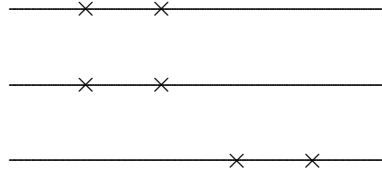


## Stochastic Models in Mathematical Genetics Problem Sheet 4

1. Three DNA sequences are observed to have the following mutation pattern at four sites, with  $\times$  indicating mutant sites.



Assuming an infinitely-many-sites model :

- (a) Sketch a gene tree that is equivalent to the sample configuration.
- (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  $\theta$ .
- (d) Find the maximum likelihood estimate of  $\theta$  from (c).

What would the estimate be based on just the number of segregating sites?

- (e) Find the joint probability density function of  $T_3, T_2$ , the times while there are three ancestors and two ancestors of the sample, conditional on the mutation pattern on the sequences.
- (f) Show that the expected time to the most recent common ancestor  $T_2 + T_3$ , conditional on the mutation pattern in the sequences, is

$$\frac{\int_0^\infty \int_0^\infty t_2^2(t_2 + t_3)^3 \exp\left(-(1 + \theta)t_2 - 3(1 + \theta/2)t_3\right) dt_2 dt_3}{\int_0^\infty \int_0^\infty t_2^2(t_2 + t_3)^2 \exp\left(-(1 + \theta)t_2 - 3(1 + \theta/2)t_3\right) dt_2 dt_3}$$

- (g) Find a formula for the mean time in (f) by evaluating the integrals.

Calculate the mean time numerically at the maximum likelihood estimate of  $\theta$ . Compare this time with the unconditional mean TMRCA.

*Q2 and Q3 are previous exam questions.*

Q2. A model of ancestry of a sample of  $n$  genes in mathematical genetics is the coalescent process, where the number of ancestral lineages back in time is a death process with death rates  $\mu_k = k(k - 1)/2$ ,  $k = n, n - 1, \dots, 2$ . Mutations also occur on ancestral lineages back in time at a rate of  $\theta/2$ .

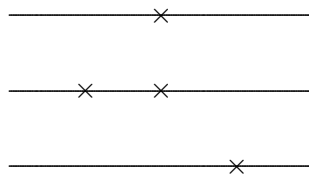
- (a) Show that the mean time to the most recent common ancestor of a sample of  $n$  genes is  $2(1 - n^{-1})$ .
- (b) Show that the probability generating function of  $M_n$ , the number of mutations occurring to ancestors of the sample, is

$$\prod_{j=1}^{n-1} \left[ 1 - \frac{\theta(z - 1)}{j} \right]^{-1}.$$

- (c) Find  $E(M_n)$  from the probability generating function in (b).
- (d) Show that  $M_n$  has an approximate Poisson distribution with mean  $\lambda = \theta \log(n)$  as  $n \rightarrow \infty$  with  $\lambda$  fixed.

- (e) If the  $n$  genes were DNA sequences with  $s$  segregating sites observed, find an estimate of  $\theta$  under the infinitely-many-sites model of mutation.
- (f) Suppose during known times  $t_1, \dots, t_2$  the numbers of mutations occurring to ancestral genes were  $a_1, \dots, a_2$ . Find the maximum likelihood estimate of  $\theta$ .

Q3. Three DNA sequences are observed to have the following mutation pattern at three sites, with  $\times$  showing mutant sites.



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

- Sketch a gene tree that is equivalent to the sample configuration of mutations.
- Sketch a coalescent tree with mutations that would give rise to such a sample.
- By arguing directly from (b) find the probability of the sample of sequences as a function of  $\theta$ .
- Find an equation that the maximum likelihood estimator  $\hat{\theta}$  satisfies, using the result in (c). (You are not required to find a solution of the equation for  $\hat{\theta}$ .)
- 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.
- Sketch the possible rooted gene trees that might produce the unrooted tree in (e).