Feb. 14, 2018. **Phylogenetic trees VI: Dating in the 21st century: clocks, & calibrations; proper use of fossils**

Reading assignments: *Tree Thinking* pp 53-58; Parham, et al. 2012. Best practices for justifying fossil calibrations, *Systematic Biology*, 61:346-359, https://doi.org/10.1093/sysbio/syr107

1. Introduction

Molecular dating methods merge temporal information from sources outside of the primary phylogenetic data in order to provide time calibration for a phylogeny with branch lengths, corrected for rate variation. These may be strict or relaxed clock methods. There are a number of popular uses of molecular dating:

Biogeography- Establishing the origin of a fauna or testing dispersal and vicariance scenarios requires having an absolute time scale.

Age of a common ancestor- Correlating the origin of clades to climate or other events requires having an absolute time scale.

Diversification rates- Variation of evolutionary rates within or between clades or trees (co-diversification) can use relative *or* absolute time scales.

Population biology- Looking at the emergence and propagation of viruses and other disease pathogens, invasive species, or other population-level questions may require relative or absolute time scales.

By definition the extant tips of the tree are the same age, i.e. the *time* that has passed from the common ancestral node is the same for all extant OTUs. However, we also know that ultrametric trees are extremely rare for real data, so the amount of *character change* along the branches differs across the tree most of the time.

Why there is no universal molecular clock:
generation time
population size
DNA repair efficiency
metabolic rates
selective pressures

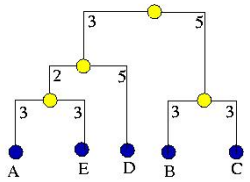
We are facing two different problems: making our tree ultrametic, and calibrating it. Even though some methods try to do both at once, they are still logically separate.

2. Searching for a clock

What is ultrametricity?

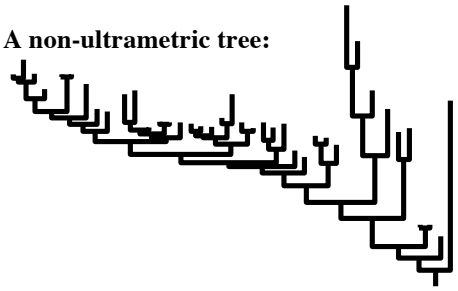
Ultrametric matrix and its tree:

	A	B	C	D	E
A		16	16	10	6
B			6	16	16
C				16	16
D					10
E					



from: http://www.diku.dk/~pawel/comp-bio/ev_trees/intro/intro/ultrametric.html

A non-ultrametric tree:



A. Determining whether your data fit a clock model

i. relative rate tests

comparing three taxa at a time, in rooted context:

ii. likelihood ratio test

Testing the Molecular Clock using a likelihood ratio test (courtesy of John Huelsenbeck)

Under the null hypothesis, the phylogeny is ultrametric (i.e., rooted and the branch lengths are constrained such that all of the tips can be drawn at a single time plane). Under the alternative hypothesis, each branch is allowed to vary independently. The alternative hypothesis invokes s - 2 additional parameters, where s is the number of sequences. The likelihood ratio test statistic is -2logL = 2(logL0 - logL1), where L0 and L1 are the likelihoods under the null and alternative hypotheses, respectively.

The significance of the likelihood ratio test statistic can be approximated using a $\chi 2$ distribution (with s - 2 degrees of freedom) or by parametric bootstrapping.

The following example shows how to perform the likelihood ratio test of the molecular clock. The data are s = 5 albumin sequences from vertebrates (a fish, frog, bird, mouse, and human). We assume the Hasegawa, Kishino, and Yano (1985) model of DNA substitution with among site rate variation described using a gamma distribution.

The maximum likelihood under the null hypothesis is logL0 = -7585.343. The best estimate of phylogeny supports the monophyly of the mammals and amniotes.

The maximum likelihood under the alternative hypothesis is logL1 = -7569.052. The likelihood under the alternative hypothesis is higher than under the null hypothesis because there are more free parameters in the substitution model (i.e., no constraints on branch lengths). The maximum likelihood estimate of phylogeny is consistent with the monophyly of mammals and amniotes (though the tree is unrooted).

The likelihood ratio test statistic is -2logL = 32.582, which is asymptotically $\chi 2$ distributed under the null hypothesis with 3 degrees of freedom. Comparing the observed value of -2logL to a $\chi 2$ with 3 df shows that the null hypothesis can be rejected at $P < 0.001$. So, we conclude the data are not clock-like.

B. If your data don't fit a clock model (and they usually don't), try smoothing the data to get an (at least locally) approximate clock. Two common methods (implemented in ***r8s*** by Mike Sanderson: https://sourceforge.net/projects/r8s/files/), both attempt to smooth the magnitude of changes in rate between neighboring branches, to give you something intermediate between the rigid clock assumption and completely unconstrained branch lengths:

i. non-parametric rate smoothing. This uses a least squares smoothing approach that penalizes rates that change too quickly from branch to neighboring branch.

ii. penalized likelihood. This is a "semi-parametric" approach that combines a ML approach with the above penalty function. The user can specify the relative weight of the penalty function and the ML component (in which parameters are being fitted as typical in ML). The parametric model has a different substitution rate for each branch.

