)}.$$

Introducing $a=K+1, b=\frac{1}{1-K}$ into x and y, the denominator of x and y can be simplified as

$$\frac{1}{(K-1)b^2-\left(a^2+ab(K-2)\right)}=\frac{(1-K)^3}{K^3(K^2-2K+1)}$$

$$=\frac{1-K}{K^3}.$$

Therefore, E may be rewritten as

$$E^{-1}=\frac{1-K}{K^3}\begin{bmatrix} 2b-a-Kb & b & \cdots & b \\ b & 2b-a-Kb & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & 2b-a-Kb \end{bmatrix}.$$

It follows that

$$Z=D^{-1}=K\cdot E^{-1}$$

$$=K\frac{1-K}{K^3}\begin{bmatrix} 2b-a-Kb & b & \cdots & b \\ b & 2b-a-Kb & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & 2b-a-Kb \end{bmatrix}$$

$$= \frac{1-K}{K^2}\begin{bmatrix} \frac{2K^2-K^3-2K+1}{(1-K)^2} & \frac{1}{1-K} & \cdots & \frac{1}{1-K} \\ \frac{1}{1-K} & \frac{2K^2-K^3-2K+1}{(1-K)^2} & \cdots & \frac{1}{1-K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-K} & \frac{1}{1-K} & \cdots & \frac{2K^2-K^3-2K+1}{(1-K)^2} \end{bmatrix}$$

$$= \frac{1-K}{K^2}\begin{bmatrix} \frac{K^2-K+1}{1-K} & \frac{1}{1-K} & \cdots & \frac{1}{1-K} \\ \frac{1}{1-K} & \frac{K^2-K+1}{1-K} & \cdots & \frac{1}{1-K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-K} & \frac{1}{1-K} & \cdots & \frac{K^2-K+1}{1-K} \end{bmatrix}$$

$$= \frac{1}{K^2}\begin{bmatrix} K^2-K+1 & 1 & \cdots & 1 \\ 1 & K^2-K+1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & K^2-K+1 \end{bmatrix}.$$

Define $F = 2Z - I - A$. F can be rewritten as

$$F = 2Z - I - A$$

$$= \frac{2}{K^2}\begin{bmatrix} K^2-K+1 & 1 & \cdots & 1 \\ 1 & K^2-K+1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & K^2-K+1 \end{bmatrix} - I - \frac{1}{K}1_{K\times K}$$

$$= \begin{bmatrix} 1-\frac{2}{K}+\frac{2}{K^2} & \frac{2}{K^2} & \cdots & \frac{2}{K^2} \\ \frac{2}{K^2} & 1-\frac{2}{K}+\frac{2}{K^2} & \cdots & \frac{2}{K^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{2}{K^2} & \frac{2}{K^2} & \cdots & 1-\frac{2}{K}+\frac{2}{K^2} \end{bmatrix} - \frac{1}{K}1_{K\times K}$$

$$= \begin{bmatrix} 1-\frac{3}{K}+\frac{2}{K^2} & \frac{2}{K^2}-\frac{1}{K} & \cdots & \frac{2}{K^2}-\frac{1}{K} \\ \frac{2}{K^2} & 1-\frac{3}{K}+\frac{2}{K^2} & \cdots & \frac{2}{K^2}-\frac{1}{K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{2}{K^2}-\frac{1}{K} & \frac{2}{K^2}-\frac{1}{K} & \cdots & 1-\frac{3}{K}+\frac{2}{K^2} \end{bmatrix}.$$

Compute the matrix product of matrices B and F as follows

$$BF = \frac{1}{K}I_K \times \begin{bmatrix} 1-\frac{3}{K}+\frac{2}{K^2} & \frac{2}{K^2}-\frac{1}{K} & \cdots & \frac{2}{K^2}-\frac{1}{K} \\ \frac{2}{K^2} & 1-\frac{3}{K}+\frac{2}{K^2} & \cdots & \frac{2}{K^2}-\frac{1}{K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{2}{K^2}-\frac{1}{K} & \frac{2}{K^2}-\frac{1}{K} & \cdots & 1-\frac{3}{K}+\frac{2}{K^2} \end{bmatrix}$$

$$= \frac{1}{K}\begin{bmatrix} 1-\frac{3}{K}+\frac{2}{K^2} & \frac{2}{K^2}-\frac{1}{K} & \cdots & \frac{2}{K^2}-\frac{1}{K} \\ \frac{2}{K^2} & 1-\frac{3}{K}+\frac{2}{K^2} & \cdots & \frac{2}{K^2}-\frac{1}{K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{2}{K^2}-\frac{1}{K} & \frac{2}{K^2}-\frac{1}{K} & \cdots & 1-\frac{3}{K}+\frac{2}{K^2} \end{bmatrix}.$$

Extracting the entry located in the very top-left corner of the matrix yields

$$v = \frac{1}{K}\left(1-\frac{3}{K}+\frac{2}{K^2}\right).$$

Solution 2.

First, calculate the eigenvalues of the matrix $1_{K\times K}$, which yields

$$\lambda_1 = K, \lambda_2 = 0, \dots, \lambda_n = 0.$$

Conduct orthogonal decomposition for $1_{K\times K}$ as

$$1_{K\times K} = R\Lambda R^T,$$

where $R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{K1} & r_{K2} & \cdots & r_{KK} \end{bmatrix}$ is an orthogonal matrix ($R^T = R^{-1}$).

Apparently, the eigenvector set of the eigenvalue K is $\{kI_{K\times 1} | k \neq 0\}$. Hence, for the orthogonal matrix R, its first column is $\begin{bmatrix} \frac{1}{\sqrt{K}} \\ \frac{1}{\sqrt{K}} \\ \vdots \\ \frac{1}{\sqrt{K}} \end{bmatrix}$, thus $R = \begin{bmatrix} \frac{1}{\sqrt{K}} & r_{12} & \cdots & r_{1K} \\ \frac{1}{\sqrt{K}} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{K}} & r_{K2} & \cdots & r_{KK} \end{bmatrix}$.

Define

$$P = \frac{1_{K\times K} - I_K}{K-1} = \frac{1}{K-1}(R\Lambda R^T - RI_K R^T) = R\left(\frac{1}{K-1}(\Lambda - I_K)\right)R^T.$$

Note

$$\frac{1}{K-1}(\Lambda - I_K) = \frac{1}{K-1}\begin{bmatrix} K & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} - \frac{1}{K-1}\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & -\frac{1}{K-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{K-1} \end{bmatrix}.$$

Thus, according to Eq. (7.37) of (Yang 2014a), we have

$$\lambda_1 = 1, \lambda_2 = \lambda_3 = \cdots = \lambda_K = \frac{1}{K-1}.$$

Hence

$$B = \text{diag}(\pi_1, \pi_2, \dots, \pi_K) = \text{diag}\left(\frac{1}{K}, \frac{1}{K}, \dots \frac{1}{K}\right) = \frac{1}{K} I_K.$$

Define

$$h = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

It follows that

$$E = B^{-\frac{1}{2}} R = \left(\frac{1}{K} I_K\right)^{-\frac{1}{2}} R = K^{\frac{1}{2}} R.$$

So

$$\begin{aligned} E^T B h &= \left(K^{\frac{1}{2}} R\right)^T B h \\ &= K^{\frac{1}{2}} R^T \times \frac{1}{K} I_K \times \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= K^{-\frac{1}{2}} R^T \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= K^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{\sqrt{K}} & r_{12} & \cdots & r_{1K} \\ \frac{1}{\sqrt{K}} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{K}} & r_{K2} & \cdots & r_{KK} \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= K^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{\sqrt{K}} \\ r_{12} \\ \vdots \\ r_{1K} \end{bmatrix}. \end{aligned}$$

Note that $E^T B h = K^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{\sqrt{K}} \\ r_{12} \\ \vdots \\ r_{1K} \end{bmatrix}$ is exactly the first column of the orthogonal matrix R. By definition, we have

$$\left(\frac{1}{\sqrt{K}}\right)^2 + r_{12}^2 + \cdots r_{1K}^2 = 1.$$

Hence,

$$\sum_{i=2}^{K} r_{1i}^2 = 1 - \frac{1}{K} = \frac{K-1}{K}.$$

Hence,

$$
\begin{aligned}
v &= \sum_{k=2}^{K} \frac{1+\lambda_k}{1-\lambda_k} (E^T B h)_k^2 \\
&= \sum_{k=2}^{K} \frac{1-\frac{1}{K-1}}{1+\frac{1}{K-1}} \times \left(\frac{1}{\sqrt{K}} \begin{bmatrix} \frac{1}{\sqrt{k}} \\ r_{12} \\ \vdots \\ r_{1K} \end{bmatrix} \right)_k^2 \\
&= \sum_{k=2}^{K} \frac{K-2}{K} \times \frac{1}{K} \times r_{1k}^2 \\
&= \frac{K-2}{K^2} \sum_{k=2}^{K} r_{1k}^2 \\
&= \frac{K-2}{K^2} \times \frac{K-1}{K} \\
&= \frac{1}{K}\left(1 - \frac{3}{K} + \frac{2}{K^2}\right).
\end{aligned}
$$

As to the efficiency,

$$
E = \frac{\pi_1(1-\pi_1)}{v} = \frac{\frac{1}{K}\left(1-\frac{1}{K}\right)}{\frac{1}{K}\left(1-\frac{3}{K}+\frac{2}{K^2}\right)} = \frac{K}{K-2}.
$$

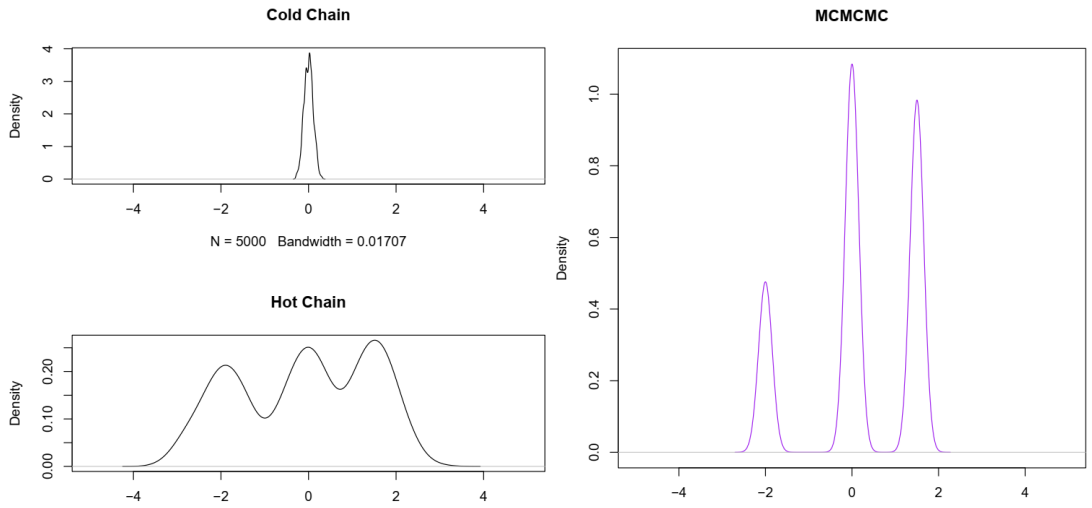
7.11* Show that if the target density is $N(0, 1)$ and the MCMC proposal density is $x'|x \sim N(x, \sigma^2)$, the acceptance proportion is given by equation (7.50).

See Problem 5.3 of (Yang 2006).

7.12* Write an MCMC program to sample from the posterior target $\pi(\theta)$ of Figure 7.15, which is a mixture of three normal distributions. Use a uniform sliding window of size 1. Record the frequencies at which the chain moves between the three regions of the parameter space corresponding to the three peaks: $(-\infty, -1)$, $(-1, 1)$, and $(1, \infty)$. Also pay attention to whether the chain visits those three regions with the correct probabilities (i.e. 0.2, 0.5, and 0.3). Then implement the parallel tempering algorithm of §7.4.1 with a second hot chain at temperature T, and see whether it helps the cold chain to mix.

Solution.

This problem involves the so-called parallel tempering MCMC method. I tentatively set the temperature of the hold chain to be 30, which achieves an acceptance ratio of chain swap of roughly 25%. The result using the script *7.12.R* is shown as follows, the swap acceptance rate being 25.2%. As advised by a prior study (Earl and Deem 2005), people should try to achieving an acceptance ratio of chain swap at around 20% by adjusting the temperature(s). The results of using a single cold chain, a single hot chain at $T = 30$, and the MCMCMC algorithm which couples the two chains, are displayed as follows.



R (*7.12.R*)

```
num_iterations <- 100000; num_chains <- 2
chains <- matrix(0, nrow = num_iterations, ncol = num_chains)
chains[1, ] <- numeric(num_chains)
temperatures <- c(1, 30)

target_distribution <- function(x) {
  weights <- c(0.2, 0.5, 0.3);  means <- c(-2, 0, 1.5);  sigma <- 0.1
  lnl <- log(sum(weights * sapply(means, function(mu) dnorm(x, mean = mu, sd = sigma))))
}
metropolis_hastings_step <- function(current, proposal, temperature, log_target) {
  log_acceptance_ratio <- (log_target(proposal) - log_target(current)) / temperature
  if (log(runif(1)) < log_acceptance_ratio) {return(proposal)} else{return(current)}
}

accepted_swap <- 0
for (i in 2:num_iterations) {
  for (chain in 1:num_chains) {
    current_value <- chains[i-1, chain]
    proposal_value <- runif(1, current_value-0.5, current_value+0.5)
    temperature <- temperatures[chain]
    chains[i, chain] <- metropolis_hastings_step(current_value, proposal_value, temperature,
target_distribution)
  }
  if (i %% 100 == 1) {
    chain1 <- 1; chain2 <- 2; current_values <- chains[i, c(chain1, chain2)]
```

```
    if (log(runif(1)) < log_acceptance_ratio_swap) {
      chains[i, c(chain1, chain2)] <- chains[i, c(chain2, chain1)]
      accepted_swap <- accepted_swap + 1
    }
  }
}
plot(density(chains[-(1:num_iterations/2), 1]), main = "MCMCMC", xlab = "", col = "purple",
xlim=c(-5,5))
cat(paste("accepted swap:", accepted_swap), "\n")
```

Chapter 8. Bayesian phylogenetics

8.1 A proposal modifies two variables x and y with their sum ($s = x + y$) fixed. Generate $x' \sim U(x - \varepsilon/2, x + \varepsilon/2)$, reflected into the range $0 < x' < s$, if necessary. Set $y' = s - x'$. Derive the proposal ratio for this move. Also explain why the proposal ratio remains the same if x' is generated from the normal or Bactrian proposals around x (§7.2.2–3). [Hint: One method is to use Theorem 2 in Appendix A. Another is to use Green's formulation and the mapping $(x, y) \leftrightarrow (x', y')$.]

8.2 Multiple peaks can be caused by conflicting prior and likelihood. Download the lizard dataset of Leaché and Mulcahy (2007) from the book's website (file name Sceloporus.nex). This is the same data as were analysed in Example 8.1, with 123 sequences and 1,606 alignment columns. Run MrBayes (Ronquist et al. 2012b) under JC69+Γ_5 with the tree topology fixed at the ML tree, and using different initial branch lengths. The default prior for branch lengths is i.i.d. exponential with mean 0.1, so that the prior mean of tree length (sum of branch lengths) is $0.1 \times (2 \times 123 - 3) = 24.3$. The ML estimate of tree length under the model is 2.2. The prior and the likelihood are thus in conflict. Try to confirm the existence of two peaks in the posterior: one higher peak with tree length $T \approx 2.2$ and log likelihood $\ell \approx -1310$ and another much lower peak at $T \approx 18.3$ with $\ell \approx -1350$. This dataset is problematic for MrBayes 3.1, which by default use 0.1 for initial branch lengths and almost always gets stuck at the lower peak. MrBayes 3.2.1, with the tree length multiplier, almost always converges to the higher peak, but can get stuck at the lower peak if the starting tree is sampled around the lower leak generated from MrBayes 3.1 (see notes in the data file).

Solution.

Unfortunately, I see the version 3.1 of MrBayes at nowhere. If you happen to know any old-fashioned Bayesian phylogenetics person who has not updated their software for a decade, please let them know this exercise.

8.3* *Fair coin paradox with no parameter*. Confirm the following either analytically or using computer simulation. Suppose a coin is fair with the probability of heads to be $\theta_0 = \frac{1}{2}$. Flip the coin n times and observe x heads. Calculate the posterior probabilities for two hypotheses: H_1: $\theta = 0.4$ and H_2: $\theta = 0.6$, by assigning equal prior probabilities for the two models. Confirm that when n is large, the posterior model probability $P_1 = \Pr(H_1|x)$ is 0 or 1, each half of the times. [Hint. There is no need to write an MCMC program, but you will need to write down the likelihood functions for the two models. To simulate in R, use runif to generate $U(0, 1)$ random numbers, and hist to plot the histogram.]

Solution.

