A Solutions Manual for Ziheng Yang's *Computational Molecular Evolution*'s Exercises

Version 1.1.1 (2023-10-12)

Sishuo Wang (Department of Microbiology, CUHK, Hong Kong SAR, China)
Jianhao Lv (School of Business, Nanfang College, China)

SW dedicates this solutions manual to his grandmother.
JL dedicates this solutions manual to CAU's honors programme.

Preface

It was in the library of the western campus of China Agricultural University (CAU) that I, a 2nd-year undergraduate, first encountered *Computational Molecular Evolution* by Ziheng Yang (Yang 2006). In the past 15 years, I have greatly benefited too much from this book and its following *Molecular Evolution: A Statistical Approach* (Yang 2014), both of which make the study of molecular phylogenetics not only easier but enjoyable. Different from most biology books, Yang 2006 and Yang 2014 offer exceptional exercises, which are wonderful resources by providing hands-on practice to reinforce the concepts and enhancing understanding. However, the solutions are notably absent, apparently limiting people especially biologists from delving deeper into the field.

In SMBE2023 in Ferrara, Italy, I met Ziheng in person for the first time. I shared with him a naïve but lovely plan to work out every problem of the books, and received much enthusiastic support and encouragement. I then invited my math-envying friend Jianhao to join this work. The way of our collaboration is as follows. I worked out all problems included in the books for the first time, then Jianhao reviewed each of the math-heavy problem and provided alternative solutions where possible, then sent it back to me, and I revised, and sent it back to Jianhao. The above was repeated until convergence. Shamelessly, I put my solution always as "Solution 1" and Jianhao's as alternative solutions, but readers will find the alternatives a better way, almost surely.

Keeping in mind that the biggest challenge for biologists to study molecular evolution is the lack of knowledge in math, this solutions manual aims at providing readers i) very detailed step-by-step solutions to each problem ii) code as R as possible iii) alternative solutions where possible. It is hopeful that most biologists well understand all the problems and solutions, but my greater aspiration is that after reading they can recollect the memories of those calculus, linear algebra, and statistics knowledge, and apply them in research.

Note that by no means are the solutions to be taken as "standard answers". Instead, I hope it can play a tiny role in bridging the divide between people with different backgrounds to work together on interesting biological questions. Just like Ziheng, Jianhao, and I, who enrolled in different programs when studying at CAU—animal sciences, math, and biology—are brought together by computational molecular evolution. Readers are highly welcome to provide feedback by github https://github.com/evolbeginner/Solutions-manual-for-CME2006-and-MESA2014 or by e-mail sishuowang{at}hotmail.ca. For citation, please refer to the online platform at github. The solutions manual is distributed under CC-BY 4.0. There may well be errors in this solutions manual, and I take full responsibility for them.

Needless to say, I particularly thank Ziheng for providing so many interesting problems and for his encouragement. I am also very grateful for my advisor Margaret Ip for supporting me performing this kind of "theoretical" work in Faculty of Medicine, CUHK.

WANG, Sishuo
Prince of Wales Hospital
2023

Chapter 1. Models of nucleotide substitution

1.1 Use the transition probabilities under the JC69 model (equation 1.3) to confirm the Chapman–Kolmogorov theorem (equation 1.4). 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 $p_{\mathrm{TT}}(t_1 + t_2) = p_{\mathrm{TT}}(t_1)p_{\mathrm{TT}}(t_2) + p_{\mathrm{TC}}(t_1)p_{\mathrm{CT}}(t_2) + p_{\mathrm{TA}}(t_1)p_{\mathrm{AT}}(t_2) + p_{\mathrm{TG}}(t_1)p_{\mathrm{GT}}(t_2)$.

Solution.

a) First calculate the right-hand side of the equation.

For $i = j$,

$$p_{ij}(t_1)p_{ji}(t_2) = \left(\frac{1}{4} + \frac{3}{4}e^{-4\lambda t_1}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4\lambda t_2}\right)$$
$$= \frac{1}{16}\left(1 + 3e^{-4\lambda t_1} + 3e^{-4\lambda t_2} + 9e^{-4\lambda(t_1+t_2)}\right).$$

For $i \neq j$,

$$p_{ij}(t_1)p_{ji}(t_2) = \left(\frac{1}{4} - \frac{1}{4}e^{-4\lambda t_1}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4\lambda t_2}\right)$$
$$= \frac{1}{16}\left(1 - e^{-4\lambda t_1} - e^{-4\lambda t_2} + e^{-4\lambda(t_1+t_2)}\right).$$

Hence

$$p_{ij}(t_1)p_{ji}(t_2) = \begin{cases} \frac{1}{16}\left(1 + 3e^{-4\lambda t_1} + 3e^{-4\lambda t_2} + 9e^{-4\lambda(t_1+t_2)}\right), & i = j \\ \frac{1}{16}\left(1 - e^{-4\lambda t_1} - e^{-4\lambda t_2} + e^{-4\lambda(t_1+t_2)}\right), & i \neq j \end{cases}.$$

It follows that

$$p_{TT}(t_1)p_{TT}(t_2) + p_{TC}(t_1)p_{CT}(t_2) + p_{TA}(t_1)p_{AT}(t_2) + p_{TG}(t_1)p_{GT}(t_2)$$
$$= \frac{1}{16}\left(1 + 3e^{-4\lambda t_1} + 3e^{-4\lambda t_2} + 9e^{-4\lambda(t_1+t_2)}\right)$$
$$+ 3 \times \frac{1}{16}\left(1 - e^{-4\lambda t_1} - e^{-4\lambda t_2} + e^{-4\lambda(t_1+t_2)}\right)$$
$$= \frac{1}{4} + \frac{3}{4}e^{-4\lambda(t_1+t_2)}$$
$$= p_{TT}(t_1 + t_2).$$

b) It follows that

$$p_{TT}(t_1)p_{TC}(t_2) + p_{TC}(t_1)p_{CC}(t_2) + p_{TA}(t_1)p_{AT}(t_2) + p_{TG}(t_1)p_{GT}(t_2)$$
$$= \left(\frac{1}{4} + \frac{3}{4}e^{-4\lambda t_1}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4\lambda t_2}\right) + \left(\frac{1}{4} - \frac{1}{4}e^{-4\lambda t_1}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4\lambda t_2}\right)$$
$$+ 2 \times \left(\frac{1}{4} - \frac{1}{4}e^{-4\lambda t_1}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4\lambda t_2}\right)$$
$$= \frac{1}{16}\left[\left(1 - e^{-4\lambda t_2} + 3e^{-4\lambda t_1} - 3e^{-4\lambda(t_1+t_2)}\right) + \left(1 - e^{-4\lambda t_1} + 3e^{-4\lambda t_2} - 3e^{-4\lambda(t_1+t_2)}\right)\right.$$
$$\left. + 2 \times \left(1 - e^{-4\lambda t_1} - e^{-4\lambda t_2} + e^{-4\lambda(t_1+t_2)}\right)\right]$$
$$= \frac{1}{4} - \frac{1}{4}e^{-4\lambda(t_1+t_2)}$$

$= p_{TC}(t_1 + t_2)$.

1.2 Derive the transition-probability matrix $P(t) = e^{Qt}$ for the JC69 model (Jukes and Cantor 1969). Set $\pi_T = \pi_C = \pi_A = \pi_G = 1/4$ and $\alpha_1 = \alpha_2 = \beta$ in the rate matrix (1.15) for the TN93 model to obtain the eigenvalues and eigenvectors of Q under JC69, using results of Subsection 1.2.3. Alternatively you can derive the eigenvalues and eigenvectors from equation (1.1) directly. Then apply equation (1.17).

Solution.

The rate matrix of the JC69 substitution model is defined as

$$Q = \begin{bmatrix} -3\theta & \theta & \theta & \theta \\ \theta & -3\theta & \theta & \theta \\ \theta & \theta & -3\theta & \theta \\ \theta & \theta & \theta & -3\theta \end{bmatrix}.$$

Note that we substitute θ for λ in the original formula. To obtain the eigenvalues and eigenvectors, let the determinant equal to zero. Denote the eigenvalue as λ. The determinant of Q is given by

$$\begin{vmatrix} -3\theta-\lambda & \theta & \theta & \theta \\ \theta & -3\theta-\lambda & \theta & \theta \\ \theta & \theta & -3\theta-\lambda & \theta \\ \theta & \theta & \theta & -3\theta-\lambda \end{vmatrix}$$

$$\underline{\underline{r_4 - r_3}} \begin{vmatrix} -3\theta-\lambda & \theta & \theta & \theta \\ \theta & -3\theta-\lambda & \theta & \theta \\ \theta & \theta & -3\theta-\lambda & \theta \\ 0 & 0 & 4\theta+\lambda & -4\theta-\lambda \end{vmatrix}$$

$$\underline{\underline{c_3 + c_4}} \begin{vmatrix} -3\theta-\lambda & \theta & 2\theta & \theta \\ \theta & -3\theta-\lambda & 2\theta & \theta \\ \theta & \theta & -2\theta-\lambda & \theta \\ 0 & 0 & 0 & -4\theta-\lambda \end{vmatrix}$$

$$\underline{\underline{r_3 - r_2}} \begin{vmatrix} -3\theta-\lambda & \theta & 2\theta & \theta \\ \theta & -3\theta-\lambda & 2\theta & \theta \\ 0 & 4\theta+\lambda & -4\theta-\lambda & 0 \\ 0 & 0 & 0 & -4\theta-\lambda \end{vmatrix}$$

$$\underline{\underline{c_2 + c_3}} \begin{vmatrix} -3\theta-\lambda & 3\theta & 2\theta & \theta \\ \theta & -\theta-\lambda & 2\theta & \theta \\ 0 & 0 & -4\theta-\lambda & 0 \\ 0 & 0 & 0 & -4\theta-\lambda \end{vmatrix}$$

$$= \lambda(4\theta + \lambda)^3.$$

Setting the above to zero, it is easy to see the eigenvalues of Q' are $\lambda_1 = 0, \lambda_2 = \lambda_3 = \lambda_4 = -4\theta$. The eigenvectors of $\lambda_2 = \lambda_3 = \lambda_4 = -4\theta$ can be obtained by solving the following.

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow a + b + c + d = 0$$

The eigenvectors of $\lambda_2 = \lambda_3 = \lambda_4 = -4\theta$ are

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \\ -a-b-c \end{bmatrix} = a\begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} + b\begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix} + c\begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}.$$

Taking $a = 1, b = c = 0$ gives $(1,0,0,-1)$.
Taking $b = 1, a = c = 0$ gives $(0,1,0,-1)$.
Taking $c = 1, a = b = 0$ gives $(0,0,1,-1)$.
Similarly, the eigenvector for $\lambda_1 = 0$ is $(1,1,1,1)$.
Hence,

$$Q = U\Lambda U^{-1},$$

where

$$U = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}, U^{-1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{3}{4} & -\frac{1}{4} \end{bmatrix}, \Lambda = \begin{bmatrix} 0 & & & \\ & -4\theta & & \\ & & -4\theta & \\ & & & -4\theta \end{bmatrix}.$$

So

$$e^{Qt} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & e^{-4\theta t} & & \\ & & e^{-4\theta t} & \\ & & & e^{-4\theta t} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \end{bmatrix}$$

$$= \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3e^{-4\theta t} & -e^{-4\theta t} & -e^{-4\theta t} & -e^{-4\theta t} \\ -e^{-4\theta t} & 3e^{-4\theta t} & -e^{-4\theta t} & -e^{-4\theta t} \\ -e^{-4\theta t} & -e^{-4\theta t} & 3e^{-4\theta t} & -e^{-4\theta t} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{4} + \frac{3}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} + \frac{3}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} + \frac{3}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} - \frac{1}{4}e^{-4\theta t} & \frac{1}{4} + \frac{3}{4}e^{-4\theta t} \end{bmatrix}.$$

Substituting λ for θ, we have

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}, i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}, i \neq j \end{cases}.$$

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

$$Q = U\Lambda U^{-1} = \begin{pmatrix} 1 & -u \\ 1 & v \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -u-v \end{pmatrix} \begin{pmatrix} v/(u+v) & u/(u+v) \\ -1/(u+v) & 1/(u+v) \end{pmatrix}, \quad (1.70)$$

so that

$$P(t) = \mathrm{e}^{Qt} = \frac{1}{u+v} \begin{pmatrix} v + u\mathrm{e}^{-(u+v)t} & u - u\mathrm{e}^{-(u+v)t} \\ v - v\mathrm{e}^{-(u+v)t} & u + v\mathrm{e}^{-(u+v)t} \end{pmatrix}. \quad (1.71)$$

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

$$P(t) = \begin{pmatrix} \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} \\ \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} \end{pmatrix}. \quad (1.72)$$

This is the binary equivalent of the JC69 model.

Solution.

First, calculate the eigen values and eigenvectors of Q.

$$\begin{vmatrix} -u-\lambda & u \\ v & -v-\lambda \end{vmatrix} = 0$$

By solving the following equation

$$\lambda^2 + (u+v)\lambda = 0$$

we get the eigenvalues and their corresponding eigenvectors:

$$\lambda_1 = 0, \qquad \lambda_2 = -u - v,$$

$$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} -u \\ v \end{bmatrix}.$$

Hence,

$$P(t) = e^{Qt}$$

$$= \begin{bmatrix} 1 & -u \\ 1 & v \end{bmatrix} e^{\begin{bmatrix} 0 & 0 \\ 0 & -(u+v)t \end{bmatrix}} \begin{bmatrix} \frac{v}{u+v} & \frac{u}{u+v} \\ -\frac{1}{u+v} & \frac{1}{u+v} \end{bmatrix}$$

$$= \frac{1}{u+v} \begin{bmatrix} 1 & -u \\ 1 & v \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & e^{-(u+v)t} \end{bmatrix} \begin{bmatrix} v & u \\ -1 & 1 \end{bmatrix}$$

$$= \frac{1}{u+v} \begin{bmatrix} v + ue^{-(u+v)t} & u - ue^{-(u+v)t} \\ v - ve^{-(u+v)t} & u + ve^{-(u+v)t} \end{bmatrix}.$$

As to the limiting distribution $\pi^{(t)}$, suppose that the system is in state 1 at time $n = 0$ with probability $\pi_1{}^{(0)}$ and in state 2 at time $n = 0$ with probability $\pi_2{}^{(0)}$, such that $\pi^{(0)} =$

$(\pi_1^{(0)}, \pi_2^{(0)})$ and $\pi_1^{(0)} + \pi_2^{(0)} = 1$. It follows that

$$\begin{aligned}\pi^{(t)} &= \pi^{(0)} P(t) \\ &= \frac{1}{u+v}[\pi_1^{(0)} \quad \pi_2^{(0)}] \begin{bmatrix} v + ue^{-(u+v)t} & u - ue^{-(u+v)t} \\ v - ve^{-(u+v)t} & u + ve^{-(u+v)t} \end{bmatrix} \\ &= \frac{1}{u+v} \begin{bmatrix} \pi_1^{(0)}\left(v + ue^{-(u+v)t}\right) + \pi_2^{(0)}\left(v + ue^{-(u+v)t}\right) \\ \pi_1^{(0)}\left(u - ue^{-(u+v)t}\right) + \pi_2^{(0)}\left(u + ve^{-(u+v)t}\right) \end{bmatrix}^T.\end{aligned}$$

By letting $t \to \infty$, $\lim_{t\to\infty} e^{-(u+v)t} = 0$. Hence, the limiting distribution π is given by

$$\pi = \lim_{t\to\infty} \pi^{(t)} = \begin{bmatrix} \frac{v}{u+v} & \frac{u}{u+v} \end{bmatrix}.$$

In case $u = v = 1$, Eq. (**Error! Reference source not found.**) is written as

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

which is the binary equivalent of JC69.

