Assignment 5 – Simulation of data on trees

We will learn in this assignment to generate data on a pre-specified tree. This method is used for many situations, such as testing computer programs, parametric bootstrapping, and very recently also for approximative Bayes methods (Majoram et al. 2004, Beaumont 2004) [citations needs checking]. The general principle for simulating data follows simple rules.

1. We need a mutation model, for example the Hasegawa-Kishino-Jano model and specify the transition rate matrix:

$$Q = \begin{pmatrix} -\mu(\pi_C + \kappa\pi_G + \pi_T) & \mu\pi_C & \kappa\mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \kappa\pi_T) & \mu\pi_G & \kappa\mu\pi_T \\ \kappa\mu\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_C + \pi_T) & \mu\pi_T \\ \mu\pi_A & \kappa\mu\pi_C & \mu\pi_G & -\mu(\pi_A + \kappa\pi_C + \pi_G) \end{pmatrix}$$

2. Start at the root of the tree with an arbitrary nucleotide, for example picked at random from the $\{A, C, G, T\}$.

3. The rate matrix tells us what are the possible rates of change starting with this arbitrary nucleotide, for example a G at the root. We only need to consider

$$\begin{pmatrix} - & - & - & - \\ - & - & - & - \\ \kappa\mu\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_C + \pi_T) & \mu\pi_T \\ - & - & - & - \end{pmatrix}.$$

   We know that the branch length from the root to the next left node is v_l and to the next right node is v_r. We only consider one branch at a time. The next mutation will arrive after an exponentially distributed waiting time.

4. We draw an exponentially distributed random number

$$t = \frac{1}{\lambda} ln(\text{uniform}(0, 1]),$$

   λ is the rate of change, for our purposes this is the rate at which the G is changing to an A or C or T. G turns into an A with rate $\kappa\mu\pi_A$. These rates can be turned into probabilities by normalizing, therefore the probability of having an A when starting with G is

$$\frac{\kappa\mu\pi_A}{\mu(\kappa\pi_A + \pi_C + \pi_T)}.$$

   The probability from G to C or T can be calculated similarly. Let's assume that we get $\text{Prob}(A|G) = 0.833, p(C|G) = 0.125, \text{Prob}(T|G) = 0.042$. We draw another uniform(0,1)

random number and if it falls between 0.0 and 0.833 G turns into A, if the random number falls between 0.833 and 0.958 it will be a C and if it falls between 0.958 and 1 it will be a T.

5. If $t > v$ we know that no change happened during time v and we advance to the node at the end of the v_i branch and restart at 3. If $t < v_i$ then we know that the nucleotide G changed to another nucleotide within v at or before position t.

6. Advance to the point t, reset λ and recalculate the probabilities dependent on the current nucleotide, evaluate a new t', if $t + t' > v$ continue at 3, otherwise redo this step.

Use this recipe for all branches up to the tips and record the found nucleotides for each tip. Then do this fro as many nucleotides as needed. at the end write the sequences at the tipnodes to a file (Nexus or plain).

Assignment

- Pick a mutation model (either K2P, F81, HKY, GTR) and generate 100 datasets and measure the number of variable sites and its standard deviation for the 12 species tree. Each sequence needs to be 1000 basepairs long. We should be able to recreate your simulations, therefore we need a random number seed, and the parameter settings, and of course the results.

- [Optional] Tabulate the result for different mutation rates, for example $\mu = \{10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$. Describe the the change of the number variable sites.

- [Optional] Tabulate the result for different length of basepairs, for example $\{10, 100, 1000, 10000, 100000, ...\}$. Describe the the change of the number variable sites.

- [Optional] Compare the number of variables sites generated by the JC and the HKY model with a $\kappa = 10$.