According to the statement of the problem, the prior for each hypothesis $P(H_1) = P(H_2) = 0.5$. It follows that

$$
\begin{aligned}
P(H_1|X=x) &= \frac{P(H_1)P(X=x|H_1)}{P(H_1)P(X=x|H_1)+P(H_2)P(X=x|H_2)} \\
&= \frac{0.5\times 0.4^x\times(1-0.4)^{n-x}\times\binom{n}{x}}{0.5\times 0.4^x\times(1-0.4)^{n-x}\times\binom{n}{x}+0.5\times 0.6^x\times(1-0.6)^{n-x}\times\binom{n}{x}} \\
&= \frac{0.4^x\times(0.6)^{n-x}}{0.4^x\times(0.6)^{n-x}+0.6^x\times(0.4)^{n-x}} \\
&= \frac{1}{1+\frac{0.6^x\times(0.4)^{n-x}}{0.4^x\times(0.6)^{n-x}}} \\
&= \frac{1}{1+\left(\frac{3}{2}\right)^x\left(\frac{2}{3}\right)^{n-x}} \\
&= \frac{1}{1+\left(\frac{3}{2}\times\frac{2}{3}\right)^x\left(\frac{2}{3}\right)^{n-2x}} \\
&= \frac{1}{1+\left(\frac{2}{3}\right)^{n-2x}}.
\end{aligned}
$$

Define a random variable $Y=P(H_1|X)=\left(1+\left(\frac{2}{3}\right)^{n-2X}\right)^{-1}$. When $x>\frac{n}{2}+5$, it can be calculated that

$$\frac{1}{1+\left(\frac{2}{3}\right)^{n-2x}}<\frac{1}{1+\left(\frac{2}{3}\right)^{-10}}=0.01.$$

Similarly, when $x<\frac{n}{2}+5$, we have

$$\frac{1}{1+\left(\frac{2}{3}\right)^{n-2x}}>\frac{1}{1+\left(\frac{2}{3}\right)^{10}}=0.99.$$

In other words, in case where the number of heads differs from the mean by five, P_1 will be fairly close to either zero or one. According to CLT, we have

$$X\sim Normal\left(\tfrac{1}{2}n,\tfrac{1}{4}n\right), \text{ as } n\to\infty.$$

Further, denote the standard normal CDF by Φ. So

$$
\begin{aligned}
P\left(\frac{n}{2}-5<X<\frac{n}{2}+5\right) &= P\left(5<X-\frac{n}{2}<5\right) \\
&= P\left(-\frac{5}{\sqrt{\frac{1}{4}n}}<\frac{X-\frac{n}{2}}{\sqrt{\frac{1}{4}n}}<\frac{5}{\sqrt{\frac{1}{4}n}}\right) \\
&= 2\Phi\left(\frac{10}{\sqrt{n}}\right)-1.
\end{aligned}
$$

Hence

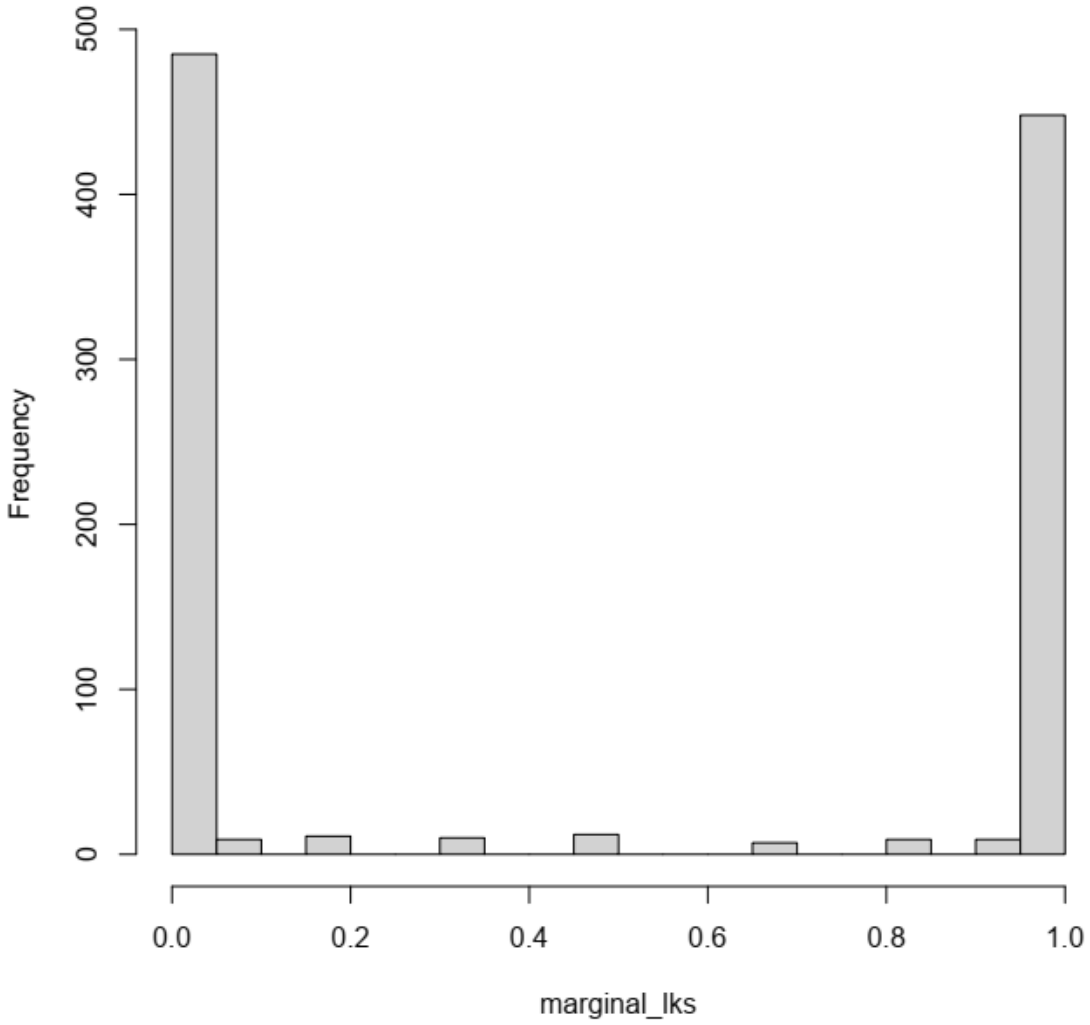
$$\lim_{n\to\infty} 2\Phi\left(\frac{10}{\sqrt{n}}\right) - 1 = 2 \times 0.5 - 1 = 0.$$

Thus, as n approaches infinity, the probability that X is between $\frac{n}{2} - 5$ and $\frac{n}{2} + 5$ tends to zero, and in such cases P_1 will be 0 or 1 each roughly half of the times. Hence, $P(H_1|X)$ behaves like a Bernoulli distributed variable with the parameter $p = 0.5$.

The R code for simulation can be found in *8.3.R*.

R (*8.3.R*)

```
get_lnl <- function(p, n, x){
    lnl <- x*log(p) + (n-x)*log(1-p)
}
##############################################
p <- 0.5
n <- 10000
n_sim <- 1000

marginal_lks <- numeric(n_sim)
for(i in 1:n_sim){
    x <- rbinom(1, n, p)
    ps <- c(0.4, 0.6)
    lnls <- sapply(ps, get_lnl, n=n, x=x)
    #print(c(lnls[2]-lnls[1], 1/(1 + exp(lnls[2] - lnls[1]))))
    lk <- 1/(1 + exp(lnls[2] - lnls[1]))
    marginal_lks[i] <- lk
}
pdf("8.3-1.pdf")
hist(marginal_lks, breaks=30)
dev.off()
```



8.4* *Fair coin paradox* (Lewis et al. 2005; Yang and Rannala 2005). Redo Problem 8.3 except that the two hypotheses now have unknown parameters: H_1: $\theta < \frac{1}{2}$ and H_2: $\theta > \frac{1}{2}$. Assign uniform priors for θ in each model. Confirm that when n is large, the posterior model probability $\Pr(H_1|x)$ behaves like a $U(0, 1)$ random number. [Hint. Given the prior $\theta \sim U(0, 1)$, the posterior $\theta|x \sim \text{beta}(x + 1, n - x + 1)$. Then $P_1 = \Pr(H_1|x) = \Pr(\theta < 1/2|x)$. Also, to simulate in R, use `runif` to generate $U(0, 1)$ random numbers, `pbeta` to calculate $\Pr(H_1|x)$, and hist to plot the histogram.]

Solution.

This very interesting problem was first formulated in (Lewis et al. 2005), but in this paper the result is shown implicitly. It is in (Yang and Rannala 2005) that an analytical result is given. However, its proof is very short and seems somehow incomplete. To help readers better understand it, I follow (Yang and Rannala 2005) to give a complete proof. Note that there might be a little confusion in the notation used at the corresponding part of the original paper (Yang and Rannala 2005) where y indicates a variable in some cases but the value it takes in other cases. This is clarified in the following solution.

According to CLT,

$$X \sim Normal\left(\frac{1}{2}n, \frac{1}{4}n\right).$$

According to the problem statement, the posterior of θ is given by

$$\theta|x \sim beta(x+1, n-x+1).$$

According to CLT, when $n \to \infty$, the above posterior converges to $Normal\left(\frac{x}{n}, \frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}\right)$.

Define a new variable $Y = \frac{X}{n}$. We have when $n \to \infty$,

$$\theta|X \sim Normal\left(Y, \frac{Y(1-Y)}{n}\right).$$

It follows that

$$P_1 = P\left(\theta < \frac{1}{2}\middle|X\right) = \Phi\left(\frac{\frac{1}{2}-Y}{\sqrt{\frac{Y(1-Y)}{n}}}\right).$$

Denote the CDF and PDF of standard normal distribution by Φ, ϕ respectively. Hence,

$$\frac{dP_1}{dy} = \left(\Phi\left(\frac{\frac{1}{2}-y}{\sqrt{\frac{y(1-y)}{n}}}\right)\right)' = \phi\left(\frac{\frac{1}{2}-y}{\sqrt{\frac{y(1-y)}{n}}}\right)\left(\frac{\frac{1}{2}-y}{\sqrt{\frac{y(1-y)}{n}}}\right)' \tag{8.1}$$

Define

$$a = \Phi^{-1}(P_1) = \frac{\frac{1}{2}-y}{\sqrt{\frac{y(1-y)}{n}}}. \tag{8.2}$$

Substituting Eq. (8.2) into Eq. (8.1), we have

$$\frac{dP_1}{dy} = \phi(a)\left(\frac{\frac{1}{2}-y}{\sqrt{\frac{y(1-y)}{n}}}\right)'. \tag{8.3}$$

Calculate the derivative of $\Phi^{-1}(P_1)$ w.r.t. y as follows

$$\left(\frac{\frac{1}{2}-y}{\sqrt{\frac{y(1-y)}{n}}}\right)' = \frac{\left(\frac{1}{2}-y\right)'\sqrt{\frac{y(1-y)}{n}} - \left(\frac{1}{2}-y\right)\left(\sqrt{\frac{y(1-y)}{n}}\right)'}{\left(\sqrt{\frac{y(1-y)}{n}}\right)^2}$$

$$= n \times \frac{-\sqrt{\frac{y(1-y)}{n}} - \left(\frac{1}{2}-y\right) \times \frac{1}{2} \times \frac{\left(y(1-y)\right)'}{n\sqrt{\frac{y(1-y)}{n}}}}{y(1-y)}$$

$$= \frac{-\sqrt{y(1-y)} - \frac{1}{2}\left(\frac{1}{2}-y\right)(1-2y)\left(y(1-y)\right)^{-\frac{1}{2}}}{y(1-y)}$$

$$= \frac{-4y(1-y)-(1-2y)^2}{4n^{-\frac{1}{2}}\left((1-y)y\right)^{\frac{3}{2}}}$$

$$= \frac{-1}{4\,n\left(\frac{(1-y)y}{n}\right)^{\frac{3}{2}}}. \tag{8.4}$$

According to Eq. (8.2), we have

$$a^2 y(1-y) = n\left(\frac{1}{2}-y\right)^2. \tag{8.5}$$

So

$$(n+a^2)y^2 - (n+a^2)y + \frac{n}{4} = 0.$$

Thus

$$\frac{y(1-y)}{n} = \frac{1}{4(n+a^2)}. \tag{8.6}$$

Plugging Eqs. (8.4) and (8.6) into Eq. (8.2), we have

$$\frac{dP_1}{dy} = \phi(a) \times \frac{-1}{4\,n\left(\frac{(1-y)y}{n}\right)^{\frac{3}{2}}}$$

$$= \phi(a) \times \frac{-1}{4n\left(\frac{1}{4(n+a^2)}\right)^{\frac{3}{2}}}$$

$$= \phi(a) \times (-2n^{-1}) \times (n+a^2)^{\frac{3}{2}}$$

$$= \phi(a) \times \left(-2\sqrt{n}\right) \times \left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}.$$

Hence

$$\left|\frac{dP_1}{dy}\right| = \phi(a) \times \left(2\sqrt{n}\right) \times \left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}.$$

Further, it can be shown that

$$f_Y(y) = \frac{1}{\sqrt{2\pi \times \frac{1}{4n}}} e^{-2n\left(y-\frac{1}{2}\right)^2} = \sqrt{\frac{2n}{\pi}} e^{-2n\left(y-\frac{1}{2}\right)^2}. \tag{8.7}$$

Plugging Eq. (8.6) into Eq. (8.5), we have

$$\left(\frac{1}{2}-y\right)^2 = \frac{a^2}{4(n+a^2)}. \tag{8.8}$$

Plugging Eq. (8.8) into Eq. (8.7) and eliminating y, we have

$$f(y(P_1)) = \sqrt{\frac{2n}{\pi}} e^{-2n\frac{a^2}{4(n+a^2)}}.$$

Hence,

$$f(P_1) = f_Y\big(y(P_1)\big) \times \left|\frac{dy}{dP_1}\right|$$

$$= \frac{\sqrt{\frac{2n}{\pi}} e^{-2n\frac{a^2}{4(n+a^2)}}}{\phi(a) \times \left(2\sqrt{n}\right) \times \left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}}$$

$$= \frac{e^{-\frac{na^2}{2(n+a^2)}}}{\phi(a) \times \sqrt{2\pi} \times \left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}} = \frac{e^{-\frac{n\left(\Phi^{-1}(P_1)\right)^2}{2\left(n+\left(\Phi^{-1}(P_1)\right)^2\right)}}}{\phi\left(\Phi^{-1}(P_1)\right) \times \sqrt{2\pi} \times \left(1+\frac{\left(\Phi^{-1}(P_1)\right)^2}{n}\right)^{\frac{3}{2}}}.$$

It thus follows that

$$\lim_{n\to\infty} f(P_1) = \lim_{n\to\infty} \frac{e^{-\frac{na^2}{2(n+a^2)}}}{\phi(a) \times \sqrt{2\pi} \times \left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}}$$

$$= \lim_{n\to\infty} \frac{e^{-\frac{na^2}{2(n+a^2)}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2}} \times \sqrt{2\pi} \times \left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}}$$

$$= \lim_{n\to\infty} \frac{e^{-\frac{na^2}{2(n+a^2)}+\frac{a^2}{2}}}{\left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}}$$

$$= \lim_{n\to\infty} \frac{e^{\frac{a^4}{2(n+a^2)}}}{\left(1+\frac{a^2}{n}\right)^{\frac{3}{2}}}$$

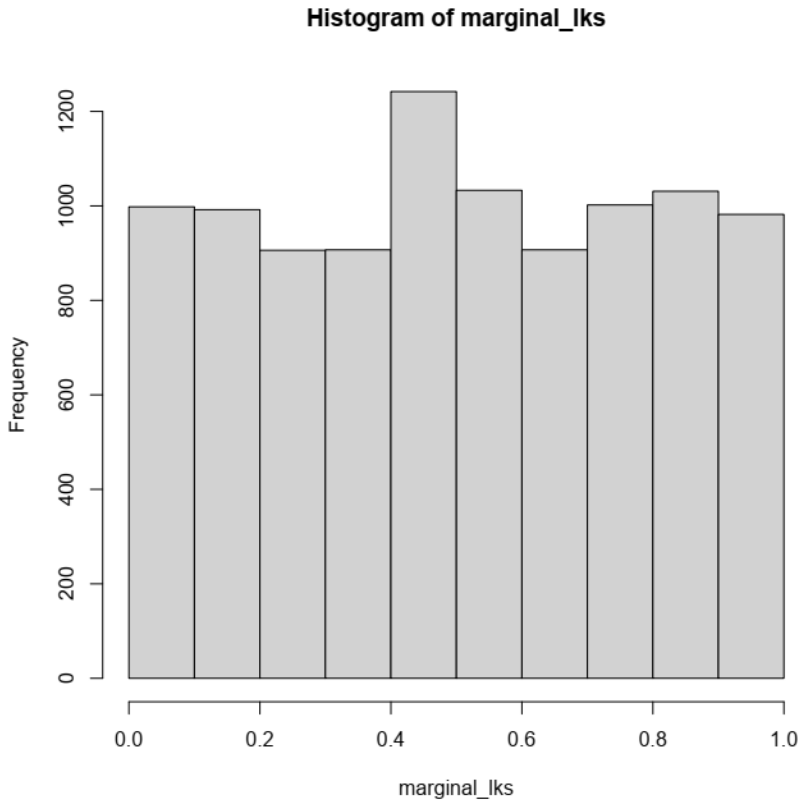
$$= 1.$$

The R code for simulation can be found in *8.4.R*.