1.4 Confirm that the two likelihood functions for the JC69 model, equations (1.42) and (1.43), are proportional and the proportionality factor is a function of n and x but not of d. Confirm that the likelihood equation, $\mathrm{d}\ell/\mathrm{d}d = \mathrm{d}\log\{L(d)\}/\mathrm{d}d = 0$, is the same whichever of the two likelihood functions is used.

Solution.

According to Eq. (1.42) in (Yang, 2006), we obtain

$$\begin{aligned}L(d;x) &= C\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4d}{3}}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4d}{3}}\right)^{n-x} \\ &= C \times 12^x \times 4^{n-x} \times \left(\frac{1}{12}\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4d}{3}}\right)\right)^x \left(\frac{1}{4}\left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4d}{3}}\right)\right)^{n-x} \\ &= 3^x 4^n C\left(\frac{1}{16} - \frac{1}{16}e^{-\frac{4d}{3}}\right)^x \left(\frac{1}{16} + \frac{3}{16}e^{-\frac{4d}{3}}\right)^{n-x}\end{aligned}$$

The answer to the second question can be easily seen from the above. Alternatively, we can for Eq. (1.42) of (Yang, 2006), set the formula

$$\frac{d\ell}{d(d)} = \left[log\left(\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4d}{3}}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4d}{3}}\right)^{n-x}\right)\right]' = 0.$$

Solving the above equation, we obtain

$$\left[log\left(\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4d}{3}}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4d}{3}}\right)^{n-x}\right)\right]' = 0$$

$$-\frac{4d}{3}e^{-\frac{4d}{3}}\left(x\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4d}{3}}\right)^{-1} - (n-x)\left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4d}{3}}\right)^{-1}\right) = 0$$

$$\frac{x}{4}\left(1 + 3e^{-\frac{4d}{3}}\right) - \frac{(n-x)}{4}\left(3 - 3e^{-\frac{4d}{3}}\right) = 0$$

$$\hat{d} = -\frac{3}{4} log\left(1 - \frac{4x}{3n}\right).$$

Similarly, for Eq. (1.43) in in (Yang, 2006), we have

$$\left[log\left(\left(\frac{1}{16} - \frac{1}{16}e^{-\frac{4d}{3}}\right)^{x}\left(\frac{1}{16} + \frac{3}{16}e^{-\frac{4d}{3}}\right)^{n-x}\right)\right]' = 0$$

$$-\frac{4d}{3}e^{-\frac{4d}{3}}\left(\left(\frac{1}{16} - \frac{1}{16}e^{-\frac{4d}{3}}\right)^{-1} + (n-x)\left(\frac{1}{16} + \frac{3}{16}e^{-\frac{4d}{3}}\right)^{-1}\right) = 0$$

$$\frac{\mathrm{x}}{16}\left(1 + 3e^{-\frac{4d}{3}}\right) - \frac{(\mathrm{n} - \mathrm{x})}{16}\left(3 - 3e^{-\frac{4d}{3}}\right) = 0$$

$$\hat{d} = -\frac{3}{4} log\left(1 - \frac{4x}{3n}\right).$$

***1.5** Suppose $x = 9$ heads and $r = 3$ tails are observed in $n = 12$ independent tosses of a coin. Derive the MLE of the probability of heads (θ). Consider two mechanisms by which the data are generated.

(a) *Binomial*. The number $n = 12$ tosses was fixed beforehand. In $n = 12$ tosses, $x = 9$ heads were observed. Then the number of heads x has a binomial distribution, with probability

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{1.73}$$

(b) *Negative binomial*. The number of tails $r = 3$ was fixed beforehand, and the coin was tossed until $r = 3$ tails were observed, at which point it was noted that $x = 9$ heads were observed. Then x has a negative binomial distribution, with probability

$$f(x|\theta) = \binom{r+x-1}{x}\theta^x(1-\theta)^{n-x}. \tag{1.74}$$

Confirm that under both models, the MLE of θ is x/n.

Solution.

(a) Set $\left(\log\left(f(x|\theta)\right)\right)'$ to zero and solve the resulting equation:

$$x(log\theta)' + (n-x)(\log(1-\theta))' = 0$$

$$\frac{x}{\theta} - \frac{(n-x)}{1-\theta} = 0$$

$$\hat{\theta} = \frac{x}{n}.$$

(b) It is easy to see $\hat{\theta} = \frac{x}{n}$ based on the result of (a) because the likelihood functions between (a) and (b) differ in only the constant term which contributes nothing to solving $\left(log\left(f(x|\theta)\right)\right)' = 0$.

Chapter 2. Models of amino acid and codon substitution

2.1 Obtain two sequences from GenBank, align the sequences and then apply the methods discussed in this chapter to estimate d_S and d_N and discuss their differences. 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, MEGA3.1 (Kumar *et al.* 2005*a*) or BAMBE (Xia and Xie 2001), followed by manual adjustments.

Solution.

Following Exercise 2.1 in (Yang, 2014), I download the sequences of the gene coding for NADH6 of *Homo sapiens* and *Pongo pygmaeus* from their mitochondrial genome sequences (Accession: X93334 and D38115 respectively). Note that to translate the CDS of human one into amino acids the reverse complement may be needed. Then, alignment the amino acid sequences using MAFFT (Katoh & Standley, 2013), and construct the DNA alignment based on the protein alignment with PAL2NAL (Suyama et al., 2006).

```
>Human
ATGATGTATGCTTTGTTTCTGTTGAGTGTGGGTTTAGTAATGGGGTTTGTGGGGT
TTTCTTCTAAGCCTTCTCCTATTTATGGGGGTTTAGTATTGATTGTTAGCGGTGTG
GTCGGGTGTGTTATTATTCTGAATTTTGGGGGAGGTTATATGGGTTTAATAGTTTT
TTTAATTTATTTAGGGGGAATGATGGTTGTCTTTGGATATACTACAGCGATGGCTA
TTGAGGAGTATCCTGAGGCATGGGGGTCAGGGGTTGAGGTCTTGGTGAGTGTTT
TAGTGGGGTTAGCGATGGAGGTAGGATTGGTGCTGTGGGTGAAAGAGTATGATG
GGGTGGTGGTTGTGGTAAACTTTAATAGTGTAGGAAGCTGAATAATTTATGAAGG
AGAGGGGTCAGGGTTCATTCGGGAGGATCCTATTGGTGCGGGGGCTTTGTATGA
TTATGGGCGTTGATTAGTAGTAGTTACTGGTTGAACATTGTTTGTTGGTGTATATA
TTGTAATTGAGATTGCTCGGGGGAATAGG
>Orangutan
ATGACATATGCTTTGTTTCTGTTGAGTGTGATTTTAGTGATGGGGTTTGTGGGGTT
TTCTTCTAAGCCCTCCCCTATTTATGGGGGTTTAGTGTTGATTATTAGTGGTGCGG
TTGGGTGTGCAGTTATTTTAAATTGTGGGGGAGGTTATATGGGTCTGGTGGTTTT
TTTAGTTTATTTAGGGGGTATGATGGTTGTTTTTGGGTATACTACGGCAATGGCTA
TTGAGGAGTATCCTGAGGCGTGAGGGTCGGGGGCTGAGGTGTTGGTGAGTGTT
CTGGTGGGGTTAGTGATGGAAGTGGGGTTGGTGTTGTGGGTGAAGGAGTGTGA
TGGGGTAGTAGTGGCGGTGAATTTTAATAGCGTAGGGAGCTGGATAATTTATGAG
GGGGAAGGGTCAGGGTTGATTCGGGAAGATCCTATTGGTGCGGGGGCTTTATAT
GACTATGGGCGTTGGTTGGTGGTGGTTACTGGTTGAACATTATTTGTTGGTGTTT
ATGTTGTAATTGAGATTGCTCGGGGTAATAGG

>Human
MMYALFLLSVGLVMGFVGFSSKPSPIYGGLVLIVSGVVGCVIILNFGGGYMGLMVF
LIYLGGMMVVFGYTTAMAIEEYPEAWGSGVEVLVSVLVGLAMEVGLVLWVKEYD
GVVVVVNFNSVGSWMIYEGEGSGFIREDPIGAGALYDYGRWLVVVTGWTLFVGV
YIVIEIARGN
```

```
>Orangutan
MTYALFLLSVILVMGFVGFSSKPSPIYGGLVLIISGAVGCAVILNCGGGYMGLVVFLV
YLGGMMVVFGYTTAMAIEEYPEAWGSGAEVLVSVLVGLVMEVGLVLWVKECDG
VVVAVNFNSVGSWMIYEGEGSGLIREDPIGAGALYDYGRWLVVVTGWTLFVGVY
VVIEIARGN
```

Then use CODEML to get the results. The following shows my settings of the file *codeml.ctl*. Note that in the control file *codeml.ctl* you need to set icode=1 which represents the codon table 1, i.e., the one used in mammalian mitochondria. Kappa is set to zero which means it is estimated rather than fixed.

```
      seqfile = DNA.paml  * sequence data filename
     treefile = HO.tree  * tree file name
      outfile = mlc           * main result file name
        icode = 1
        noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 2  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise
      seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 0 * 0 : 1/61 each, 1:F1X4, 2:F3X4, 3:codon table
                   * 4:F1x4MG, 5:F3x4MG, 6:FMutSel0, 7:FMutSel
        model = 0
      NSsites = 0

        clock = 0  * 0:no clock, 1:global clock; 2:local clock
       aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a

    fix_kappa = 0 * 1: kappa fixed, 0: kappa to be estimated
        kappa = 1  * initial or fixed kappa
    fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
        omega = 1.5  * initial or fIf yoixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 10  * # of categories in dG of NSsites models

        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 0  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)
   Small_Diff = 1e-8
```

Bash

```
$ mafft AA.fas > AA.aln
$ pal2nal.pl    AA.aln DNA.fas -output paml -codontable 2 > DNA.paml
$ codeml codeml.ctl
```

The estimates for d_S, d_N, d_{1B}, d_{2B}, d_{3B}, d_S^*, and d_N^* by using different models are displayed as follows which are directly obtained from CODEML's output.

Fequal:

```
d123[*] =  0.03464  0.02501  0.18682  average  0.08216
    [B] =  0.18752  0.18752  0.19864  average  0.19123
accept  =  0.18475  0.13335  0.94049

w =  0.13335 dN =  0.02550 dS =  0.19123 d4 =  0.20092 (92.8 four-fold sites)
             dN*=  0.02244 dS*=  0.25946 S* = 131.53 N* = 390.47
```

F1×4:

```
d123[*] =  0.05377  0.02258  0.44606  average  0.17414
    [B] =  0.41723  0.40331  0.46944  average  0.43000
accept  =  0.12888  0.05598  0.95018

w =  0.05598 dN =  0.02407 dS =  0.43000 d4 =  0.47199 (75.7 four-fold sites)
             dN*=  0.01984 dS*=  0.67516 S* = 122.90 N* = 399.10
```

F3×4:

```
d123[*] =  0.05131  0.02617  0.43320  average  0.17023
    [B] =  0.59851  0.61221  0.44745  average  0.55272
accept  =  0.08572  0.04275  0.96814

w =  0.04275 dN =  0.02363 dS =  0.55272 d4 =  0.54748 (92.1 four-fold sites)
             dN*=  0.02284 dS*=  0.60783 S* = 131.52 N* = 390.48
```

F61:

```
d123[*] =  0.10059  0.02769  0.97232  average  0.36687
    [B] =  1.15159  1.28790  0.98952  average  1.14300
accept  =  0.08735  0.02150  0.98262

w =  0.02150 dN =  0.02458 dS =  1.14300 d4 =  1.18640 (88.5 four-fold sites)
             dN*=  0.02284 dS*=  1.38220 S* = 132.11 N* = 389.89
```

FMutSel:

```
d123[*] =  0.05746  0.02827  0.34757  average  0.14443
    [B] =  0.39129  0.47924  0.36540  average  0.41197
accept  =  0.14684  0.05900  0.95120

w =  0.05900 dN =  0.02431 dS =  0.41197 d4 =  0.36797 (88.5 four-fold sites)
             dN*=  0.02246 dS*=  0.50441 S* = 132.11 N* = 389.89
```

***2.2** Are there really three nucleotide sites in a codon? How many synonymous and nonsynonymous sites are in the codon TAT (use the universal code)?

Solution.

The following assumes $\kappa = 1$. The three codon positions can each change to three other nucleotides, so the codon TAT which codes for Tyr, has nine immediate neighbors, namely TAC (Tyr), TAA (*), TAG (*), TTT (Phe), TCT (Ser), TGT (Cys), CAT (His), AAT (Asn), GAT (Asp). Among these nine, TAA and TAG are stop codon, TAC codes for the same amino acid as the original one TAT, while the remaining six code for amino acids different from TAT.

Hence, the number of synonymous sites is equal to $3 \times \frac{1}{9} = \frac{1}{3}$ and that of nonsynonymous sites is equal to $3 \times \frac{6}{9} = 2$ for the codon TAT.

2.3 Behaviour of LWL85 and related methods under the *two-fold and four-fold mixture regular code*. Imagine a genetic code in which a proportion γ of codons are four-fold degenerate while all other codons are two-fold degenerate. (If $\gamma = 48/64$, the code would encode exactly 20 amino acids.) Suppose that neutral mutations occur according to the K80 model, with transition rate α and transversion rate β, with $\alpha/\beta = \kappa$. The proportion of nonsynonymous mutations that are neutral is ω. The numbers of nondegenerate, two-fold, and four-fold degenerate sites in a codon are $L_0 = 2$, $L_2 = 1 - \gamma$, and $L_4 = \gamma$. Over time interval t, the numbers of transitional and transversional substitutions at the three degeneracy classes are thus $A_0 = \alpha t\omega$, $B_0 = 2\beta t\omega$, $A_2 = \alpha t$, $B_2 = 2\beta t\omega$, $A_4 = \alpha t$, $B_4 = 2\beta t$. (a) Show that the LWL85 method (equation 2.10) gives

$$d_S = \frac{3(\kappa + 2\gamma)\beta t}{1 + 2\gamma},$$
$$d_N = \frac{3(\kappa + 3 - \gamma)\beta t\omega}{4 - \gamma}, \tag{2.29}$$

with the ratio $d_N/d_S = \omega[(\kappa + 3 - \gamma)(1 + 2\gamma)]/[(4 - \gamma)(\kappa + 2\gamma)]$, which becomes $\omega(\kappa + 3)/(4\kappa)$ if $\gamma = 0$ (so that the code is the two-fold regular code) and ω if $\gamma = 1$ (so that the code is the four-fold regular code). (b) Show that both LPB93 (equation 2.11) and LWL85m (equation 2.12) give $d_S = (\alpha + 2\beta)t$ and $d_N = d_S\omega$. (Comment: under this model, LWL85 gives d_S^* and d_N^*, distances using the physical-site definition.)

