Dating phylogenetic trees

1 Dating phylogenetic trees

1.1 Modeling subsitution rates

When building a tree, every reconstruction methods is outputting branch length (bl) that are a function of the rate of substitution (μ) and the time of evolution (t)

$$bl = \mu * t$$

If we want to estimate the divergence time t of each node, we need to separate the two parameters from each branch length.
To do so, we need to model how μ might vary between every branches in the tree. By knowing μ, it is easy to get t

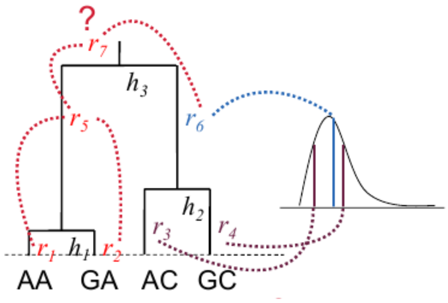
$$\frac{bl}{\mu} = t$$

How do we model μ? There are a few possibilities:

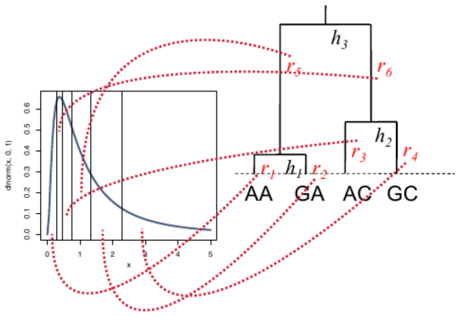
strict clock we can assume that μ is identical in every branches. By setting $\mu = 1$, all nodes will become proportional to t only.

local clocks We can try to find lineages where μ is identical, and then let μ vary between different lineages.

autocorrelated relaxed clock μ is a biological property, so it is possible that μ is evolving alongside other characteristics of the species. Therefore, we can assume that although μ is not identical in every branches, there is a correlation between μ of the ancestral and descendants lineages. The rate of the ancestral branch gives us the mean of a distribution (usually log-normal), and the rates of the descendants are drawn at random from this distribution



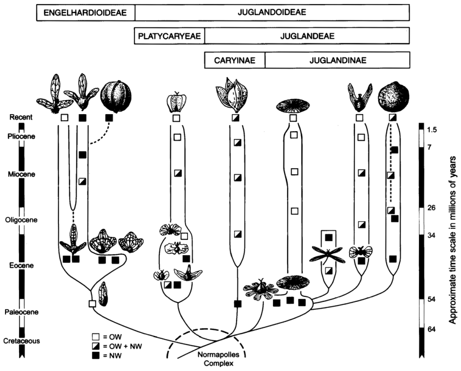
generalized relaxed clock The assumption of a correlated μ between lineages can be rejected. We are then left with modeling μ using statistical distribution such as the lognormal. The rates are drawn from this distribution and assigned to the different branches



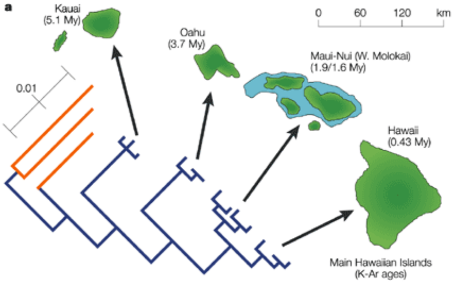
1.2 Calibrating the tree

The second aspect when dating phylogenetic trees has to do with calibrating the tree. Knowing the rate of substitution μ is one thing, but this will give us a relative time scale. We need to have external information on the absolute dates of some nodes in the tree to be able to translate the relative divergence times into absolute ones.
We basically do that by imposing some constraints on some interior nodes. The constraints can be of two sorts:

- fossils: usually give us a minimum time for a node. We know the node is as old as the fossil, but it could be older. Some new methods have been recently developed that allow fossils to be integrated in the prior for the tree (method is called Fossilized Birth-Death; e.g. Heath et al. 2014 PNAS). We will however not use this today.



- geological events: can give us a maximum time for a node. For example, a speciation event happening in an oceanic island constraint the age of the speciation to be at most as old as the origin of the island


a
Kauai (5.1 My)
Oahu (3.7 My)
0 60 120 km
Maui-Nui (W. Molokai) (1.9/1.6 My)
0.01
Hawaii (0.43 My)
Main Hawaiian Islands (K-Ar ages)

What is important to remember is that these constraints are very important to get accurate divergence times. The more you have, and the more evenly they are distributed along the tree, the better will the estimation be.

