# Stochastic Models in Mathematical Genetics
## Problem Sheet 2

Class 2, Michaelmas Term 2020

1. 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 $n$ sequences. 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 $\Pi_n$ 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 $\Pi_n$, and hence derive a moment estimator of $\theta$ based on $\Pi_n$.

   (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 $\theta$;
   (i) based on the number of segregating sites
   (ii) based on $H_{200}$.
   If the data were from humans with an effective gene population size of $N = 20,000$, and assumed generation time of 20 years, and the sequences were of length 3,000 bases, what are the estimates of the mutation rate per base per year?

2. The data set below shows the segregating sites in a sample of DNA sequences.

| Site | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lineage | | | | | | | | | | | | | | | | | | | Freq |
| *a* | A | G | G | A | A | T | C | C | T | C | T | T | C | T | C | T | T | C | 2 |
| *b* | A | G | G | A | A | T | C | C | T | T | T | T | C | T | C | T | T | C | 2 |
| *c* | G | A | G | G | A | C | C | C | T | C | T | T | C | C | C | T | T | T | 1 |
| *d* | G | G | A | G | A | C | C | C | C | C | T | T | C | C | C | T | T | C | 3 |
| *e* | G | G | G | A | A | T | C | C | T | C | T | T | C | T | C | T | T | C | 19 |
| *f* | G | G | G | A | G | T | C | C | T | C | T | T | C | T | C | T | T | C | 1 |
| *g* | G | G | G | G | A | C | C | C | T | C | C | C | C | C | C | T | T | T | 1 |
| *h* | G | G | G | G | A | C | C | C | T | C | C | C | T | C | C | T | T | T | 1 |
| *i* | G | G | G | G | A | C | C | C | T | C | T | T | C | C | C | C | C | T | 4 |
| *j* | G | G | G | G | A | C | C | C | T | C | T | T | C | C | C | C | T | T | 8 |
| *k* | G | G | G | G | A | C | C | C | T | C | T | T | C | C | C | T | T | C | 5 |
| *l* | G | G | G | G | A | C | C | C | T | C | T | T | C | C | C | T | T | T | 4 |
| *m* | G | G | G | G | A | C | C | T | T | C | T | T | C | C | C | T | T | C | 3 |
| *n* | G | G | G | G | A | C | T | C | T | C | T | T | C | C | T | T | T | C | 1 |

Ward, R. H., Frazier, B. L., Dew, K. and Paabo, S. (1991) Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Nat. Acad. Sci. USA* **88** 8720-8724.

(a) Assuming the infinitely-many-sites model estimate $\theta$, the mutation parameter by using:

(i) the number of segregating sites,

(ii) the number of alleles in the sample.

(b) Use Gusfield's algorithm to construct a gene tree from the sequences, taking the root sequence as lineage $k$. Sketch the gene tree.

3. The data set below shows the segregating sites in a sample of 50 DNA sequences. Suppose $b$ is the ancestor sequence of the sample.

| Site | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Lineage | | | | | | | | Freq |
| *a* | A | G | T | C | T | G | G | 1 |
| *b* | A | G | C | T | C | G | G | 40 |
| *c* | A | G | T | C | C | G | G | 6 |
| *d* | G | A | T | C | C | A | A | 1 |
| *e* | A | A | T | C | C | A | A | 1 |
| *f* | A | A | T | T | C | A | A | 1 |

(a) Show that sites 4 and 7 are incompatible with the assumption of mutation occurring only once at each site.

(b) Omitting sequence $f$ use Gusfield's algorithm to construct a gene tree from the sequences. Sketch the gene tree. Show how sequence $f$ may have arisen from the tree by an additional mutation at site 4.

4. In a sample of DNA sequences $n-1$ sequences are identical containing $m \geq 1$ mutations since the most recent common ancestor of the sample. The sample also contains one sequence identical to the most recent common ancestor. Show that the probability of obtaining the sample is

$$2(n-2)!\left(\frac{\theta}{2(1+\theta)}\right)^m \prod_{j=1}^{n-1} \frac{1}{j+\theta}$$

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

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| – | – | – | – | X | – | – | – | – | X | – | – | – | – | – | – | – | – | – | – | – | – |
| – | – | – | – | X | – | – | – | – | X | – | – | – | – | – | – | – | – | – | – | – | – |
| – | – | – | – | – | – | – | – | – | – | – | – | – | – | X | – | – | – | X | – | – | – | – |

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

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

(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.

Questions with a * are more challenging.