The coalscence and phylogeny

Peter Beerli

November 21, 2005

1 The coalescent and different population sizes through time

Biologist would prefer to estimate all possible parameters based from a single sample today, but some of these phenomena are difficult because we need data that shows variation to see the events. A difficult problem is the estimation of population bottlenecks in the past. In principle it is very simple to establish the model for such a bottleneck but in practise it needs large amounts of data. Bottleneck calculation are easy when we know the time and duration of a bottleneck, Figure 3 shows an example of a strong bottleneck, where one can see that during the bottleneck the population size is very small and all ineages coalesce almost immediately, and only few lineages are older than the bottleneck. The formula of the simple coalescent

$$\text{Prob}(G|\Theta) = \prod_{k=2}^{n} \left[\exp\left(-u_k \frac{k(k-1)}{\Theta} \right) \frac{2}{\Theta} \right]$$

breaks now down into 3 main parts and 2 transitions, see figure. During the transition looking backwards in time, there is certainty between the last coalescent and the time where the population shrinks, that no coalescent happened:

$$\begin{aligned} \text{Prob}(G|\Theta_A, \Theta_B, \Theta_C) = & \prod_{k=2}^{n_A} \left[\exp\left(-u_k \frac{k(k-1)}{\Theta_A} \right) \frac{2}{\Theta_A} \right] \\ & \times \prod_{k=n_A+1}^{n_B} \left[\exp\left(-u_k \frac{k(k-1)}{\Theta_B} \right) \frac{2}{\Theta_B} \right] \\ & \times \prod_{k=n_B+1}^{n} \left[\exp\left(-u_k \frac{k(k-1)}{\Theta_C} \right) \frac{2}{\Theta_C} \right] \end{aligned}$$

n_A are the maximal number of lineages in population A, n_B are the maximal number of lineages in population B, etc.