2 BEAST

BEAST is a program to reconstruct and date phylogenetic trees. It is using MCMC algorithm to do two steps at once: i) search the tree space to find the topologies plausible for your data and ii) estimate the rate of evolution and divergence time of each branch. By doing this, it will integrate the uncertainty coming from the model of evolution, topology reconstruction and estimation of rates of evolution into the divergence dates returned.

2.1 Setting up BEAST command file

BEAST uses xml format (a kind of extended version of html) for its command file. As it is not easy to write xml directly, the authors have released a simply program, BEAUTi, to write the commands.

1. Import the concatenated data file that you used for MrBayes into BEAUTi using the *File > Import NEXUS...*. If you don't have the file, you can find it here.

2. Unlink the partitions in the same way it was done in MrBayes using the *Partitions* tab. You should unlink the *Subs. Models*, but not the *Clock Models* because we assume here that all genes have similar rate distribution between branches.

3. Define the calibration nodes by using the *Priors* tab. Each calibration point refers to a node represented by a fossil. Here, we thus define clades that have to be monophyletic in the trees selected by MCMC. But keep in mind that by doing this, we restrict the prior on tree topologies by setting a probability of zero to the ones not showing monophyly for these clades.

   - ingroup: all the clownfish
   - african clade: *A. allardi*, *A. bicinctus*, *A. omanensis* and *A. nigripes* dating back between 3.3 to 5.3 Myr. Use a log-normal distribution as a prior with *Log(Mean) = 0*, *Log(Stdev) = 0.5* and *Offset = 3.0*

   Make sure that you click the monophyly box for each of these three groups.

4. Set the evolutionary models to the one you used in the MrBayes analyses using the *Site* tab. Take care to unfix the mean substitution rate if you use the codon partitions for *cytB* and set the molecular clock model to "uncorrelated log-normal" in the *Clocks* tab.
   Questions: Why do we unfix the mean substitution rate? What does this mean in the analysis we will be doing?

5. In the *Trees* tab, set the priors on the topologies to "Speciation: Yule Process". Then set the priors on the african clade and the endemic species defined above to a log-normal distribution. Select appropriate mean, standard deviation and offset parameters to fit the fossil calibrations. These two dates are known from fossil data. Let all the other priors as default.

6. Just ignore the *Operators* tab. We will leave these values as default.

7. Set the MCMC options using the *MCMC* tab. Use two million generations for the MCMC chain. Note that BEAST can only run one chain and not several like MrBayes.

Generate the xml command file by pressing *Generate XML...*.

2.2 Running BEAST

Run BEAST by inputting the saved xml file like this:

```
beast my_file.xml
```

2.3 Analysing the results

1. Open the file ending by *.log* into Tracer to check if you ran the MCMC chain long enough to reach equilibrium for the different parameters estimated. You should also check the ESS statistics, which measure the number of independent samples that the trace is equivalent to. ESSs lower than 100 (marked in red by Tracer) is not good. It means that the trace contained a lot of correlated samples and thus may not represent the posterior distribution well.

2. If the chain has been ran long enough, answer the following questions

   - How old is the root of the tree. Give the mean as well as the 95% highest posterior density (HPD).
   - How fast does the different DNA regions evolve?
   - What source of error does this estimate include?
   - Is the rate of evolution significantly different in different lineages?

3. Use the program TreeAnnotator to summarize the information found in the set of plausible trees sampled. For this, set a burnin period based on the trace shown by Tracer, a posterior probability limit of 0.5, a target tree type as Maximum clade credibility, a node heights as Mean heights. Press *Run* to execute the analysis.

4. Open the resulting file in FigTree. Turn on the "Node Bars" and "Branch Labels" and check what kind of information is displayed.

5. What is the marginal posterior estimate and HPD for the *A. ocellaris* and *A. percula* split?

6. Open the BEAST xml file in a text editor and find the <patterns> element in the xml file. It should look like this:

   ```
   <patterns id="patterns" from="1">
     <alignment idref="alignment"/>
   </patterns>
   ```

   Add an attribute called "to" with value "100" at the end of the patterns tag. Re-run the analysis with half the data now. How do the posterior clade probabilities change? How do the divergence time estimates change?

7. Run again Beauti, but this time by leaving out the fossil calibration prior for the endemic species. Start a new analysis with BEAST. Is the posterior distribution obtained for the missing calibration congruent with the prior used before? If not, what could cause this? Try to explain the impact of this on your dating analysis?

8. Run again Beauti and remove all calibration points. This time, selects a fixed clock as the clock model for the analysis and the ingroup date to the one obtained in the other analyses. Run BEAST and check how different are the age estimates. Do they make sense?