
Do not turn this page until you are told that you may do so

1. In a coalescent tree $\mathcal{T}$ of n sequences, for $j = n, n-1, ..., 2$ the times T_j while j ancestors of the sample remain are independent and T_j has an exponential distribution with rate parameter $\binom{j}{2}$. At the times of coalescence events, pairs of edges are chosen at random, and coalesce. Mutations occur on edges of the coalescent tree at rate $\theta/2$, according to the infinite-sites model.

(a) (13 marks)

(i) At a time in the past with k ancestral lineages, let $Z_1, Z_2, \ldots, Z_k$ be the number of sequences descended from each respective labelled lineage. Prove that

$$P\left(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k\right) = \binom{n-1}{k-1}^{-1}$$

where $z_1, z_2, \ldots, z_k$ are integers greater than zero and $z_1 + z_2 + \ldots + z_k = n$.

(ii) Deduce that

$$\mathbb{P}\left(Z_1 = b\right) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}},\ b = 1, 2, \ldots, n-k+1.$$

What is the distribution of the number of sample members that carry a mutation occurring while 2 ancestors remain in $\mathcal{T}$?

(iii) By considering a mutation in a small region $[x, x+\delta x)$, show that the probability q_{nb} that the number of leaves in $\mathcal{T}$ carrying a given mutation is b is given by

$$q_{nb} = \frac{\sum_{k=2}^{n} \mathbb{P}\left(Z_1 = b\right) k\mathbb{E}\left(T_k^n\right)}{\sum_{k=2}^{n} k\mathbb{E}\left(T_k^n\right)}.$$

(b) (12 marks) In a sample of $n = 4$ DNA sequences labelled a to d, 4 mutant sites segregate in the following configuration:

Sequence\Site	1	2	3	4
a	1	1	1	0
b	1	1	1	0
c	1	0	0	1
d	0	0	0	0

At each site the ancestral allele is denoted by 0, and the model of the beginning of the question holds.

(i) By first sketching a gene tree, or otherwise, sketch a possible coalescent tree for the sample. When mutations 1 and 4 occur, respectively, what possible values are there for the number of ancestors to the sample?

(ii) While j ancestors remain in a coalescent tree, write down the probability that two specific lineages k and l coalesce, and the probability that a specific lineage k mutates.

(iii) Show that the likelihood of the observed data as a function of the mutation rate θ is equal (up to a constant independent of θ) to:

$$\frac{\theta^4}{54(1+\theta)^2(2+\theta)^3(3+\theta)} \times \left(\frac{3}{3(2+\theta)} + \frac{1}{4(3+\theta)}\right).$$

(iv) A researcher applies a commonly used Monte-Carlo approach to estimate the likelihood. Within this approach, coalescent trees producing the observed data are sampled randomly so that backwards in time, with probability 2/3 the first event in the tree is a coalescence between the ancestors of sequences a and b. Show that the *actual* probability, conditional on the observed data, that this coalescence is the first event exceeds 2/3 for any θ.

2. In a dataset of n sequences, suppose that the infinitely-many-sites mutation model holds.

(a) (8 marks) The history of the n sequences is described by a coalescent tree, whose times $T_n, T_{n-1}, \ldots T_2$ while $n, n-1 \ldots, 2$ ancestors remain are independent with exponential distributions of rates $\binom{n}{2}, \binom{n-1}{2}, \ldots \binom{2}{2}$ respectively. Mutations occur at rate $\theta/2$ along each edge of the tree.

Show that the total number M_n of mutations on the tree has probability generating function

$$\mathbb{E}\big(z^{M_n}\big) = \prod_{j=1}^{n-1} \frac{j}{j + \theta\,(1-z)}.$$

Derive the expectation of M_n, and its behaviour as $n \to \infty$.

(b) (6 marks) A sample of $n = 7$ DNA sequences shows the following 7x8 incidence matrix, ordering mutations from left to right along the DNA sequence. The ancestral type is denoted by zero at each site.

Sequence\Site	1	2	3	4	5	6	7	8
a	1	0	1	0	0	0	1	0
b	0	1	0	0	1	1	0	0
c	0	0	1	1	0	0	0	0
d	1	0	1	0	0	0	1	1
e	0	0	1	0	0	0	0	0
f	0	1	0	0	0	1	0	0
g	0	0	1	0	0	0	1	0

Use Gusfield's algorithm to draw a rooted gene tree for the sample.