Solution.

(a) Refer to Eq. (2.10) in (Yang, 2006):

$$d_S = \frac{L_2A_2 + L_4d_4}{\frac{L_2}{3} + L_4}, d_N = \frac{L_2B_2 + L_0d_0}{\frac{2}{3}L_2 + L_0}.$$

According to the given information, we obtain $\alpha = \kappa\beta$, $A_0 = \alpha t\omega$, $B_0 = 2\beta t\omega$, $A_2 = \alpha t$, $B_2 = 2\beta t\omega$, $A_4 = \alpha t$, $B_4 = 2\beta t$, $L_0 = 2$, $L2 = 1 - \gamma$, and $L_4 = \gamma$. Hence,

$$d_0 = A_0 + B_0 = \alpha t\omega + 2\beta t\omega$$
$$d_4 = A_4 + B_4 = \alpha t + 2\beta t.$$

Plugging the above into Eq. (2.10) in (Yang, 2006), we have

$$d_S = \frac{(1-\gamma)\alpha t + \gamma(\alpha t + 2\beta t)}{\frac{1-\gamma}{3} + \gamma}$$
$$= 3 \times \frac{(1-\gamma)\kappa\beta t + \gamma(\kappa\beta t + 2\beta t)}{1 + 2\gamma}$$
$$= \frac{3(\kappa + 2\gamma)\beta t}{1 - 2\gamma},$$

and

$$d_N = \frac{2\beta t\omega(1-\gamma) + 2(\alpha t\omega + 2\beta t\omega)}{\frac{2}{3}(1-\gamma) + 2}$$

$$= \frac{3\big(\beta t\omega(1-\gamma) + (\kappa\beta t\omega + 2\beta t\omega)\big)}{4-\gamma}$$
$$= \frac{3(\kappa + 3 - \gamma)\beta t\omega}{4-\gamma}.$$

(b) As to LPB93, following Eq. (2.11) in (Yang, 2006), we obtain

$$d_S = \frac{L_2A_2 + L_4A_4}{L_2 + L_4} + B_4$$
$$= \frac{(1-\gamma)\alpha t + \gamma\alpha t}{(1-\gamma)+\gamma} + 2\beta t$$
$$= \alpha t + 2\beta t,$$

and

$$d_N = A_0 + \frac{L_0B_0 + L_2B_2}{L_0 + L_2}$$
$$= \alpha t\omega + \frac{2(2\beta t\omega) + (1-\gamma)\times 2\beta t\omega}{2+1-\gamma}$$
$$= \omega(\alpha t + 2\beta t).$$

Hence, it is easy to see that $d_N = \omega d_S$.

As to LWL85m, we follow Eq. (2.12) in (Yang, 2006) to calculate d_S and d_N as

$$d_S = \frac{L_2A_2 + L_4d_4}{\rho L_2 + L_4}$$
$$= \frac{(1-\gamma)\alpha t + \gamma(\alpha t + 2\beta t)}{\frac{\alpha t}{\alpha t + 2\beta t}(1-\gamma) + \gamma}$$
$$= \frac{(\alpha t + 2\beta t)(\alpha t + 2\gamma\beta t)}{\alpha t(1-\gamma) + (\alpha t + 2\beta t)\gamma}$$
$$= \alpha t + 2\beta t,$$

and

$$d_N = \frac{L_2B_2 + L_0d_0}{(1-\rho)L_2 + L_0}$$
$$= \frac{(1-\gamma)2\omega\beta t + 2(\omega\alpha t + 2\omega\beta t)}{\left(1 - \frac{\alpha t}{\alpha t + 2\beta t}\right)(1-\gamma) + 2}$$
$$= \omega(\alpha t + 2\beta t)\frac{(1-\gamma)\beta t + (\alpha t + 2\beta t)}{(1-\gamma)\beta t + (\alpha t + 2\beta t)}$$
$$= \omega d_S.$$

Note that in the above ρ is given as $\rho = \frac{A_4}{A_4 + B_4}$ according to Eq (2.13) in (Yang, 2006).

Chapter 4. Maximum likelihood methods

***4.1** Collapsing site patterns for likelihood calculation under the JC69 model. Under JC69, the probability of data at a site depends on whether the nucleotides are different in different species, but not on what the nucleotides are. For example, sites with data TTTC, TTTA, AAAG all have the same probability of occurrence. Show that if such sites are collapsed into patterns, there is a maximum of $(4^{s-1} + 3 \times 2^{s-1} + 2)/6$ site patterns for s sequences (Saitou and Nei 1986).

Solutions.

Three solutions are provided. The first employs mathematical induction, the second one builds upon the first but adopts a linear algebra perspective, and the final approach solves the problem purely combinatorically.

Solution 1.

At the beginning, it is important to illustrate the different patterns under model JC69 with the following example when there are three species ($s = 3$). In total, the $4^3 = 64$ combinations of nucleotides collapse into the following five patterns denoted as aaa, abb, bab, bba, abc.

111, where all three species have the same nucleotide:

TTT, CCC, AAA, GGG

122, where the second and third species have the same nucleotide but the first species has a different one.

CTT, ATT, GTT, TCC, ACC, GCC, TAA, CAA, GAA, TGG, CGG, AGG

212, where the first and third species have the same nucleotide but the second species has a different one.

221, where the first and second species have the same nucleotide but the third species has a different one.

123, where all three species have different nucleotides.

TCG, GCT, GAT, GTA, CTG, ACG, ACT, CGT, GTC, TGC, GAC, CGA, GCA, TGA, AGC, TAC, ATG, ATC, CTA, TCA, AGT, CAT, TAG, CAG

Now, denote $f(s) = \frac{4^{s-1}+3\times 2^{s-1}+2}{6}$, the formula given in the problem. Denote E_s as the set of the different patterns for s species, A_s as the set where each element consists of only a single state, B_s as the set where each element consists of exactly two states, C_s as the set where each element consists of three states, and D_s as the set where each element consists of three states. It is easy to see that A_s, B_s, C_s, D_s are disjoint sets. Equivalently,

$$E_s = A_s \cup B_s \cup C_s \cup D_s, \text{with } A_s \cap \mathrm{B_s} = A_s \cap \mathrm{C_s} = \cdots = C_s \cap \mathrm{D_s} = \emptyset.$$

In the above example, $A_3 = \{111\}, B_3 = \{122, 212, 221\}, C_3 = \{123\}$.

1) It is easy to verify that the given formula holds for $s = 1, 2$.

2) When $s = 3$, there are the following five patterns aaa (all the same), baa, aba, aab (one different from the others), and abc (all different), satisfying $f(3) = \frac{4^{3-1}+3\times 2^{3-1}+2}{6} = 5$.

3) When $s = 4$, there are 15 patterns (see below), well meeting the given formula $f(4) =$

$\frac{4^{4-1}+3\times 2^{4-1}+2}{6} = 15$. Further divide the 15 patterns into three disjoint sets according to the number of states that each element has as follows

$$E_4 = A_4 \cup B_4 \cup C_4 \cup D_4,$$

where $A_4 = \{1111\}$, $B_4 = \{1111, 1121, 1211, 2111, 1122, 1212, 1221\}$, $C_4 = \{1123, 1213, 1231, 2311, 2131, 2113\}$, $D_4 = \{1234\}$. We also have the numbers of elements in the above sets $|A_4| = 1, |B_4| = 7, |C_4| = 7$.

4) As to $s = 5$, we have the following result based on the above pattern for $s = 4$.

Divide the set E_5 into four disjoint sets according to the number of states of its elements as follows

$A_5 = \{11111\}$,
$B_5 = (\{2\} \times A_4) \cup (\{1,2\} \times B_4)$,
$C_5 = (\{3\} \times B_4) \cup (\{1,2,3\} \times C_4)$,
$D_5 = (\{4\} \times C_4) \cup (\{1,2,3,4\} \times D_4)$,

and apparently we have $|A_5| = |A_4| = 1$, $|B_5| = 1 + 2 \times |B_4| = 15$, $|C_5| = 1 \times |B_4| + 3 \times |C_4| = 28$, $|D_5| = 1 \times |C_4| + 4 \times |D_4| = 7$. So,

$$|E_5| = 1 + 15 + 28 + 7 = 51 = \frac{4^{5-1} + 3 \times 2^{5-1} + 2}{6} = f(5).$$

5) As to $s = 6$, likewise, we can have the following result based on the above patterns for $s = 5$.

Again, divide set E_5 into three disjoint sets according to the number of states of the elements as follows

$A_6 = \{111111\}$,
$B_6 = (\{2\} \times A_5) \cup (\{1,2\} \times B_5)$,
$C_6 = (\{3\} \times B_5) \cup (\{1,2,3\} \times C_5)$,
$D_6 = (\{4\} \times C_5) \cup (\{1,2,3,4\} \times D_5)$,

and we have $|A_6| = 1$, $|B_6| = 1 + 2 \times |B_5| = 31$, $|C_6| = 1 \times |B_5| + 3 \times |C_5| = 99, |D_6| = 1 \times |C_5| + 4 \times |D_5| = 56$. So,

$$|E_6| = 1 + 31 + 99 + 56 = 187 = \frac{4^{6-1} + 3 \times 2^{6-1} + 2}{6} = f(6).$$

You of course can do similar stuff for $s = 7,8, ...$, but at this point, it may be clear enough to see the following

$$\begin{cases} A_s = \{1\} \times \{A_{s-1}\}, \\ B_s = (\{2\} \times A_{s-1}) \cup (\{1,2\} \times B_{s-1}) \\ C_s = (\{3\} \times B_{s-1}) \cup (\{1,2,3\} \times C_{s-1}) \\ D_s = (\{4\} \times C_{s-1}) \cup (\{1,2,3,4\} \times D_{s-1}) \end{cases} \tag{4.1}$$

Hence, the total number of patterns for any given s can be calculated as

$$\begin{aligned} f(s) &= |E_s| \\ &= |A_s| + |B_s| + |C_s| + |D_s| \\ &= |A_{s-1}| \times 1 + (1 + |B_{s-1}| \times 2) + (|B_{s-1}| + |C_{s-1}| \times 3) + (|C_{s-1}| + |D_{s-1}| \times 4) \\ &= |A_{s-1}| + |B_{s-1}| \times 3 + (|C_{s-1}| + |D_{s-1}|) \times 4 \end{aligned}$$

where $|A_s| \equiv 1$ for any s.

As follows we apply mathematical induction. It is obvious from the above that $f(s)$ holds for $s = 1,2,3,4$. Suppose $f(s) = \frac{4^{s-1}+3\times 2^{s-1}+2}{6}$ holds for all values up to some k.

Compare $f(k+1)$ and $f(k)$, and by eliminating the items containing $|C_{k-1}|$ or $|D_{k-1}|$, we can further show that

$$\begin{aligned} f(k+1) &= 4 \times (|A_{k-1}| + |B_{k-1}| + |C_{k-1}| + |D_{k-1}|) - (|B_{k-1}| \times 2 + 3) \\ &= 4f(k) - (2|B_{k-1}| + 3). \end{aligned}$$

Now our focus shifts to $|B_k|$.The elements of $|B_k|$ consist of two sources. One is $\{\underbrace{aa\ldots a}_{k \text{ times}}\}$, which is a singleton set. The other is composed of those with exactly two states from B_{k-1}. Consider a few different values for s:

$$\begin{aligned} |B_4| &= 7 \times 1 \\ |B_5| &= 1 + 7 \times 2 \\ |B_6| &= 1 + (1 + 7 \times 2) \times 2 \\ |B_7| &= 1 + 2 \times \big(1 + 2 \times (1 + 7 \times 2)\big). \end{aligned}$$

It is not difficult to see the following

$$|B_k| = 2^{k-4} \times 7 + 2^{k-4} - 1,$$

thus

$$|B_{k-1}| = 2^{k-5} \times 7 + 2^{k-5} - 1,$$

where $s \geq 4$. Introducing the expression of $|B_{k-1}|$ to $f(k+1)$, we have

$$\begin{aligned} f(k+1) &= 4 \times \frac{4^{k-1} + 3 \times 2^{k-1} + 2}{6} - 2 \times \big(2^{k-5} \times 7 + 2^{k-5} - 1\big) - 3 \\ &= \frac{2 \times 4^{k-1} + \big(6 \times 2^{k-1} + \big(-48 \times 2^{k-5} + 1\big)\big)}{3} \\ &= \frac{4^k + \big(6 \times 2^k - 3 \times 2^5 \times 2^{k-5}\big) + 2}{6} \qquad (both\ top\ and\ bottom \times 2) \\ &= \frac{4^k + 3 \times 2^k + 2}{6}. \end{aligned}$$

Solution 2.

Denote $a_s = |A_s|, b_s = |B_s|, c_s = |C_s|, d_s = |D_s|, e_s = |E_s|$.

Starting from Eq. (4.1), we can re-write it in the form of matrix as

$$\begin{bmatrix} a_{s+1} \\ b_{s+1} \\ c_{s+1} \\ d_{s+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} a_s \\ b_s \\ c_s \\ d_s \end{bmatrix}.$$

Further denote

$$W = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 4 \end{bmatrix}.$$

Apparently,

$$(a_1, b_1, c_1, d_1)^T = (1,0,0,0)^T.$$

It follows that

$$\begin{bmatrix} a_{s+1} \\ b_{s+1} \\ c_{s+1} \\ d_{s+1} \end{bmatrix} = W \begin{bmatrix} a_s \\ b_s \\ c_s \\ d_s \end{bmatrix} = W^2 \begin{bmatrix} a_{s-1} \\ b_{s-1} \\ c_{s-1} \\ d_{s-1} \end{bmatrix} = \cdots = W^s \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We can do Eigenvalue decomposition for W as follows

$$W = P\Lambda P^{-1}$$

where $\Lambda = \text{diag}(1,2,3,4)$, $P = \begin{bmatrix} -6 & 0 & 0 & 0 \\ 6 & 2 & 0 & 0 \\ -3 & -2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$, $P^{-1} = \begin{bmatrix} -1/6 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ -1/2 & -1 & -1 & 0 \\ 1/6 & 1/2 & 1 & 1 \end{bmatrix}$.

