Stochastic Models in Mathematical Genetics
MSc Problem Sheet 1

Michaelmas Term 2020

1. In a coalescent tree of three haplotypes, let T_3, T_2 be the times respectively while there are three and two ancestors of the haplotypes. Derive a formula for the density of the time to the most recent common ancestor $W = T_3 + T_2$.

2. Four haplotypes can have two possible different unlabelled coalescent trees. Sketch the two trees and work out the respective probabilities of their occurrence.

3. Simulate five coalescent trees of ten haplotypes and sketch them. If $T_{10}, \ldots, T_2$ are times while there are $10, 9, \ldots, 2$ edges in the coalescent tree, then a distribution identity useful for simulation is
$$T_j = -\binom{j}{2}^{-1} \log U_j, \quad j = 10, \ldots, 2,$$
where $U_{10}, U_9, \ldots, U_2$ are independent uniform random variables on (0,1).

4. Let $T_n, T_{n-1}, \ldots, T_2$ be the times while there are $n, n-1, \ldots, 2$ ancestors of a sample of n genes. In a coalescent model these times are distributed as independent exponential random variables with means $2/n(n-1), \ldots, 2/2(2-1)$. Mutations occur along the edges of the coalescent tree as a Poisson process of rate $\theta/2$, conditional on edge lengths.

   (a) Derive formulae for the mean and variance of the time $T_n + \cdots + T_2$ to the most recent common ancestor of the sample.

   (b) Show that the probability generating function of the number of mutations M_n on the coalescent tree is
   $$\prod_{j=1}^{n-1} \left(1 - \frac{(z-1)\theta}{j}\right)^{-1}.$$

   (c) Using the probability generating function find a formula for
   (i) $P(M_2 = m)$
   (ii) $P(M_3 = m)$
   for $m = 0, 1, \ldots$.

   (d) In a sample of DNA sequences, suppose that every mutation happens at a unique site. This is called the infinitely-many-sites model, and means the number of segregating sites in any sample is the same as the number of mutations in the history of that sample. Let d_{ij} be the number of sites which differ between sequences i and j, $i \neq j$ and Π_n be the average of the pairwise site differences defined by
   $$\Pi_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} d_{ij}.$$
   Find the expected value of Π_n, and hence derive a moment estimator of θ based on Π_n.

(e) Define the total site heterozygosity as

$$H_n = \sum_{i=1}^{s} \frac{2r_i(n - r_i)}{n(n-1)},$$

where r_i is the number of sequences with a mutation at site i and s is the number of segregating sites. Show that $\Pi_n = H_n$.

(f) A sample of 200 DNA sequences has 18 segregating sites and $H_{200} = 1.9$. Assuming the infinitely-many-sites model, find two estimates of θ;

(i) based on the number of segregating sites

(ii) based on H_{200}.