3. Calibrating the clock

Calibration involves setting the age of one or several nodes in the tree using a point estimate, mean value, or probability distribution, allowing for the estimation of the age of other, uncalibrated nodes. Some folks simply import dates from prior dating analyses or "known" rates for the same gene based on a published estimate. However, this multiplies the error and uncertainty -- it's better to come up with a calibration from an analysis.

A. Three ways that have been used to estimate the age of a node:

i. a fossil (see below for details) -- gives a *minimum* age for a node.

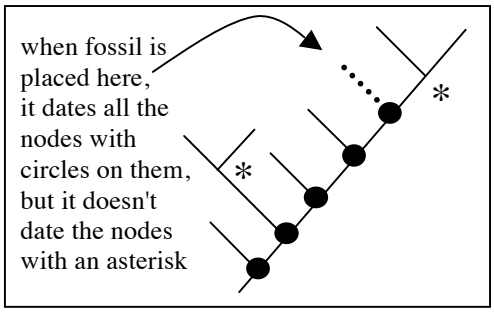
ii. availability of necessary habitat (e.g. origin of the paramo vegetation type)-- gives a *maximum* age for a node (maybe).

iii. geographic vicariance event (based on the vicarious distribution of sister groups in relation to well-dated geological events; obviously shouldn't be used when the biogeographic history of the group is the question being addressed) -- neither a *maximum* or *minimum* age for a node.

B. How to use a fossil to date a node? Some principles:

i. You never find a taxon in the fossil record, or a lineage; you find remains of an organism displaying *characters*. These characters need to be phylogenetically analyzed using the principles talked about earlier, in relation to other fossils and extant organisms in the group.

ii. Therefore, a fossil can never be compared to a strictly molecular phylogeny (unless it has preserved molecular data!); all relevant morphological characters need to have been analyzed and incorporated in the phylogenetic reconstruction.



iii. When a fossil can be placed using synapomorphies as sister to some other lineage, that other lineage (and the node connecting them) must be at least as old as the fossil. Nodes deeper must also have been in existence by that time. This is the important principle of ***equal age of sister groups***.

iv. Uncertainty: Dating of the fossil itself has a certain error rate, which is relatively symmetrical. But the uncertainty about dating of the *node* is highly skewed; at best the fossil gives the minimum age, but the maximum age in principle goes back to the origin of life. The richer the fossil record, the better is the chance to constrain the maximum age somewhat.

C. Several sources of uncertainty with fossils:

i. *Phylogenetic placement* is often based on overall similarity or simply taxon ID. Without a phylogenetic analysis there is no way to know the clade indicated really is a relative. We need integrated molecular-morphological analyses, and need to consider the possibility of multiple placements of the fossil.

ii. *Determining the fossil's age*.

What method and how precise?
Has there been a review of the dating of the fossil or strata
Has the geologic scale changed?
Correlated vs. direct dating

iii. The number, age and placement of multiple fossils considered to be same "taxon." Problem arise with over-determined specimens, and differing taxonomic concepts

D. In the best possible case you will have multiple fossils to establish hard minimum dates, i.e. *the youngest possible age of the oldest known fossil*, that are 1. based on real documented specimens, 2. have a confirmed identification, 3. consistently and explicitly dated using a standard geologic time scale, and 4. placed in the phylogeny with the same set of OTUs as those with molecular data.

4. Algorithmic approaches

A. Autocorrelation vs. non-autocorrelated rates

Autocorrelated- Descendant branches draw a rate from a distribution with a mean given by the ancestral branch.

Non-autocorrelated- rates for each branch are drawn independently from an identical distribution.

B. Use of prior trees or simultaneous estimation.

Many dating studies estimate the tree using a time-independent method followed by the estimation of divergence dates using a molecular clock (strict or relaxed).

-Independence of models and unlinking assumptions
-relatively computationally easy and fast

Some examples:
Local clock models (PAML, QDate) – clock-like within a given clade, but variable between clades.

Ad hoc heuristic rate smoothing (PAML)

Non-parametric rate smoothing (r8s) - simultaneously estimate unknown divergence times and smooth the rapidity of rate change along lineages

Penalized likelihood (r8s) Bayesian approach and penalized likelihood both have in common that they smooth or minimize rate variation over evolutionary time by means of an autocorrelated process.

C. Simultaneous estimation of the tree, its branch lengths and time estimates (e.g. BEAST).

-confidence intervals

-takes into account phylogenetic uncertainty, given that there is information for estimating dates of divergence, this could be used to give better estimates of phylogeny

-flexible clock models

-computationally difficult

-tip-dating vs. node-dating

5. Conclusions

A. For many questions in evolutionary biology you don't need absolute time; relative time will do (e.g., ordering of nodes in time). So, don't go out on a limb trying to manufacture clocks unless you need them.

B. If you do need to calibrate a clock, you want to have as many calibration points (preferably fossils), as local to your questions, as possible.

C. The fossils need to be embedded in the phylogenetic analysis, i.e., placed by the characters they bear, as with all taxa (tip-dating).

D. If you have enough calibration points, you don't need an actual molecular clock (or even a manufactured one) to answer many questions.

E. As always, carefully consider what questions you want to address first, then select your approach; for every positive hypothesis, be sure you have a null hypothesis.