Hence,

$$\begin{bmatrix} a_{s+1} \\ b_{s+1} \\ c_{s+1} \\ d_{s+1} \end{bmatrix} = P\Lambda^s P^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -6 & 0 & 0 & 0 \\ 6 & 2 & 0 & 0 \\ -3 & -2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2^s & 0 & 0 \\ 0 & 0 & 3^s & 0 \\ 0 & 0 & 0 & 4^s \end{bmatrix} \begin{bmatrix} -1/6 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ -1/2 & -1 & -1 & 0 \\ 1/6 & 1/2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -6 & 0 & 0 & 0 \\ 6 & 2 & 0 & 0 \\ -3 & -2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2^s & 0 & 0 \\ 0 & 0 & 3^s & 0 \\ 0 & 0 & 0 & 4^s \end{bmatrix} \begin{bmatrix} -\frac{1}{6} \\ \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{6} \end{bmatrix}$$

$$= \begin{bmatrix} -6 & 0 & 0 & 0 \\ 6 & 2 & 0 & 0 \\ -3 & -2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{6} \\ 2^{s-1} \\ -\frac{3^s}{2} \\ \frac{4^s}{6} \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 2^s - 1 \\ \frac{1}{2} - 2^s + \frac{3^s}{2} \\ 2^{s-1} - \frac{3^s}{2} + \frac{4^s}{6} - \frac{1}{6} \end{bmatrix}.$$

So,

$$f(s+1) = 1 + (2^s - 1) + \left(\frac{1}{2} - 2^s + \frac{3^s}{2}\right) + \left(2^{s-1} - \frac{3^s}{2} + \frac{4^s}{6} - \frac{1}{6}\right)$$

$$= \frac{4^s + 3 \times 2^s + 2}{6}.$$

Hence,

$$f(s) = \frac{4^{s-1} + 3 \times 2^{s-1} + 2}{6}.$$

Solution 3.

This solution uses a combinatorics way. Let X denote the set of all sequences composed of n types of nucleotides. Let Y_i denote the set of sequences where the i-th type of nucleotide does not appear. Let Z denote the set of sequences where all m types of nucleotides appear. Let us take an example. For cases where $m = 3$ and $n = 3$, we have

$$X = \{111,112,113,\dots,333\},$$
$$Y_1 = \{222,223,232,\dots,333\},$$
$$Y_2 = \{111,113,131,\dots,333\},$$
$$Y_3 = \{111,112,121,\dots,222\},$$
$$Z = \{123,132,213,231,312,321\},$$

where $1,2,3$ denote the different states (note again that in this example $m = 3$). It is easy to see

$$|X| = 27, |Y_1| = |Y_2| = |Y_3| = 8, |Z| = 6.$$

In general, we have

$$|X| = m^n,$$
$$|Z| = |X| - |Y_1 \cup Y_2 \cup \dots \cup Y_m|,$$
$$|Y_1| = |Y_2| = \cdots |Y_m| = (m-1)^n,$$
$$|Y_1 \cap Y_2| = |Y_1 \cap Y_3| = \cdots = |Y_{m-1} \cap Y_m| = (m-2)^n,$$
$$|Y_1 \cap Y_2 \cap Y_3| = |Y_1 \cap Y_2 \cap Y_4| = \cdots = |Y_{m-2} \cap Y_{m-1} \cap Y_m| = (m-3)^n,$$
$$\vdots$$
$$|Y_1 \cap Y_2 \cap \dots \cap Y_m| = (m-m)^n = 0.$$

Apply the inclusion-exclusion principle. It follows that

$$\begin{aligned}
&|Y_1 \cup Y_2 \cup \dots \cup Y_m| \\
&= \sum_{i=1}^{m} |Y_i| - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} |Y_i \cap Y_j| + \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{k=j+1}^{m} |Y_i \cap Y_j \cap Y_k| + \cdots \\
&\qquad + (-1)^m |Y_1 \cap Y_2 \cap \dots \cap Y_m| \\
&= \binom{m}{1}(m-1)^n + (-1)^1 \binom{m}{2}(m-2)^n + (-1)^2 \binom{m}{3}(m-3)^n + \cdots \\
&\qquad + (-1)^{m-1} \binom{m}{m}(m-m)^n \\
&= \sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k}(m-k)^n.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
|Z| &= |X| - |Y_1 \cup Y_2 \cup \dots \cup Y_m| \\
&= m^n - \sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k}(m-k)^n
\end{aligned}$$

$$= -(-1)^{0-1}\binom{m}{0}(m-0)^n - \sum_{k=1}^{m}(-1)^{k-1}\binom{m}{k}(m-k)^n$$

$$= -1 \times \left(\sum_{k=0}^{m}(-1)^{k-1}\binom{m}{k}(m-k)^n\right)$$

$$= \sum_{k=0}^{m}(-1)^{k}\binom{m}{k}(m-k)^n.$$

Now, denote in a sequence of length n there are $a_{m,n}$ combinations that have exactly m states (types of nucleotides). Equivalently,

$$a_{m,n} = \sum_{k=0}^{m}(-1)^{k}\binom{m}{k}(m-k)^n.$$

Denote $b_{m,n}$ as the number of patterns with given m and n. Note that $b_{m,n}$ differs from $a_{m,n}$ in that it refers to the number of patterns instead of combinations. As an example, the combination "TTT" and "CCC" refer to the same pattern which can be denoted by "111". Hence,

$$b_{m,n} = \frac{a_{m,n}}{m!}.$$

According to the context, the number of patterns for four types of nucleotides thus $m = 1,2,3,4$ may be calculated as

$$\sum_{m=1}^{4} b_{m,n} = \sum_{m=1}^{4}\frac{a_{m,n}}{m!}.$$

Hence,

$$a_{1,1} = \binom{1}{0}(1-0)^n = 1,$$

$$a_{2,1} = \binom{2}{0}(2-0)^n - \binom{2}{1}(2-1)^n = 2^n - 2,$$

$$a_{3,1} = \binom{3}{0}(3-0)^n - \binom{3}{1}(3-1)^n + \binom{3}{2}(3-2)^n = 3^n - 3\times 2^n + 3,$$

$$a_{4,1} = \binom{4}{0}(4-0)^n - \binom{4}{1}(4-1)^n + \binom{4}{2}(4-2)^n - \binom{4}{3}(4-3)^n$$

$$= 4^n - 4\times 3^n + 6\times 2^n - 4.$$

Therefore, the total number of patterns can be calculated as

$$\sum_{m=1}^{4}\frac{a_{m,n}}{m!} = \frac{1}{1!} + \frac{2^n-2}{2!} + \frac{3^n - 3\times 2^n + 3}{3!} + \frac{4^n - 4\times 3^n + 6\times 2^n - 4}{4!}$$

$$= 1 + 2^{n-1} - 1 + \frac{3^n - 3\times 2^n + 3}{6} + \frac{4^{n-1} - 3^n + 3\times 2^{n-1} - 1}{6}$$

$$= \frac{3\times 2^n + 3^n - 3\times 2^n + 3 + 4^{n-1} - 3^n + 3\times 2^{n-1} - 1}{6}$$

$$= \frac{4^{n-1} + 3 \times 2^{n-1} + 2}{6}.$$

***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 $d\ell/dt = 0$. Show that the transformed parameter $z = e^{-4/3t}$ is a solution to the following quintic equation:

$$36z^5 + 12(6 - 3f_0 - f_1)z^4 + (45 - 54f_0 - 42f_1)z^3 + (33 - 60f_0 - 36f_1)z^2 + (3 - 30f_0 - 2f_1)z + (3 - 12f_0 - 4f_1) \equiv 0. \quad (4.30)$$

Solution.

Define the likelihoods of observing a site with i substitutions as L_i. Let a, b, and c indicate the proportion of the site with 1, 2, and 3 different sites in the alignment respectively. Hence,

$$\ell = n[\mathrm{a} \cdot \log(L_1) + \mathrm{b} \cdot \log(L_2) + \mathrm{c} \cdot \log(L_3)]. \quad (4.2)$$

Recall Eq. (1.3) in (Yang, 2006)

$$P(t) = \begin{bmatrix} p_{0(t)} & p_{1(t)} & p_{1(t)} & p_{1(t)} \\ p_{1(t)} & p_{0(t)} & p_{1(t)} & p_{1(t)} \\ p_{1(t)} & p_{1(t)} & p_{0(t)} & p_{1(t)} \\ p_{1(t)} & p_{1(t)} & p_{1(t)} & p_{0(t)} \end{bmatrix}, \text{with } \begin{cases} p_{0(t)} = \frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3}t} \\ p_{1(t)} = \frac{1}{4} - \frac{1}{4} e^{-\frac{4}{3}t} \end{cases}.$$

Substitute z for $e^{-\frac{4}{3}t}$, it follows that

$$L_0 = \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3}t}\right)^3 + \frac{3}{4}\left(\frac{1}{4} - \frac{1}{4} e^{-\frac{4}{3}t}\right)^3 = \frac{1}{4}\left(\frac{1}{4} + \frac{3}{4} z\right)^3 + \frac{3}{4}\left(\frac{1}{4} - \frac{1}{4} z\right)^3.$$

Differentiate the log-likelihood of L_0 w.r.t. z, we have

$$\frac{d \log(L_0)}{dz} = \frac{\frac{1}{4}\left(\left(\frac{1}{4} + \frac{3}{4} z\right)^3\right)' \left(\frac{3}{4} z\right)' + \frac{3}{4}\left(\left(\frac{1}{4} - \frac{1}{4} z\right)^3\right)' \left(-\frac{1}{4} z\right)'}{\frac{1}{4}\left(\frac{1}{4} + \frac{3}{4} z\right)^3 + \frac{3}{4}\left(\frac{1}{4} - \frac{1}{4} z\right)^3} = \frac{\frac{9}{32}(z^2 + z)}{\frac{3}{32} z^3 + \frac{3}{64} z^2 + \frac{1}{64}} = \frac{18\, z\, (1 + z)}{6\, z^3 + 9\, z^2 + 1} \quad (4.3)$$

Likewise,

$$L_1 = \frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}t}\right)^2\left(\frac{1}{4}-\frac{1}{4}e^{-\frac{4}{3}t}\right)+\frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4}{3}t}\right)\left(\frac{1}{4}-\frac{1}{4}e^{-\frac{4}{3}t}\right)^2+\frac{1}{2}\left(\frac{1}{4}-\frac{1}{4}e^{-\frac{4}{3}t}\right)^3$$
$$= \frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}z\right)^2\left(\frac{1}{4}-\frac{1}{4}z\right)+\frac{1}{4}\left(\frac{1}{4}+\frac{3}{4}z\right)\left(\frac{1}{4}-\frac{1}{4}z\right)^2+\frac{1}{2}\left(\frac{1}{4}-\frac{1}{4}z\right)^3$$

Hence,

$$\frac{d\log(L_1)}{dz} = \frac{2\,(1\,-\,3\,z)z}{-\,2\,z^3+z^2+1} \tag{4.4}$$

For $L2$,

$$L_2 = \frac{3}{4}\left(\frac{1}{4}+\frac{1}{4}e^{-\frac{4}{3}t}\right)\left(\frac{1}{4}-\frac{1}{4}e^{-\frac{4}{3}t}\right)^2+\frac{1}{4}\left(\frac{1}{4}-\frac{1}{4}e^{-\frac{4}{3}t}\right)^3$$
$$= \frac{3}{4}\left(\frac{1}{4}+\frac{3}{4}z\right)\left(\frac{1}{4}-\frac{1}{4}z\right)^2+\frac{1}{4}\left(\frac{1}{4}-\frac{1}{4}z\right)^3,$$

and so

$$\frac{d\log(L_2)}{dz} = \frac{-6\,z}{-\,2\,z^2+z+1} \tag{4.5}$$

By plugging Eqs. (4.3-4.5) into Eq. (4.2), we obtain the log-likelihood of the whole alignment given in the problem as follows (also note that $c = 1-a-b$):

$$\ell = n\left(a\frac{18\,z\,(1\,+\,z)}{6\,z^3+9\,z^2+1}+b\frac{2\,(1\,-\,3\,z)z}{-\,2\,z^3+z^2+1}+(1-a-b)\frac{-6\,z}{-\,2\,z^2+z+1}\right)$$

$$= n\times\frac{-2\,z\,\left(\begin{array}{c}-3\,+\,12\,a\,+\,4\,b\,-\,3\,z\,+\,30\,a\,z\,+\,2\,b\,z\,-\,33\,z^2\,+\,60\,a\,z^2\,+\,36\,b\,z^2\,-\\ 45\,z^3\,+\,54\,a\,z^3\,+\,42\,b\,z^3\,-\,72\,z^4\,+\,36\,a\,z^4\,+\,12\,b\,z^4\,-\,36\,z^5\end{array}\right)}{(-1\,+\,z)(1\,+\,2\,z)(1\,+\,z\,+\,2\,z^2)(1\,+\,9\,z^2\,+\,6\,z^3)}.$$

Setting the above to zero, it is easy to note that it is equivalent to solving the following equation w.r.t. z by setting the numerator to zero, that is to set

$$2\,z\,(36\,a\,z^4+\,54\,a\,z^3+\,60\,a\,z^2+\,30\,a\,z\,+\,12\,a\,+\,12\,b\,z^4+\,42\,b\,z^3+\,36\,b\,z^2$$
$$+\,2\,b\,z\,+\,4\,b\,-\,36\,z^5-\,72\,z^4-\,45\,z^3-\,33\,z^2-\,3\,z\,-\,3)=0$$

Combining the similar terms, we obtain the following

$$-36z^5+(36a-72+12b)z^4+(54a+42b-45)z^3+(60a+36b-33)z^2$$
$$+(30a+2b-3)z+(12a+4b-3)=0,$$

which, if substituting f_0 for a and f_1 for b, is exactly the quintic equation given in the problem. Therefore, $z = e^{-\frac{4}{3}t}$ is a solution to the quintic equation given in the problem.

R

```
> library(Ryacas0)
> z <- Sym("z")
> Simplify(deriv(log(1/4*(1/4+3/4*z)^3+3/4*(1/4-1/4*z)^3), z)) #L0
> Simplify(deriv(log(1/4*(1/4+3/4*z)^2*(1/4-1/4*z)+1/4*(1/4+3/4*z)*(1/4-
1/4*z)^2+1/2*(1/4-1/4*z)^3), z)) #L1
> Simplify(deriv(log(3/4*(1/4+3/4*z)*(1/4-1/4*z)^2+1/4*(1/4-1/4*z)^3), z)) #L2
```

4.3 Calculate the probabilities of sites with data *xxyy*, *xyyx*, and *xyxy* in four species for the unrooted tree of Fig. 4.13, using two branch lengths p and q under a symmetrical substitution model for binary characters (Exercise 1.3). 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).

Solution.

