

BCB/GDCB/STAT/COM S 568 Spring 2010

Notes

March, 2010

Derivation of the Jukes-Cantor model for nucleotide substitution.

Consider a site in an alignment of two related nucleotide sequences subject to mutational changes during a time period from time t to time $t + \Delta t$. Let the proportion of aligned sites that are different between the two sequences at time t be p_t and the proportion of identical sites $q_t = 1 - p_t$. The Jukes-Cantor model for nucleotide substitution assumes equal probability for a given nucleotide to change into any other. Let $\alpha\Delta t$ be the probability that nucleotide X mutates to nucleotide Y in time period Δt . Then, according to the model, the probability of a nucleotide to remain unchanged in Δt is $1 - \gamma\Delta t$, where $\gamma = 3\alpha$.

Now let us evaluate the changes across the alignment of the two sequences. Given q_t , we can derive $q_{t+\Delta t}$ as follows:

$$q_{t+\Delta t} = q_t(1 - \gamma\Delta t)^2 + (1 - q_t)[2\alpha\Delta t(1 - \gamma\Delta t) + 2(\alpha\Delta t)^2].$$

The first term in the above equation accounts for the sites where the two sequences are identical at time t and no change occurs in either sequence, with probability $(1 - \gamma\Delta t)^2$ (changes in either sequence are assumed independent events). The second term accounts for the sites where the two sequences differ at time t but either one of the two sequences mutates to match the unchanged nucleotide of the other (occurring with probability $2\alpha\Delta t(1 - \gamma\Delta t)$) or both sequences change in parallel to one of the two other nucleotides (occurring with probability $2(\alpha\Delta t)^2$). Assuming that Δt is suitably small such that the probability of two mutation events during this time period is vanishingly small, we may ignore all terms involving $(\Delta t)^2$ to give

$$q_{t+\Delta t} = q_t(1 - 2\gamma\Delta t) + (1 - q_t)2\left(\frac{\gamma}{3}\right)\Delta t,$$

where we have substituted $\frac{\gamma}{3}$ for α . Rearrangement gives

$$\frac{q_{t+\Delta t} - q_t}{\Delta t} = -\frac{8}{3}\gamma q_t + \frac{2}{3}\gamma$$

Interpreting t as the time since the two sequences evolved independently from a common ancestor, we arrive at the differential equation

$$\frac{dq}{dt} = -\frac{8}{3}\gamma q_t + \frac{2}{3}\gamma, \quad q_0 = 1$$

with solution

$$q_t = \frac{3}{4}e^{-\frac{8}{3}\gamma t} + \frac{1}{4}.$$

The evolutionary distance of the two sequences may be defined as $d = 2\gamma t$ (an estimate of the total number of substitutions incurred by the sequences in time t). Substituting and recalling $p_t = 1 - q_t$ gives

$$p_t = \frac{3}{4}(1 - e^{-\frac{4}{3}d})$$

and thus

$$d = -\frac{3}{4} \ln(1 - \frac{4}{3}p_t).$$

Note that d is only defined when p_t (the fraction of changed sites at time t) is less than $\frac{3}{4}$. Given the simplifying assumptions of the model, the model applies at most in the range where q_t (the fraction of unchanged sites) goes from 1 initially to the 0.25 value that represents the fraction expected by chance in unrelated sequences.