STAT 517: Stochastic Modeling II **Winter 2023**

Homework 02 - Partial solutions

Instructor: Abel Rodriguez *TA: Apara Venkat*

Note: This is only a high-level sketch of the solutions. If you catch any typos or have questions, post them in the discussion board.

(a) We have the Q matrix

$$Q = \begin{bmatrix} - & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \gamma\pi_T \\ \beta\pi_A & \beta\pi_G & \gamma\pi_C & - \end{bmatrix},$$

where the diagonals are such that rows sum to 0. It is easy to verify that $v = \pi$ satisfies $Q^\top v = 0$.

(b) Easy to verify the time reversibility condition, $q^*_{ij} = q_{ij}$ where $q^*_{ij} = \pi_j q_{ji}/\pi_i$ for all i, j.

(c) Can be simulated by using code from Lab 2. Just run the code for those time periods indicated by the branch length and use the final state for the observed node sequence.

(d) I will write down the likelihood for just one DNA base $X = X^{(1)}$ (since the bases are independent and identical, just multiply all likelihoods). Suppose the root node is X_0, the left child is X_1, the right child is X_2, and the two grandchildren are X_3, X_4 (left to right). Then the likelihood can be written as

$$\mathcal{L}(\theta; X_0, \ldots, X_4) = P_{X_0,X_1}(1) P_{X_0,X_2}(0.65) P_{X_2,X_3}(0.35) P_{X_2,X_4}(0.3)$$

where $\theta = (\pi_A, \pi_G, \pi_C, \alpha, \gamma)$ are the parameters (note π_T can be calculated from other $\pi_\cdot$'s). We have $P_{ij}(t) = e^{Qt}$, which can be estimated as well. Now just run a numerical solver, like R's `optim`, to maximize the log-likelihood.

(e) There are a few ways to proceed with this. One is just marginalizing over X_0, X_2. If we assume the process is already stationary, then

$$\mathcal{L}'(\theta; X_1, X_3, X_4) = \sum_{X_2} \sum_{X_0} \pi_{X_0} P_{X_0,X_1}(1) P_{X_0,X_2}(0.65) P_{X_2,X_3}(0.35) P_{X_2,X_4}(0.3).$$

Now, just use the same numerical solver as in part (d).

Another method is to use the EM algorithm. Initialize $\theta := \theta^{(0)}$. This contains two steps

(i) E-step:

$$Q(\theta \mid \theta^{(t-1)}) = \mathbb{E}_{X_0,X_2|X_1,X_3,X_4,\theta^{(t-1)}} \left[\ell(\theta^{(t-1)}; X_0, \ldots, X_4) \right]$$

(ii) M-step:

$$\theta^{(t)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta \mid \theta^{(t-1)})$$

We repeat this until some numerical convergence criterion for $\theta^{(t)}$ is satisfied.