Stochastic Models in Mathematical Genetics
Problem Sheet 3

Class 3, Michaelmas Term 2020

1. (This is a previous exam question.) A model of ancestry of a sample of n haplotypes in mathematical genetics is the coalescent process, where the number of ancestral lineages back in time is a death process with death rates $\mu_k = k(k-1)/2$, $k = n, n-1, \ldots, 2$. Mutations also occur on ancestral lineages back in time at a rate of $\theta/2$.

   (a) Show that the mean time to the most recent common ancestor of a sample of n haplotypes is $2(1 - n^{-1})$.

   (b) Show that the probability generating function of S_n, the number of mutations occurring to ancestors of the sample, is
   $$\prod_{j=1}^{n-1} \left[1 - \frac{\theta(z-1)}{j}\right]^{-1}.$$

   (c) Find $E(S_n)$ from the probability generating function in (b).

   (d) Show that S_n has an approximate Poisson distribution with mean $\lambda = \theta \log(n)$ as $n \to \infty$ with λ fixed.

   (e) If the n haplotypes were DNA sequences with s segregating sites observed, find an estimate of θ under the infinitely-many-sites model of mutation.

   (f) Suppose during known times $t_n, \ldots, t_2$ the numbers of mutations occurring to ancestral sequences were $a_n, \ldots, a_2$. Find the maximum likelihood estimate of θ.

2. (This is a previous exam question.) Three DNA sequences are observed to have the following mutation pattern at three sites, with X showing mutant sites.

–	–	–	–	–	–	–	X	–	–	–	–	–	–	–
–	–	–	X	–	–	–	X	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–	–	–	X	–	–	–

   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 sequences 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.

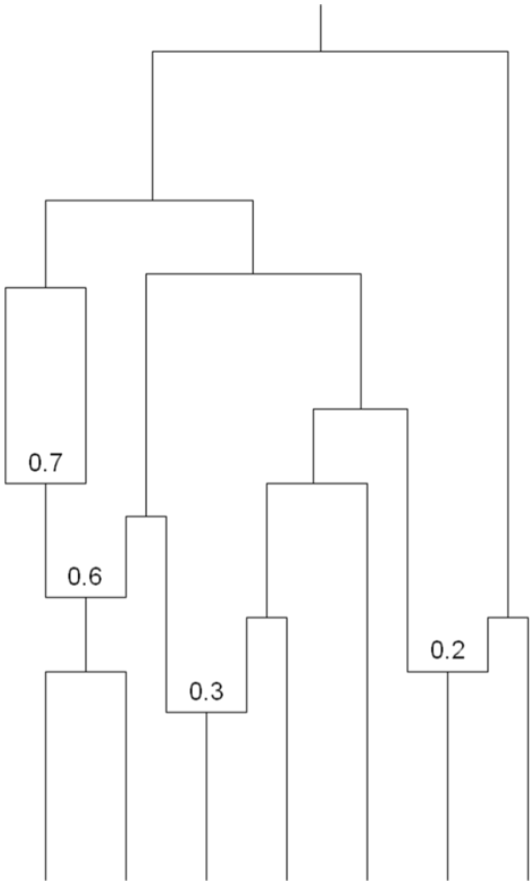
   (c) By arguing directly from (b) find the probability of the sample of sequences as a function of θ.

   (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).

3. 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.

4. 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.

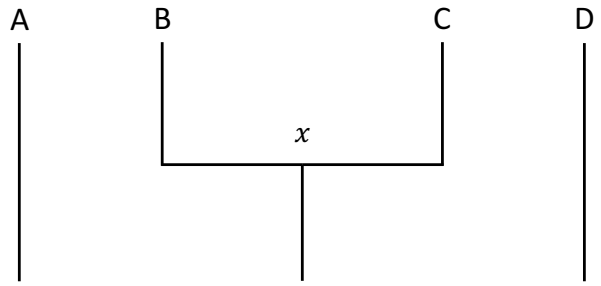
Site	1	2	3	4	5	6
Position	0.1	0.15	0.3	0.6	0.7	0.9
Sequence1	1	0	1	0	0	0
Sequence2	1	0	0	1	0	0
Sequence3	0	0	0	1	0	0
Sequence4	0	1	0	1	0	0
Sequence5	0	1	0	0	0	1
Sequence6	0	0	0	0	1	1
Sequence7	0	0	0	0	1	0

(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.

(c) (Harder) Show that an ancestral recombination graph for the sample can be generated with 3 recombination events. Deduce that 3 is the true minimum number of recombination events in the sample history.

5. Consider a sample of $n = 3$ individuals. Suppose also that a single recombination event occurs, at position $0 < x < 1$, in the history of the sample, while $n = 3$ ancestors remain, so that the lower part of the ancestral recombination graph (ARG) is as follows:



No further recombination occurs. Assume also that all pairs of lineages are equally likely to coalesce at any given time, and that coalescence events involving > 2 lineages never occur.

(a) Sketch two possible ancestral recombination graphs (ARGs) for the sample.

(b) By considering the regions $[0, x)$ and $(x, 1]$, explain why there are at most two different marginal trees along $[0, 1]$.

(c) Show that if the ancestral lineages B and C coalesce before either coalesces with another lineage, there is only a single marginal tree along all of $[0, 1]$.

(d) By considering possible ARGs, show that the probability that lineages B and C coalesce before either coalesces with another lineage is $4/18 = 2/9$.

(e) Under the infinite-sites model, declare the recombination event as *potentially detectable* if by placing two mutations suitably on the ancestral recombination graph, $R_m = 1$. For potentially detectable events, where in $[0, 1]$ must these mutations occur, in order to give $R_m = 1$.

(f) Deduce that the probability the recombination events is potentially detectable is at most $7/9$.

(g) Show that this upper bound is not tight, by finding an ARG where the recombination event is not potentially detectable but where the condition of (d) is not met.

(h) (Harder). By considering possible ARGs, show the actual probability the recombination event is potentially detectable is $4/9$.

6. 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. Solve this system of equations 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.

7. Consider the expected number E_j of recombination events in the history of a sample of size j for a coalescent with recombination rate ρ. Based on the random walk formulation of question 6, show that this satisfies the equation system

$$E_j = \frac{j-1}{\rho+j-1}E_{j-1} + \frac{\rho}{\rho+j-1}E_{j+1} + \frac{\rho}{\rho+j-1}.$$

What is the boundary condition on this system? By integration of an appropriate quantity by parts, show that if we set $E_j = \rho\int_0^1 \frac{1-(1-x)^{j-1}}{x}e^{\rho x}dx$, then E_j satisfies the required equation system. What happens to the expected number of events in the ARG for sample of size 2, as ρ becomes large?

8*. Consider the expected number of recombination events in the "small" ARG (ancestral recombination graph) for a sample of size n, when the overall recombination rate is ρ. By considering recombination events in only ancestral material, and in the full ARG (Q 7), show that his expectation, $\mathbb{E}[R_n^{\mathrm{S}}]$ satisfies

$$\rho\int_0^1 \frac{1-(1-x)^{n-1}}{x}dx \leq \mathbb{E}[R_n^{\mathrm{S}}] \leq \rho\int_0^1 \frac{1-(1-x)^{n-1}}{x}e^{\rho x}dx.$$

Deduce that as $n \to \infty$, the number of recombination events in the full ARG occurring outside ancestral material remains bounded.

Questions with a * are more challenging.