Genome 541 Spring, 2004
Computational Molecular Biology J. Felsenstein

Homework no. 7
Due Friday, May 7

Write a program to

- Given a value of the population size N (perhaps 10^6 or so) of a diploid species, use Kingman's coalescent to simulate a tree for a sample of 15 copies of a gene. Recall that Kingman showed that the time back to a coalescent, 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. 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 500 bases, use a Jukes-Cantor model with a value of μ (the mutation rate per site per generation) that you provide, to simulate the evolution of sequences along the tree. 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, each branch, the probability of a change along the branch, choose for each site, independently, whether it shows a net change along that branch, and decide which of the three other bases to change to if there is a change. Each site evolves independently, although on the same tree.

- Show me the tree and the values of N and of μ, and the resulting sequences.

Be sure to keep $4N\mu$ well below 1 (perhaps 0.01 to 0.001) to be realistic.

A few notes on simulation. You will need a way to draw a random uniform fraction from 0 to 1. A random exponential variate with mean 1 is minus the natural logarithm of a uniform fraction. A random exponential with mean A is just A times that. To decide whether an event of probability p is to happen, draw a uniform variate – if it is less than p, the event is to happen.

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