Introductory seminar on
"Mathematical Population Genetics"
winter term 2017/18

Prof. Joachim Hermisson
Ilse Höllinger

4 Genetic drift, neutral theory and the coalescent

4.1 Heterozygosity in a Wright-Fisher model with a finite number of alleles

The lecture notes deal with heterozygosity in the infinite alleles model or infinite sites model, where each mutation creates a new allele. The detailed results were obtained for the two-alleles model.
We consider a Wright-Fisher population of N haploid individuals.

1. Derive the recurrence equation for heterozygosity H_t at time t and give the equation for H at equilibrium under mutation and drift for an arbitrary number of alleles n, where $\forall(i,j) \in \{1,\ldots,n\}, i \neq j, \mu_{ij} = \frac{\mu}{(n-1)}$, so that $\sum_{j \neq i} \mu_{ij} = \mu$.
2. *Think about:* Would this model be relevant at the single nucleotide level? What is the appropriate value for n and what are the model assumptions?

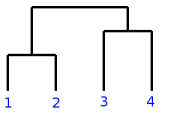
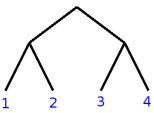
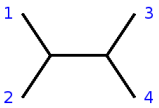
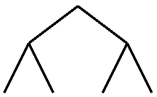
****

Hint: Use the fact that mutation rates are small to approximate the solutions and ignore all terms of order μ^2 in the derivations.

4.2 Topologies of coalescent trees

We use the following definitions:

- A *coalescent tree* is a *labelled* and *time-ordered* history.
- A *rooted tree* is a *labelled* but *not time-ordered* history.
- An *unrooted tree* is a *labelled*, *not time-ordered* history without root.
- A *rooted topology* is an *unlabelled*, *not time-ordered* history.

	coalescent tree	rooted tree	unrooted tree	rooted topology
				
rooted	yes	yes	no	yes
time-ordered	yes	no	no	no
labelled	yes	yes	yes	no

Two coalescent trees are called topologically equivalent if they share the same rooted topology

1. Draw all possible topologically different coalescent trees for a sample of size $n = 5$.
2. Using a forward branching process, give the probability of these given topologies.
3. Propose a recursion to compute the number of topologies for a sample of size n. *Hint:* Define a left and right subtree by "splitting" the tree at the root and relate the number of possible topologies of the entire tree to the number of possible topologies of these two subtrees.
4. How many possible topologies exist for $n = 8$?

4.3 Molecular clock

Based on geological data, the split between two species of fruit flies, *Drosophila differens* and *Drosophila sylvestris*, occupying two different Hawaiian islands, has been estimated to have occurred 2.5 million years ago. In the wild, for fruit flies ten generations per year seem realistic. Based on experiments, the per nucleotide and per generation mutation rate in Drosophila has been estimated to $\mu = 3.5 \cdot 10^{-9}$.

- Give a 'naive' expection, based on an infinite-sites mutation model, of how many differences between the two species can be expected per 1000 nucleotides?
- Do you think it is reasonable to use this substitution rate as the basis of a molecular clock? Check how realistic this 'naive' expectation is: Use the Poisson distribution to give a probability for multiple mutational hits at a single site. For simplicity, assume that mutation rates between all for nucletides, A,C,G,T are equal.
- How do the probabilities for multiple hits change your expectation for the number of segregating polymorphic sites between the two species? You can ignore more than 3 hits at the same site.

4.4 Variance of the number of segregating sites

In the lecture, we have learned that the expected value for S_n, the number of segregating sites in a sample of size n is

$$\mathbb{E}(S_n) = \theta \sum_{i=1}^{n-1} \frac{1}{k}$$

Compute the variance $\mathbb{V}(S_n)$. As for the expected value in the lecture notes assume continuous time for your calculations.

Hints:

- You can express $\mathbb{E}(S_n^2)$ as $\int_0^\infty \mathbb{E}(S_n^2 | L_n = t) f_{L_n}(t)\, dt$, where $f_{L_n}(t)$ is the density of the distribution of the total length of a coalescent tree with n tips (similar of course for $\mathbb{E}(S_n)$).
- It is often efficient to partition the coalescent tree into $n-1$ independent parts with $k = 2, \ldots, n$ lineages. These parts of the tree can then be treated independently.