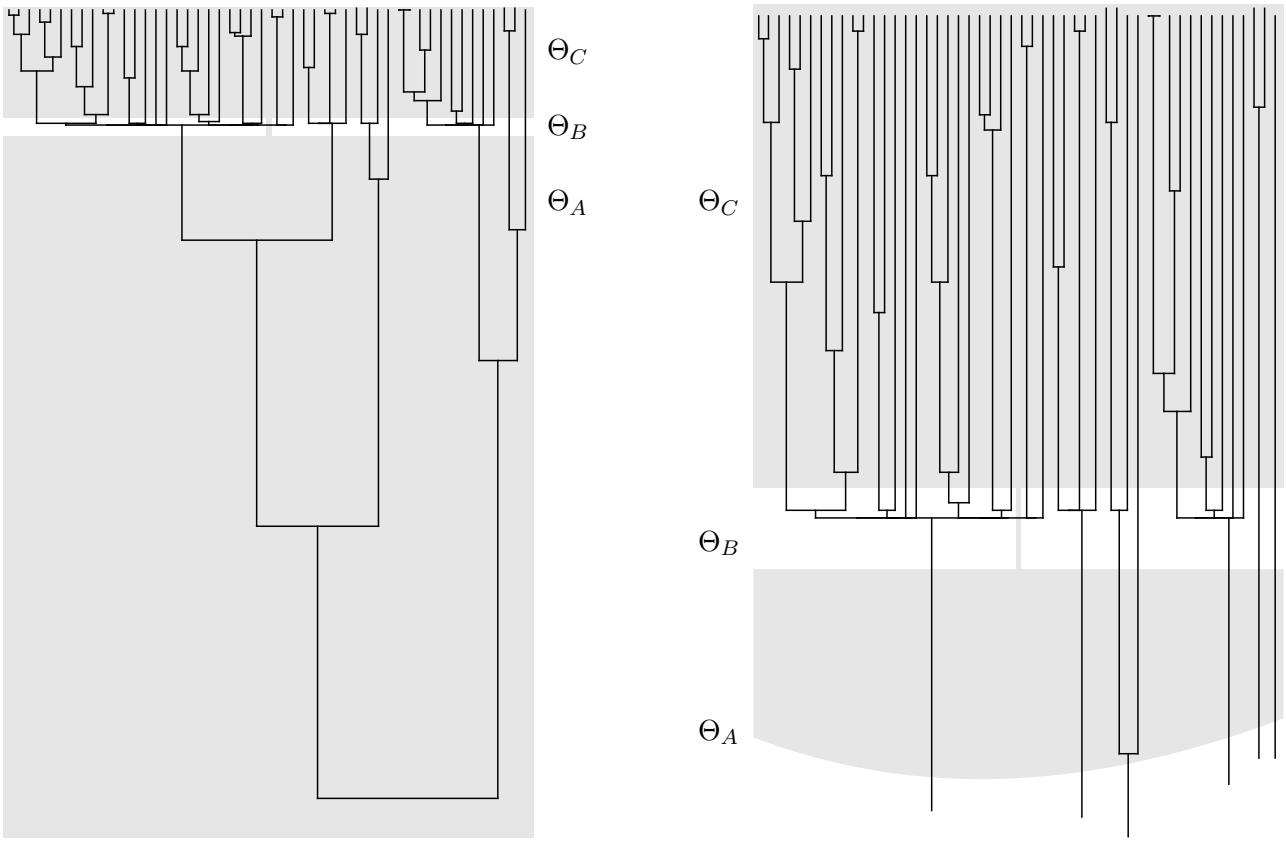

Figure 1: Population size change: The gray area gives the population size at different times (top is present, bottom is past). The right graph is an enlargement of the top part of the graph on the left.

2 Addition of speciation events

IN principle adding speciation to a coalscent framwork is not difficult, we need to add parameters such as species divergence time and acnestral population size. Inferences taking into account migration after the split might complicate the inference process, because now we nee to control the different prior distributions in such a way that the gene tree and the species tree are always compatible. In a framework where all lineages change at time τ this is more difficult than a system where one would allow a distribution of the speciation event and lineages switch from the ancestral species to a daughter species with some probability thatn increases towards today (the sample date) where we know that all lineages from species A are in species A, but since we do not know anything

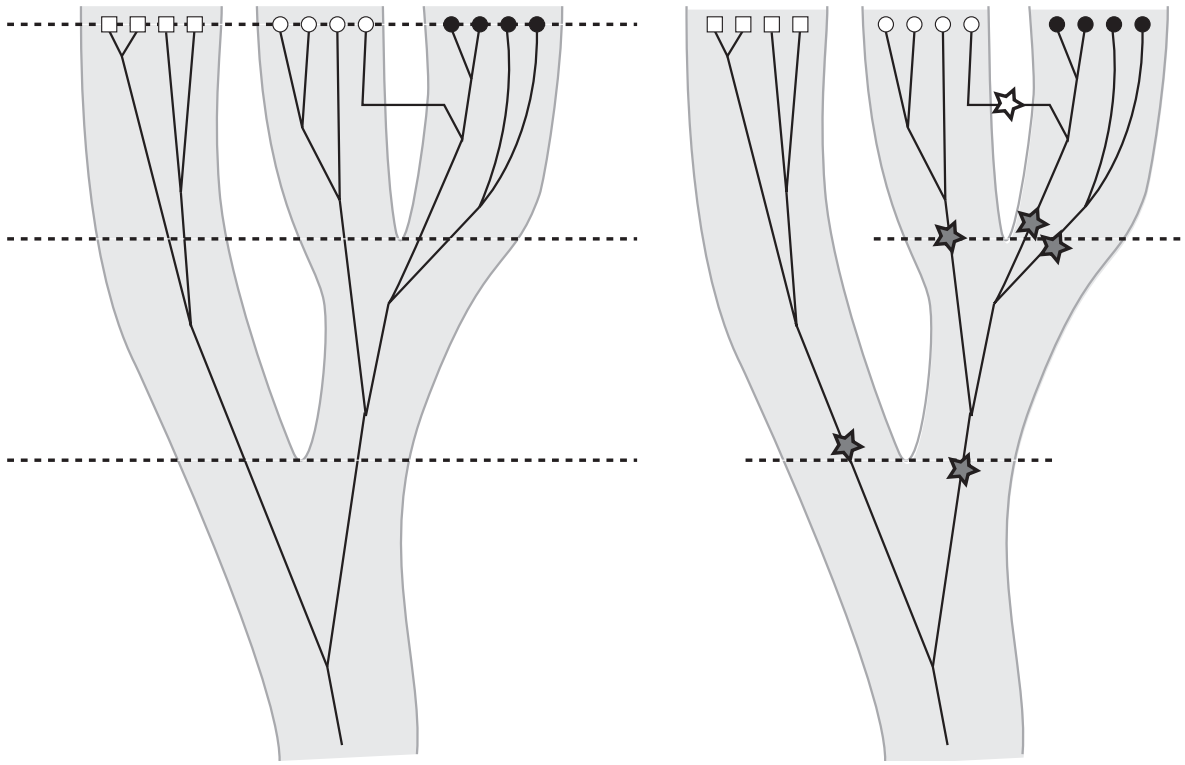

Figure 2: Two "different" approaches to speciation events

about the past we do not really know with certainty whether a specific lineage in species A was species A since many million of years (of course common sense might assume that). In any case, the latter approach would allow for some uncertainty of the speciation process for each lineage and is easier to model than all lineages switch at once. Formula for speciation events are not really very different from the bottleneck example, but simple have a single ancestor population and two or more independent (if there is no migration) populations today. You might guess that there are many different scenarios we can model with such an approach, most have been never tried, setting up simulation machinery so that we can split or merge populations or split unevenly (one or more lineage), with or without migration, or with a smooth migration rate decline over time, or changes in population size.

