Homework #7, Spring 2013

Version 2.3 corrected May 17

Background

You are interested in estimating the date of divergence between two isolated populations of the Scots pine, *Pinus sylvestris*. These populations were reported by Naydenov *et al.* (2007 see http://www.biomedcentral.com/1471-2148/7/233/ if you are curious). Following the protocols of that paper, you collect DNA from haploid tissue of the plants. Unfortunately, the federal government's decision to enact budget cuts via sequestration results in a dramatic cut to your project's lab budget. You must proceed using only genome sequencing of one haploid genome from Spain and two haploid genomes (obtained from separate plants) from Turkey.

You are willing to consider a simplified scenario in which:

- each of the current populations is panmictic (no substructuring);
- the current populations have the same effective population size as each other;
- the populations are descended from an ancestral population that has the same effective population;
- the population divergence was a distinct, instantaneous event τ generations ago. In other words there was no messy period of substructure or migration between the diverging ancestral populations. In one generation there in a common ancestral population, and in the next generation there were two equally sized and disconnected daughter populations.

Data

You conduct shotgun sequencing and assemble 250,672 regions in which you trust the sequencing reads (you have some threshold # of reads that have to be stacked over a site to give you confident in the base calls) and for which you have a base call for each of the three haploid samples.

From each aligned block of sequence, you randomly select one site and code it as a 0/1 column in an alignment. In your coding scheme '0' is the base that is found at the site in the Spanish sample. '1' indicates a different base. There are no sites in which there are 3 different bases. Thus, if both of the Turkish samples have a '1' it means that they shared a nucleotide that was different from the one found in the Spanish sample. As you might expect, most of these sites are not polymorphic; in fact, 248,124 sites display the same nucleotide in each of the samples.

In tabular form your full data set is:

Sample	Pattern written as columns in "SNP coding"			
Spanish	0	0	0	0
Turkish-1	0	1	0	1
Turkish-2	0	0	1	1
Count for each pattern	248,124	781	803	964

Model and parameters

We'll consider a model in which *a priori* state '0' and state '1' are equally likely. Let

- μ be the mutation rate which is expressed per site per generation (you can think of it as the proportion of sites in the genome that you would expect to see mutated in one generation of these plants);
- N_e be the effective population of each population (the Spanish pop., the Turkish pop., and the common ancestral pop.); and
- τ be the divergence time for the 2 populations (expressed in number of generations).

Bryant *et al.* (2012) have developed a general approach for analyzing this sort of data.

The key to calculating likelihoods for this sort of model is to break down the events by population and think of the events in each population as conditionally independent of each other. They express the likelihood via a set of partial likelihoods. A partial likelihood gives the probability of a subset of each data pattern, conditional upon a latent variable. The probability statements pertain to each SNP; they describe the probability of seeing each one of the four distinct SNP patterns given sampling of a random position in the genome.

Bryant *et al.* characterize a state as a pair of numbers (n_{*b}, r_{*b}). n_{*b} represents the number of lineages at the present at the most recent time point for population $*$. b stands for 'bottom' of the branch (David Bryant is a mathematician and mathematicians draw their trees upside down with the bottom being the most recent time). r_{*b} is the number of lineages with state 0. In our problem, the $*$ will be either S for Spain, T for Turkey, or A for the common ancestral time point (so n_{Ab} is the number of lineages in a gene genealogy at the point of population divergence).

In our coding of the data, the current Spanish population would be coded as $(n_{Sb} = 1, r_{Sb} = 1)$ for each of the 250,672 sites.

In the population from Turkey our data reveal lots of examples of the population being described as $(n_{Tb} = 2, r_{Tb} = 2)$, almost one thousand examples of the $(n_{Tb} = 2, r_{Tb} = 0)$ outcome, and over one thousand examples of the $(n_{Tb} = 2, r_{Tb} = 1)$ state. Note that n_{Tb} is part of the study design (it is always 2), not a random outcome determined by the evolutionary process. However, r_{Tb} is an observation that the model seeks to explain.

It would probably be helpful for you to recode the data as a table of different r_{Tb} values and the count of how many times each occurs. This representation is a sufficient statistic.