R

```
get_lnl <- function(p, n, x){
   lnl <- x*log(p) + (n-x)*log(1-p)
}
integrand <- function(k) exp(get_lnl(k,n=n,x=x))
```

```
##############################################
p <- 0.5
n <- 1000
n_sim <- 1000

marginal_lks <- numeric(n_sim)

for(i in 1:n_sim){
   x <- rbinom(1, n, p)
   i1 <- integrate(integrand, 0,1/2)
   i2 <- integrate(integrand, 1/2,1)
   marginal_lk <- i1$value/(i1$value+i2$value)
   marginal_lks[i] <- marginal_lk
}

pdf("8.4-1.pdf")
hist(marginal_lks, breaks=30)
dev.off()
```



8.5* *Two equally wrong exponential models*. Conduct a computer simulation to explore the posterior model probability in large datasets, when two equally wrong exponential models are compared. Suppose we take a sample of size $n = 10$ from the exponential distribution with mean $\mu_0 = \log 4$, and use the data to compare two models H_1: $\mu_1 = 1$, and H_2: $\mu_2 = 2$. Those two models are equally wrong. [Note that the two models are equally wrong if $\mu_0 = \log\left\{\frac{\mu_2}{\mu_1}\right\}\Big/\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)$.] We assign the prior $\pi_1 = \pi_2 = \frac{1}{2}$ for the two models. Calculate the posterior model probability $P_1 = \Pr(H_1|x)$. Simulate 10,000 datasets and plot the histogram of the 10,000 P_1 values. Repeat the analysis using $n = 100$ and 1,000 to see the impact of the sample size. [Hint. In R, use rexp to generate exponential random variables.]

Solutions.

As follows two solutions are provided, the first showing that the expectation of the two probabilities are equal, and the second adopting a way similar to Problem 8.4's solution. Note that the two ways of solving the problem is somehow different in how to define "equally wrong models", and our feeling is that the way it is defined in Solution 2 may be closer to what Ziheng hopes to seek.

Solution 1.

Denote the variable X as the mean of $X_1, X_2, \ldots X_n$. According to the problem statement,

$$P(H_1|X = x) = \frac{0.5 \times \mu_1^{-1}\, \mathrm{e}^{-\mu_1^{-1}x}}{0.5 \times \mu_1^{-1}\mathrm{e}^{-\mu_1^{-1}x} + 0.5 \times \mu_2^{-1}\mathrm{e}^{-\mu_2^{-1}x}}$$
$$= \frac{1}{1 + \frac{\mu_1}{\mu_2}\mathrm{e}^{\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)x}}.$$

Denote k as any positive number. It follows that

$$\begin{cases} P(H_1|\mu_0 + k) + P(H_2|\mu_0 + k) = 1 \\ P(H_1|\mu_0 - k) + P(H_2|\mu_0 - k) = 1 \end{cases}. \tag{8.1}$$

Consider the more general case where $\mu_0 = \log\left(\frac{\frac{\mu_2}{\mu_1}}{\mu_1^{-1} - \mu_2^{-1}}\right)$. Therefore, we have

$$P(H_1|\mu_0 + k) + P(H_1|\mu_0 - k)$$
$$= \frac{1}{1 + \frac{\mu_1}{\mu_2}\mathrm{e}^{(\mu_1^{-1} - \mu_2^{-1})\left(\frac{\log\left(\frac{\mu_2}{\mu_1}\right)}{\mu_1^{-1} - \mu_2^{-1}} + k\right)}} + \frac{1}{1 + \frac{\mu_1}{\mu_2}\mathrm{e}^{(\mu_1^{-1} - \mu_2^{-1})\left(\frac{\log\left(\frac{\mu_2}{\mu_1}\right)}{\mu_1^{-1} - \mu_2^{-1}} - k\right)}}$$
$$= \frac{1}{1 + \frac{\mu_1}{\mu_2}\mathrm{e}^{\log\left(\frac{\mu_2}{\mu_1}\right) + (\mu_1^{-1} - \mu_2^{-1})k}} + \frac{1}{1 + \frac{\mu_1}{\mu_2}\mathrm{e}^{\log\left(\frac{\mu_2}{\mu_1}\right) - (\mu_1^{-1} - \mu_2^{-1})k}}$$
$$= \frac{1}{1 + \frac{\mu_1}{\mu_2}\frac{\mu_2}{\mu_1}e^{(\mu_1^{-1} - \mu_2^{-1})k}} + \frac{1}{1 + \frac{\mu_1}{\mu_2}\frac{\mu_2}{\mu_1}e^{-(\mu_1^{-1} - \mu_2^{-1})k}}$$
$$= \frac{1}{1 + e^{(\mu_1^{-1} - \mu_2^{-1})k}} + \frac{1}{1 + e^{-(\mu_1^{-1} - \mu_2^{-1})k}}$$

$$= \frac{1}{1+e^{(\mu_1^{-1}-\mu_2^{-1})k}} + \frac{e^{(\mu_1^{-1}-\mu_2^{-1})k}}{e^{(\mu_1^{-1}-\mu_2^{-1})k}\cdot 1+e^{(\mu_1^{-1}-\mu_2^{-1})k}\cdot e^{-(\mu_1^{-1}-\mu_2^{-1})k}}$$
$$= 1.$$

Actually, at the third-to-last step, if you are familiar with logistic regression or some basics of neural network, you should be able to recognize that it is in the form of sigmoid function $\sigma\big(f(x)\big)$ where $f(x) = -(\mu_1^{-1} - \mu_2^{-1})k$. Then it is easy to find the result according to a well-known property of sigmoid function which states that $\sigma\big(f(x)\big) + \sigma\big(-f(x)\big) = 1$.

By simultaneously solving Eq. 8.1 and $P(H_1|\mu_0 + k) + P(H_1|\mu_0 - k) = 1$ as calculated above, we have

$$\begin{cases} P(H_1|X = \mu_0 + k) = P(H_2|X = \mu_0 - k) \\ P(H_1|X = \mu_0 - k) = P(H_2|X = \mu_0 + k) \end{cases},$$

or equivalently

$$\begin{cases} P(H_1|X = x) = P(H_2|2\mu_0 - x) \\ P(H_2|X = x) = P(H_1|2\mu_0 - x) \end{cases}.$$

According to CLT,

$$X \sim Normal\left(\mu_0, \frac{\mu_0^2}{n}\right), \text{ as } n \to \infty.$$

Define a new variable $Y = \frac{X-\mu_0}{\sqrt{\frac{\mu_0^2}{n}}}$. Thus, Y follows standard normal distribution as $n \to \infty$. Hence, the expectation of $P(H_1|X)$ can be calculated as

$$E\big(P(H_1|X)\big) \approx \int_{-\infty}^{\infty} \left|\frac{dy}{dx}\right| \times \phi\left(\frac{x-\mu_0}{\sqrt{\frac{\mu_0^2}{n}}}\right) \times P(H_1|X = x)\,\mathrm{dx}$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\frac{\mu_0^2}{n}}} \times \phi\left(\frac{\mu_0 - x}{\sqrt{\frac{\mu_0^2}{n}}}\right) \times P(H_2|X = 2\mu_0 - x)\,\mathrm{dx}$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\frac{\mu_0^2}{n}}} \times \phi\left(\frac{(2\mu_0 - x) - \mu_0}{\sqrt{\frac{\mu_0^2}{n}}}\right) \times P(H_2|X = 2\mu_0 - x)dx$$
$$\approx E\big(P(H_1|X)\big).$$

Solution 2.

According to the problem statement, we have

$$P(H_1|X_1 = x_1, \dots, X_n = x_n) = \frac{\left(\frac{1}{\mu_1}\right)^n e^{-\frac{1}{\mu_1}(x_1+\cdots+x_n)}}{e^{-\frac{1}{\mu_1}(x_1+\cdots+x_n)} + e^{-\frac{1}{\mu_2}(x_1+\cdots+x_n)}}$$

$$= \frac{1}{1 + \left(\frac{\mu_1}{\mu_2}\right)^n e^{\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)(x_1 + \cdots + x_n)}}$$

$$= \frac{1}{1 + e^{\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)(x_1 + \cdots + x_n) + n\log\left(\frac{\mu_1}{\mu_2}\right)}}.$$

As given in the problem statement,

$$\mu_0 = \frac{\log\left(\frac{\mu_2}{\mu_1}\right)}{\frac{1}{\mu_1} - \frac{1}{\mu_2}}.$$

So

$$n \log\left(\frac{\mu_1}{\mu_2}\right) = -\mu_0 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right).$$

Substituting $n \log\left(\frac{\mu_2}{\mu_1}\right)$ for $\mu_0 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)$ into the above, we have

$$P(H_1 | X_1 = x_1, \ldots, X_n = x_n) = \frac{1}{1 + e^{\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)(x_1 + \cdots + x_n - n\mu_0)}}.$$

Define $Y_n = X_1 + \cdots + X_n$. According to CLT, apparently

$$Y_n \dot{\sim} Normal(n\mu_0, n\mu_0^2).$$

Define $Z_n = P(H_1 | X_1, \ldots, X_n)$, we have

$$Z_n = \frac{1}{1 + e^{\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)(X_1 + \cdots + X_n - n\mu_0)}}$$

$$= \frac{1}{1 + e^{\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)(Y_n - n\mu_0)}}.$$

Because $Y_n - n\mu_0 \dot{\sim} Normal(0, n\mu_0^2)$, using a similar logic in Problem 8.3 above, it is not difficult to see that it holds that the two models given in the problem statement are equally wrong.

8.6 *Phylogenetic reconstruction using Bayesian inference*. Use the same data of Problem 4.6 to infer the phylogeny using the Bayesian method under the same substitution models, in comparison with the likelihood analysis. You can use MRBAYES to run tree search under those models.

Solution.

I choose to use RevBayes (Hohna et al. 2016), the so-called new generation of MrBayes (MrBayes is no longer updated). The following takes JC69 as an example (see "JC69.revbayes"). More information can be found at RevBayes' tutorial website at https://revbayes.github.io/tutorials/ctmc. As to the alignment, "sample.fasta" given in the folder "data/" of the current chapter is the same as the one used in Problem 4.6 ("rbcL.nogaps_amb.fas").

RevBayes

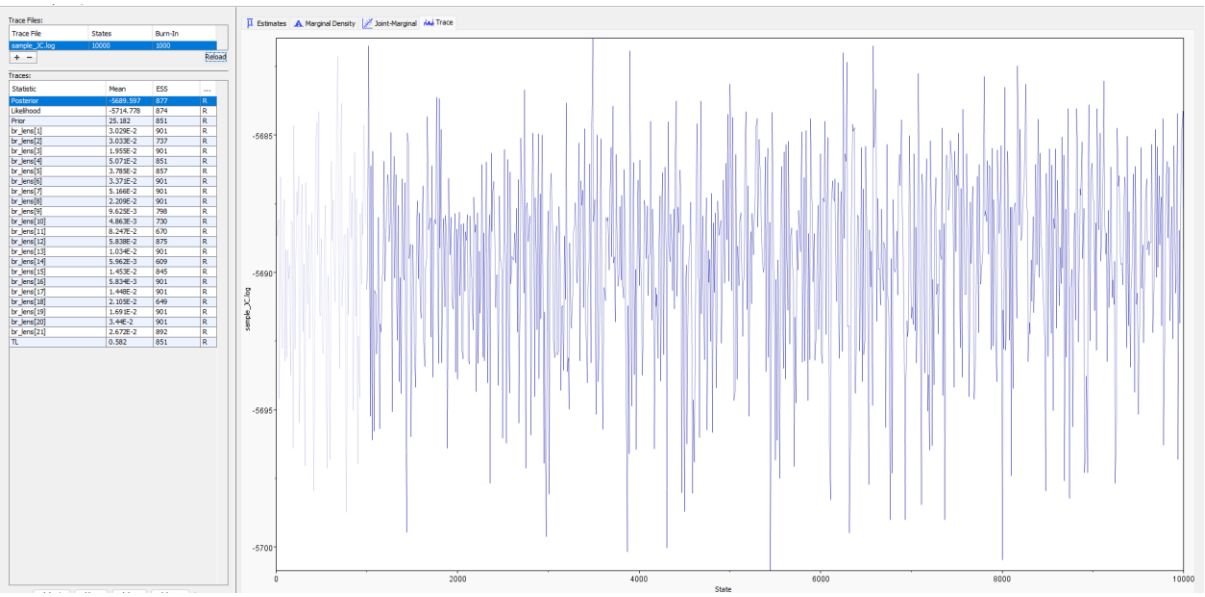
```
# load alignment
data <- readDiscreteCharacterData("sample.fasta")
# taxon info
num_taxa <- data.ntaxa()
num_branches <- 2 * num_taxa - 3
taxa <- data.taxa()
# moves
moves    = VectorMoves()
monitors = VectorMonitors()
# transition rate matrix
Q <- fnJC(4)
# topology prior
topology ~ dnUniformTopology(taxa)
# NNI+SPR
moves.append( mvNNI(topology, weight=num_taxa) )
moves.append( mvSPR(topology, weight=num_taxa/10.0) )
for (i in 1:num_branches) {
  br_lens[i] ~ dnExponential(10.0)
  moves.append( mvScale(br_lens[i]) )
}
TL := sum(br_lens)
# phylogeny
psi := treeAssembly(topology, br_lens)

# set up everything
seq ~ dnPhyloCTMC(tree=psi, Q=Q, type="DNA")
seq.clamp(data)
mymodel = model(Q)
monitors.append( mnModel(filename="output/JC69/sample_JC.log", printgen=10) )
monitors.append( mnFile(filename="output/JC69/sample_JC.trees", printgen=10, psi) )
monitors.append( mnScreen(printgen=100, TL) )
mymcmc = mcmc(mymodel, monitors, moves)

# run mcmc
mymcmc.burnin(1000,100)
mymcmc.run(generations=2000)
```

```
q()
```

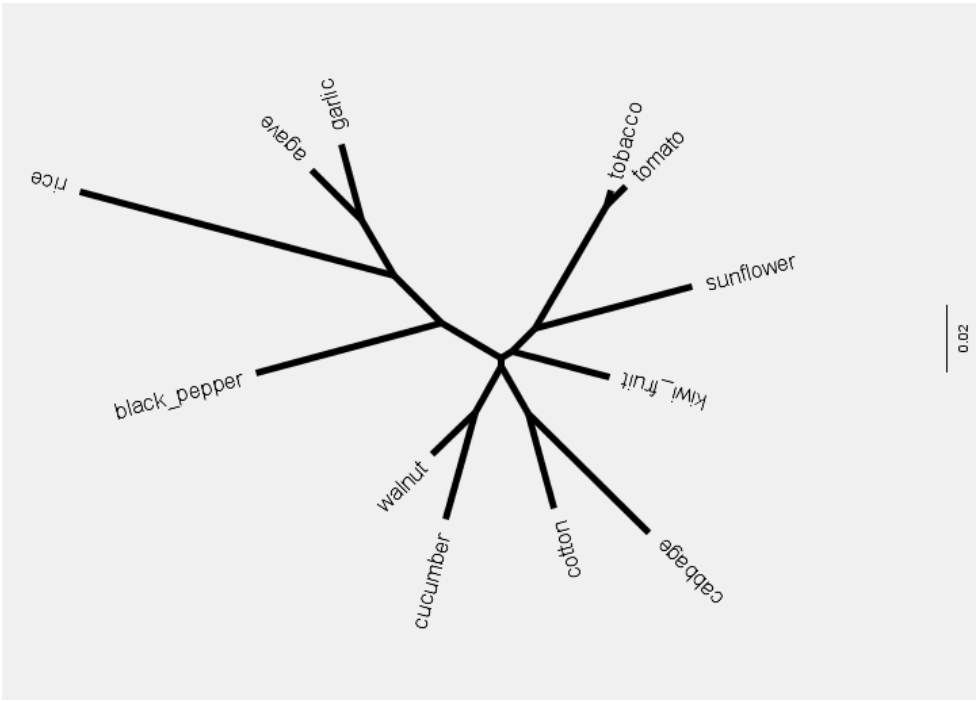
We can further summarize the result by the following command where the MAP tree is generated.

```
RevBayes
# summarize
treetrace = readTreeTrace("output/JC69/sample_JC.trees ", treetype="non-clock")
map_tree = mapTree(treetrace,"output/JC69/sample_JC_MAP.tree")
```

Now, open the file “output/JC69/sample_JC.log” in Tracer. You should see something like the following.



The MAP tree is displayed as follows.



Another script, “GTR+G5.revbayes” is used for tree construction under the model GTR+G5. Specifically, to enable the gamma rate variation in RevBayes, it is needed to specify the following.

RevBayes

alpha ~ dnUniform(0.0, 10)
sr := fnDiscretizeGamma(alpha, alpha, 5) # discrete gamma w 5 categories
moves.append(mvScale(alpha, weight=2.0))

Bonus question.

Compare the estimated parameters by RevBayes with those by IQ-Tree or raxml-ng in Problem 4.6 under the same model. What do you see?

Chapter 9. Coalescent theory and species trees

9.1 Write a small simulation program to generate a sample of n (= 10, say) DNA sequences from a population. Use the JC69 mutation model. Simulate L = 5 loci, with l = 1,000 sites at each locus. For example, implement Algorithms 9.1 or 9.2.