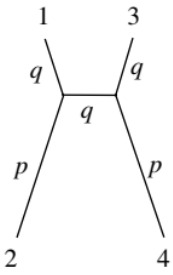
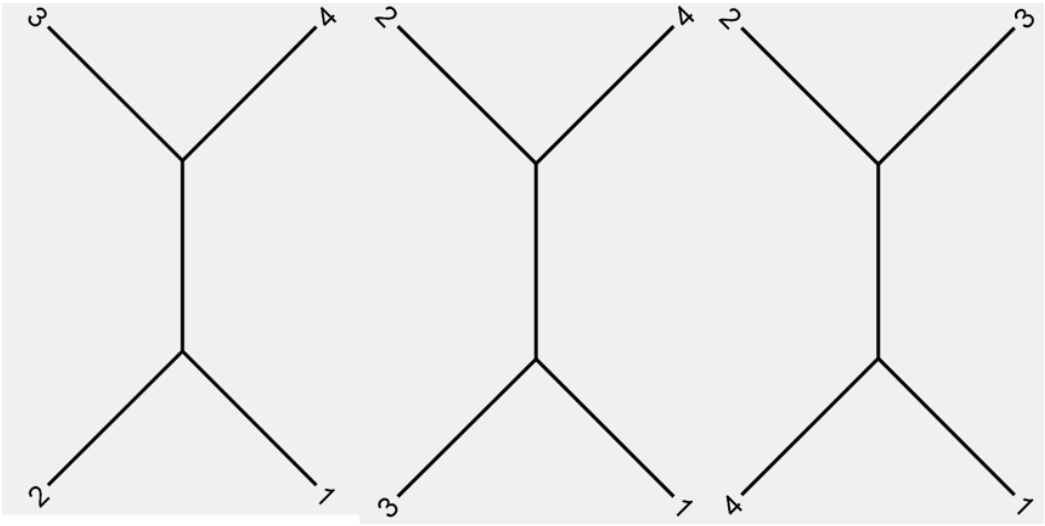


Fig. 4.13 A tree of four species with two branch lengths p and q, defined as the probability that any site is different at the two ends of the branch. For a binary character, this probability is $p = (1-e^{-2t})/2$, where t is the expected number of character changes per site (see Exercise 1.3 in Chapter 1).

As to a four-tip tree, there are three possible tree topologies. They are depicted as *xxyy*, *xyxy*, and *xyyx*, which correspond to ((1,2),3,4), ((1,3),2,4), ((1,4),2,3), respectively.



For tree reconstruction using the most parsimonious (MP) method, only sites with the patterns where two sites have the same nucleotide while the other two have another same nucleotide such as TTCC or TATA, are informative. In other words, a site in the alignment with any other patterns such as AAAA or ATGG do not provide any information to distinguish between any of the possible trees because all of them are equally likely. There have to be at least two nucleotides each occurring twice at least. See Section 3.4 in (Yang, 2006) for more details.

Denote the state at node i as S_i. Hence,

$$P(S_0) = P(S_1) = \ldots = P(S_5) = 0.5.$$

The probability of observing the pattern $xxyy$ by the MP method can be calculated as

$$P(xxyy) = P(S_1 = A, S_2 = A, S_3 = B, S_4 = B) + P(S_1 = B, S_2 = B, S_3 = A, S_4 = A)$$

$$
\begin{aligned}
&= 0.5 \sum_{s_0 \in \{A,B\}} \big(P(S_1 = A, S_2 = A, S_3 = B, S_4 = B | S_0 = s_0) \\
&\qquad + P(S_1 = B, S_2 = B, S_3 = A, S_4 = A | S_0 = s_0)\big) \\
&= 2 \times 0.5 \\
&\qquad \times \big(P(S_1 = A, S_2 = A, S_3 = B, S_4 = B | S_0 = A) \\
&\qquad + P(S_1 = B, S_2 = B, S_3 = A, S_4 = A | S_0 = A)\big) \\
&= P(S_1 = A, S_2 = A, S_3 = B, S_4 = B | S_0 = A) \\
&\qquad + P(S_1 = B, S_2 = B, S_3 = A, S_4 = A | S_0 = A) \\
&= P(S_1 = A, S_2 = A, S_3 = B, S_4 = B, S_5 = A | S_0 = A) \\
&\qquad + P(S_1 = A, S_2 = A, S_3 = B, S_4 = B, S_5 = B | S_0 = A) \\
&\qquad + P(S_1 = B, S_2 = B, S_3 = A, S_4 = A, S_5 = A | S_0 = A) \\
&\qquad + P(S_1 = B, S_2 = B, S_3 = A, S_4 = A, S_5 = B | S_0 = A) \\
&= (1-q)(1-p)(1-q)qp + (1-q)(1-p)q(1-q)(1-p) \\
&\qquad + qp(1-q)(1-q)(1-p) + qpqpq \\
&= 2p^2q^2 - p^2q + (q^3 - 2q^2 + q).
\end{aligned}
$$

Likewise, calculate the probability of obtaining the pattern *xyxy* as

$$
\begin{aligned}
P(xyxy) &= P(S_1 = A, S_2 = B, S_3 = A, S_4 = B) + P(S_1 = B, S_2 = A, S_3 = B, S_4 = A) \\
&= 2 \times 0.5 \\
&\qquad \times \big(P(S_1 = A, S_2 = B, S_3 = A, S_4 = B | S_0 = A) \\
&\qquad + P(S_1 = B, S_2 = A, S_3 = B, S_4 = A | S_0 = A)\big) \\
&= (1-q)p(1-q)(1-q)p + (1-q)pqq(1-p) + q(1-p)(1-q)q(1-p) \\
&\qquad + q(1-p)q(1-q)p \\
&= 2(p^2\, q^2) - \, 3(p^2 q) + \, p^2 + \, (q^2 - \, q^3),
\end{aligned}
$$

and the probability of obtaining the pattern *xyyx* as

$$
\begin{aligned}
P(xyyx) &= P(S_1 = A, S_2 = B, S_3 = B, S_4 = A) + P(S_1 = B, S_2 = A, S_3 = A, S_4 = B) \\
&= (1-q)p(1-q)q(1-p) + (1-q)pq(1-q)p + q(1-p)(1-q)(1-q)p \\
&\qquad + q(1-p)qq(1-p) \\
&= 2(p^2 q^2) - \, p^2 q \, + \, \big(-4(pq^2) + \, 2(pq)\big) + \, q^3.
\end{aligned}
$$

According to the context of the problem, calculate the difference between $P(xxyy)$ and $P(xyxy)$ as

$$
\begin{aligned}
&P(xxyy) - P(xyxy) \\
&\quad = \big(2p^2q^2 - p^2q + (q^3 - 2q^2 + q)\big) - \big(2(p^2\, q^2) - \, 3(p^2 q) + \, p^2 + \, (q^2 - \, q^3)\big) \\
&\quad = (2q-1)(p^2 - q + q^2).
\end{aligned}
$$

Note that from Exercise 1.3 in (Yang, 2006) the range of p and q are both $[0, +\infty)$. Specifically, it is discerned that the minimum value of *p* is rendered as 0, whereas the upper limit of p converges to 1/2 with the progression of time (*t*) towards infinity $\max(p) = \lim_{t\to\infty} \frac{1}{2}(1 - e^{-2t}) = \frac{1}{2}$ (the same for q). Accordingly, $2q - 1 < 0$. Hence, if and only if $p^2 - q + q^2 > 0$ thus $q(1-q) < p^2$, $P(xxyy) < P(xyxy)$ holds.

R code
> library(Ryacas0) > p<-Sym("p"); q<-Sym("q")

```
> f1 <- Simplify( ryacas(expression((1-q)*(1-p)*(1-q)*q*p+(1-q)*(1-p)*q*(1-q)*(1-
p)+q*p*(1-q)*(1-q)*(1-p)+q*p*q*p*q)) )
> f2<- Simplify( ryacas(expression((1-q)*p*(1-q)*(1-q)*p+(1-q)*p*q*q*(1-p)+q*(1-p)*(1-
q)*q*(1-p)+q*(1-p)*q*(1-q)*p)) )
> f3<-Simplify( ryacas(expression((1-q)*p*(1-q)*q*(1-p)+(1-q)*p*q*(1-q)*p+q*(1-p)*(1-
q)*(1-q)*p+q*(1-p)*q*q*(1-p))) ) # not used in this exercise
> f1-f2
```

Chapter 5. Bayesian methods

5.1 (a) In the example of testing for infection in Subsection 5.1.2, suppose that a person tested negative. What is the probability that he has the infection (b) Suppose a person was tested twice and found to be positive both times. What is the probability that he has the infection?

Solution.

a) According to the context, A denotes the event that a person is infected, and B denotes the event that a person tests positive. According to the Bayes' theorem, it follows that

$$\begin{aligned} P(A|\bar{B}) &= \frac{P(A)P(\bar{B}|A)}{P(\bar{B})} \\ &= \frac{P(A)\big(1-P(B|A)\big)}{1-P(B)} \\ &= \frac{P(A)\big(1-P(B|A)\big)}{1-\Big(P(A)P(B|A)+\big(1-P(A)\big)P(B|\bar{A})\Big)}. \end{aligned}$$

Introducing $P(B|A)=0.99,\ P(B|\bar{A})=0.02,\ P(A)=0.001$ as given in Section 5.1.2 of (Yang, 2006), we have

$$P(A|\bar{B}) = \frac{0.001\times(1-0.99)}{1-(0.001\times 0.99-0.999\times 0.02)} = 1.021419e-05$$

b) Denote C as the event that a person tests positive twice. It follows that

$$\begin{aligned} P(A|C) &= \frac{P(A)P(C|A)}{P(C)} \\ &= \frac{P(A)P(B|A)P(B|A)}{P(A)P(C|A)+P(\bar{A})P(C|\bar{A})} \\ &= \frac{P(A)P(B|A)P(B|A)}{P(A)P(B|A)P(B|A)+P(\bar{A})P(B|\bar{A})P(B|\bar{A})} \\ &= \frac{0.001\times 0.99^2}{0.001\times 0.99^2+0.999*0.02^2} \\ &= 0.7103718. \end{aligned}$$

5.2 *Criticism of unbiasedness*. Both likelihood and Bayesian proponents point out that strict adherence to unbiasedness may be unreasonable. For the example of Subsection 5.1.2, consult any statistics textbook to confirm that the expectation of the sample frequency x/n is θ under the binomial model and $\theta(n-1)/n$ under the negative binominal model. Thus the unbiased estimator of θ is $x/n = 9/12$ under the binomial and $x/(n-1) = 9/11$ under the negative binomial. Unbiasedness thus violates the likelihood principle. Another criticism of unbiased estimators is that they are not invariant to reparametrization; if $\hat{\theta}$ is an unbiased estimator of θ, $h(\hat{\theta})$ will not be an unbiased estimator of $h(\theta)$ if h is not a linear function of θ.

Solutions.

We provide two solutions to (b). The first solves it combinatorically, while the second is uses n approach based on infinite series. Note that the answer to (b) as given in the problem might be wrong (see below).

Solution 1.

a) Under the model of binomial distribution:

$$
\begin{aligned}
E(X) &= \sum_{i=0}^{n} i \cdot P(X = i) \\
&= \sum_{i=1}^{n} i \binom{n}{i} \theta^i (1-\theta)^{n-i} \\
&= n\theta \sum_{i=1}^{n} \binom{n-1}{i-1} \theta^{i-1} (1-\theta)^{(n-1)-(i-1)} \\
&= n\theta \sum_{j=0}^{n-1} \binom{n-1}{j} \theta^{j} (1-\theta)^{(n-1)-j} \\
&= n\theta (1-\theta+\theta)^{n-1} \\
&= n\theta.
\end{aligned}
$$

Hence, $E\left(\frac{X}{n}\right) = \theta$ and accordingly x/n is an unbiased estimator of θ.

b) Under the model of negative binomial distribution, given the context, the number of heads $x = 9$ is fixed and the number of tails, denoted as Y, follows a negative binomial distribution NB(9, θ). Hence, the number of total tosses can be denoted as $N = Y + 9$. In the example given in the problem, $N = 12$.

It should be noted however that an unbiased estimator of θ is $\tilde{\theta} = \frac{x-1}{N-1}$. In other words, $E(\tilde{\theta}) = E\left(\frac{x-1}{n-1}\right) = \theta$ (*Negative Binomial Distribution*, 2023). This is different from the "answer" given in the problem. The following shows why $\tilde{\theta} = \frac{x-1}{N-1}$ instead of $\frac{x}{N-1}$ is an unbiased estimator. The expectation of $\tilde{\theta} = \frac{x-1}{N-1}$ may be calculated as follows:

$$
\begin{aligned}
E(\tilde{\theta}) &= \sum_{n=x}^{\infty} \frac{x-1}{n-1} \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x} \\
&= \sum_{i=0}^{\infty} \frac{x-1}{x+i-1} \binom{x+i-1}{x-1} \theta^x (1-\theta)^i \qquad (n = x+i) \\
&= \sum_{i=0}^{\infty} \frac{x-1}{x+i-1} \binom{x+i-1}{i} \theta^x (1-\theta)^i
\end{aligned}
$$

By plugging $\binom{x+i-1}{i} = \frac{x+i-1}{i} \binom{x+i-2}{i-1}$ into the above, we have

$$E(\tilde{\theta}) = \sum_{i=0}^{\infty} \frac{x-1}{x+i-1} \frac{x+i-1}{i} \binom{x+i-2}{i-1} \theta^x (1-\theta)^i$$

$$= \sum_{i=0}^{\infty} \frac{x-1}{i} \frac{(x+i-2)!}{(i-1)!\,(x-1)!} \theta^x (1-\theta)^i$$

$$= \sum_{i=0}^{\infty} \frac{1}{i(i-1)!} \frac{(x+i-2)!}{(x-2)!} \theta^x (1-\theta)^i$$

$$= \sum_{i=0}^{\infty} \frac{(x+i-2)!}{i!\,(x-2)!} \theta^x (1-\theta)^i$$

$$= \sum_{i=0}^{\infty} \binom{x+i-2}{i} \theta^x (1-\theta)^i$$

$$= \theta \sum_{i=0}^{\infty} \binom{x+i-2}{i} \theta^{x-1} (1-\theta)^i$$

$$= \theta \sum_{i=0}^{\infty} \binom{x'+i-1}{i} \theta^{x'} (1-\theta)^i \qquad (x' = x-1)$$

$$= \theta \times 1$$

$$= \theta.$$

Hence, unbiased estimator of θ under the negative binomial model when 3 failures and 9 successes are observed should be $\frac{9-1}{12-1} = \frac{8}{11}$.

Solution 2.