When we want to calculate the probability of a particular r_{Tb} site, the general approach is to treat the number of relevant lineages at the population divergence time as a latent variable n_{Ab}, and to treat the number of these lineages that are in state '0' at the population divergence point as another latent variable, r_{Ab}

$$\mathbb{P}(r_{Sb} = 1, r_{Tb} \mid \tau, \mu, N_e) = \sum_{n_{Ab}} \sum_{r_{Ab}} \mathbb{P}(r_{Sb} = 1, r_{Tb} | \mu, \tau, N_e, r_{Ab}, n_{Ab}) \mathbb{P}(n_{Ab} \mid N_e, \tau) \mathbb{P}(r_{Ab} \mid N_e, \mu, n_{Ab}) \quad (1)$$

That is the correct general formulation, but it is hard to figure out how to calculate $\mathbb{P}(r_{Tb} | \mu, \tau, N_e, r_{Ab}, n_{Ab})$ directly. The probability of a particular number of '0' alleles in the Turkish population (out of the 2 individuals) is only indirectly related to the number of ancestors with state '0' in the common ancestor at population splitting. It would be nice to think about the number of lineages at the "top" of the Turkish population (the top of the branch that leads to the current Turkish population). Similarly, the probability of generating a '0' at the end of the Spanish population depends on whether the ancestor of that gene copy had state '0' or '1'. This suggests:

$$\begin{aligned} \mathbb{P}(r_{Tb} \mid \tau, \mu, N_e) = \sum_{n_{Ab}} \sum_{r_{Ab}} \sum_{n_{Tt}} \sum_{r_{St}} \sum_{r_{Tt}} \quad & \Big[\mathbb{P}(r_{Sb} = 1 \mid r_{St}, \mu, \tau) \mathbb{P}(r_{Tb} | \mu, \tau, N_e, r_{Tt}, n_{T_t}) \\ & \times \mathbb{P}(r_{St}, r_{Tt} \mid n_{Tt}, r_{Ab}) \mathbb{P}(n_{Tt} \mid n_{Ab}) \\ & \times \mathbb{P}(n_{Ab} \mid N_e, \tau) \mathbb{P}(r_{Ab} \mid N_e, \mu, n_{Ab}) \Big] \end{aligned} \quad (2)$$

Bryant *et al.* give us all of the pieces that we need to calculate the likelihood in this way.

The events along the lineage leading to the Spanish population

The easiest case to consider is the Spanish population. For each SNP, we have sampled one gene copy so $n_{S,b} = 1$. That site must have had an ancestor at the top of the Spanish branch so $n_{S,t} = 1$ with probability 1. However, a mutation could have occurred along the branch. So, $r_{S,t}$ could be 0 or 1. Bryant *et al.* use the notation similar to $F_{*t}[r_{*t}]$ to describe probabilities that pertain to the state of a population at the top of a branch. It turns out that, for our sampling scheme of one sample from Spain:

$$F_{St}[r_{St} = 1] = \mathbb{P}\left(r_{Sb} = 1 \mid \tau, \mu, r_{St} = 1\right) = \frac{1 + e^{-2\tau\mu}}{2} \quad (3)$$

$$F_{St}[r_{St} = 0] = \mathbb{P}\left(r_{Sb} = 1 \mid \tau, \mu, r_{St} = 0\right) = \frac{1 - e^{-2\tau\mu}}{2} \quad (4)$$

If you sketch these probabilities out as a function of time, you'll see that they make sense. If you consider a tiny branch length (small τ) and you know that there was $r_{St} = 1$ at the top of the branch, then the probabily of seeing $r_{Sb} = 1$ at the bottom of the branch is almost 1 (because there is almost no chance for a mutation to have occurred).

The events along the lineage leading to the Turkish population

The case of the Turkish population is more complex. We have sampled two gene copies (for each single nucleotide locus) from the current population. We do not know if those two gene copies trace their ancestry back to the common ancestral population via a shared ancestor or 2 distinct ancestors. In notation, our sampling design guarantees that $n_{Tb} = 2$, but n_{Tt} could be 1 or 2. And, given the number of ancestors at the top of the branch, we do not know how many had state 0. As indicated in equation (2), we will break this down by making the probability statements conditional on specific values of n_{Tt} and r_{Tt} and summing over all possibilities.