Solution.

If you were as lazy as me, you could simply use the following R code for the purpose. Note that the R packages ape (Paradis and Schliep 2019) and Phangorn (Schliep 2011) need to be installed.

R (9.1.R)

```
#! /usr/bin/env Rscript
library(ape)
library(phangorn)
coalescent_tree <- rcoal(10)
simSeq(coalescent_tree, type="DNA", l=1000) # need to further check if the default model is JC69.
```

9.2 Two species A and B, each of population size N, separated τ generations ago. We sample two sequences a_1 and a_2 from species A and one sequence b from species B. Derive the probability that the gene tree has the topology $((a_1, a_2), b)$.

Solution.

Denote the probability that the gene tree has the topology $\big((a_1, a_2), b\big)$ as P_{same} and the probability of other gene tree topologies as P_{diff}. According to the problem statement, we have

$$P_{diff} = P_{no\_coal} \times P_{joining},$$

where P_{no_coal} is the probability that sequences a_1 and a_2 do not coalesce in population A, and $P_{joining}$ indicates the probability that random joining of all gene sequences in population AB. It is not difficult to see the following

$$P_{no\_coal} = e^{-\frac{\tau}{2N}},$$

$$P_{joining} = \frac{2}{3}.$$

So

$$P_{same} = 1 - P_{diff} = 1 - \frac{2}{3}e^{-\frac{\tau}{2N}}.$$

9.3* ML estimation of the species tree for three species given the gene tree at one locus, with one sequence from each species. Use the gene tree $G = ((A, B), C)$, with node ages t_0 and t_1, of Figure 9.22a as given data to evaluate the likelihood for the species

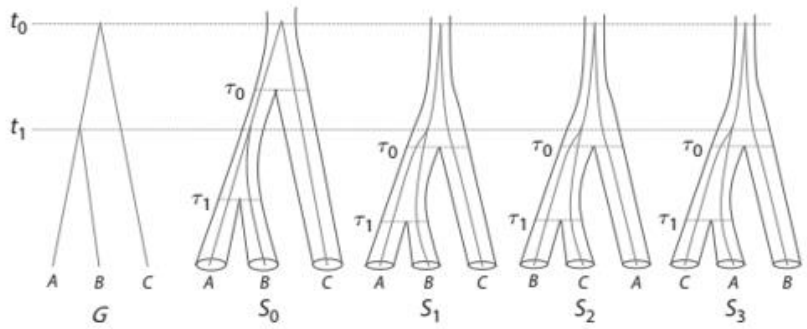


Fig. 9.22 Estimation of the species tree for three species using a gene tree for one locus, with one sequence from each species. (**a**) The gene tree, with topology $G = ((A, B), C)$ and node ages t_0 and t_1, is the given data. (**b**) The species trees with their parameters. Species tree S_0 involves parameters τ_0 and τ_1, under the constraints $\tau_1 \le t_1 \le \tau_0 \le t_0$, while each of species trees S_1, S_2, and S_3 involves parameters τ_0 and τ_1, with $\tau_1 \le \tau_0 \le t_1 \le t_0$. Each of the species tree also involve two population size parameters θ_0 and θ_1, which are not shown. The Maximum Tree algorithm assumes $\theta_0 = \theta_1 = \theta$. Note that species trees S_0 and S_1 have the same topology.

trees of Figure 9.22b, under the assumption that all populations have the same θ. Treat species trees S_0 and S_1 separately even though they have the same tree topology. Show that the ML estimate of the species tree is S_0, with $\hat{\tau}_0 = t_0$ and $\hat{\tau}_1 = t_1$. [Hint: Write down the likelihood function for species tree S_0, which is the multispecies coalescent density for the gene tree, $f(G, t_0, t_1 | S_0, \tau_0, \tau_1, \theta)$, and maximize it by adjusting τ_0, τ_1, and θ under the constraints $\tau_1 \le t_1 \le \tau_0 \le t_0$. Then repeat the analysis for species trees S_1, S_2, and S_3.]

Solution.

According to Eq. (9.45) in (Yang 2014a), and because $\theta_0 = \theta_1 = \theta$, we have

$$f(G, t | S, \Theta) = \left(\frac{2}{\theta}\right)^C e^{-\frac{2}{\theta}T},$$

where C is the number of coalescent events on the gene tree, T is the so-called "total per-lineage-pair coalescent time" summed over all populations and all gene trees, and $\Theta = (S, \tau_0, \tau_1, \theta)$. Denote the "total per-lineage-pair coalescent time" for species tree S_k as T_k. According to the statement of the problem, we have $C = 2$ for all species trees, and

$$T_0 = (t_1 - \tau_1) + (t_0 - \tau_0),$$
$$T_1 = T_2 = T_3 = (\tau_0 - \tau_1) + 3(t_1 - \tau_0) + (t_0 - t_1).$$

The logic is to calculate the maximum likelihood under four species trees S_0, S_1, S_2, S_3 one by one and compare their values.

a)

As to the species trees S_1, S_2, S_3, we have

$$f(G,t|S_k,\tau_0,\tau_1,\theta)=\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}T_1}=\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))}$$

where $k=1,2,3$. Thus, we are looking for

$$(\hat{\tau}_0,\hat{\tau}_1,\hat{\theta})=\underset{\tau_0,\tau_1,\theta}{\operatorname{argmax}}\left\{\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))}|\tau_1\le\tau_0\le t_1\le t_0,\theta>0\right\}.$$

Rewrite the likelihood function as

$$\begin{aligned}f(G,t|S_k,\tau_0,\tau_1,\theta)&=\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))}\\&=\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}(2t_1+t_0-2\tau_0-\tau_1)}.\end{aligned}$$

Define $T^*=2t_1+t_0-2\tau_0-\tau_1$.

Set

$$\frac{\partial\left(\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}T^*}\right)}{\partial\theta}=0,$$

and by solving the above, we obtain $\hat{\theta}=T^*$.

Thus, the maximum of $f(G,t|S_k,\tau_0,\tau_1,\theta)=\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}(2t_1+t_0-2\tau_0-\tau_1)}$ can be achieved when $\theta=T^*=2t_1+t_0-2\tau_0-\tau_1$. Accordingly, we have

$$f(G,t|S_k,\tau_0,\tau_1,\theta=T^*)=\left(\frac{2}{2t_1+t_0-2\tau_0-\tau_1}\right)^2 e^{-2}.$$

Because of the constraint $\tau_1\le\tau_0\le t_1\le t_0$, it can be seen that the maximum is achieved when $\tau_0=\tau_1=t_1$. Hence, we have

$$\begin{aligned}&\max\left\{\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))}\middle|\tau_1\le\tau_0\le t_1\le t_0,\theta>0\right\}\\&=f(G,t|S_k,\tau_0=\mathrm{t}_1,\tau_1=\mathrm{t}_1,\theta=T^*)\\&=\left(\frac{2}{t_0-t_1}\right)^2 e^{-2},\end{aligned}$$

where $k=1,2,3$.

b)

As to species tree S_0,

$$\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}T_0}=\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((t_1-\tau_1)+(t_0-\tau_0))},$$

s.t. $\tau_1\le t_1\le\tau_0\le t_0,\theta>0$.

In other words, we are looking for

$$(\hat{\tau}_0, \hat{\tau}_1, \hat{\theta}) = \underset{\tau_0,\tau_1,\theta}{\operatorname{argmax}} \left\{ \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((t_1-\tau_1)+(t_0-\tau_0))} | \tau_1 \le t_1 \le \tau_0 \le t_0, \theta > 0 \right\}.$$

According to a), it is already known that the maximum of $f(x)$ is achieved when $\hat{\theta} = T^* = (t_1 - \tau_1) + (t_0 - \tau_0)$. Considering the constraint $\tau_1 \le t_1 \le \tau_0 \le t_0$, it is obvious that setting $\tau_0 = t_0, \tau_1 = t_1, \theta = (t_1 - \tau_1) + (t_0 - \tau_0) = 0$, the maximum of the likelihood function is achieved at $\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((t_1-t_1)+(t_0-t_0))} = \left(\frac{2}{\theta}\right)^2 \to \infty$.

Based on a) and b), when $S = S_0, \tau_0 = t_0, \tau_1 = t_1, \theta = (t_1 - \tau_1) + (t_0 - \tau_0) = 0$, the maximum of the likelihood is achieved.

9.4* *Singularity in the likelihood function in estimation of the species tree given the gene tree.* As in Problem 9.3, use the gene tree G with node ages t_0 and t_1 of Figure 9.22a to evaluate the likelihood for the species trees of Figure 9.22b, but allowing different population size parameters θ_0 and θ_1. Show that the likelihood for species tree S_0 can become infinite when $\tau_1 \to t_1, \theta_1 \to 0$, with $\tau_0 > \tau_1$ and $\theta_0 > 0$. In this particular case, species trees S_1–S_3 do not show singularity in the likelihood function, but in the general case of more than three species and sequences, multiple species trees can have infinite likelihood. Also convince yourself that singularity cannot occur if the likelihood is calculated using the sequence alignment (i.e. Yang 2002b, Equation 8) rather than the coalescent times on a gene tree.

Solution.

a)

According to the problem statement, the likelihood given the species tree S_0 is given as

$$f(G, t|S_0, \Theta) = \frac{2}{\theta_0}\frac{2}{\theta_1} e^{-\frac{2}{\theta_1}(t_1-\tau_1)-\frac{2}{\theta_0}(t_0-\tau_0)}$$
$$= \frac{2}{\theta_0} e^{-\frac{2}{\theta_0}(t_0-\tau_0)} \times \frac{2}{\theta_1} e^{-\frac{2}{\theta_1}(t_1-\tau_1)}.$$

According to the problem statement, let $\tau_1 \to t_1$ and $\theta_1 = k(t_1 - \tau_1) \to 0$. Thus, when $\tau_1 \to t_1$, we have

$$f(G, t|S_0, \Theta) = \frac{2}{\theta_0} e^{-\frac{2}{\theta_0}(t_0-\tau_0)} \times \frac{2}{\theta_1} e^{-\frac{2}{k}}$$

which approaches infinite.

b)

For the three wrong species trees S_1, S_2, S_3, the likelihood is given as

$$f(G, t|S_1, \Theta) = f(G, t|S_2, \Theta) = f(G, t|S_3, \Theta)$$
$$= e^{-\frac{2}{\theta_1}(\tau_0-\tau_1)} \times \frac{2}{\theta_0} e^{-\frac{2}{\theta_0}\times 3(t_1-\tau_0)} \times \frac{2}{\theta_0} e^{-\frac{2}{\theta_0}(t_0-t_1)}$$

$$= e^{-\frac{2}{\theta_1}(\tau_0-\tau_1)} \times \left(\frac{2}{\theta_0}\right)^2 e^{-\frac{2}{\theta_0}(3t_1-3\tau_0+t_0-t_1)}.$$

Because $t_1 \geq \tau_0$, it is easy to see that the expression $3t_1 - 3\tau_0 + t_0 - t_1$ is bounded below $t_0 - t_1$. It is not difficult to realize that the maximum of $f(G, t|S_1, \Theta)$ is achieved when its first term $f(G, t|S_1, \Theta)$ and second term $\left(\frac{2}{\theta_0}\right)^2 e^{-\frac{2}{\theta_0}(3t_1-3\tau_0+t_0-t_1)}$ both independently reach its maximum. The following show that this maximum can be achieved.

Because $\theta_1 > 0$, the first term of $f(G, t|S_1, \Theta)$, which is $e^{-\frac{2}{\theta_1}(\tau_0-\tau_1)}$, reaches the maximum whenever $\tau_0 = \tau_1$. As to the second term $\left(\frac{2}{\theta_0}\right)^2 e^{-\frac{2}{\theta_0}(3t_1-3\tau_0+t_0-t_1)}$ which does not involve the unknown τ_1, its log-likelihood can be calculated as

$$\ell = \log\left(\left(\frac{2}{\theta_0}\right)^2 e^{-\frac{2}{\theta_0}(3t_1-3\tau_0+t_0-t_1)}\right)$$
$$= 2\log\left(\frac{2}{\theta_0}\right) - \frac{2}{\theta_0}(3t_1 - 3\tau_0 + t_0 - t_1).$$

Noting that τ_0 has to meet the requirement $\tau_0 \leq t_1$, it is not difficult to realize that the maximum of ℓ is reached when $\tau_0 = t_1$, as also evident in Problem 9.3 and Section 9.4.3.2 of (Yang 2014a).

Accordingly, by setting it to zero and calculating the derivative w.r.t to θ_0, we obtain

$$\hat{\theta}_0 = 3t_1 - 3\hat{\tau}_0 + t_0 - t_1$$
$$= t_0 - t_1.$$

Hence, the maximum of the likelihood function is given as

$$f(G, t|S_1, \tau_0 = \tau_1 = t_1, \theta_0 = t_0 - t_1) = e^{-\frac{2}{\theta_1}(\tau_0-\tau_1)} \times \left(\frac{2}{\theta_0}\right)^2 e^{-\frac{2}{\theta_0}(t_0-t_1)}$$
$$= 1 \times \left(\frac{2}{t_0 - t_1}\right)^2 e^{-\frac{2}{t_0-t_1}(t_0-t_1)}$$
$$= \left(\frac{2}{t_0 - t_1}\right)^2 e^{-2}.$$

Thus, the singularity problem does not exist for the three wrong three-species trees S_1, S_2, S_3.

9.5 Examine the probability densities $f_0(t)$ and $f_1(t)$ of Figure 9.15 for the divergence time t between two sequences at a locus under the symmetrical IM model for two species (SIM2s). Derive the densities $f_0(t)$ and $f_1(t)$ for the limiting cases of $M = 0$ and ∞.

Solutions.

We provide as follows two solutions to b) and c), the first being the most straightforward solution using eigendecomposition, the other using a clever way to simplify the calculation.

Solution 1

a)

Referring back to Eq. (9.50) of (Yang 2014a), the PDF of t conditioned on the parameters Θ is given by

$$f_0(t|\Theta) = \begin{cases} P_{aS_{11}}(t) \times \frac{2}{\theta_1} + P_{aS_{22}(t)} \times \frac{2}{\theta_2}, t < \tau \\ \left(P_{aS_{11}}(\tau) + P_{aS_{12}}(\tau) + P_{aS_{22}}(\tau)\right) \times \frac{2}{\theta_a} e^{-\frac{2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases}, \quad (9.1)$$

where the initial state a is either S_{11} or S_{22} or in other words the two sequences are sampled from the same species, and

$$f_1(t|\Theta) = \begin{cases} P_{aS_{11}}(t) \times \frac{2}{\theta_1} + P_{aS_{22}(t)} \times \frac{2}{\theta_2}, t < \tau \\ \left(P_{aS_{11}}(\tau) + P_{aS_{12}}(\tau) + P_{aS_{22}}(\tau)\right) \times \frac{2}{\theta_a} e^{-\frac{2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases}, \quad (9.2)$$

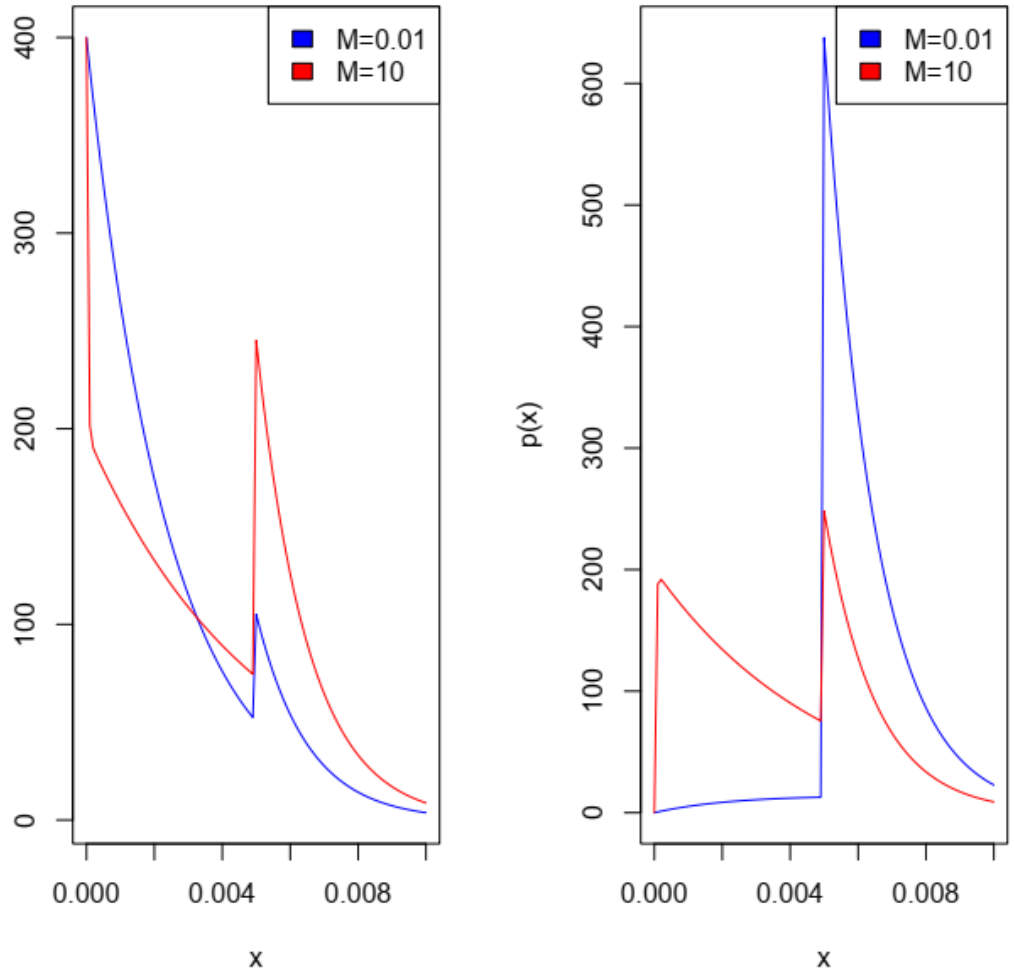
where the initial state a is S_{12} or in other words the two sequences are sampled from two species.

