Answers – Lecture 3
Reconstruction of phylogenetic trees

Question 1
Question 1a (without `R`*)*
$2m - 2$ edges.

Question 1b (without `R`*)*
$m - 1$ hidden nodes.

Question 2
Question 2a
Let X_0 and X_1 denote the random variables representing the 'binary nucleotide' at the locus of the common ancestor and the first present day organism, respectively. Note that for convenience, the index t is dropped (as only one generation separates the common ancestor and the present day species). Similarly, as only one locus is considered, the index j is dropped.

Making use of the fact that the stationary distribution is $(\frac{1}{2}, \frac{1}{2})^T$ (which follows from the fact that the symmetry of the transition matrix $\mathbf{P}$, see lecture slides), we have:

$$\begin{aligned} P(X_0 \,|\, X_1) &= P(X_1 = 0 \,|\, X_0 = 0)\, P(X_0 = 0)/P(X_1 = 0) \\ &= (1-\alpha) * \frac{1}{2} / [P(X_1 = 0 \,|\, X_0 = 0)\, P(X_0 = 0) + P(X_1 = 0 \,|\, X_0 = 1)\, P(X_0 = 1)] \\ &= (1-\alpha) * \frac{1}{2} / [(1-\alpha) * \frac{1}{2} + \alpha * \frac{1}{2}] \;=\; (1-\alpha). \end{aligned}$$

An alternative solution follows directly from the reversibility of the Markov process (this requires checking whether the detailed balance equations hold).

Question 2b
The loci are independent, thus: $(1-\alpha)^2$.

Question 2c
The loci are independent. Hence, it suffices to study only one. Now note that:

$$\mathbf{P}^t \;=\; \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

This may be verified by evaluating $\mathbf{P}^2 = \mathbf{P} \cdot \mathbf{P}$, and considering $\mathbf{P}^3 = \mathbf{P}^2 \cdot \mathbf{P} = \mathbf{P} \cdot \mathbf{P}$, and so on.

Then:

$$\begin{aligned} P(X_1^{(t)}=0, X_2^{(t)}=1) &= \sum_{x_0=0}^{1} P(X_1^{(t)}=0, X_2^{(t)}=1 \,|\, X_0^{(0)}=x_0)\, P(X_0^{(0)}=x_0) \\ &= \sum_{x_0=0}^{1} P(X_1^{(t)}=0 \,|\, X_0^{(0)}=x_0)\, P(X_2^{(t)}=1 \,|\, X_0^{(0)}=x_0)\, P(X_0^{(0)}=x_0) \\ &= \sum_{x_0=0}^{1} \frac{1}{2}\frac{1}{2}\frac{1}{2} \;=\; \frac{1}{4}. \end{aligned}$$

Put everything together:

$$\begin{aligned} & P(X_{1,1}^{(t)} \neq X_{1,2}^{(t)})\, P(X_{2,1}^{(t)} = X_{2,2}^{(t)}) + P(X_{1,1}^{(t)} = X_{1,2}^{(t)})\, P(X_{2,1}^{(t)} \neq X_{2,2}^{(t)}) \\ &= 2\, P(X_{1,1}^{(t)} \neq X_{1,2}^{(t)})\, P(X_{2,1}^{(t)} = X_{2,2}^{(t)}) \\ &= 2\,[P(X_{1,1}^{(t)}=0, X_{1,2}^{(t)}=1) + P(X_{1,1}^{(t)}=1, X_{1,2}^{(t)}=0)] \\ &\quad \times [P(X_{2,1}^{(t)}=0, X_{2,2}^{(t)}=0) + P(X_{2,1}^{(t)}=1, X_{2,2}^{(t)}=1)] \\ &= 2\,(\frac{1}{4}+\frac{1}{4})\,(\frac{1}{4}+\frac{1}{4}) = \frac{1}{2}. \end{aligned}$$

Question 3
The Kimura has a uniform stationary distribution, and the transition matrix is symmetric. Hence, the detailed balance equations are satisfied.

First determine stationary distribution of the extended model simply assess whether the detailed balance equations: $\varphi_i p_{ij} = \varphi_j p_{ji}$ hold by substitution. E.g.:

$$\begin{aligned} \varphi_1 p_{13} &= \frac{\gamma\delta(\beta+\gamma)}{(\gamma+\delta)(\alpha+\beta+2\gamma)} \\ &\stackrel{?}{=} \frac{\delta\gamma(\alpha+\delta)}{(\gamma+\delta)(\alpha+\beta+2\delta)} \;=\; \varphi_3 p_{31}. \end{aligned}$$

Work through the details and verify that the detailed balance equation do not hold for all acceptable choices of the parameters.

Question 4
Throughout the answers below: Let $X_t^{(1)}$ and $X_t^{(2)}$ denote the random variables representing the nucleotide at the locus of organisms 1 and 2 at time t. The random variable $X_0^{(ca)}$ represents the nucleotide at the locus in the common ancestor.