First, we'll consider the case of a SNP in which both samples from Turkey have state 0. Thus $n_{Tb} = 2$ and $r_{Tb} = 2$. Bryant *et al.* use $\theta = 4N_e\mu$ as a notational convenience to discuss the coalescence in terms of expected number of mutations seperating two randomly chosen genes drawn from a population (note that θ is just a combination of our existing parameters). They also find it helpful to calculate some odd looking combination of probabilities:

$$\begin{aligned} F_{Tt}[n_{Tt}=2, r_{Tt}=2 \mid r_{Tb}=2] &= \mathbb{P}\left(r_{Tb}=2 \mid \tau,\theta,\mu,n_{Tt}=2,r_{Tt}=2\right)\mathbb{P}(n_{Tt}=2 \mid \theta,\tau) \\ &= \frac{1}{4}\left(e^{\frac{-2\tau}{\theta}} + 2e^{\frac{-2\tau(1+\theta\mu)}{\theta}} + e^{\frac{-2\tau(1+2\theta\mu)}{\theta}}\right) \end{aligned} \tag{5}$$

$$\begin{aligned} F_{Tt}[n_{Tt}=2, r_{Tt}=1 \mid r_{Tb}=2] &= \mathbb{P}\left(r_{Tb}=2 \mid \tau,\theta,\mu,n_{Tt}=2,r_{Tt}=1\right)\mathbb{P}(n_{Tt}=2 \mid \theta,\tau) \\ &= \frac{1}{4}\left(e^{\frac{-2\tau(1+2\theta\mu)}{\theta}}\right)\left(e^{4\tau\mu}-1\right) \end{aligned} \tag{6}$$

$$\begin{aligned} F_{Tt}[n_{Tt}=2, r_{Tt}=0 \mid r_{Tb}=2] &= \mathbb{P}\left(r_{Tb}=2 \mid \tau,\theta,\mu,n_{Tt}=2,r_{Tt}=0\right)\mathbb{P}(n_{Tt}=2 \mid \theta,\tau) \\ &= \frac{1}{4}\left(e^{\frac{-2\tau}{\theta}} - 2e^{\frac{-2\tau(1+\theta\mu)}{\theta}} + e^{\frac{-2\tau(1+2\theta\mu)}{\theta}}\right) \end{aligned} \tag{7}$$

$$F_{Tt}[n_{Tt}=1, r_{Tt}=1 \mid r_{Tb}=2] = \mathbb{P}\left(r_{Tb}=2 \mid \tau,\theta,\mu,n_{Tt}=1,r_{Tt}=1\right)\mathbb{P}(n_{Tt}=1 \mid \theta,\tau) \tag{8}$$

$$= \frac{1}{4}\left(-e^{\frac{-2\tau}{\theta}} + 2e^{-2\tau\mu} - 2e^{\frac{-2\tau(1+\theta\mu)}{\theta}} + \left(2+2\theta\mu - e^{\frac{-2\tau(1+2\theta\mu)}{\theta}}\right)/\left(1+2\theta\mu\right)\right)$$

$$F_{Tt}[n_{Tt}=1, r_{Tt}=0 \mid r_{Tb}=2] = \mathbb{P}\left(r_{Tb}=2 \mid \tau,\theta,\mu,n_{Tt}=1,r_{Tt}=0\right)\mathbb{P}(n_{Tt}=1 \mid \theta,\tau) \tag{9}$$

$$= \frac{1}{4}\left(-e^{\frac{-2\tau}{\theta}} - 2e^{-2\tau\mu} + 2e^{\frac{-2\tau(1+\theta\mu)}{\theta}} + \left(2+2\theta\mu - e^{\frac{-2\tau(1+2\theta\mu)}{\theta}}\right)/\left(1+2\theta\mu\right)\right)$$

these will help us calculate the probability of a data pattern as we sum over alternative outcomes of the coalescent process.

