Stochastic Models in Mathematical Genetics
MSc Problem Sheet 3

Michaelmas Term 2020

1. Three DNA sequences are observed to have the following mutation pattern at three sites, with X showing mutant sites.

$$\begin{array}{cccccccccccccc} - & - & - & - & - & - & - & X & - & - & - & - & - & - & - \\ - & - & - & X & - & - & - & X & - & - & - & - & - & - & - \\ - & - & - & - & - & - & - & - & - & - & - & X & - & - & - \end{array}$$

   Assume an infinitely-many-sites model of mutation with parameter θ. In this model mutations occur on the edges of a coalescent tree of a sample of genes at rate $\theta/2$.

   (a) Sketch a gene tree that is equivalent to the sample configuration of mutations.

   (b) Sketch a coalescent tree with mutations that would give rise to such a sample.

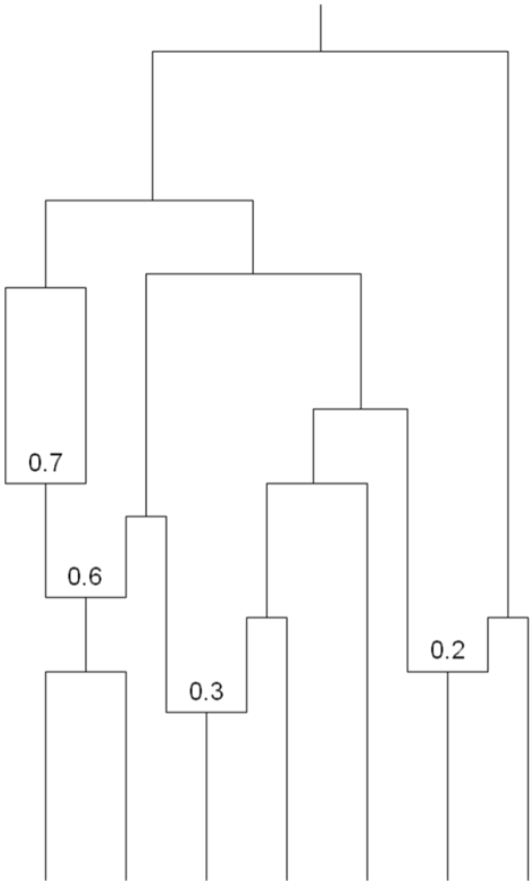
   (c) By arguing directly from (b) find the probability of the sample of sequences as a function of θ.

   (d) Find an equation that the maximum likelihood estimator $\widehat{\theta}$ satisfies, using the result in (c). (You are not required to find a solution of the equation for $\widehat{\theta}$.)

   (e) If the ancestral type at mutant sites was unknown an unrooted gene tree should be considered. If this was the case for the sequences shown above, sketch the unrooted gene tree.

   (f) Sketch the possible rooted gene trees that might produce the unrooted tree in (e).

2. Consider the following ancestral recombination graph for a sample of size 7, in a region of the genome modelled as $[0, 1]$:



(a) Sketch the marginal trees at 0.15, 0.55, 0.75.

(b) How many recombination events occur in genetic material which is ancestral to the sample, for these data?

(c) Under the infinite-sites model, suppose mutations occur at positions 0.1, 0.58, and 0.8 somewhere on the genealogy. Why do you already know the marginal trees at these sites?

(d) Show that the possible sets of mutation carriers in the sample are identical for the mutations at positions 0.58 and 0.8. In what sense are the marginal trees at these positions equivalent?

(e) Identify for which pairs of sites among $(0.1, 0.58)$, $(0.1, 0.8)$, and $(0.58, 0.8)$, it is possible for the four gamete test to reveal recombination.

3. Consider the following dataset for seven DNA sequences. Assume mutation occurs according to the infinite-sites model, and that the ancestral type is <u>not</u> known.

Site	1	2	3	4	5	6
Position	0.1	0.15	0.3	0.6	0.7	0.9
Sequence1	1	0	1	0	0	0
Sequence2	1	0	0	1	0	0
Sequence3	0	0	0	1	0	0
Sequence4	0	1	0	1	0	0
Sequence5	0	1	0	0	0	1
Sequence6	0	0	0	0	1	1
Sequence7	0	0	0	0	1	0

(a) Calculate Hudson's R_m for these data, based on the pairwise incompatibility test for all pairs of sites.

(b) By considering the variation patterns at sites 1, 2, and 4 respectively, or otherwise, show that R_m is less than the true minimum number of recombination events in the sample history. Produce a matrix of recombination bounds, and deduce a lower bound on the value of the best possible "haplotype bound" R_h for these data.

4. Show that the changes in the number of ancestral edges j in the coalescent with recombination with recombination rate ρ, occur as a random walk with the probability of moves to the right and left given by $\frac{\rho}{\rho+j-1}$ and $\frac{j-1}{\rho+j-1}$, respectively. Define p_j^m to be the probability that, starting from position j, the walk hits position 1 before position m. Show that

$$p_j^m = \frac{\rho}{\rho + j - 1} p_{j+1}^m + \frac{j-1}{\rho + j - 1} p_{j-1}^m$$

and state the boundary conditions for the system.

5. (Harder) Solve the system of equations in Question 4 to obtain p_j^m (as a ratio of sums), and deduce for any sample size n, a most recent common ancestor is certain to be reached backward in time.