(c) (11 marks) The incidence matrix of part (b) is extended by adding rows, corresponding to DNA sequences from additional individuals. This extended incidence matrix now indicates recombination events in the sample history. For each pair of sites i and j, for $1 \le i < j \le 8$, a non-negative lower bound R_{ij} on the number of recombination events between these sites is obtained. Define W to be the minimum number of recombination events, between sites 1 and 8, required to satisfy these bounds simultaneously.

(i) Describe one approach that could be used to obtain the bounds R_{ij}.

(ii) H_M^{18} is calculated using the recursion

$$H_M^{11} = 0; \qquad H_M^{1j} = \max\left\{H_M^{1k} + R_{kj} : \; k = 1, 2, \ldots, j-1\right\}.$$

Explain why $H_M^{18} \le W$.

(iii) Define $r_j = H_M^{1j} - H_M^{1(j-1)}$for $j = 2, 3, \ldots, 8$. Show that by placing r_j recombination events between sites $j-1$ and j for each j, each bound R_{ij} is satisfied. Deduce that $H_M^{18} = W$.

(iv) If $R_{15} = 4$, $R_{23} = 1$, $R_{24} = 2$, $R_{27} = 4$, $R_{45} = 1$, $R_{46} = 2$, $R_{67} = 1$, $R_{ij} = 0$ otherwise, obtain H_{18}. Show that there are multiple possible recombination event placements that are minimal.

3. (a) (13 marks) In an ancestral recombination graph for n sequences, while there are j ancestors of the sample, backward in time recombination events occur as a Poisson process of rate $\rho j/2$, and coalescence events occur as a Poisson process of rate $j(j-1)/2$. The process terminates the first time the number of ancestors reaches $j = 1$.

(i) Define E_n to be the number of recombination events in the ancestral recombination graph. Show that for $n = 2, 3, \ldots$, E_n satisfies the equation system:

$$E_n = \frac{n-1}{n-1+\rho} E_{n-1} + \frac{\rho}{n-1+\rho} E_{n+1} + \frac{\rho}{n-1+\rho}. \tag{1}$$

Give a boundary condition on this system.

(ii) By rearranging so that the left hand side of (1) becomes $E_{n+1} - E_n$, or otherwise, show that a possible solution is

$$E_n = \rho \int_0^1 \frac{1-(1-x)^{n-1}}{x} e^{\rho x} dx. \tag{2}$$

(iii) Given equation (2) gives E_n, by considering $\rho \int_0^1 \frac{1-(1-x)^{n-1}}{x} dx$ obtain the asymptotic behaviour of E_n as $n \to \infty$.

(b) (12 marks) Consider a time-homogeneous diffusion process X_t ($t \geq 0$) on $[0, 1]$ with infinitesimal mean $b(x) \equiv 0$ and infinitesimal variance $a(x) = x(1-x)$, $0 \leq x \leq 1$.

(i) The *generator* $\mathcal{L}$ of a Markov process is defined as the functional operator

$$\mathcal{L}(f)(x) = \frac{d}{dt}(\mathbb{E}[f(X_t \mid X_0 = x])|_{t=0} = \lim_{t \to 0} \frac{\mathbb{E}_{X_0=x}[f(X_t)] - f(x)}{t}.$$

Write down the generator of X_t.

(ii) Define $h(x)$, $0 < x < 1$ to be the probability that X_t reaches 1 (fixation) given $X_0 = x$. Write down and solve a differential equation for $h(x)$.

(iii) For a twice continuously differentiable function $g : [0, 1] \to \mathbb{R}$, a fixed time $s > 0$ and $0 \leq x \leq 1$ define $u(s, x) = \mathbb{E}[g(X_s \mid X_0 = x)] = \mathbb{E}_{X_0=x}[g(X_s)]$. By conditioning on X_t, show that for any $t > 0$

$$u(s + t, x) = \mathbb{E}_{X_0=x}[u(s, X_t)].$$

(iv) Deduce that u satisfies the partial differential equation:

$$\frac{\partial u(s,x)}{\partial s} = \frac{1}{2} x(1-x) \frac{\partial^2 u(s,x)}{\partial x^2}.$$

(v) In the particular case where $g(x) = 2x(1-x)$, called the heterozygosity, show that $u(s, x) = 2x(1-x)e^{-s}$ solves the partial differential equation of part (b)(iv).