When we consider a calculating the probability of a different SNP pattern, the likelihood has a different set of terms. For genomic sites in which the two samples from Turkey have different states $r_{Tb} = 1$, and the probabilities become:

$$\begin{aligned} F_{Tt}[n_{Tt} = 2, r_{Tt} = 2 \mid r_{Tb} = 1] &= \mathbb{P}\left(r_{Tb} = 1 \mid \tau, \theta, \mu, n_{Tt} = 2, r_{Tt} = 2\right) \mathbb{P}(n_{Tt} = 2 \mid \theta, \tau) \\ &= \frac{1}{2} e^{\frac{-2\tau(1+2\theta\mu)}{\theta}} \left(e^{4\tau\mu} - 1\right) \end{aligned} \tag{10}$$

$$\begin{aligned} F_{Tt}[n_{Tt} = 2, r_{Tt} = 1 \mid r_{Tb} = 1] &= \mathbb{P}\left(r_{Tb} = 1 \mid \tau, \theta, \mu, n_{Tt} = 2, r_{Tt} = 1\right) \mathbb{P}(n_{Tt} = 2 \mid \theta, \tau) \\ &= \frac{1}{2} e^{\frac{-2\tau(1+2\theta\mu)}{\theta}} \left(e^{4\tau\mu} + 1\right) \end{aligned} \tag{11}$$

$$\begin{aligned} F_{Tt}[n_{Tt} = 2, r_{Tt} = 0 \mid r_{Tb} = 1] &= \mathbb{P}\left(r_{Tb} = 1 \mid \tau, \theta, \mu, n_{Tt} = 2, r_{Tt} = 0\right) \mathbb{P}(n_{Tt} = 2 \mid \theta, \tau) \\ &= \frac{1}{2} e^{\frac{-2\tau(1+2\theta\mu)}{\theta}} \left(e^{4\tau\mu} - 1\right) \end{aligned} \tag{12}$$

$$\begin{aligned} F_{Tt}[n_{Tt} = 1, r_{Tt} = 1 \mid r_{Tb} = 1] &= \mathbb{P}\left(r_{Tb} = 1 \mid \tau, \theta, \mu, n_{Tt} = 1, r_{Tt} = 1\right) \mathbb{P}(n_{Tt} = 1 \mid \theta, \tau) \\ &= -\frac{1}{2} e^{\frac{-2\tau}{\theta}} + \left(2\theta\mu + e^{\frac{-2\tau(1-2\theta\mu)}{\theta}}\right) / (2 + 2\theta\mu) \end{aligned} \tag{13}$$

$$\begin{aligned} F_{Tt}[n_{Tt} = 1, r_{Tt} = 0 \mid r_{Tb} = 1] &= \mathbb{P}\left(r_{Tb} = 1 \mid \tau, \theta, \mu, n_{Tt} = 1, r_{Tt} = 0\right) \mathbb{P}(n_{Tt} = 1 \mid \theta, \tau) \\ &= -\frac{1}{2} e^{\frac{-2\tau}{\theta}} + \left(2\theta\mu + e^{\frac{-2\tau(1-2\theta\mu)}{\theta}}\right) / (2 + 2\theta\mu) \end{aligned} \tag{14}$$

Note that, because of the symmetry in our modeling of states '0' and '1', we only have three distinct formulae for these five combinations of probabilities.

Events in the ancestral population

Dealing with coalescence and mutation in the common ancestral population seem like it will be very tricky. We know that there was one ancestor of the gene copy sampled from Spain at the time of speciation, but the two samples from the Turkish population could be descendants of two distinct gene copies at the time of speciation, or the descendants of just one individual. Furthermore we do not know the nucleotide for the 2 or 3 ancestral lineages that gave rise to our sample.

Fortunately, Bryant *et al.* figured out that we can combine the F_{St} and F_{Tt} information as follows:

$$F_{Ab}[n_{Ab}, r_{Ab} \mid r_{Tb}] = \sum_{z=0}^{\min(1, r_{Ab})} \left[F_{St}[r_{St} = z] F_{Tt}[n_{Tt} = n_{Ab} - 1, r_{Tt} = r_{Ab} - z \mid r_{Tb}] \binom{n_{Ab} - 1}{r_{Ab} - z} \Big/ \binom{n_{Ab}}{r_{Ab}} \right]$$

where I have not denoted the fact that $F_{Ab}[n_{Ab}, r_{Ab} \mid r_{Tb}]$ depends on the parameter values.