Question 4a Using conditional independence, one obtains:

$$\begin{aligned} P(X_1^{(1)}=\mathtt{G}, X_1^{(2)}=\mathtt{G} \,|\, X_0^{(ca)}=\mathtt{G}) &= P(X_1^{(1)}=\mathtt{G} \,|\, X_0^{(ca)}=\mathtt{G}) P(X_1^{(2)}=\mathtt{G} \,|\, X_0^{(ca)}=\mathtt{G}) \\ &= (1-3\alpha)^2. \end{aligned}$$

Question 4b Condition on the locus of the common ancestor (and use the fact that the events at the loci in the present day organisms are independent conditional on the locus of the common ancestor).

$$\begin{aligned} P(X_1^{(1)} = \mathtt{G}, X_1^{(2)} = \mathtt{G}) &= \sum_{x_{ca}} P(X_1^{(1)} = \mathtt{G}, X_1^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}) P(X_0^{(ca)} = x_{ca}) \\ &= \sum_{x_{ca}} P(X_1^{(1)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}) P(X_1^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}) P(X_0^{(ca)} = x_{ca}) \\ &= \frac{1}{4} \sum_{x_{ca}} P(X_1^{(1)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}) P(X_1^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}) \\ &= \frac{1}{4} [3\alpha^2 + (1-3\alpha)^2] \end{aligned}$$

Question 4c Simply using the definition of conditional probability:

$$\begin{aligned} P(X_1^{(1)} = \mathtt{G} \,|\, X_1^{(2)} = \mathtt{G}) &= P(X_1^{(1)} = \mathtt{G}, X_1^{(2)} = \mathtt{G}) / P(X_1^{(2)} = \mathtt{G}) \\ &= [3\alpha^2 + (1-3\alpha)^2], \end{aligned}$$

where we have used the answer to 4b and the fact that ...
Question 4d Again, conditional independence gives:

$$P(X_2^{(1)} = \mathtt{G}, X_2^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = \mathtt{G}) = P(X_2^{(1)} = \mathtt{G} \,|\, X_0^{(ca)} = \mathtt{G}) P(X_2^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = \mathtt{G}).$$

Further note that (using the total probability law):

$$\begin{aligned} P(X_2^{(1)} = \mathtt{G} \,|\, X_0^{(ca)} = \mathtt{G}) &= \sum_{x_1} P(X_2^{(1)} = \mathtt{G}, X_1^{(1)} = x_1 \,|\, X_0^{(ca)} = \mathtt{G}) \\ &= \sum_{x_1} P(X_2^{(1)} = \mathtt{G} \,|\, X_1^{(1)} = x_1) P(X_1^{(1)} = x_1 \,|\, X_0^{(ca)} = \mathtt{G}) \\ &= (1-3\alpha)^2 + 3\alpha^2. \end{aligned}$$

Substitute this in the first equation to obtain the answer.

Question 4e Condition on the locus of the common ancestor:

$$\begin{aligned} P(X_2^{(1)} = \mathtt{G}, X_2^{(2)} = \mathtt{G}) &= \sum_{x_{ca}} P(X_2^{(1)} = \mathtt{G}, X_2^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}) P(X_0^{(ca)} = x_{ca}) \\ &= \frac{1}{4} \sum_{x_{ca}} P(X_2^{(1)} = \mathtt{G}, X_2^{(2)} = \mathtt{G} \,|\, X_0^{(ca)} = x_{ca}). \end{aligned}$$

The remaining probability is simplified analogously to the previous question, as conditional on the locus of the common ancestor the events in the present day organisms are independent.

Question 4f Simply using the definition of conditional probability:

$$\begin{aligned} P(X_2^{(1)} = \mathtt{G} \,|\, X_2^{(2)} = \mathtt{G}) &= P(X_2^{(1)} = \mathtt{G}, X_2^{(2)} = \mathtt{G}) / P(X_2^{(2)} = \mathtt{G}) \\ &= 4P(X_2^{(1)} = \mathtt{G}, X_2^{(2)} = \mathtt{G}). \end{aligned}$$

Now substitute answer of the previous question.

Note: in the last three subquestions an alternative strategy may to be to first calculate $\mathbf{P}^2$. This may simplify things above.

Question 5

Question 5a See lecture slides for the spectral decomposition of the Jukes-Cantor model, from which one obtains:

$$P(X_t^{(1)} = \mathtt{A} \,|\, X_0^{(ca)} = \mathtt{A}) \;=\; \frac{1}{4} + \frac{3}{4}(1-4\alpha)^t.$$

Question 5b

$$P(X_t^{(2)} = \mathtt{T} \,|\, X_0^{(ca)} = \mathtt{A}) \;=\; \frac{1}{4} - \frac{1}{4}(1-4\alpha)^t.$$