(b) Denote $(x)_n = x \times (x-1) \times \cdots \times (x-n+1)$. This is known as the falling and rising factorial (Graham et al., 1994). Hence,

$$E(\tilde{\theta}) = \sum_{n=x}^{\infty} \frac{x-1}{n-1} \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x}$$

$$= \sum_{n=x}^{\infty} \frac{x-1}{n-1} \times \frac{(n-1)_{x-1}}{(x-1)!} \times \theta^x (1-\theta)^{n-x}$$

$$= \frac{x-1}{(x-1)!} \theta^x \sum_{n=x}^{\infty} \frac{(n-1)_{x-1}}{n-1} (1-\theta)^{n-x}$$

$$= \frac{\theta^x}{(x-2)!} \sum_{n=x}^{\infty} (n-2)_{x-2} (1-\theta)^{n-x}.$$

Consider the summation function

$$f(t) = \sum_{n=x}^{\infty} (n-2)_{x-2} t^{n-x}.$$

It follows that

$$f(t) = \sum_{n=x}^{\infty} \frac{d^{x-2}}{dt^{x-2}}(t^{n-2})$$

$$= \sum_{k=x-2}^{\infty} \frac{d^{x-2}}{dt^{x-2}}(t^k).$$

Denote

$$g(y) = \frac{d^{x-2}}{dt^{x-2}}(t^y).$$

It is easy to see that

$$g(0) = g(1) = \cdots = g(k-3) = 0.$$

Also, noting that $0 < t < 1 - \theta < 1$, the geometric series $\sum_{k=0}^{\infty} t^k$ can be calculated as

$$\sum_{k=0}^{\infty} t^k = \lim_{n\to\infty}(1 + t^1 + t^2 + \cdots t^n)$$

$$= \frac{1}{1-k}.$$

So $f(t)$ may be re-written as

$$f(t) = \sum_{k=0}^{\infty} \frac{d^{x-2}}{dt^{x-2}}(t^k)$$

$$= \frac{d^{x-2}}{dt^{x-2}}\left(\sum_{k=0}^{\infty} t^k\right) \qquad \text{(sum} - \text{differentiation order swapping)}$$

$$= \frac{d^{x-2}}{dt^{x-2}}\left(\frac{1}{1-t}\right)$$

$$= -1 \times (-1) \times (-2) \times \cdots \times (2-x) \times (t-1)^{1-x}.$$

Noting that

$$(-1) \times (-2) \times \cdots \times (2-x) = (-1)^{2-x} \times (x-2) \times (x-1) \times \cdots \times 1$$
$$= (-1)^{2-x}(x-2)!$$

and that

$$(t-1)^{1-x} = (-1)^{1-x} \times (1-t)^{1-x},$$

we have

$$f(t) = -1 \times (-1)^{2-x}(x-2)! \times (-1)^{1-x} \times (1-t)^{1-x}$$
$$= (x-2)!\,(1-t)^{1-x}.$$

So

$$f(1-\theta) = \sum_{n=x}^{\infty} (n-2)_{x-2}(1-\theta)^{n-x} = (x-2)!\,\theta^{1-x},$$

$$f(\theta) = (x-2)!\,(1-\theta)^{1-x}.$$

Plugging this into $E(\tilde{\theta}) = \frac{\theta^x}{(x-2)!}\sum_{n=x}^{\infty}(n-2)_{x-2}(1-\theta)^{n-x}$ (see above), we have

$$E(\tilde{\theta}) = \frac{\theta^x}{(x-2)!}(x-2)!\,\theta^{1-x}$$
$$= \theta.$$

***5.3** Suppose the target density is $N(\theta, 1)$, and the MCMC uses the sliding-window proposal with normal proposals, with the jump kernel $x^* \sim N(x, \sigma^2)$. Show that the acceptance proportion (the proportion at which the proposals are accepted) is (Gelman *et al.* 1996)

$$P_{\text{jump}} = \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right). \tag{5.43}$$

Solutions.

As follows three solutions are provided. The first is the most complicated in calculation but might be the most straightforward to come up with, the second much simplifies the first by a smart trick, while the last one is mainly based on some nice properties of normal distribution.

Solution 1.

Without loss in generality, assume the original normal distribution to be approximated by MCMC is the standard normal distribution $Normal(0,1)$. According to the Metropolis-Hastings algorithm, the acceptance rate for any θ is

$$A(\theta';\theta) = \min\left(1, \frac{\pi(\theta')g(\theta|\theta')}{\pi(\theta)g(\theta'|\theta)}\right).$$

Because the proposal distribution g is the same for all θ, the above can be re-written as

$$\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right).$$

For any x randomly sampled from the standard normal distribution, its acceptance rate is equal to the following

$$P = \iint_{f(y)\geq f(x)} f(x)q(y|x)dydx + \iint_{f(y)<f(x)} f(x)q(y|x)\frac{f(y)}{f(x)}dydx$$

$$= \iint_{f(y)\geq f(x)} f(x)q(y|x)dydx + \iint_{f(y)<f(x)} f(y)q(y|x)dydx$$

$$= \iint_{f(y)>f(x)} f(x)q(y|x)dydx + \iint_{f(y)<f(x)} f(y)q(y|x)dydx + \iint_{f(y)=f(x)} f(y)q(y|x)dydx,$$

where $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2}}, f(y) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2}}$, and $q(y|x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-y)^2}{2\sigma^2}}$.

Denote the following

$$I_1 = \iint\limits_{f(y)>f(x)} f(x)q(y|x)dydx\,,$$

$$I_2 = \iint\limits_{f(y)<f(x)} f(y)q(y|x)\,dydx,$$

$$I_3 = \iint\limits_{f(y)=f(x)} f(y)q(y|x)\,dydx.$$

Apparently,

$$I_3 = 0.$$

Hence,

$$P = I_1 + I_2 + I_3 = I_1 + I_2. \tag{5.1}$$

As to I_2: Because of symmetry of the density, without loss of generality, consider only x that are positive. So I_2 can be written as

$$I_2 = \iint\limits_{f(y)<f(x)} q(y|x)f(y)\,dydx$$

$$= 2\times\left(\frac{1}{2\pi\sigma}\int_0^{+\infty}\int_x^{+\infty} e^{\frac{-x^2-(\sigma^2+1)y^2+2\sigma xy}{2\sigma^2}}dydx + \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_{-\infty}^{-x} e^{\frac{-x^2-(\sigma^2+1)y^2+2\sigma xy}{2\sigma^2}}dydx\right).$$

Denote

$$I_{21} = \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_x^{+\infty} e^{\frac{-x^2-(\sigma^2+1)y^2+2\sigma xy}{2\sigma^2}}dydx,$$

$$I_{22} = \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_{-\infty}^{-x} e^{\frac{-x^2-(\sigma^2+1)y^2+2\sigma xy}{2\sigma^2}}dydx.$$

We have

$$I_2 = 2(I_{21} + I_{22}). \tag{5.2}$$

To calculate I_{21}, apply the following linear transformation

$$\begin{bmatrix}u\\v\end{bmatrix} = \begin{bmatrix}1 & -1\\0 & \sigma\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix},$$

so that we have

$$(x-y)^2 + (\sigma y)^2 = u^2 + v^2,$$

and

$$u = x - y, v = \sigma y$$

$$x = u + \frac{v}{\sigma}, y = \frac{v}{\sigma}.$$

Calculate the determinant of the Jacobian as follows

$$|J| = det\begin{bmatrix}\frac{dx}{du} & \frac{dx}{dv}\\ \frac{dy}{du} & \frac{dy}{dv}\end{bmatrix} = det\begin{bmatrix}1 & \frac{1}{\sigma}\\ 0 & \frac{1}{\sigma}\end{bmatrix} = \frac{1}{\sigma}.$$

For I_{21}, the original integral area is the region enclosed by $x = 0$ and $y = x$, thus $0 <$

$x < y$. It is easy to see that $0 < u + \frac{1}{\sigma}v < \frac{v}{\sigma}$, thus $0 > u > -\frac{v}{\sigma}$. This corresponds to the region enclosed by $u = 0$ and $u = -\frac{1}{\sigma}v$ (thus $v = -\sigma u$). So I_{21} may be rewritten as

$$I_{21} = \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_x^{+\infty} e^{\frac{-x^2-(\sigma^2+1)y^2+2\sigma xy}{2\sigma^2}} dydx$$

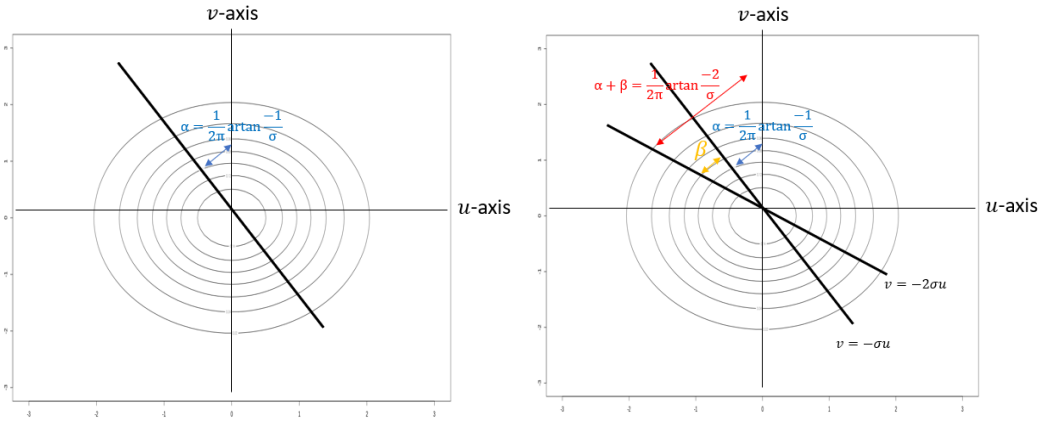
$$= \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_{-\frac{v}{\sigma}}^{0} e^{-\frac{u^2+v^2}{2\sigma^2}} |J| dudv$$

$$= \frac{1}{2\pi\sigma^2}\int_0^{+\infty}\int_{-\frac{v}{\sigma}}^{0} e^{-\frac{u^2+v^2}{2\sigma^2}} dudv$$

$$= \int_0^{+\infty} \frac{1}{\sqrt{(2\pi)\sigma^2}} e^{-\frac{v^2}{2\sigma^2}} dv \int_{-\frac{v}{\sigma}}^{0} \frac{1}{\sqrt{(2\pi)\sigma^2}} e^{-\frac{u^2}{2\sigma^2}} du$$

$$= \frac{\tan^{-1}\left(\frac{1}{\sigma}\right)}{2\pi} \tag{5.3}$$

The last step is because as indicated by the following contour graph of the probability density of a standard bivariate distribution (left), it can be seen that $\int_0^{+\infty} \frac{1}{\sqrt{(2\pi)\sigma^2}} e^{-\frac{v^2}{2\sigma^2}} dv \int_{-\frac{v}{\sigma}}^{0} \frac{1}{\sqrt{(2\pi)\sigma^2}} e^{-\frac{u^2}{2\sigma^2}} du$ is equal to the proportion of $\frac{\tan^{-1}\left(\frac{1}{\sigma}\right)}{2\pi}$ of the probability covered by a standard bivariate normal distribution the latter of which equals 1.



Alternatively, we can apply the following transformation from Cartesian coordinates to polar coordinates $u = rcos\theta$ and $v = rsin\theta$, so it follows that

$$I_{21} = \frac{1}{2\pi\sigma^2}\int_0^{+\infty}\int_{-\frac{v}{\sigma}}^{0} e^{-\frac{u^2+v^2}{2\sigma^2}} dudv$$

$$= \frac{1}{2\pi\sigma^2}\int_0^{+\infty} re^{-\frac{r^2}{2\sigma^2}} dr \int_{\tan^{-1}\left(-\frac{1}{\sigma}\right)}^{0} d\theta$$

$$= \frac{1}{2\pi\sigma^2}\int_0^{+\infty} re^{-\frac{r^2}{2\sigma^2}}dr \int_0^{\tan^{-1}\left(\frac{1}{\sigma}\right)} d\theta$$

$$= \frac{1}{2\pi\sigma^2} \times \frac{\Gamma(1)}{2 \times \frac{1}{2} \times (\frac{1}{\sigma})^2} \cdot \tan^{-1}\left(\frac{1}{\sigma}\right)$$

$$= \frac{\tan^{-1}\left(\frac{1}{\sigma}\right)}{2\pi}.$$

As to I_{22}, the original integral area is the region enclosed $x = 0, y = -x$, which corresponds to the region enclosed by $\mu = -\frac{v}{\sigma}$ and $\mu = -\frac{2v}{\sigma}$ (the right panel in the above figure). According to the sum formula for tangent $\tan(\alpha + \beta) = \frac{tan\alpha + \tan\beta}{1 - tan\alpha \tan\beta}$, it is easy to see that

$$\tan\left(\frac{1}{\sigma} + \beta\right) = -\frac{2}{\sigma} = \frac{-\frac{1}{\sigma} + \tan\beta}{1 + \frac{1}{\sigma}\tan\beta}.$$

By solving the above, we get

$$\tan\beta = -\frac{\sigma}{\sigma^2 + 2}.$$

Hence,

$$I_{22} = \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_{-\infty}^{-x} e^{\frac{-x^2-(\sigma^2+1)y^2+2\sigma xy}{2\sigma^2}}dydx$$

$$= \frac{1}{2\pi\sigma}\int_0^{+\infty}\int_{-\frac{\sigma}{\sigma^2+2}v}^{0} e^{-\frac{u^2+v^2}{2\sigma^2}}|J|dudv$$

$$= \frac{1}{2\pi\sigma^2}\int_0^{+\infty}\int_{-\frac{\sigma}{\sigma^2+2}v}^{0} e^{-\frac{u^2+v^2}{2\sigma^2}}dudv$$

$$= \frac{1}{2\pi\sigma^2}\int_0^{+\infty} e^{-\frac{(x-y)^2}{2\sigma^2}}\int_{-\frac{\sigma}{\sigma^2+2}v}^{0} e^{-\frac{y^2}{2\sigma^2}}dydx$$

$$= \frac{\tan^{-1}\left(\frac{\sigma}{\sigma^2+2}\right)}{2\pi}$$

Note that the last step in the above can be solved by using the same trick applied in deriving Eq. (5.3). Alternatively, it is also possible to use the transformation from Cartesian coordinates to polar coordinates, so that

$$I_{22} = \frac{1}{2\pi\sigma^2}\int_0^{+\infty} e^{-\frac{r^2}{2\sigma^2}}rdr \int_{\tan^{-1}\left(-\frac{\sigma}{\sigma^2+2}\right)}^{0} d\theta$$

$$= \frac{1}{2\pi\sigma^2}\int_0^{+\infty} e^{-\frac{r^2}{2\sigma^2}} r dr \int_0^{\tan^{-1}\left(\frac{\sigma}{\sigma^2+2}\right)} d\theta$$

$$= \frac{1}{2\pi\sigma^2} \times \frac{\Gamma(1)}{2 \times \frac{1}{2} \times (\frac{1}{\sigma})^2} \cdot \tan^{-1}\left(\frac{\sigma}{\sigma^2+2}\right)$$

$$= \frac{\tan^{-1}\left(\frac{\sigma}{\sigma^2+2}\right)}{2\pi}.$$

Further, apply the sum formula for arctangent

$$\arctan(A) + \arctan(B) = \arctan\left(\frac{A+B}{1-AB}\right).$$

It follows that

$$\begin{aligned} I_2 &= I_{21} + I_{22} \\ &= 2 \times \left(\frac{\tan^{-1}\left(\frac{1}{\sigma}\right)}{2\pi} + \frac{\tan^{-1}\left(\frac{\sigma}{\sigma^2+2}\right)}{2\pi} \right) \\ &= \frac{1}{\pi}\tan^{-1}\left(\frac{\frac{1}{\sigma} + \frac{\sigma}{\sigma^2+2}}{1 - \frac{1}{\sigma}\cdot\frac{\sigma}{\sigma^2+2}} \right) \\ &= \frac{1}{\pi}\tan^{-1}\left(\frac{2\sigma^2+2}{\sigma^3+\sigma} \right) \\ &= \frac{1}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right). \end{aligned} \tag{5.4}$$

Similarly, **as to I_1,** without loss of generality, considering $x > 0$, I_1 can be re-written as

$$I_1 = \iint\limits_{f(y)>f(x)} f(x)q(y|x)dydx$$

$$= 2 \times \frac{1}{\sigma} \iint\limits_{f(y)>f(x)} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-y)^2}{2\sigma^2}} dydx$$

$$= \frac{2}{\sigma}\int_0^{+\infty}\int_{-x}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-y)^2}{2\sigma^2}} dydx$$

$$= \frac{2}{\sigma}\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \int_{-x}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-y)^2}{2\sigma^2}} dy.$$

Apply the following linear transformation

$$\begin{bmatrix} z \\ w \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma} & -\frac{1}{\sigma} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

so that

$$x = w + \sigma z, y = w.$$

The determinant of the Jacobian is calculated as

$$|J| = \det\begin{bmatrix} \frac{dx}{dz} & \frac{dx}{dw} \\ \frac{dy}{dz} & \frac{dy}{dw} \end{bmatrix} = \begin{vmatrix} \sigma & 1 \\ 0 & 1 \end{vmatrix} = \sigma.$$

The above can thus be re-written as

$$\begin{aligned} I_1 &= \frac{2}{\sigma}\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw \int_{-x}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}} |J| dz \\ &= \sigma \cdot \frac{2}{\sigma} \cdot \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw \int_0^{\frac{2w}{\sigma}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2}} dz \\ &= 2\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw \int_0^{\frac{2w}{\sigma}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2}} dz \\ &= \frac{1}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right). \end{aligned} \tag{5.5}$$

Finally, by plugging Eqs. (5.3-0) into Eq. (5.1), we obtain

$$P_{jump} = I_1 + I_2 = \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right).$$

Solution 2.