If a particular set of n, r values is not possible, you simply use 0 for the probability. For example it is not possible of there to be 2 lineages in state '0' at the top of the Turkish lineage but only a total of one lineage. Hence $Tt[n_{Tt} = 1, r_{Tt} = 2 \mid r_{Tb}] = 0$.

You would need to calculate this summation for the values:

- $n_{Ab} = 2, r_{Ab} = 0$
- $n_{Ab} = 2, r_{Ab} = 1$
- $n_{Ab} = 2, r_{Ab} = 2$
- $n_{Ab} = 3, r_{Ab} = 0$
- $n_{Ab} = 3, r_{Ab} = 1$
- $n_{Ab} = 3, r_{Ab} = 2$
- $n_{Ab} = 3, r_{Ab} = 3$

because these are all of the possible combinations of feasible latent states.

What is $F_{Ab}[n_{Ab}, r_{Ab} \mid r_{Tb}]$?

$$F_{Ab}[n_{Ab}, r_{Ab} \mid r_{Tb}] = \mathbb{P}(r_{Sb} = 1, r_{Tb} | \mu, \tau, N_e, r_{Ab}, n_{Ab}) \mathbb{P}(n_{Ab} \mid N_e, \tau)$$

Note that includes probability statements about the allelic states (the r values) but also the number of lineages (the n values).

Finally, we get to the likelihood (and we realize that we need one last quantity):

$$\mathbb{P}(r_{Sb} = 1, r_{Tb}) = \sum_{n=2}^{3} \sum_{r=0}^{n} [F_{Ab}[n, r \mid r_{Tb}] \mathbb{P}(\mathbf{R} = r_{Ab} | \mathbf{N} = n_{Ab})] \quad (15)$$

Here $\mathbf{N}$ and $\mathbf{R}$ are the random variables that represent total number of ancestors ($\mathbf{N}$) and ancestors with state 0 (called $\mathbf{R}$) in existence at the time of speciation. Fortunately, Bryant *et al.* provide calulations that tell us:

$$\mathbb{P}(\mathbf{R} = 0 | \mathbf{N} = 2) = \mathbb{P}(\mathbf{R} = 2 | \mathbf{N} = 2) = (1 + \theta\mu)/(2 + 4\theta\mu) \quad (16)$$

$$\mathbb{P}(\mathbf{R} = 1 | \mathbf{N} = 2) = (2\theta\mu)/(2 + 4\theta\mu) \quad (17)$$

$$\mathbb{P}(\mathbf{R} = 0 | \mathbf{N} = 3) = \mathbb{P}(\mathbf{R} = 3 | \mathbf{N} = 3) = (2 + \theta\mu)/(4 + 8\theta\mu) \quad (18)$$

$$\mathbb{P}(\mathbf{R} = 1 | \mathbf{N} = 3) = \mathbb{P}(\mathbf{R} = 2 | \mathbf{N} = 3) = (3\theta\mu)/(4 + 8\theta\mu) \quad (19)$$

Assignment

In addition to the steps defined in the "preview" of the homework

- What are your priors?
- Implement the model. What is the 95% credible interval for τ according to your analyses?
- Implement a model-jumping MCMC that considers the hypothesis that $\tau = 0$ (which is the hypothesis that the Spanish and Turkish populations of these pines are still freely exchanging genes that at a rate that makes them essentially one population). Report your prior and posterior for that hypothesis.

Appendix

In case you look at the Bryant *et al.* paper, I thought that I should document how I altered their model and notation.

- They describe their model in term of red and green alleles, but we'll use the '0' or '1' notation.
- They treat the rate of $0 \to 1$ mutations as different from the rate of $1 \to 0$ mutations (u and v), but we will constrain both events to occur at rate μ.
- They use T and B as superscripts to indicate top and bottom of a population branch. I was afraid that would look like raising a number to a power, so I made them subscripts. I made them lower case so T could be used for Turkish.
- Their notation is generic for populations indexed by a number y. I'm just using special names S, T, and A for the Spanish, Turkish, and ancestral populations.