Homework #9
Selected solutions

6.5 Exercises

20. The formula for e^{Qt} for the Jukes-Cantor model in equation (6.11) and its predecessor can be used to understand the effect of an infinite amount of mutation, by letting $t \to \infty$.

    a. If $\alpha > 0$, then $\lim_{t\to\infty} e^{-\frac{4}{3}\alpha t} = 0$, because e^x has an asymptote at 0 as $x \to -\infty$.

    b. For Jukes-Cantor, the Markov matrix e^{Qt} is determined by the parameter $a = \frac{3}{4}(1 - e^{-\frac{4}{3}\alpha t})$. By the results of (a), as $t \to \infty$, $a \to \frac{3}{4}$. So $1 - a = 0.25$, and $\frac{a}{3} = 0.25$ as well.
    This is what we would expect, as we know from analyses on the last homework that as time progresses, the rows of the Matrix converge to the stable distribution.

    c. Mathematically, because if $\alpha = 0$, then $a = 0$ for all t, and we obtain the identity in (b) rather than the given matrix. Biologically, this would correspond to the rate of mutation being 0, which is an uninteresting case.

22. a. If you choose notation so that the two JC parameters are a_1 and a_2 then when you compute the product you get a JC Markov matrix with parameter $a = a_1 + a_2 - \frac{4}{3}a_1a_2$.

    b. If the Hamming distance were additive, then we would have that $a_3 = a_1 + a_2$. Instead, we obtain a smaller number since $\frac{4a_1a_2}{3} < 0$.

    c. If a_1 and a_2 are small, then the product $a_1a_2 \to 0$ quickly and so $a_3 \approx a_1 + a_2$.

23. One space-saving way to do this is to notice that the product of two K3ST Markov matrices commute and therefore the product is also symmetric. (Quick proof: Since A, B are simultaneously diagonalizable, they commute. Since A, B are symmetric, their product is $AB = A^TB^T = B^TA^T = (AB)^T$.) Then you can just commute four entries of the product to get the correct formulas. They are not elegant.

24. In a), checking for eigenvalues $0, -2(\beta+\delta), -2(\beta+\gamma), -2(\gamma+\delta)$ is straightforward matrix vector product.

27. Let u denote a column vector with all entries 1. Then

$$\mathbf{p}M = u^T \operatorname{diag}(\mathbf{p})M = u^TM^T \operatorname{diag}(\mathbf{p}) = (Mu)^T \operatorname{diag}(\mathbf{p}) = u^T \operatorname{diag}(\mathbf{p}) = \mathbf{p},$$

    and $\mathbf{p}$ is a stable base distribution for M.

30 b. Since the expression given is the sum of the off-diagonal entries weight by $\mathbf{p}$, that expression can be thought of as the total mutation rate. As we discussed in class, because of parameter non-identifiability issues it is natural to set this to one as a normalization so that branch lengths t are measured in units of expected number of substitutions per site.

7.4 Exercises

1. The Hamming distance $\hat{a} = \frac{5}{40} = 0.1250$. Thus,

$$d_{JC}(S0, S1) = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{a}\right) \approx 0.1367.$$

2. a. The Hamming distance $\hat{a}$ is $\frac{41}{400} = 0.1025$, so $d_{JC} \approx 0.1102$.

    b. The proportion of transitions is $\hat{b} = \frac{14}{400} = 0.035$, and the proportion of transversions is $\hat{e} = \frac{27}{400} = 0.0675$. Thus, the K2P distance is

$$d_{K2P} = -\frac{1}{2}\log(1 - 2\hat{b} - \hat{e}) - \frac{1}{4}\log(1 - 2\hat{e}) = 0.1102.$$

c. The answers are not identical, but close.

3. a. From the data, $\hat{a} = \frac{77}{400} = 0.1925$, so

$$d_{JC} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{a}\right) \approx 0.2225.$$

Furthermore, $\hat{b} = \frac{58}{400} = 0.145$ while $\hat{e} = \frac{19}{400} = 0.0475$. Thus,

$$d_{K2P} = -\frac{1}{2}\log(1 - 2\hat{b} - \hat{e}) - \frac{1}{4}\log(1 - 2\hat{e}) \approx 0.2308.$$

b. The K2P distance is likely to be the better estimate, because the data are from an underlying K2P model. Distances computed from more complex models are likely to overfit the data.

c. The Kimura 3-parameter distance for Exercise 2 is 0.1105, while the log-det distance is 0.1106. For Exercise 3, the K3ST distance is 0.2309, and the log-det distance is 0.2338. These distances are not as apppropriate, because they're over-parameterizations of a simple system.