Hence, the R code to calculate and plot the PDF of $f_0(t|\Theta)$ and $f_1(t|\Theta)$ given the parameters $\theta = 0.005, \theta_a = 0.003, \tau = 0.005$ is shown as follows. The results of only $M = 0.01$ and $M = 10$ are shown. Note that m is determined based on $m = \frac{4M}{\theta}$ [see Section 9.4.4.2 of (Yang 2014a)].

R (9.5.R)

```
#! /usr/bin/env Rscript
library(expm)
g <- function(m,theta){
   matrix(c(-2*m-2/theta,m,0,0,0, 2*m,-2*m,2*m,0,0, 0,m,-2*m-2/theta,0,0, 2/theta,0,0,-m,m,
0,0,2/theta,m,-m), ncol=5)
}
h0 <- function(t){
   theta<-0.005; m<-4*M/theta
   sum(sapply(c(1,3), function(x){ init * 2/theta * expm(g(m=m,theta=theta)*t)[1:3,x]}))
}
h1 <- function(t) {
   theta <- 0.005; theta_a <- 0.003; m <- 4*M/theta; m_a <- 4*M/theta_a
   dexp(t-tau, 2/theta_a) * sum(sapply(c(1,2,3), function(x){ init *
expm(g(m=m,theta=theta)*tau)[1:3,x]}))
}
h <- function(t, tau=0.005) if(t<tau){h0(t)} else{h1(t)}


##############################################
```

```
pdf("9.5.pdf")
par(mfrow=c(1,2)); tau <- 0.005
inits <- data.frame(c(0.5,0,0.5), c(0,1,0)) # (0.5,0,0.5) -> f0, (0,1,0) -> f1
for(i in 1:ncol(inits)){
    init <- inits[,i]
    p <- Vectorize(h)
    M <- 0.01
    curve(p, from=0, to=2*tau, col="blue")
    integrate(p, lower=0, upper=1)
    M <- 100000
    curve(p, from=0, to=2*tau, add=T, col="red")
    legend(x = "topright", legend=c("M=0.01", "M=10"),
           fill = c("blue","red")
    )
}
dev.off()
```



b)

Consider the case where $m = 0$. According to the problem statement, the rate matrix of the IM model is defined as

$$Q = \begin{bmatrix} -2(m+\frac{1}{\theta}) & 2m & 0 & \frac{2}{\theta} & 0 \\ m & -2m & m & 0 & 0 \\ 0 & 2m & -2(m+\frac{1}{\theta}) & 0 & \frac{2}{\theta} \\ 0 & 0 & 0 & -m & m \\ 0 & 0 & 0 & m & -m \end{bmatrix}.$$

We perform the eigendecomposition of Q as follows

$$Q(m=0) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{2}{\theta} & 0 \\ 0 & 0 & 0 & 0 & -\frac{2}{\theta} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}.$$

Therefore, the transition probability matrix $P(t)$ is given as

$$P(t, m=0) = U \times e^{\text{diag}\{0,0,0,-\frac{2}{\theta},-\frac{2}{\theta}\}t} \times U^{-1}$$

$$= \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} e^0 & 0 & 0 & 0 & 0 \\ 0 & e^0 & 0 & 0 & 0 \\ 0 & 0 & e^0 & 0 & 0 \\ 0 & 0 & 0 & e^{-\frac{2}{\theta}t} & 0 \\ 0 & 0 & 0 & 0 & e^{-\frac{2}{\theta}t} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} e^{-\frac{2}{\theta}t} & 0 & 0 & 1 - e^{-\frac{2}{\theta}t} & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & e^{-\frac{2}{\theta}t} & 0 & 1 - e^{-\frac{2}{\theta}t} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

We denote the probability that the initial state is S_{11} as $p_1(0)$ and the probability that the initial state is S_{22} as $p_3(0)$. When the initial state is either S_{11} or S_{22}, we have

$$P_{aS_{11}}(t) = [p_1(0) \quad p_3(0)] \begin{bmatrix} p_{11}(t) \\ p_{31}(t) \end{bmatrix} = p_1(0) \times \frac{1}{2} e^{-\frac{2}{\theta}t},$$

$$P_{aS_{12}}(t) = [p_1(0) \quad p_3(0)] \begin{bmatrix} p_{12}(t) \\ p_{32}(t) \end{bmatrix} = 0,$$

$$P_{aS_{22}}(t) = [p_1(0) \quad p_3(0)] \begin{bmatrix} p_{13}(t) \\ p_{33}(t) \end{bmatrix} = p_3(0) \times \frac{1}{2} e^{-\frac{2}{\theta}t}.$$

According to the context, $p_1(0) + p_3(0) = 1$. Hence, according to Eq. (9.1), we obtain

$$f_0(t|\mathrm{m} = 0, \theta, \theta_a) = \begin{cases} \frac{2}{\theta} e^{-\frac{2}{\theta}t}, t < \tau \\ e^{-\frac{2}{\theta}\tau} \times \frac{2}{\theta_a} e^{\frac{-2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases}$$

which can further be simplified as

$$f_0(t|\mathrm{m} = 0, \theta, \theta_a) = \frac{2}{\theta} e^{-\frac{2}{\theta}t}. \tag{9.3}$$

Clearly, $f_0(t|\mathrm{m} = 0, \theta, \theta_a)$ specifies an exponential distribution with rate of $\frac{2}{\theta}$.

Then, denote the probability that the initial state is S_{12} as $p_2(0)$. When the initial state is S_{12}, $p_2(0) = 1$. Thus, it is easy to see the following

$$P_{S_{12}S_{11}}(t) = 0,$$

$$P_{S_{12}S_{12}}(t) = 1,$$

$$P_{S_{12}S_{22}}(t) = 0.$$

So according to Eq. (9.2),

$$f_1(t|\mathrm{m} = 0, \theta, \theta_a) = \begin{cases} 0, t < \tau \\ \frac{2}{\theta_a} e^{\frac{-2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases}. \tag{9.4}$$

which specifies an exponential distribution with the rate parameter $\frac{2}{\theta_a}$ and setoff at τ.

c)

According to the context, $m \to \infty$.

We need to first derive the transition probability matrix $P(t)$. The rate matrix Q may be re-written as a block matrix

$$Q = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where $A = \begin{bmatrix} -2\left(m+\frac{1}{\theta}\right) & 2m & 0 \\ m & -2m & m \\ 0 & 2m & -2\left(m+\frac{1}{\theta}\right) \end{bmatrix}, B = \begin{bmatrix} \frac{2}{\theta} & 0 & 0 \\ 0 & 0 & \frac{2}{\theta} \end{bmatrix}^T, C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D = \begin{bmatrix} -m & m \\ m & -m \end{bmatrix}$.

To calculate the eigenvalue λ of Q, we set the following

$$|Q - \lambda| = |A - \lambda||D - \lambda|$$

to zero.

Further, it can be calculated

$$|D - \lambda| = \begin{vmatrix} -m-\lambda & m \\ m & -m-\lambda \end{vmatrix} = (-m-\lambda)^2 - m^2,$$

$$|A - \lambda| = \begin{vmatrix} -2\left(m+\frac{1}{\theta}\right)-\lambda & 2m & 0 \\ m & -2m-\lambda & m \\ 0 & 2m & -2\left(m+\frac{1}{\theta}\right)-\lambda \end{vmatrix}$$

$$= \begin{vmatrix} -2\left(m+\frac{1}{\theta}\right)-\lambda & 2m & 0 \\ 2m & -2m-\lambda & m \\ 0 & 0 & -2\left(m+\frac{1}{\theta}\right)-\lambda \end{vmatrix}$$

$$= -\left(2\left(m+\frac{1}{\theta}\right)+\lambda\right)\left(\left(2m+\lambda+\frac{1}{\theta}\right)^2 - \left(\left(\frac{1}{\theta}\right)^2 + 2m^2\right)\right).$$

Accordingly, the five eigenvalues should be the roots of any of the following three equations

$$-2\left(m+\frac{1}{\theta}\right)+\lambda = 0,$$

$$2m+\lambda+\frac{1}{\theta} = \pm\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2},$$

$$(-m-\lambda)^2 - m^2 = 0.$$

Hence,

$$\lambda_1 = -2\left(m+\frac{1}{\theta}\right),$$

$$\lambda_2 = \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - 2m - \frac{1}{\theta},$$

$$\lambda_3 = -\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - 2m - \frac{1}{\theta},$$

$$\lambda_4 = 0,$$

$$\lambda_5 = -2m.$$

Accordingly, the eigenvalues are

$$v1 = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 1 \\ \frac{1}{\theta} + \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} \\ \hline 2m \\ 1 \\ 0 \\ 0 \end{bmatrix}, v_3 = \begin{bmatrix} 1 \\ \frac{1}{\theta} - \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} \\ \hline 2m \\ 1 \\ 0 \\ 0 \end{bmatrix} v_4 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, v_5 = \begin{bmatrix} -1 \\ 0 \\ 1 \\ -1 \\ 1 \end{bmatrix}.$$

Hence,

$$U = \begin{bmatrix} -1 & 1 & 1 & 1 & -1 \\ 0 & \frac{\frac{1}{\theta} + \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2}}{2m} & \frac{\frac{1}{\theta} - \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2}}{2m} & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

Define

$$\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} = a \tag{9.5}$$

Rewrite U as a block matrix

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix},$$

where $U_{11} = \begin{bmatrix} -1 & 1 & 1 \\ 0 & \frac{a+\frac{1}{\theta}}{2m} & \frac{\frac{1}{\theta}-a}{2m} \\ 1 & 1 & 1 \end{bmatrix}, U_{12} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}, U_{21} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, U_{22} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$

Hence,

$$U_{11}^{-1} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{a\theta - 1}{4at} & \frac{m}{a} & \frac{a\theta - 1}{4at} \\ \frac{a\theta + 1}{4at} & -\frac{m}{a} & \frac{a\theta + 1}{4at} \end{bmatrix}, U_{22}^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

and accordingly,

$$-U_{11}^{-1}U_{12}U_{22}^{-1} = -\begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{a\theta - 1}{4at} & \frac{m}{a} & \frac{a\theta - 1}{4at} \\ \frac{a\theta + 1}{4at} & -\frac{m}{a} & \frac{a\theta + 1}{4at} \end{bmatrix} \times \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \times \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{-a\theta - 2m\theta + 1}{4a\theta} & \frac{-a\theta - 2m\theta + 1}{4a\theta} \\ \frac{-a\theta + 2m\theta - 1}{4a\theta} & \frac{-a\theta + 2m\theta - 1}{4a\theta} \end{bmatrix}.$$

Hence,

$$U^{-1} = \begin{bmatrix} U_{11}^{-1} + U_{11}^{-1}U_{12}(U_{22} - U_{21}U_{11}^{-1}U_{12})^{-1}U_{21}U_{11}^{-1} & -U_{11}^{-1}U_{12}(U_{22} - U_{21}U_{11}^{-1}U_{12})^{-1} \\ -(U_{22} - U_{21}U_{11}^{-1}U_{12}) & (U_{22} - U_{21}U_{11}^{-1}U_{12})^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} U_{11}^{-1} & -U_{11}^{-1}U_{12}U_{22}^{-1} \\ \mathbf{0} & U_{22}^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{a\theta - 1}{4a\theta} & \frac{m}{a} & \frac{a\theta - 1}{4a\theta} & \frac{-a\theta - 2m\theta + 1}{4a\theta} & \frac{-a\theta - 2m\theta + 1}{4a\theta} \\ \frac{a\theta + 1}{4a\theta} & -\frac{m}{a} & \frac{a\theta + 1}{4a\theta} & \frac{-a\theta + 2m\theta - 1}{4a\theta} & \frac{-a\theta + 2m\theta - 1}{4a\theta} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Applying eigendecomposition, we obtain

$$P(t) = e^{Qt} = Ue^{\Lambda t}U^{-1},$$

where $\Lambda = \begin{bmatrix} -2\left(m + \frac{1}{\theta}\right) & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - 2m - \frac{1}{\theta} & 0 & 0 & 0 \\ 0 & 0 & -\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - 2m - \frac{1}{\theta} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2m \end{bmatrix}$, and t indicates the time.

Define the following

$$\begin{cases} x = e^{-2\left(m + \frac{1}{\theta}\right)t} \\ y = e^{\left(\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - 2m - \frac{1}{\theta}\right)t} \\ z = e^{\left(-\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - 2m - \frac{1}{\theta}\right)t} \\ w = e^{-2mt} \end{cases} \tag{9.6}$$

According to the context or Eqs. (9.1-9.2), to calculate $f(t|\Theta)$, we only need to know the following submatrix of $P(t)$

$$P_{3\times3}(t) = \begin{bmatrix} p_{11}(t) & p_{12}(t) & p_{13}(t) \\ p_{21}(t) & p_{22}(t) & p_{23}(t) \\ p_{31}(t) & p_{32}(t) & p_{33}(t) \end{bmatrix},$$

which can be analytically calculated to be

$$\begin{bmatrix} \frac{-y+z+2ax\theta+ay\theta+az\theta}{4a\theta} - \frac{y-z}{4a\theta} & \frac{m(y-z)}{a} & \frac{-y+z-2ax\theta+ay\theta+az\theta}{4a\theta} \\ \frac{-y+z+a^2y\theta^2-a^2z\theta^2}{4a\theta} & \frac{y-z+ay\theta+az\theta}{2a\theta} & \frac{-y+z+a^2y\theta^2-a^2z\theta^2}{4a\theta} \\ \frac{-y+z-2ax\theta+ay\theta+az\theta}{4a\theta} & \frac{m(y-z)}{a} & \frac{-y+z+2ax\theta+ay\theta+az\theta}{4a\theta} \end{bmatrix}$$

where $a = \sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2}$ as defined in Eq. (9.5), and x, y, z, w are defined in Eq. (9.6).

Now, we need to calculate the limit of $p_{ij}(t)$ one by one, but before that, for simplicity we need to calculate the value of x, y, z as m approaches the infinite as follows

$$\lim_{m\to\infty} x = \lim_{m\to\infty} e^{-2\left(m+\frac{1}{\theta}\right)t} = 0,$$

$$\lim_{m\to\infty} z = \lim_{m\to\infty} e^{\left(-\sqrt{\left(\frac{1}{\theta}\right)^2+4m^2}-2m-\frac{1}{\theta}\right)t} = 0,$$

$$\lim_{m\to\infty} y = e^{t\times\lim_{m\to\infty}\left(\sqrt{\left(\frac{1}{\theta}\right)^2+4m^2}-\left(2m+\frac{1}{\theta}\right)\right)}$$

$$= e^{t\times\lim_{m\to\infty}\left(\frac{\left(\frac{1}{\theta}\right)^2+4m^2-\left(2m+\frac{1}{\theta}\right)^2}{\sqrt{\left(\frac{1}{\theta}\right)^2+4m^2}+\left(2m+\frac{1}{\theta}\right)}\right)}$$

$$= e^{t\times\lim_{m\to\infty}\left(-\frac{4}{\theta}\times\frac{1}{\sqrt{\left(\frac{1}{\theta m}\right)^2+4}+\left(2+\frac{1}{\theta m}\right)}\right)}$$

$$= e^{-\frac{t}{\theta}}.$$

For the calculation of $\lim_{m\to\infty} y$ above, note that $\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2} - \left(2m + \frac{1}{\theta}\right)$ is in the form of $\infty - \infty$.