In Solution 1 we see that $I_1 = I_2$ and that calculating I_1 is easier than I_2. As follows, it can be shown that $I_1 = I_2$ is not a coincidence, and as such, it implies that we can greatly simplify the calculation by computing only I_1 and then multiplying it by two to obtain P.

$$\begin{aligned} I_1 &= \iint\limits_{f(y)>f(x)} f(x)q(y|x)dydx \\ &= \iint\limits_{|y|>|x|} f(x)q(y|x)dydx \\ &= \iint\limits_{|x|>|y|} f(y)q(x|y)dxdy \qquad \text{(swapping } x \text{ and } y\text{)} \\ &= \iint\limits_{|x|>|y|} f(y)q(x|y)dydx \\ &= I_2. \end{aligned}$$

The following steps are the same as Solution 1, but there will not be any need to calculate I_2 which is more difficult to calculate than I_1 any more.

Solution 3.

Denote

$$X \sim Normal(0,1),$$
$$Y \sim Normal(X, \theta).$$

Denote also

$$f_{XY}(X=x,Y=y)=\pi(x)q(y|x).$$

According to the above, we have

$$\begin{aligned}P_{jump}&=E\big(\alpha(X,Y)\big)\\&=E\left(\min\left(1,\frac{f_{XY}(X=y,Y=x)}{f_{XY}(X=x,Y=y)}\right)\right)\\&=2\times P\big(f_{XY}(X,Y)<f_{XY}(X,Y)\big).\end{aligned}$$

It can also be shown that

$$\begin{aligned}f_{XY}(X,Y)&=f_X(X)f_Y(Y|X)\\&=\frac{1}{\sqrt{2\pi}}e^{-\frac{(X-\theta)^2}{2}}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y-X)^2}{2\sigma^2}}.\end{aligned}$$

Likewise,

$$f_{XY}(Y,X)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y-\theta)^2}{2}}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{(X-Y)^2}{2\sigma^2}}.$$

Hence,

$$\begin{aligned}P_{jump}&=2P\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(X-\theta)^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y-X)^2}{2\sigma^2}}<\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y-\theta)^2}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(X-Y)^2}{2\sigma^2}}\right)\\&=2P\left(-\frac{(X-\theta)^2}{2}-\frac{(Y-X)^2}{2\sigma^2}<-\frac{(Y-\theta)^2}{2}-\frac{(X-Y)^2}{2\sigma^2}\right)\\&=2P((X-\theta)^2>(Y-\theta)^2)\\&=2P\big((X-Y)(X+Y-2\theta)>0\big).\end{aligned}$$

According to the context, we have

$$\begin{aligned}X-1&\sim Normal(0,1),\\Y-X&\sim Normal(0,\sigma^2).\end{aligned}$$

Also, X and Y are independent variables. Apply the following transformation

$$\begin{cases}U=X-\theta\\V=\dfrac{Y-X}{\sigma}.\end{cases}$$

Equivalently,

$$\begin{cases}X=U+\theta\\Y=\sigma V+U+\theta.\end{cases}$$

So

$$\begin{aligned}P_{jump}&=2P(-\sigma V(\sigma V+2U)>0)\\&=2P(V(\sigma V+2U)<0)\\&=2\big(P(V<0,\sigma V+2U>0)+P(V>0,\sigma V+2U<0)\big)\\&=2\left(P\left(V<0,U>-\frac{\sigma V}{2}\right)+P\left(V>0,U<-\frac{\sigma V}{2}\right)\right).\end{aligned}$$

Using corresponding results from Solution 1, it is clear that

$$P\left(V<0,U>-\frac{\sigma V}{2}\right)=P\left(V>0,U<-\frac{\sigma V}{2}\right)=\frac{\tan^{-1}\left(\frac{2}{\sigma}\right)}{2\pi}.$$

Thus

$$P_{jump} = 2 \times 2 \times \frac{\tan^{-1}\left(\frac{2}{\sigma}\right)}{2\pi} = \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right).$$

5.4 Write a program to implement the MCMC algorithm of Subsection 5.3.2 to estimate the distance between the human and orangtutan 12s rRNA genes under the JC69 model. Use any programming language of your choice, such as BASIC, Fortran, C/C++, Java, or Mathematica. Investigate how the acceptance proportion changes with the window size w. Also implement the proposal of equation (5.34). (Hint: use the logarithms of the likelihood and prior in the algorithm to avoid numerical problems.)

Solution.

Download the mitochondrial genome sequences for human (D38112) and orangutan (NC_001646), and extract the 12s rRNA gene. Align them using your favorite software. However, I am lazy enough to take the precalculated values provided in Section 5.3.2 of (Yang, 2006). Accordingly, the number of aligned sites n = 948, the number of different sites x = 90. Consider a uniform proposal and a flat prior of θ.

(a) According to Eq. (5.32) of (Yang, 2006), the new state θ^* is drawn from a uniform distribution $U\left(\theta - \frac{w}{2}, \theta + \frac{w}{2}\right)$. The likelihood function is given as Eq. (1.42) of (Yang, 2006):

$$L(\theta; x) = f(x|\theta) = \binom{n}{x}\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{n-x}.$$

Because a flat prior of θ is adopted (see above), the acceptance ratio can be simplified as

$$\alpha = \min\left(1, \frac{L(\theta^*; x)}{L(\theta; x)}\right).$$

The MCMC sampling may be performed using the following R code (*5.4a.R*).

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))
}

do_mcmc <- function(w, nsample, n, x){
    ds <- numeric(nsample)
    accepts <- numeric(nsample)
    d <- 0.8; lnl <- lnl_JC69(d, n, x)
    for(i in 2:nsample){
        d_new <- runif(1, d-w/2, d+w/2)
        d_new <- ifelse(d_new>=0, d_new, d)
        lnl_new <- lnl_JC69(d_new, n, x)
        alpha <- min(1, exp(lnl_new-lnl))
        if(runif(1) < alpha){
            d <- d_new
            lnl <- lnl_new
```

```
            accepts[i] <- 1
        }
        ds[i] <- d
    }
    return(list(ds=ds, accepts=accepts))
}


########################################
x <- 90; n<-948
w <- 0.3; nsample <- 100000

for(w in seq(0.02,0.5,0.04)){
    res <- do_mcmc(w=w, nsample=nsample, n=n, x=x)
    ds <- res$ds; accepts <- res$accepts
    burnin <- round(nsample/2)
    ds_after_burnin <- ds[burnin:nsample]
    accepts_after_burnin <- accepts[burnin:nsample]
    acceptance_ratio <-
length(accepts_after_burnin[accepts_after_burnin==1])/length(accepts_after_burnin)
    cat(w, mean(ds_after_burnin), var(ds_after_burnin), acceptance_ratio, '\n', sep="\t")
}
```

The result is displayed as follows.

w	**mean**	**acceptance ratio**
0.02	0.102988	0.8202036
0.06	0.1028488	0.5233295
0.1	0.1027127	0.3461731
0.14	0.1028259	0.250555
0.18	0.1027715	0.1940761
0.22	0.1026357	0.200516
0.26	0.1028359	0.2371753
0.3	0.1031249	0.2712746
0.34	0.1025325	0.2963541
0.38	0.102375	0.3228735
0.42	0.10298	0.3376332
0.46	0.1024573	0.351973
0.5	0.1024287	0.3653127

(b) The proportional shrinking and expanding algorithm is introduced in Section 5.4.4 in (Yang, 2006). The proposal ratio is $\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} = \frac{\frac{1}{\epsilon|\theta|}}{\frac{1}{\epsilon|\theta^*|}} = c$. Accordingly, the acceptance ratio is calculated as

$$\alpha = \min\left(1, \frac{L(\theta^*; x)}{L(\theta; x)} \times c\right),$$

where $c = e^{\epsilon(r-0.5)}$ and $r \sim U(0,1)$.

For simplicity, I do not show the R code but interested readers can use the script *5.4b.R* to generate the result highly similar to the following.

epsilon	mean	acceptance ratio
0.1	0.103152	0.906742
0.2	0.103107	0.815444
0.3	0.102691	0.728265
0.4	0.102586	0.650007
0.5	0.102893	0.579668
0.6	0.102608	0.51731
0.7	0.102648	0.454111
0.8	0.102537	0.417932
0.9	0.102746	0.376893
1	0.102605	0.343573
1.1	0.10277	0.310374
1.2	0.102795	0.284594
1.3	0.102775	0.262255
1.4	0.102722	0.246135
1.5	0.102799	0.226476
1.6	0.102709	0.217176
1.7	0.102803	0.201936
1.8	0.102752	0.188296
1.9	0.102789	0.182356
2.0	0.10292	0.171957

5.5 Modify the program above to estimate two parameters under the JC69 model: the substation rate $\mu = 3\lambda$ and the time of species divergence T, instead of the distance $\theta = 3\lambda \times 2T$. Consider one time unit as 100 million years, and assign an exponential prior $f(T) = (1/m)e^{-T/m}$ for T with mean $m = 0.15$ (15 million year for human–orangutan divergence) and another exponential prior with mean 1.0 for rate μ (corresponding to a prior mean rate of about 1 substitution per 100 million years). Use two proposal steps, one updating T and another updating μ. Change the prior to examine the sensitivity of the posterior to the prior.

Solution.

Run the R script *5.5.R*, I get the following results.

w	Mean (μ)	Mean (T)	Acceptance ratio (μ)	Acceptance ratio (T)
0.1	0.131414	0.5706318	0.3962921	0.7791444
0.2	0.1527619	0.5607569	0.3201336	0.5778084
0.3	0.1207071	0.6603784	0.3180136	0.5381892

0.4	0.09718148	0.8470733	0.3612328	0.5214896
0.5	0.1159755	1.055233	0.3804924	0.4951101
0.6	0.162647	0.5278965	0.3419932	0.3243335
0.7	0.1050719	1.279745	0.4064319	0.448031
0.8	0.1161608	0.9025814	0.4069719	0.3772325
0.9	0.1459897	0.6342053	0.400032	0.3217936
1.0	0.1325197	0.6303445	0.4133117	0.297614

5.6 Modify the program of Exercise 5.4 to estimate the sequence distance under the K80 model. Use the exponential prior $f(\theta) = (1/m)\mathrm{e}^{-\theta/m}$ with mean $m = 0.2$ 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 Subsection 1.4.2.

Solution.

Refer to Section 1.4.2 of (Yang, 2006) to get the number of transitions $n_S = 84$, and the number of transversions $n_V = 6$. The log-likelihood function is given by Eq. (1.48) in (Yang, 2006) as

$$\ell(d, k|n_S, n_V) = (n - n_S - n_V)\log\left(\frac{p_0}{4}\right) + n_S \log\left(\frac{p_1}{4}\right) + n_V \log\left(\frac{p_2}{4}\right),$$

where

$$p_0(t) = \frac{1}{4} + \frac{1}{4}e^{-\frac{4d}{\kappa+2}} + \frac{1}{2}e^{-\frac{2d(\kappa+1)}{\kappa+2}},$$

$$p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-\frac{4d}{\kappa+2}} - \frac{1}{2}e^{-\frac{2d(\kappa+1)}{\kappa+2}},$$

$$p_2(t) = \frac{1}{4} - \frac{1}{4}e^{-\frac{4d}{(\kappa+2)}},$$

according to Eq. (1.10) of (Yang, 2006).

I fix the width of the sliding-window proposal for k to be 10. Interestingly, in a paper published 11 years after (Yang, 2006) was published, the authors performed very detailed MCMC analysis of estimates of the parameters on the same data set (Nascimento et al., 2017), but note that they used different priors for d and k from those used in the present example. The MCMC sampling may be performed using the script *5.6.R*. My result is summarized as follows.

w.d	**Mean(*d*)**	**Mean(*k*)**	**Acceptance ratio (*k*)**	**Acceptance ratio (*d*)**
0.02	0.1053183	20.72473	0.775045	0.7758448
0.06	0.1045369	20.14857	0.5208958	0.5464907
0.1	0.1049112	19.81965	0.3491302	0.3629274
0.14	0.1044798	19.97007	0.2615477	0.275145
0.18	0.104955	21.21373	0.2037592	0.2127574
0.22	0.1043603	20.79053	0.1871626	0.1979604
0.26	0.1055806	19.95754	0.2021596	0.2079584

0.3	0.1044757	19.67916	0.2385523	0.2413517
0.34	0.1047287	20.57366	0.2475505	0.2557489
0.38	0.104306	19.94364	0.275145	0.2729454
0.42	0.1045011	20.4294	0.2909418	0.2853429
0.46	0.1043686	20.83632	0.30014	0.2945411
0.5	0.104588	20.28651	0.3107379	0.3119376

To compare with the MLE, use the following R code which uses the function *optim* to obtain the MLE. The result is $\hat{k} = 0.1045327,\ \hat{d} = 30.8064718$. You can also refer to the estimates given in Section 1.4.2 of (Yang, 2006).

R

