Population Genetic Analyses of Genomic Data: Assignment 2

1 Exercises

1.1 Phylogenetics

Consider the following way of representing a rooted tree:

1	3	1.5	10	3.0
3	12	1.8	8	4.5
4	14	1.1	15	1.1
5	1	2.2	7	3.5
7	8	4.7	2	4.7
10	4	1.9	11	3.0
12	6	2.7	13	2.7

where each line represents an internal node in the tree: the first column is the node label, and subsequent pairs of columns represent daughter nodes by their label and branch length. So for example, node 1 is the parent of nodes 3 and 10, connected to them by branches of length 1.5 and 3.0 respectively.

a) Write a program which reads a data structure in this format and computes, for each internal node, the numbers of internal nodes and leaves below it.

b) Modify your program to calculate the branch length from each leaf to the root (this will also entail inferring which node is the root). Hence evaluate whether the tree is consistent with an evolutionary tree in which branch lengths represent times.

c) Make your program take an option which prints out a relabelled version of the tree in the same format, in which the nodes are numbered so that every daughter node has a lower number than its parent.

1.2 Time-dependent selection

An organism has a total of 1000 genes, each of which may be fixed in the state '1' or '0'. The magnitude of the fitness advantage of the allele $a \in \{0, 1\}$ in gene i, denoted f_i^a, is drawn from an exponential distribution with parameter 100. Initially the '1' state is beneficial, so that $f_i^1 > 0$ and $f_i^0 = -f_i^1$. Mutation in the population is slow, such that it may be assumed that no more than one gene is polymorphic at any given time. Given a mutation in a gene, the probability of the fixation of that mutation is given by

$$\frac{1 - e^{-2f_i^a}}{1 - e^{-2Nf_i^a}} \tag{1}$$

where $a \in \{0, 1\}$ is the new allele and N is the population size. We suppose that the lifespan of each polymorphism (i.e. time to fixation or death of the new mutation) takes

$$\min\left(10, \frac{2\log(N-1)}{919 f_i^a}\right) \tag{2}$$

units of time, where we suppose that $N = 100$.

a) Construct a simulation of the above system. Make a plot of the statistic n_1, equal to the number of genes in the state '1'. This value should tend to an equilibrium-like state over time.

b) Once your system has reached equilibrium, begin collecting statistics from the system. Once you have sufficient data, make a plot of the distribution of the fitnesses of mutations which fix in the population. What do you observe?

c) We now suppose that the environment in which the population exists changes over time. Changes in the environment occur infrequently according to a Poisson distribution with rate $\tau = 5 \times 10^{-6}$ per unit time. Upon a change in the environment, the fitnesses of each allele are instantaneously reversed, such that f_i^1 becomes $-f_i^1$ and f_i^0 becomes $-f_i^0$. Simulate the behaviour of the population under these conditions, and again plot n_1 and the distribution of the fitnesses of mutations which fix in the population. What do you observe?

d) What happens to the distribution of the fitnesses of mutations that fix in the population as τ becomes large? Explain your answer.

e) Studies of genomic data from wild *Drosophila* populations show results which suggest an unusually large proportion of beneficial mutations. Explain this result. Are these populations becoming fitter over time?

1.3 Ancestral inference

What can we say about the time to the most recent common ancestor from which any pair of the MPhil students from the class of 2018-19 share inherited DNA?

a) What theory would you use to estimate this time, and what parameters are required? Suggest some values for these parameters.

b) What additional factors might modify this estimate? Estimate how they would affect it and by how much.

c) Can you come up with an estimate and range?

1.4 Neanderthal introgression

Based on autosomal genome DNA analysis it is believed that Neanderthals hybridised with modern humans around the time they left Africa 60,000 years ago, resulting on a small fraction (1-2%) of all non-African present day human autosomal genome sequence being derived from Neanderthals. However, in large numbers (say 1,000,000) of non-African mitochondrial genomes none of them match the Neanderthal mitochondrial genome.

a) Assuming that there has been no selection, estimate the probability of seeing no Neanderthal mitochondria in modern people if there was 1% contribution from Neanderthals to the out-of-Africa population 60,000 years ago. State what assumptions you have made in your model and what approximations if any in your calculation.

b) What is the equivalent probability if there was 10% replacement? Discuss why this test has surprisingly little power even with large numbers of modern people.

c) What biological factors might be important in determining the actual likelihood of observing Neanderthal mitochondria in modern humans which were not included in your model, and qualitatively what effect they would have.