Hence,

$$\lim_{m\to\infty} p_{32}(t) = \lim_{m\to\infty} p_{12}(t) = \lim_{m\to\infty} \frac{m(y-z)}{a}$$

$$= \lim_{m\to\infty} (y-z) \times \lim_{m\to\infty} \frac{m}{\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2}}$$

$$= \frac{1}{2} e^{-\frac{t}{\theta}}.$$

As to $p_{22}(t)$, we calculate it as

$$\lim_{m\to\infty} p_{22}(t) = \lim_{m\to\infty} \frac{y-z+ay\theta+az\theta}{2a\theta}$$

$$= \lim_{m\to\infty} \frac{y-z}{2\theta\sqrt{\left(\frac{1}{\theta}\right)^2 + 4m^2}} + \lim_{m\to\infty} \frac{a\theta(y+z)}{2a\theta}$$

$$= \frac{1}{2}\lim_{m\to\infty}(y+z)$$
$$= \frac{1}{2}e^{-\frac{t}{\theta}}.$$

As to $p_{21}(t), p_{23}(t)$, we have

$$\lim_{m\to\infty} p_{21}(t) = \lim_{m\to\infty} p_{23}(t) = \lim_{m\to\infty} \frac{(a^2\theta^2-1)(y-z)}{8am\theta^2}$$
$$= \lim_{m\to\infty} \frac{4m^2\theta^2}{8\sqrt{4m^2+\frac{1}{\theta^2}m\theta^2}}(y-z)$$
$$= \frac{1}{4}e^{-\frac{t}{\theta}}.$$

Likewise, we have

$$\lim_{m\to\infty} p_{21}(t) = \lim_{m\to\infty} p_{23}(t) = \frac{1}{4}\times \lim_{m\to\infty}(y-z) = \frac{1}{4}e^{-\frac{t}{\theta}},$$

and

$$\lim_{m\to\infty} p_{11}(t) = \lim_{m\to\infty} p_{33}(t) = \lim_{m\to\infty}\left(\frac{a\theta(2x+y+z)}{4a\theta} - \frac{y-z}{4a\theta}\right)$$
$$= \lim_{m\to\infty}\frac{1}{4}(2x+y+z) - \frac{\lim_{m\to\infty}(y-z)}{\lim_{m\to\infty}4a\theta}$$
$$= \frac{1}{4}e^{-\frac{t}{\theta}}.$$

Further, noting $p_{31}(t) = p_{13}(t) = p_{11}(t) - x$, and $\lim_{m\to 0} x = 0$ (see above), we obtain

$$\lim_{m\to 0} p_{31}(t) = \lim_{m\to 0} p_{31}(t) = \lim_{m\to 0}(p_{11}(t)-x) = \frac{1}{4}e^{-\frac{t}{\theta}}.$$

Hence, we have

$$P_{3\times 3}(t, m\to\infty) = \begin{bmatrix} p_{11}(t) & p_{12}(t) & p_{13}(t) \\ p_{21}(t) & p_{22}(t) & p_{23}(t) \\ p_{31}(t) & p_{32}(t) & p_{33}(t) \end{bmatrix}$$
$$= e^{-\frac{t}{\theta}} \times \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}.$$

According to the context, when the initial state is either S_{11} or S_{22}, thus $p_1(0) + p_3(0) = 1$, we have

$$P_{aS_{11}}(t) = [p_1(0) \quad p_3(0)]\begin{bmatrix} p_{11}(t) \\ p_{31}(t) \end{bmatrix} = \frac{1}{4}e^{-\frac{1}{\theta}t},$$
$$P_{aS_{12}}(t) = [p_1(0) \quad p_3(0)]\begin{bmatrix} p_{12}(t) \\ p_{32}(t) \end{bmatrix} = \frac{1}{2}e^{-\frac{1}{\theta}t},$$

$$P_{aS_{22}}(t) = [p_1(0) \quad p_3(0)] \begin{bmatrix} p_{13}(t) \\ p_{33}(t) \end{bmatrix} = \frac{1}{4} e^{-\frac{1}{\theta}t}.$$

According to Eq. (9.1), it thus follows that

$$f_0(t|\mathrm{m} \to \infty, \theta, \theta_a) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta}t}, t < \tau \\ e^{-\frac{1}{\theta}\tau} \times \frac{2}{\theta_a} e^{\frac{-2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases}. \quad (9.7)$$

When the initial state is either S_{12} thus $p_2(0) = 1$, we have

$$P_{aS_{11}}(t) = \frac{1}{4} e^{-\frac{1}{\theta}t},$$

$$P_{aS_{12}}(t) = \frac{1}{2} e^{-\frac{1}{\theta}t},$$

$$P_{aS_{22}}(t) = \frac{1}{4} e^{-\frac{1}{\theta}t}.$$

According to Eq. (9.2), it thus follows that

$$f_1(t|\mathrm{m} \to \infty, \theta, \theta_a) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta}t}, t < \tau \\ e^{-\frac{1}{\theta}\tau} \times \frac{2}{\theta_a} e^{\frac{-2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases}. \quad (9.8)$$

Solution 2.

b)

When $m = 0$, the rate matrix is given as

$$Q(m = 0) = \begin{bmatrix} -\frac{2}{\theta} & 0 & 0 & \frac{2}{\theta} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{2}{\theta} & 0 & \frac{2}{\theta} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Suppose that the initial state is S_{11}. Then, it either visits the absorbing state S_1 stays in the initial state S_{11}, and the time it visits S_1 follows $\exp\left(\frac{2}{\theta}\right)$. The same applies to the case where the initial state is S_{22}. Hence,

$$f_0(t|\mathrm{m} = 0, \theta, \theta_a) = \frac{2}{\theta} e^{-\frac{2}{\theta}t},$$

which is exactly the same as what we obtain in Eq. (9.3) in Solution 1.

Likewise, we obtain the time of coalescence when the initial state is S_{22} as an exponential distribution with a setoff at τ as

$$f_1(t|\mathrm{m} = 0, \theta, \theta_a) = \begin{cases} 0, t < \tau \\ \frac{2}{\theta_a} e^{\frac{-2}{\theta_a}(t-\tau)}, t \geq \tau' \end{cases}$$

the same as what we obtain in Eq. (9.4) in Solution 1.

c)

Consider the case $m \to \infty$. That said, no matter which initial state it is, the Markov chain with the following transition rate matrix $\begin{bmatrix} -2m & 2m & 0 \\ m & -2m & m \\ 0 & 2m & -2m \end{bmatrix}$ immediately reaches equilibrium as soon as the chain starts. Hence, it is easy to obtain that the equilibrium frequency is

$$\left(\pi_{S_{11}}, \pi_{S_{12}}, \pi_{S_{12}}\right) = (0.25, 0.5, 0.25)$$

by noting the following

$$[\pi_{S_{11}} \quad \pi_{S_{12}} \quad \pi_{S_{12}}] \begin{bmatrix} -2m & 2m & 0 \\ m & -2m & m \\ 0 & 2m & -2m \end{bmatrix} = \mathbf{0},$$

$$\pi_{S_{11}} + \pi_{S_{12}} + \pi_{S_{22}} = 1.$$

After the species divergence time τ, because i) the transition rates from S_{11} to S_{22} i.e., q_{13} and from S_{22} to S_{11} i.e., q_{31} are both zero, and ii) the transition rate from S_{11} or S_{22} to S_{12} i.e., q_{12}, q_{32} is half of the opposite, i.e., q_{21}, q_{23}, it can be inferred that rate it visits S_1 or S_2 will be decreased from $\frac{2}{\theta}$ to $\frac{1}{\theta}$. Hence, before the species divergence time τ, the time it coalesces follows an exponential distribution Exp($\frac{2}{\theta}$).

Hence, it is not difficult to see the following: $f_0(t|\mathrm{m} = \infty, \theta, \theta_a)$ and $f_0(t|\mathrm{m} = \infty, \theta, \theta_a)$ are the same and they should be the PDF of a truncated exponential distribution with rate of $\frac{1}{\theta}$ after time τ, and another exponential distribution with rate of $\frac{2}{\theta_a}$ and with a setoff at τ and a scaling factor $e^{-\frac{1}{\theta}\tau} = 1 - \int_0^\tau \frac{1}{\theta} e^{-\frac{1}{\theta}t} dt$ (the probability that no coalescence occurs after species split) before time τ.

That said, the PDF of $f_0(t|\mathrm{m} = \infty, \theta, \theta_a)$ and $f_0(t|\mathrm{m} = \infty, \theta, \theta_a)$ are given as

$$f_0(t|\mathrm{m} = 0, \theta, \theta_a) = f_1(t|\mathrm{m} \to \infty, \theta, \theta_a) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta}t}, t < \tau \\ e^{-\frac{1}{\theta}\tau} \times \frac{2}{\theta_a} e^{\frac{-2}{\theta_a}(t-\tau)}, t \geq \tau \end{cases},$$

which is exactly the same as what we obtain in Eqs. (9.7-9.8).

9.6 Use the four nuclear autosomal loci from the two species/populations of butterflies (*Heliconius demeter* and *H. eratosignis*) to delimit species (Dasmahapatra et al. 2010; Zhang et al. 2011). The four loci are Mpi (9 sequences, 496 bp), Tektin (9 sequences, 733 bp), Rp15 (15 sequences, 713 bp), and Ef1a (18 sequences, 766 bp). Examine the effect of the number of loci by analysing 1, 2, 3, or 4 loci. Use different priors on τs and θs to run BPP to evaluate the impact of the prior.

Solution.

I use BP&P v3.4a (Yang 2015) for this analysis. Note that for species delimitation, users should set speciesdelimitation = 1 and speciestree = 0 in the control file, thus the mode "A10". The results of using one to four loci are displayed as follows where the posterior probabilities of the two models are given. Note that Model 0 denotes the hypothesis that all belong to the same species while Model 1 denotes the hypothesis that the species delimitation should be as indicated in the "Imap" file with species D and E.

Number of loci	**Model 0**	**Model 1**
1	0	1.0
2	0.15510	0.84490
3	1.0	0
4	0.06150	0.93850

Chapter 10. Molecular clock and estimation of species divergence times

10.1 Use the mitochondrial 12S rRNA genes from the apes (Horai et al. 1995) of Example 10.1 to test the molecular clock hypothesis under different substitution models to examine the sensitivity of the test to the assumed model. Use the following models to conduct the LRT: JC69, HKY85, GTR, JC69 + Γ_5, HKY85 + Γ_5, and GTR + Γ_5. Note that the phylogeny is shown in Figure 10.2.

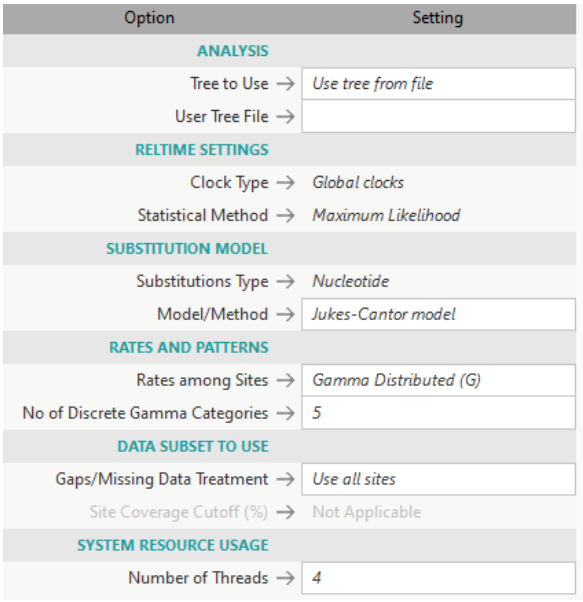
Solution.

Download the mitochondrial genome sequences of *Homo sapiens*, *Pan troglodytes*, *Gorrila gorilla*, *Pongo pygmaeus*, and *Pan paniscus* from NCBI Genbank according to the accessions indicated in the study (Horai et al. 1995), and have the 12s rRNA sequences extracted. Perform a sequence alignment using your favourite software, and name the file as *Horai1995.aln*.

Write to the file *Horai1995.nwk* the following Newick-formatted tree:

```
(Pongo_pygmaeus,(Homo_sapiens,((Pan_troglodytes,Pan_paniscus),Gorrila_gorrila)));
```

I choose to use MEGA (Kumar et al. 2018) to perform the molecular test.



Option	Setting
ANALYSIS	
Tree to Use →	*Use tree from file*
User Tree File →	
RELTIME SETTINGS	
Clock Type →	*Global clocks*
Statistical Method →	*Maximum Likelihood*
SUBSTITUTION MODEL	
Substitutions Type →	*Nucleotide*
Model/Method →	*Jukes-Cantor model*
RATES AND PATTERNS	
Rates among Sites →	*Gamma Distributed (G)*
No of Discrete Gamma Categories →	*5*
DATA SUBSET TO USE	
Gaps/Missing Data Treatment →	*Use all sites*
Site Coverage Cutoff (%) →	Not Applicable
SYSTEM RESOURCE USAGE	
Number of Threads →	*4*

For the model GTR, remember to choose the "CUSTOM" model and then input "012345". The results are displayed as follows.

Model	**lnL without clock**	**lnL with clock**	***P*-value**
JC69	-2073.96	-2076.89	0.12
HKY85	-1945.50	-1948.03	0.13
GTR	-1931.43	-1933.83	0.19
JC69+G5	-2064.74	-2067.14	0.19
HKY85+G5	-1930.118	-1931.544	0.42
GTR+G5	-1921.48	-1922.98	0.39

Note that there are $(2 \times 5 - 3) - 4 = 3$ constrained parameters when a global clock is assumed hence a degree of freedom of 3. Use "pchisq(2*delta_lnL,3,lower.tail=F)" in R to obtain the *P*-value.

Model	**lnL without clock**	**lnL with clock**	***P*-value**
JC69	-2073.96	-2089.85	5.82e-07
HKY85	-1945.50	-1961.55	4.97e-07
GTR	-1931.43	-1947.75	3.85e-07

10.2 Run a Bayesian MCMC program (such as MCMCTREE) to estimate the species divergence times using the data of Steiper et al. (2004) (see §10.4.7.1). Treat the five loci as independent loci with different rate trajectories. Compare the results with those of Table 10.1, which were obtained by concatenating the five loci as one partition. Are the CIs any narrower? Note that the tree is shown in Figure 10.7.

Solution.

See the data set Steiper2004 from C10 in (Yang 2014b). We take the global clock as an example. Open *mcmctree.clock2.ctl*, the control file for MCMCTree.

1. HKY+G5: use HKY with five rate categories of discrete gamma distribution. In Line 5, change to the alignment file separated as five loci *Steiper2004ABCDE.txt*. In Line 8, set *ndata* = 5. Run MCMCTree.
2. JC69. On top of *mcmctree.clock2.ctl*, edit Line 13 to change model = 0 and Line 14 to set alpha = 0. Run MCMCTree.

The results, as compared to those shown in Table 10.1 in (Yang 2014a), are summarized as follows (STR: strict rate or global rate; IR: independent rate; AR: autocorrelated rate). The posteriors estimated by treating the five genes as independent loci should yield narrower CI intervals under independent or correlated rate models (but not the global clock because for global clock applying independent loci does not change any parameter).

		Concatenated alignment			**5 independent loci**		
	Prior	STR	IR	AR	STR	IR	AR
JC69							
t(root)	34 (8, 60)	32 (28, 37)	33 (24, 45)	32 (25, 40)	32 (28, 37)	33 (27, 40)	33 (27, 40)
t(ape)	6.5 (5, 8)	5.7 (5.0, 6.6)	5.9 (5.0, 7.6)	6.0 (5.0, 7.7)	5.7 (4.9, 6.6)	5.7 (4.9, 7.1)	5.9 (4.9, 7.5)
t(monkey)	6.5 (5, 8)	7.1 (6.1, 8.0)	7.0 (5.3, 8.0)	6.1, 8.1	7.1 (6.0, 8.0)	7.1 (5.8, 8.1)	7.1 (5.6, 8.0)
HKY+G5							
t(root)	34 (8, 60)	33 (29, 38)	34 (24, 46)	33 (26, 41)	33 (28, 37)	34 (28, 41)	34 (27, 41)
t(ape)	6.5 (5, 8)	5.7 (4.9, 6.6)	5.9 (5.0, 7.7)	6.0 (5.0, 7.8)	5.6 (4.9, 6.5)	5.7 (4.9, 7.1)	5.9 (5.0, 7.5)
t(monkey)	6.5 (5, 8)	7.1 (6.1, 8.1)	7.0 (5.3, 8.0)	6.9 (5.2, 8.0)	7.1 (6.1, 8.1)	7.1 (5.7, 8.0)	7.0 (5.5, 8.1)

10.3 Use the mitochondrial data (codon positions 1 and 2) of §10.4.7.2 to examine the impact of priors on posterior divergence time estimation. Change the priors for the substitution rate (μ) and for the rate-drift parameter (ν) to examine the robustness of posterior time estimates.

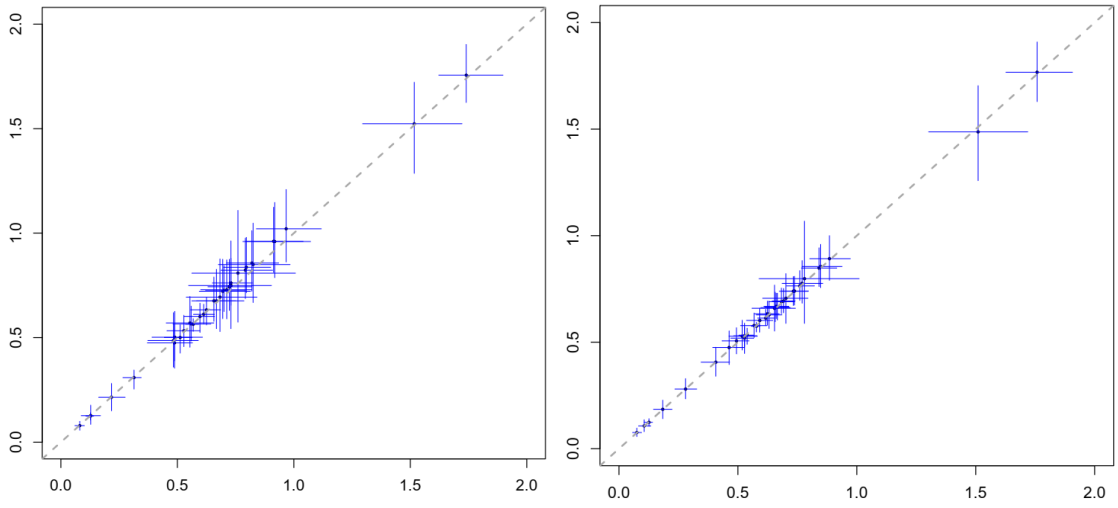
Solution.