3 Estimation of divergence time

For phylogenetic inference we often use only one individual of a species, a praxis that will vanish hopefully sooner than later. If the species have very small overall population size than we often can ignore the coalescent in the past population size at all because, all individuals from the species

today have coalesced before the speciation event (looking backward in time). Such species pairs are reciprocally monophyletic, although such pairs are easy concerning the branching pattern the estimation of species divergence time is rather difficult and needs multiple loci, the more the better, the locus with the shallowest divergence time will be a lower bound on the speciation time. The problem is the population size of the ancestral population Θ, if the population is small the difference between gene divergence T (scaled by mutation rate it is D) and species divergence τ (γ) is small. but for recent speciation events the coalescence time in the ancestral population can be a sizable fraction of the the gene divergence time. Even when we know the gene divergence time without error and even when we know the exact ancestral population size there is still a sizable error attached to the species divergence time, here is an example of this magnitude assuming that the two species are reciprocally monophyletic and so only two lineages will enter the ancestral population:

$$\begin{aligned}\sigma^2(t_S) &= \sigma^2(t) + \sigma^2(2N) \\ \sigma^2(\tau_S) &= \sigma^2(t\mu) + \sigma^2(2N\mu) \\ &= \sigma^2(\frac{\Theta}{2}) = \left(\frac{\Theta}{2}\right)^2 = \frac{\Theta^2}{4}\end{aligned}$$

where $\Theta = 4N\mu$. Even in the best of all worlds we would have still a larger error on the divergence time. For example, if we have a gene divergence time of 0.001 expected mutations per generation and site and a population size of 0.1 and 0.00001, respectively, we will get estimates of species divergence time and coefficient of variation of the speciation time

$$\begin{aligned}t_S &= t_G - \Theta \\ CV(t_S) &= \frac{(\frac{\Theta^2}{4})^{-2}}{t_G - \Theta} \\ t_s &= 0.001 - 0.1 \\ CV(t_S|t_G = 0.001, \Theta = 0.1) &= \frac{0.01/2}{0.001 - 0.1} \\ t_s &= 0.001 - 0.00001 \\ CV(t_S|t_G = 0.001, \Theta = 0.00001) &= \frac{0.00001/2}{0.001 - 0.00001}\end{aligned}$$

We can see that the outcomes are rather different depending on the ancestral population size. The large ancestral population size is incompatible with with the supposed species divergence being smaller than the gene divergence time and so the CV returns a nonsensical result. With small population size the divergence time can postdate the gene divergence.

If we have more than two lineages such calculations gets rather difficult.

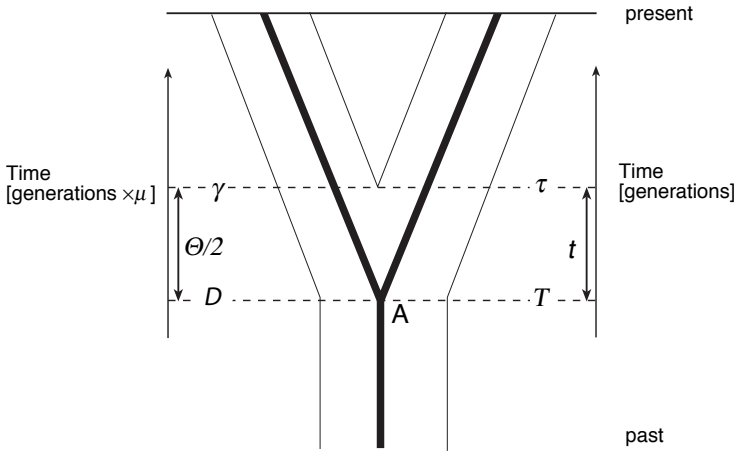

Figure 3: Estimation of divergence time – what is the problem?

Study questions

1. Explain why the speciation time cannot be at the same time as the gene divergence. Give an example, too.

2. Can you reason under what condition the ancestral population size is more important than changes in evolutionary rates between the species?

3. If you have many loci, can you come up with a simple method to get a rough estimate of the speciation time, the ancestral population size?