Question 5c

$$\begin{aligned}
P(X_t^{(1)} = \mathtt{A},\, X_t^{(2)} = \mathtt{T}) &= \sum_{x_{ca}} P(X_t^{(1)} = \mathtt{A},\, X_t^{(2)} = \mathtt{T} \,|\, X_0^{(ca)} = x_{ca}) P(X_0^{(ca)} = x_{ca}) \\
&= \frac{1}{4} \sum_{x_{ca}} P(X_t^{(1)} = \mathtt{A} \,|\, X_0^{(ca)} = x_{ca}) P(X_t^{(2)} = \mathtt{T} \,|\, X_0^{(ca)} = x_{ca}) \\
&= \frac{1}{4}\{2\,[\frac{1}{4} + \frac{3}{4}(1-4\alpha)^t][\frac{1}{4} - \frac{1}{4}(1-4\alpha)^t] + 2\,[\frac{1}{4} - \frac{1}{4}(1-4\alpha)^t]^2\} \\
&= \frac{1}{32}\{(1+3u)(1-u) + (1-u)^2\} \\
&= \frac{1}{16}(1+u)(1-u) \\
&= \frac{1}{16}(1-u^2)
\end{aligned}$$

where $u = (1-4\alpha)^t$.

Question 5d
For $\alpha = \frac{1}{8}$, we have:

$$P(X_t^{(1)} = \mathtt{A},\, X_t^{(2)} = \mathtt{T}) \;=\; \frac{1}{16}[1 - (\frac{1}{2})^{2t}].$$

This probability increases with t. Hence, the species separated a long time ago.

Question 5e
Higher in a non-coding region, as this is less well-preserved.

Question 6
To be inserted.

Question 7
From the lecture notes we have:

$$L(\mathbf{X}) \;=\; \prod_{j=1}^{p} \prod_{k_1,k_2 \in \{\mathtt{A},\mathtt{C},\mathtt{T},\mathtt{G}\}} [\pi_{k_1} \mathbf{P}^{2t}_{k_1,k_2}]^{I_{\{X_{1j}=k_1, X_{2j}=k_2\}}}.$$

Define the following quantities:

$$\begin{aligned}
A &= \sum_{j=1}^{p} I_{\{X_{1j}=X_{2j}\}} \\
B &= \sum_{j=1}^{p} I_{\{X_{1j}\neq X_{2j},\text{ and both purines}\}} \\
&\quad + \sum_{j=1}^{p} I_{\{X_{1j}\neq X_{2j},\text{ and both pyrimidines}\}} \\
C &= \sum_{j=1}^{p} I_{\{X_{1j}\neq X_{2j},\ X_{1j}\text{ purine},\ X_{1j}\text{ pyrimidine}\}} \\
&\quad + \sum_{j=1}^{p} I_{\{X_{1j}\neq X_{2j},\ X_{1j}\text{ pyrimidine},\ X_{1j}\text{ purine}\}}
\end{aligned}$$

Note that $p = A + B + C$.

The likelihood may now be written as (using the spectral decompositions):

$$\begin{aligned}
L(\mathbf{X}) \;=\; & \frac{1}{4}\big[\frac{1}{4} + \frac{1}{4}(1-4\beta)^t + \frac{1}{2}(1-2\alpha-2\beta)^t\big]^A \\
& \times \big[\frac{1}{4} + \frac{1}{4}(1-4\beta)^t - \frac{1}{2}(1-2\alpha-2\beta)^t\big]^B \\
& \times \big[\frac{1}{4} - \frac{1}{4}(1-4\beta)^t\big]^C,
\end{aligned}$$

where we have used the fact that the stationary distribution is uniform. Or, when $u = (1-4\beta)^t$ and $v = (1-2\alpha-2\beta)^t$:

$$L(\mathbf{X}) \;=\; \frac{1}{4}\big(\frac{1}{4}\big)^p (1+u+2v)^A\,(1+u-2v)^B\,(1-u)^C.$$

Now take the logarithm:

$$\log[L(\mathbf{X})] \;\propto\; A\log(1+u+2v) + B\log(1+u-2v) + C\log(1-u).$$

Equate the first order partial derivatives to zero:

$$\begin{aligned}
\frac{\partial \log[L(\mathbf{X})]}{\partial u} &= \frac{A}{1+u+2v} + \frac{B}{1+u-2v} - \frac{C}{1-u} \;=\; 0, \\
\frac{\partial \log[L(\mathbf{X})]}{\partial u} &= \frac{2A}{1+u+2v} - \frac{2B}{1+u-2v} \;=\; 0.
\end{aligned}$$

The second equation gives:

$$v \quad = \quad \frac{1}{2}\frac{A-B}{A+B}(1+u).$$

The first equation, when combined with the second, yields:

$$\frac{2A}{1+u+2v} \quad = \quad \frac{C}{1-u}.$$

Or, rewritten: $2A(1-u) = C(1+u+2v)$. Finally, combine the two solutions and solve for u:

$$\hat{u} \quad = \quad \frac{2\frac{A}{B} - \left[1+\frac{A-B}{A+B}\right]}{2\frac{A}{B} + \left[1+\frac{A-B}{A+B}\right]} \quad = \quad \frac{2A}{A+B} \quad = \quad 1 - 2\frac{C}{p},$$

and

$$\hat{v} \quad = \quad \frac{1}{2}\frac{A-B}{p-C}(1+1-2\frac{C}{p}) \quad = \quad \frac{A-B}{p}.$$