```
> mle <- optim(par=c(0.13,2), lnl_K80, n=948, ns=84, nv=6)
> print(mle$par)
```

Chapter 9. Simulating molecular evolution

9.1 Write a small simulation program to study the *birthday problem*. Suppose that there are 365 days in a year and that one's birthday falls on any day at random. Calculate the probability that at least two people out of a group of $k = 30$ people have the same birthday (that is, they were born on the same day and month but not necessarily in the same year). Use the following algorithm. (The answer is 0.706.)

1. Generate $k = 30$ birthdays, by taking 30 random draws from $1, 2, \ldots, 365$.
2. Check whether any two birthdays are the same.
3. Repeat the process 10^6 times and calculate the proportion of times in which two out of 30 people have the same birthday.

Solution:

It is easy to see the result is given by

$$1 - \prod_{i=1}^{30} \frac{365 - i + 1}{365} = 0.7063162.$$

The R code for simulation is as follows.

R
> n<-10^6 > a<-sapply(1:n, function(x){length(unique(sample(1:365,30,T))) == 30}) > print(length(a[a==T])/n)

9.2 Monte Carlo integration (Subsection 5.3.1). Write a small program to calculate the integral $f(x)$ in the Bayesian estimation of sequence distance under the JC69 model, discussed in Subsection 5.1.2. 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

$$f(x) = \int_0^\infty f(\theta) f(x|\theta) \mathrm{d}\theta \simeq \frac{1}{N} \sum_{i=1}^{N} f(x|\theta_i). \tag{9.8}$$

Note that the likelihood $f(x|\theta_i)$ may be too small to represent in the computer, so scaling may be needed. One way is as follows. Compute the maximum log likelihood $\ell_m = \log\{f(x|\hat{\theta})\}$, where $\hat{\theta} = 0.1015$ is the MLE. Then multiply $f(x|\theta_i)$ in equation (9.8) by a big number $e^{-\ell_m}$ so that they are not all vanishingly small before summing them up; that is,

$$\sum_{i=1}^{N} f(x|\theta_i) = e^{\ell_m} \cdot \sum_{i=1}^{N} \exp\left(\log\{f(x|\theta_i)\} - \ell_m\right). \tag{9.9}$$

Solution.

$$f(x|\theta) = \left(\frac{3}{4} - \frac{3}{4} e^{-\frac{4\theta}{3}}\right)^x \left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4\theta}{3}}\right)^{n-x}$$

The MLE of θ can be easily calculated by setting $f(x|\theta)'$ to zero and solving the equation.

$$x log\left(\left(\frac{3}{4}-\frac{3}{4}e^{-\frac{4\theta}{3}}\right)\right)' + (n-x)\log\left(\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4\theta}{3}}\right)\right)' = 0$$

According to Problem 1.5, we already know that the solution to the above equation is given by $\hat{\theta} = -\frac{3}{4} log\left(1-\frac{4x}{3n}\right)$. Plugging $x = 90$ and $n = 948$ into it, we calculate $\hat{\theta}$ = 0.1015.

The R code for the Monte Carlo integral is given as follows. Note that in an exponential distribution mean of 0.2 means the density is rate parameter $\lambda = \frac{1}{0.2} = 5$, and accordingly the probability density function is $f(x) = 5e^{-5x}$. My result is $5.180803 \times 10^{-131}$ which is slightly less than $f(x|\hat{\theta}) = 6.30386e \times 10^{-130}$.

R

```
> N<-10^6
> r<-rexp(N,1/0.2)
> theta_ml <- 0.1015
> mean(exp(log((3/4-3/4*exp(-4*r/3) )^x * (1/4+3/4*exp(-4*r/3) )^(n-x))-theta_ml) *
exp(theta_ml))
```

Alternatively, $\hat{\theta}$ can be numerically estimated in R with the following code.

R

```
> f <- function(t){-( x*log(3/4-(3/4)*t) + (n-x)*log((1/4+3*t/4)) )}
> log(optim(0.2, f)$par)*(-3/4)
```

9.3 Write a small simulation program to study the optimal sequence divergence when two sequences are compared to estimate the transition/transversion rate ratio κ under the K80 model. 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 1000 replicate data sets under the K80 model and analyse it under the same model to estimate d and κ using equation (1.11). Calculate the mean and variance of the estimate $\hat{\kappa}$ across replicate data sets. Each data set consists of a pair of sequences, which can be generated using any of the three approaches discussed in Subsection 9.5.1.

Solution.

According to Eq. (1.9) and Eq. (1.10) in (Yang, 2006), the transition probability matrix is given by

$$P(t) = \begin{bmatrix} p_0(t) & p_1(t) & p_2(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_2(t) \\ p_2(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_2(t) & p_1(t) & p_0(t) \end{bmatrix},$$

where $p_0(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$, $p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}$ and $p_2(t) = \frac{1-p_0(t)-p_1(t)}{2} = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}$. Applying the re-parametrization $d = (\alpha + 2\beta)t$ and $\kappa = \frac{\alpha}{\beta}$ we have

$$p_0(t) = \frac{1}{4} + \frac{1}{4}e^{-\frac{4d}{\kappa+2}} + \frac{1}{2}e^{-\frac{2d(\kappa+1)}{\kappa+2}},$$

$p_1(t) = \frac{1}{4} + \frac{1}{4}e^{-\frac{4d}{\kappa+2}} - \frac{1}{2}e^{-\frac{2d(\kappa+1)}{\kappa+2}}$,

$$p_2(t) = \frac{1}{4} - \frac{1}{4}e^{-\frac{4d}{(\kappa+2)}}.$$

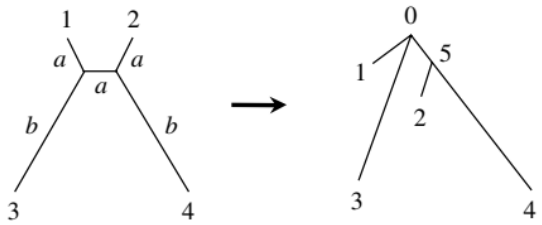
Here, we apply the third method mentioned in Chapter 9 of (Yang, 2006) by **multinomial sampling** with the following R code.

R

```
generate_pair_sequence_k80 <- function(d,k,l){
    p0<-1/4+1/4*exp(-4*d/(k+2))+1/2*exp(-2*d*(k+1)/(k+2))
    p1<-1/4+1/4*exp(-4*d/(k+2))-1/2*exp(-2*d*(k+1)/(k+2))
    p2<-1/4-1/4*exp(-4*d/(k+2))

    P<-matrix(c(p0,p1,p2,p2,p1,p0,p2,p2,p2,p2,p0,p1,p2,p2,p1,p0),ncol=4,byrow=T)
    freq<-rep(0.25,4)
    A <- freq*P

    dna <- c("T","C","A","G")
    dna_pair <- expand.grid(dna,dna)
    dna_pair_char<-paste(dna_pair[,1], dna_pair[,2], sep="")

    sample(dna_pair_char, l, prob=A, replace=T);
}
estimate_d_k_from_seq <- function(sequence){
    n_s <- 0; n_v <- 0
    transition <- c("TC", "CT", "AG", "GA")
    transversion <- c("TA", "TG", "CA", "CG", "AC", "AT", "GC", "GT")
    for(i in names(table(sequence))){
        num <- as.numeric(unname(table(sequence)[i]))
        if(i %in% transition){
            n_s <- n_s + num
        } else if(i %in% transversion){
            n_v <- n_v + num
        }
    }
    s <- n_s/length(sequence)
    v <- n_v/length(sequence)
    d_hat = -1/2*log(1-2*s-v) - 1/4*log(1-2*v)
    k_hat = 2*log(1-2*s-v)/log(1-2*v) - 1
    return(c(d_hat,k_hat))
```

```
}

> for(d in seq(0.01,2,0.01)){
      k_hats <- sapply(1:1000, function(i){
            sequence <- generate_pair_sequence_k80(d=d,k=2,l=500)
            x <- estimate_d_k_from_seq(sequence)
            return(x[2])
      })
      k_hats <- k_hats[!is.na(k_hats) & !is.infinite(k_hats)]
      cat(mean(k_hats), var(k_hats), "\n")
}
```

9.4 Long-branch attraction by parsimony. Use the JC69 model to simulate data sets on a tree of four species (Fig. 9.3a), with two different branch lengths $a = 0.1$ and $b = 0.5$. Simulate 1000 replicate data sets. For each data set, count the sites with the three site patterns *xxyy*, *xyxy*, *xyyx*, and determine the most parsimonious tree. To simulate a data set, reroot the tree at an interior node, as in, say, Fig. 9.3(b). Generate



(a) Unrooted tree (b) Rooted tree

Fig. 9.3 (a) A tree of four species, with three short branches (of length a) and two long branches (with length b) for simulating data to demonstrate long-branch attraction. (b) The same tree rerooted at an ancestral node for simulation.

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 the approach of multinomial sampling. Consider a few sequence lengths, such as 100, 1000, and 10 000 sites.

Solution.

Denote the state (nucleotide) at node i as S_i which can take values $s_i \in \{T, C, A, G\}$. Under the model JC69, the probability of observing $S_1 = s_1, \dots, S_4 = s_4$ where may be calculated as

$$P(S_1 = s_1, S_2 = s_2, S_3 = s_3, S_4 = s_4)$$
$$= \sum_{s_7} P(S_0 = s_0)P(S_1|S_0 = s_0)P(S_3|S_0 = s_0)P(S_2, S_4|S_0 = s_0)$$

$$= 0.25 \sum_{s_0} \left(P(S_1|S_0 = s_0)P(S_3|S_0 = s_0)\right) \left(\sum_{s_5} P(S_5|S_0)P(S_2|S_5 = s_5)\, P(S_4|S_5 = s_5) \right).$$

Use the following R code to, for each of the total sequence length of 100, 1000, and 10000 bases, conduct 1000 simulations of sequence evolution. Then, the most parsimonious (MP) tree(s) inferred from each simulation are summarized. In case where 2 and 3 MP trees are inferred, each is counted as $1/2$ and $1/3$ respectively. The following R code counts the three patterns *xxyy*, *xyxy*, and *xyyx*. See Section 3.4 in (Yang, 2006) or Problem 4.3 in the book for more details in MP tree reconstruction.

R

```
sample2 <- function(d, base, nsamples=1){
    p0 <- 1/4 + 3/4*exp(-4*d/3)
    p1 <- (1-p0)/3
    bases <- c(base, setdiff(BASES, base))
    sample(bases, nsamples, prob=c(p0,rep(p1,3)), replace=T)
}

sim_seq <- function(l){
    seq <-    matrix(0, nrow=4, ncol=l)
    dimnames(seq) = list(paste0('S',1:4), paste0('site',1:l))
    for (i in 1:l){
        for(s0 in sample(BASES, 1, prob=rep(0.25,4), replace=T)){
            list2env( setNames(as.list(sample2(d=0.1, s0, nsamples=2)), paste0('s',
c(1,5))), envir = .GlobalEnv )
            s3 <- sample2(d=0.5, s0)
            s2 <- sample2(d=0.1, s5)
            s4 <- sample2(d=0.5, s5)
            seq[1,i] <- s1
            seq[2,i] <- s2
            seq[3,i] <- s3
            seq[4,i] <- s4
        }
    }
    return(seq)
}

determine_mp <- function(x){
    if(identical(as.integer(table(x)), as.integer(c(2,2)))){
        if(x[1] == x[3]){
            'xxyy'
        } else if(x[1] == x[2]){
            'xyxy'
        } else if(x[1] == x[4]){
```

```
                    'xyyx'
                }
        } else{
                return(NA)
        }
}

> BASES <- c("T", "C", "A", "G")

> for(l in c(100,1000,10000)){
        c_total <- numeric(3)
        names(c_total) <- c('xxyy', 'xyxy', 'xyyx')
        for(n in 1:100){
                seq_matrix <- sim_seq(l=l)
                pattern <- apply(seq_matrix, 2, determine_mp)
                count <- table(pattern[!is.na(pattern)])
                mp_index <- which(count == max(count))
                for(i in names(count[mp_index]) ){
                        c_total[i] <- c_total[i] + 1/length(mp_index)
                }
        }
        print(c_total)
}
```

The result is as follows. *xxyy* means (1,3),(2,4), *xyxy* means (1,2),(3,4), and *xyyx* means (1,4),(2.3).

MP / **tree Length**	***xxyy***	***xyxy***	***xyyx***
100	61.833333	33.333333	4.833333
1000	88.5	11.5	0
10000	100	0	0

9.5 A useful test of a new and complex likelihood program is to generate a few data sets of very long sequences under the model and then analyse them under the same model, to check whether 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 Exercise 9.4 to generate a few data sets of 10^6, 10^7, or 10^8 sites and analyse them using a likelihood program (such as PHYLIP, PAUP, or PAML) under the same JC69 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 data sets; save your important work before this exercise in case of a computer crash.

Solution.

Instead of using any software, use the following R code (**9.5.R**) for branch length estimation. It is based on the famous Felsenstein's pruning algorithm (Felsenstein, 1973). See also Section 4.2.2 in (Yang, 2006) for a detailed explanation of the algorithm. To use the R script, please be aware that the following packages *getopt*, *parallel*, *matrixStats*, *seqinr*, *expm*, *ape, phangorn* may need to be installed.

Write the three possible tree topologies (1,2),(3,4), (1,3),(2,4), (1,4),(2.3) into three files *1.nwk*, *1-2.nwk*, and *1-3.nwk*, respectively. Now, run *9.5.R* by fixing the tree topology for each of the three one by one and compare the log-likelihood. The MLEs of the branch lengths of the tree with the highest likelihood are then compared with the branch lengths used in simulation.

Simulate a sequence of 10^6 nucleotide sites using the script written for Exercise 9.4 and name the simulated alignment as *1.nwk*. Then, for each of the three possible tree topologies, run the following in Shell.

Bash

```
$ for(tree in 1.nwk 1-2.nwk 1-3.nwk){
      echo $tree
      Rscript 9.5.R -t 1.nwk -s 1.aln --cpu 10
      echo
}
```

The log likelihood of using the three tree topologies are respectively -4417007.5, -4433276.3, and -4433276.4. Hence, the first topology, thus the correct topology, is indicated as the best tree. The branch length MLEs of using this tree topology are highly similar to the parameters used in simulation:

```
0.09944458 0.49980742 0.09127108 0.10035842 0.49797523
```

Also confirm this using IQ-Tree (Minh et al., 2020) by the following command in Shell where the MLEs of branch lengths can be found in the file *iqtree.treefile* and the log-likelihood is indicated in the file *iqtree.log* (-4417007.476).

Bash

```
$ iqtree -s 1.aln -m JC -redo -pre iqtree
```

References

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, *22*(3), 240–249. https://doi.org/10.1093/sysbio/22.3.240

Graham, R., Knuth, D., & Patashnik, O. (1994). *Concrete Mathematics*. Addison-Wesley Longman Publishing Co., Inc.75 Arlington Street, Suite 300 Boston, MAUnited States.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nascimento, F. F., dos Reis, M., & Yang, Z. H. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, *1*(10), 1446–1454. https://doi.org/10.1038/s41559-017-0280-x

Negative binomial distribution. (2023). https://en.wikipedia.org/wiki/Negative_binomial_distribution

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl315

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press. http://abacus.gene.ucl.ac.uk/CME/

Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press. http://abacus.gene.ucl.ac.uk/MESA/