

Homework no. 5  
Due Sunday, May 9

Write a program that

- Given a value of the population size  $N$  (perhaps  $10^6$  or so) of a diploid species, uses Kingman's coalescent to simulate a tree for a sample of 15 copies of a gene. Recall that Kingman showed that the time (in generations) back to a coalescence, when there are presently  $k$  sequences, is drawn from an exponential distribution with expectation  $4N/(k(k-1))$  generations, and when that coalescence occurs, it is between two random lineages. (How do you draw a random variate from an exponential? If you draw a uniform random fraction  $R$  from 0 to 1, and then take  $-\ln(R)$ , that is exponentially distributed with mean 1. If you multiply that by  $4N/(k(k-1))$  it will be the exponential variate we need). Set up the tree in memory (recording with it the branch lengths in generations) and then
- Starting at the bottom with a random sequence of 1000 bases use a Jukes-Cantor model with a value of  $\mu$  (the mutation rate per site per generation) that you provide, to simulate the evolution of sequences along the tree. Use a value of  $\mu$  that makes  $4N\mu = 0.01$  (this is higher than actual values). You start at the bottom of the tree with a random sequence which has equal probabilities of all four bases. You will need to compute, for each branch, the probability of a change along the branch, then choose for each site, independently, whether or not it shows a net change along that branch, and decide which of the three other bases to change to if there is a change. Note that the probability of a change can be computed from the formula from the Jukes-Cantor model (not the distance formula but the probability of change formula). Each site evolves independently, although on the same tree.
- Show me the tree and the values of  $N$  and of  $\mu$ , and the resulting sequences. The way to print out the sequences for this exercise is to show the first sequence, and then for each of the other 14 sequences, print out a period if it is the same at that site as the first sequence, otherwise print out the base. This makes the differences in the sequences immediately visible.

A word about random events: To decide whether an event of probability  $p$  is to happen, draw a uniform variate from 0 to 1 – if it is less than  $p$ , the event happens.

(Just for your own information, you might run your UPGMA program on the Jukes-Cantor distances computed from the sequences. Does the tree look exactly the same as the one that was used for the simulation?)