We run the analysis for only the independent rate (IR; clock2) and correlated rate (AR; clock3) model. For global rate, obviously nothing would change. I change the gamma priors for rate as "rgene_gamma = 1 10" and set the gamma priors for sigma2 as "sigma2_gamma = 10 1" in the file *mcmctree.ctl*. After MCMCTree run is finished, I use the following script to summarize the result.

Bash

```
$ ~/lab-tools/dating/CI/get_bl_CI_interval.sh FigTree.tre --minmax --format mcmctree --header > time-ci.tbl
$ Rscript ~/lab-tools/dating/graph/compare_age.R -i ori/time-ci.tbl -j change_rgene_sigma/time-ci.tbl -m 0,2 -c blue -o ci-compare.pdf
```

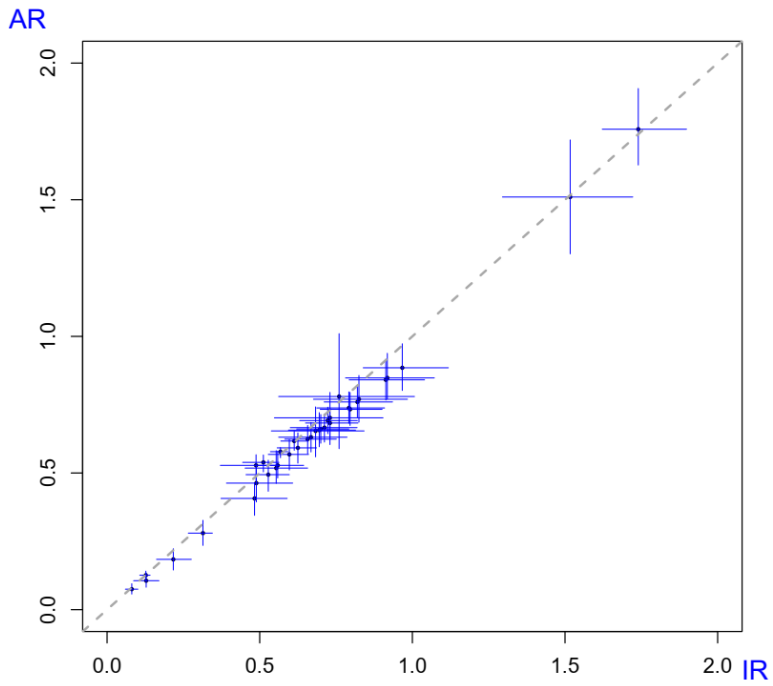
The results are displayed in the following graph for IR (left) and AR (right), where the mean and 95% HPD interval are indicated between using the original priors and the changed priors.



10.4 Use the mitochondrial data (codon positions 1 and 2) of §10.4.7.2 to compare the independent-rates and correlated-rate models for relaxing the molecular clock.

Solution.

It looks that IR model yields slightly larger time estimates in some nodes.

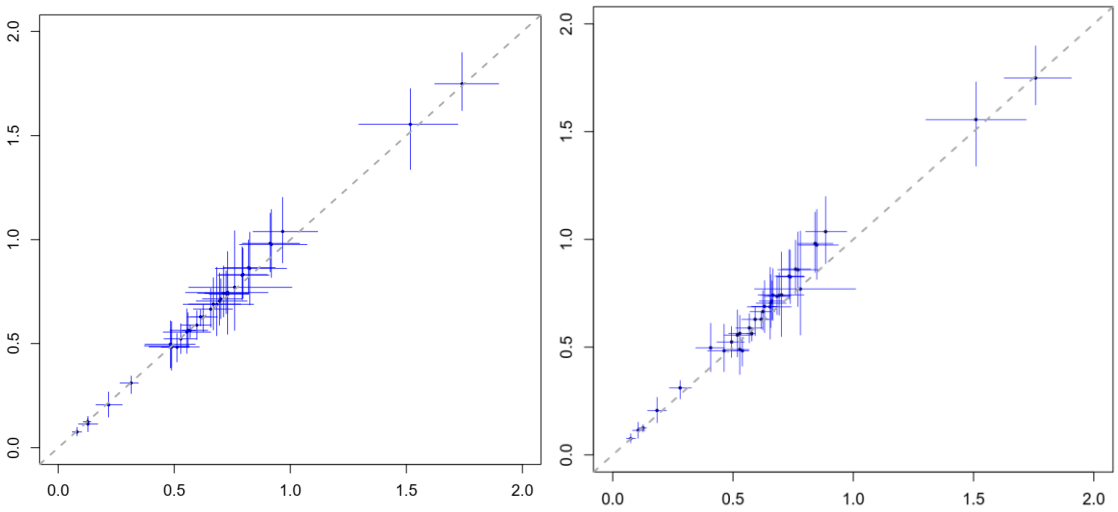


10.5 Use the protein sequence data (36 species, 3,630 amino acid sites) of §10.4.7.2 to estimate the species divergence times under the independent and correlated-rate models, and compare results with those of Problems 10.3 and 10.4. You can use the MCMCTREE program to run the MCMC, with the approximation method (dos Reis and Yang 2011) for likelihood calculation under the MTMAM model (Yang et al. 1998).

Solution.

Use the following Bash code to summary the results and to generate the graphs where the x-axis indicates the posterior times obtained by DNA and the y-axis indicates those obtained with protein sequences.

Bash
$ for i in 2 3; do cd clock${i}_AA/ori/; ~/lab-tools/dating/CI/get_bl_CI_interval.sh FigTree.tre --minmax --format mcmctree --header > time-ci.tbl; cd ../../; cd clock${i}_DNA/ori/; ~/lab-tools/dating/CI/get_bl_CI_interval.sh FigTree.tre --minmax --format mcmctree --header > time-ci.tbl; cd ../../; Rscript ~/lab-tools/dating/graph/compare_age.R -i clock${i}_DNA/ori/time-ci.tbl -j clock${i}_AA/ori/time-ci.tbl -c blue -o clock${i}_DNAvsAA.pdf -m 0,2; done



Left: IR (clock2); Right: AR (clock3)

Chapter 11. Neutral and adaptive protein evolution

11.1 Re-analyse the MHC data of §11.4.4 under the site models M1a, M2a, M7, and M8, but using the mutation–selection model of Yang and Nielsen (2008) to accommodate codon frequencies instead of F3×4 used in §11.4.4. First run model M0 (one-ratio) to obtain MLEs of branch lengths. Then run the site models with the branch lengths fixed at the M0 estimates. Compare results with those of Table 11.3 and Figure 11.4 (see also Yang and Swanson 2002).

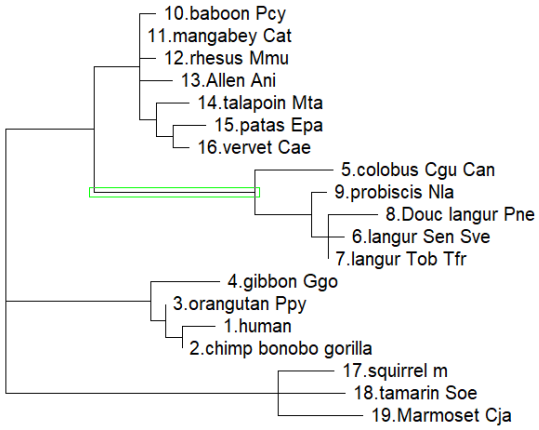
Solution.

Download the files *bigmhc.** from (Yang 2014b). Using PAML v4.10.3, my result is pretty similar to Fig. 11.4 in (Yang 2014a), the only small difference being positively selected sites. See the folder *data/11.4/*.

11.2 Use the branch and branch-site models to analyse the lysozyme genes from 24 primates of Messier and Steward (1997) to detect positive selection on the branch ancestral to the colobine monkeys. Use the F3×4 model of codon usage. (See results for the branch model in Yang 1998a.)

Solution.

According to the problem statement, label the following branch as "#1" in the tree, which is the so-called foreground lineage in the analysis.



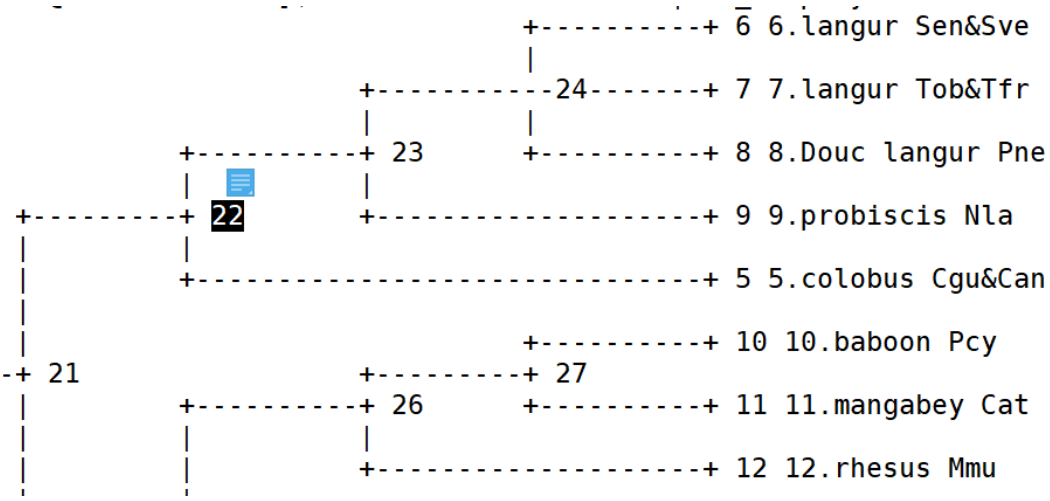
Then, run codeml under the branch model (model=2, NSsites=0) and branch-site (model=2, NSsites=2) model. The positively selected sites identified by the NEB and BEB methods are given as follows [but note that BEB is more reliable and is suggested to use; see (Yang 2014a)]. More detailed results are given in *data/11.2/B-mlc* and *data/11.2/BS-mlc*.

Naive Empirical Bayes (NEB) Positive sites for foregrou			Bayes Empirical Bayes (BEB) Positive sites for foregrou		
14	R	0.917	14	R	0.859
21	R	0.912	21	R	0.858
23	I	0.904	23	I	0.853
37	G	0.562	37	G	0.510
41	R	0.754	41	R	0.710
50	R	0.741	50	R	0.704
62	R	0.582	62	R	0.564
87	D	0.932	87	D	0.869
126	Q	0.728	126	Q	0.710

11.3 Use nucleotide and codon models to reconstruct ancestral sequences using the lysozyme genes from 24 primates to identify the likely amino acid changes on the branch ancestral to the colobine monkeys. Compare the results with the amino acids identified on the same branch by the branch-site test of Problem 11.2. For the nucleotide-based analysis, use the HKY85 + C model of Table 4.3 in §4.3.3, which accounts for different substitution rates, different transition/transversion rate ratios and different base compositions at the three codon positions.

Solution.

Run codeml. Identify that Branch 2 is the branch that connects Node 22 (the clade of colobine monkeys) and its parent (Node 21) in the file *rst*.



The nucleotide/codon changes on Branch 2 are shown as follows.

```
Branch 2:   21..22 (n= 9.0 s= 1.0 Branch 2:   21..22 (n= 9.0 s= 1.0)

     41 G 0.997 -> A 0.997          14 AGA (R 0.997) -> AAA (K) 0.996
     62 G 0.997 -> A 0.997          21 AGG (R 0.998) -> AAG (K) 0.997
     67 A 0.997 -> G 0.997          23 ATC (I 0.998) -> GTC (V) 0.998
    110 A 0.628 -> G 0.998          37 GAT (D 0.634) -> GGT (G) 0.997
    121 C 0.999 -> G 0.998          41 CAA (Q 0.992) -> GAA (E) 0.996
    148 C 0.989 -> G 0.998          50 CAA (Q 0.987) -> GAA (E) 0.999
    185 A 0.628 -> G 0.998          62 CAC (H 0.627) -> CGC (R) 0.999
    259 G 0.997 -> A 0.997          87 GAT (D 0.995) -> AAT (N) 0.995
    319 A 0.999 -> C 0.999         107 AGA (R 1.000) -> CGA (R) 1.000
    376 C 0.999 -> A 0.526         126 CAA (Q 1.000) -> AAA (K) 0.555
```

11.4 (Yang and dos Reis 2011) Conduct a computer simulation to examine the false positive rate of the branch-site test, and examine how long the sequence should be for the asymptotic null distribution to be reliable. Note that according to theory, the test statistic $2\Delta\ell$ should be 0 in half of the datasets and χ^2_1 distributed in the other half if many datasets are simulated under the null model (see §11.5.1). Use the EVOLVER program in the PAML package to simulate 1,000 replicate datasets, on the unrooted tree of Figure 11.7, with all branch lengths equal to 0.5 substitutions per codon, and with branch α as the foreground branch. Assume $\kappa = 2$ for the transition/transversion rate ratio and equal codon frequencies ($\pi_j = 1/61$ for codon j). Generate data under the null model of the branch-site test, with $p_0 = 0.5$, $p_1 = 0.3$, $\omega_0 = 0.5$, and $\omega_1 = \omega_2 = 1$ (see Table 11.4). Then analyse each dataset under the null and alternative models to calculate the test statistic $2\Delta\ell$. Calculate the proportion of datasets in which $2\Delta\ell = 0$ and the proportion of datasets in which $2\Delta\ell > 2.71$. (Note that the latter proportion is the false positive rate of the test at the 5% level when the null distribution is the 1:1 mixture of 0 and χ^2_1.) Use different sequence lengths, such as $N = 50,\ 100,\ 200,\ 500,$ and 1000.

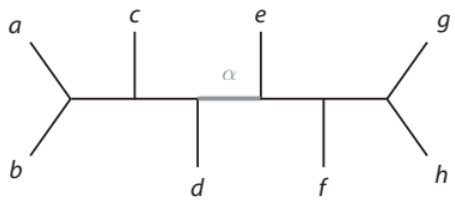


Fig. 11.7 A tree of eight species for simulating data under the branch-site model (Problem 11.4). The foreground branch *a* is highlighted.

Solution.

Note that

$$P(Class = 0) = p_0 = 0.5,$$
$$P(Class = 1) = p_1 = 0.3,$$
$$P(Class = 2a) = \frac{(1 - p_0 - p_1)p_0}{p_0 + p_1} = 0.125,$$
$$P(Class = 2b) = \frac{(1 - p_0 - p_1)p_1}{p_0 + p_1} = 0.075.$$

To simulate codon sequences under the branch-site model, we need to use the program evolverNSbranchsites from PAML, which however is not compiled in the default mode of installation. One has to type the following in Bash to compile evolver to do codon sequence simulation under the branch-site model (as well as the branch and site models if necessary). The control files used to simulate sequences under the null hypothesis are provided in the folder "C11/data/11.4/".

```
Bash
cd ./paml/src/
cc -O2 -DCodonNSbranches -o evolverNSbranches evolver.c tools.c -lm
cc -O2 -DCodonNSsites -o evolverNSsites evolver.c tools.c -lms
cl -O2 -DCodonNSbranchsites -o evolverNSbranchsites evolver.c tools.c -lm
```

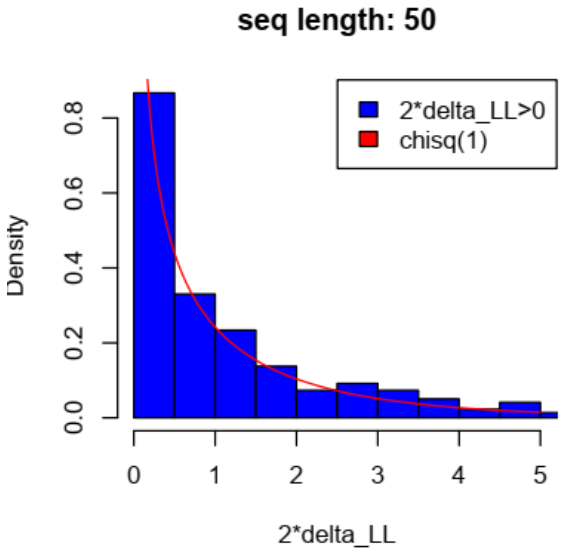
After simulation of sequence evolution, to set the control file codeml.ctl for the unconstrained analysis, one needs to make sure $\text{model} = 2, \text{Nsites} = 2, \text{fix\_omega} = 0$. To set it for the constrained analysis (under the null hypothesis), one needs to change the above settings to $\text{model} = 2, \text{Nsites} = 2, \text{fix\_omega} = 1, \text{omega} = 1$. The control files as well as the results are provided in the folder "C11/data/11.4/".

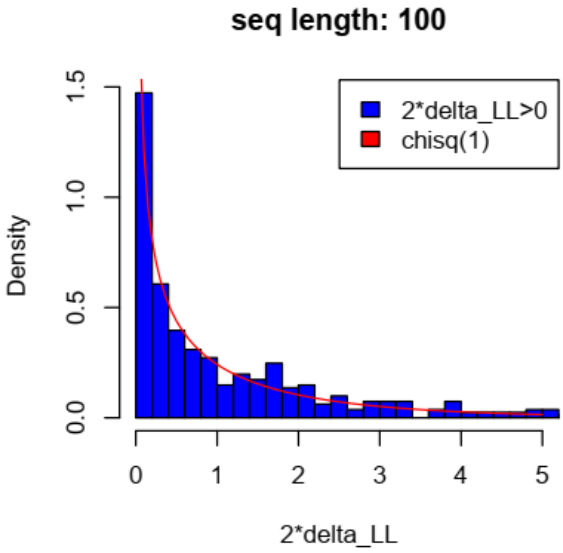
The result is summarized as follows. Note that datasets where $2\Delta L < 0$ are discarded.

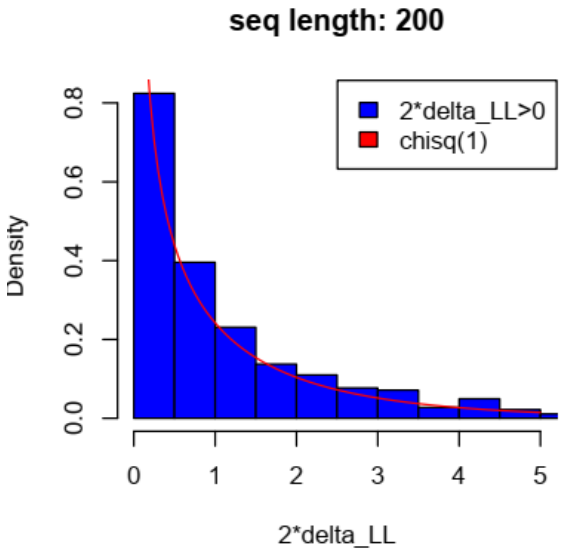
Sequence length	**Proportion of datasets where $2\Delta L = 0$**	**Proportion of datasets where $2\Delta L > 2.71$**
50	0.4941995	0.08352668
100	0.556044	0.06483516
200	0.6013143	0.05257393
500	0.6015284	0.06004367

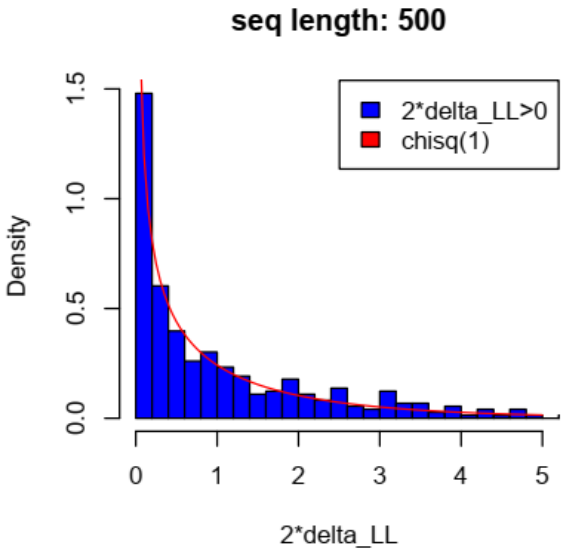
As follows the histograms of the $2\Delta\ell$, the likelihood ratio test statistic, for the branch-site test are plotted against the curve of χ_1^2 density. Note that datasets with $2\Delta L = 0$ are not included in the plot and that sequence length indicates the number of **codons** instead of that of nucleotides.

The result is generally supportive of the conclusions of (Yang and Dos Reis 2011), but note the differences in the proportions of datasets where $2\Delta L = 0$ and where $2\Delta L = 0$ which suggests the impact of tree topologies and branches to test.


seq length: 50
2*delta_LL>0
chisq(1)
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
2*delta_LL


seq length: 100
2*delta_LL>0
chisq(1)
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
5
2*delta_LL


seq length: 200
2*delta_LL>0
chisq(1)
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
2*delta_LL


seq length: 500
2*delta_LL>0
chisq(1)
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
5
2*delta_LL

Chapter 12. Simulating molecular evolution

12.1 If $x_0 = 11, x_n = (37x_{n-1} \bmod 1000)$, find $x_1, x_2, x_3, \ldots, x_{10}$ (see equation (12.1)).

Solution.

Use the following R code. The first 10 pseudo-random numbers generated in this way are

```
11   407 15059  2183  6771 28527 19499 18463 17131  4847
```

R

```
> x<-numeric(10)
> x[1] <- 11
> for(i in 2:10){x[i] <- 37*x[i-1] %% 1000}
> print(x)
```

12.2* *Memory-less property of exponential waiting time*. Suppose random variable X has the exponential distribution with rate λ or mean $1/\lambda$. Show that $\Pr\{X > a + x \mid X > a\} = \Pr\{X > x\}$. This result may be paraphrased as follows. Suppose the waiting time until the bus arrives is an exponential variable with mean $1/\lambda = 10$ minutes. Then, given that we have waited for $a = 9$ minutes, the extra time we have to wait for the bus to arrive is still an exponential variable with a mean of 10 minutes. [Hint. Note that $\Pr\{X > x\} = e^{-\lambda x}$.]

Solution.

For exponential distribution,

$$P(X > x) = 1 - \int_0^{\infty} e^{-\lambda x} dx = e^{-\lambda x}.$$

Apply the Bayes' theorem. It follows that

$$\begin{aligned} P(X > a + x | X > a) &= \frac{P(X > a + x)P(X > a | X > a + x)}{P(X > a)} \\ &= \frac{P(X > a + x)}{P(X > a)} \\ &= \frac{e^{-\lambda(a+x)}}{e^{-\lambda a}} = e^{-\lambda x} = P(X > x). \end{aligned}$$

12.3 *Monte Carlo integration* (§6.4.5). Write a small program to calculate the integral $f(x)$ in the Bayesian estimation of sequence distance under the JC69 model, discussed in Example 6.4. The data are $x = 90$ differences out of $n = 948$ sites. Use the exponential prior with mean 0.2 for the sequence distance θ. Generate $N = 10^6$ or 10^8 random variables from the exponential prior: $\theta_1, \theta_2, \ldots, \theta_N$, and calculate

$$f(x) = \int_0^\infty f(\theta) f(x|\theta)\, d\theta \simeq \frac{1}{N} \sum_{i=1}^{N} f(x|\theta_i). \tag{12.42}$$

Note that the likelihood $f(x|\theta_i)$ may be too small to represent in the computer, so scaling may be needed. One way to do this is as follows. Compute the maximum log likelihood $\ell_m = \log\{f(x|\hat{\theta})\}$, where $\hat{\theta} = 0.1015$ is the maximum likelihood estimate (MLE). Then multiply $f(x|\theta_i)$ in equation (12.42) by $e^{-\ell_m}$ before taking the sum:

$$e^{-\ell_m} \times \sum_{i=1}^{N} f(x|\theta_i) = \sum_{i=1}^{N} \exp\left(\log\{f(x|\theta_i)\} - \ell_m\right). \tag{12.43}$$

See Problem 9.2 of (Yang 2006).

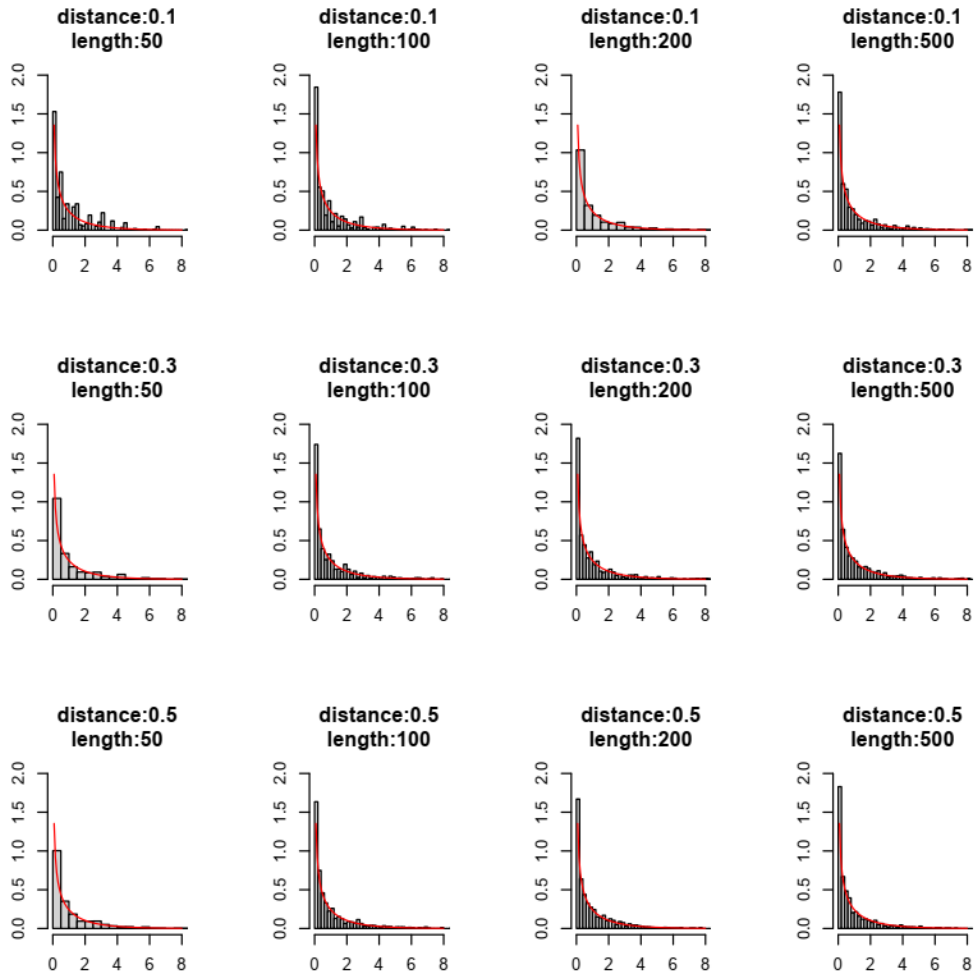
12.4 What is the optimal sequence divergence when a pair of sequences are used to estimate the transition/transversion rate ratio κ under the K80 model? Intuitively very similar sequences will have little information while very divergent sequences will have too much noise, so the optimum sequence divergence should be intermediate. Write a small simulation program to study the optimal sequence divergence. Each dataset consists of a pair of sequences, which can be generated using any of the three approaches discussed in §12.6.1. Alternatively you can use a simulation program such as SEQ-GEN or EVOLVER to simulate the datasets. Assume $\kappa = 2$ and use a sequence length of 500 sites. Consider several sequence distances, say, $d = 0.01, 0.02, \ldots, 2$. For each d, simulate 1,000 replicate datasets under K80 and analyse them under the same model to estimate d and κ using equation (1.12). Calculate the mean and variance of the estimate $\hat{\kappa}$ across replicate datasets. Calculate the standard deviation of $\hat{\kappa}$ and plot it against t.

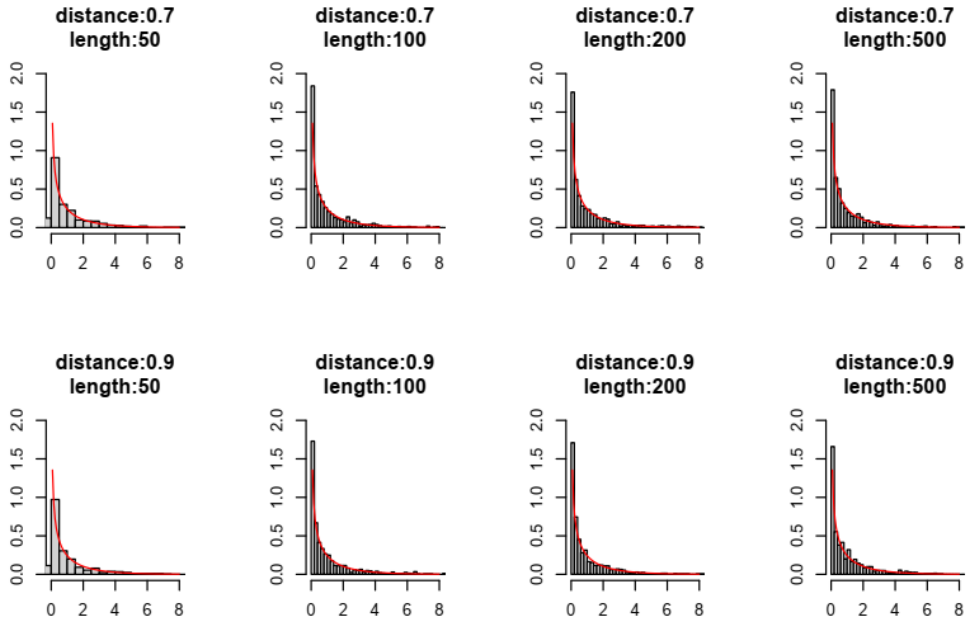
See problem 9.3 of (Yang 2006).

12.5 Use a computer simulation to examine the null distribution of the likelihood ratio test (LRT) statistic for comparing JC69 and K80 using a pair of sequences. Simulate 10,000 replicate datasets as in Problem 12.4 under the JC69 model, and, for each of them, calculate the log likelihood values under the two models (ℓ_0 and ℓ_1) and $2\Delta\ell = 2(\ell_1 - \ell_0)$. See equations (1.47) and (1.52). Then construct a histogram (for example, using the R function `hist`). Compare it with the χ^2 distribution with one degree of freedom. Use different sequence distances ($d = 0.1, 0.5,$ and 1, say) and sequence lengths ($l = 50, 100, 200,$ and 500, say) to examine their impact.

Solution.

I use EVOLER to simulate 1000 replicates of sequences under JC69 with sequence distances from 0.1 to 1.0 (every 0.2) and sequence lengths l =50, 100, 200, and 500. Then run BASEML with both JC69 and K80. Summarize the results using the R script *12.5.R*. You can use the following Bash command to get the following histogram. It looks that with a larger d and longer sequence, the distribution of 2Δ better approximates χ_1. The codes are available at *12.5.sh* and *12.5.R*.





12.6 *Long-branch attraction by parsimony*. Use the JC69 model to simulate datasets on a tree of four species (Figure 12.5a), with two different branch lengths $a = 0.1$ and $b = 0.5$ (in expected number of substitutions per site). Simulate 1,000 replicate datasets. For each dataset, count the sites with the three site patterns *xxyy*, *xyxy*, and *xyyx*, and determine the most parsimonious tree. To simulate a dataset, reroot the tree at an interior node as in Figure 12.5b, say. Generate a sequence for the root (node 0) by random sampling of the four nucleotides, and then evolve the sequence along the five branches of the tree. You may also use a program such as SEQ-GEN or EVOLVER. Consider a few sequence lengths, such as 100, 1,000, and 10,000 sites. Calculate the proportion of datasets in which parsimony recovers the true tree.

See Problem 9.4 of (Yang 2006).

12.7 A useful test of a new and complex likelihood program is to generate a few very large datasets under the model and then analyse them under the same model, to confirm that the MLEs are close to the true values used in the simulation. As MLEs are consistent, they should approach the true values when the sample size (sequence length) becomes larger and larger. Use the program written for Problem 12.6 (or SEQ-GEN or EVOLVER) to generate one or two datasets of 10^6, 10^7, or 10^8 sites under JC69 and analyse them using a likelihood program (such as PAML, PHYML, or RAxML) under the same model, to see whether the MLEs of branch lengths are close to the true values. Beware that some programs may demand a lot of resources to process large datasets.

See Problem 9.5 of (Yang 2006).

References

Earl DJ, Deem MW. 2005. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7:3910–3916.

Hohna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* 65:726–736.

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. U. S. A.*

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35:1547–1549.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and bayesian phylogenetic inference. *Syst. Biol.* 54:241–253.

Paradis E, Schliep K. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian Estimation of Divergence Times. *Mol. Biol. Evol.* 28:2161–2172.

Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3:217–223.

Schliep KP. 2011. Phangorn: phylogenetic analysis in {R}. *Bioinformatics* [Internet] 27:592–593. Available from: https://academic.oup.com/bioinformatics/article/27/4/592/198887

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press Available from: http://abacus.gene.ucl.ac.uk/CME/

Yang Z. 2014a. Molecular Evolution: A Statistical Approach. Oxford University Press Available from: http://abacus.gene.ucl.ac.uk/MESA/

Yang Z. 2014b. MESA Dataset. Available from: http://abacus.gene.ucl.ac.uk/MESA/Yang2014.MESA.data.tgz

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr. Zool.* 61:854–865.

Yang Z. 2022. Errata I of Yang 2014 Molecular Evolution: A Statistical Approach. Available from: http://abacus.gene.ucl.ac.uk/MESA/Yang2014.MESA.Corrections.pdf

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* 54:455–470.

Yang Z, Dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* 28:1217–1228.