Lecture 16 – Molecular Clocks

I. Introduction

Perhaps one of the most appealing, yet most abused, hypotheses in molecular evolution is the molecular clock – that the rate of molecular evolution is roughly constant across time.

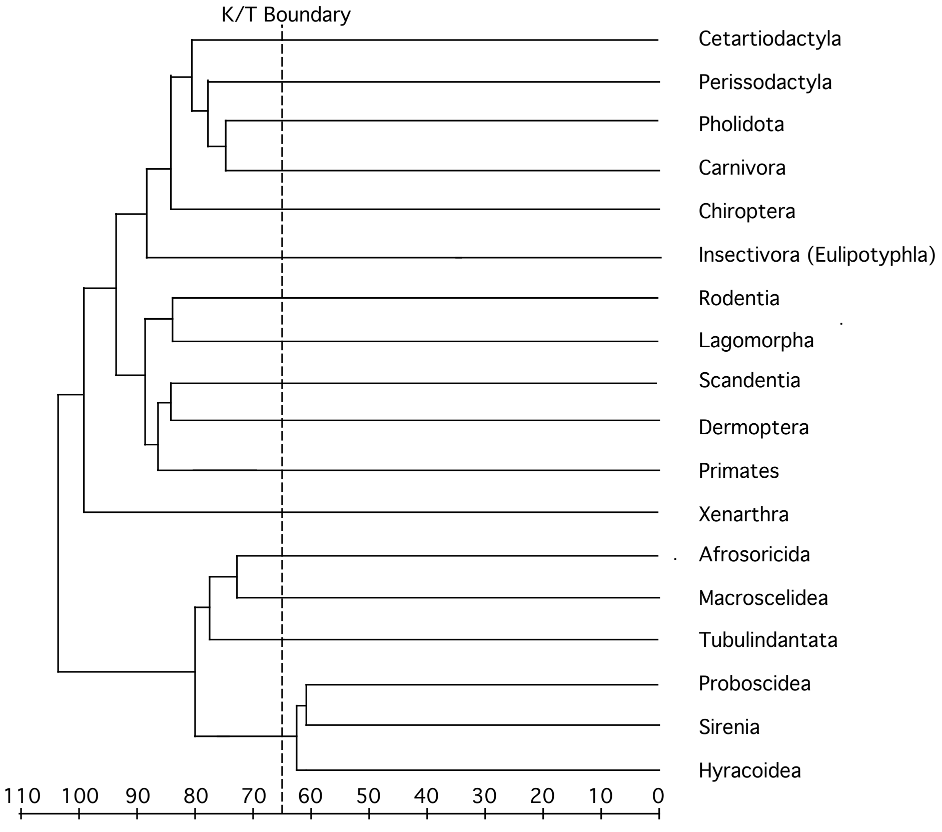
This was first postulated by Zuckerkandl & Pauling (1965), and it received support in 1983 from Kimura's neutral theory of molecular evolution. If the clock hypothesis holds, we have a chance to date divergence events.

Because branch length = (rate) x (time), if we can come up with a good estimate of rates, by say reference to the fossil record, we can date branching events (although usually without reference to uncertainty in fossil dates).

Here's an example from Mammals.

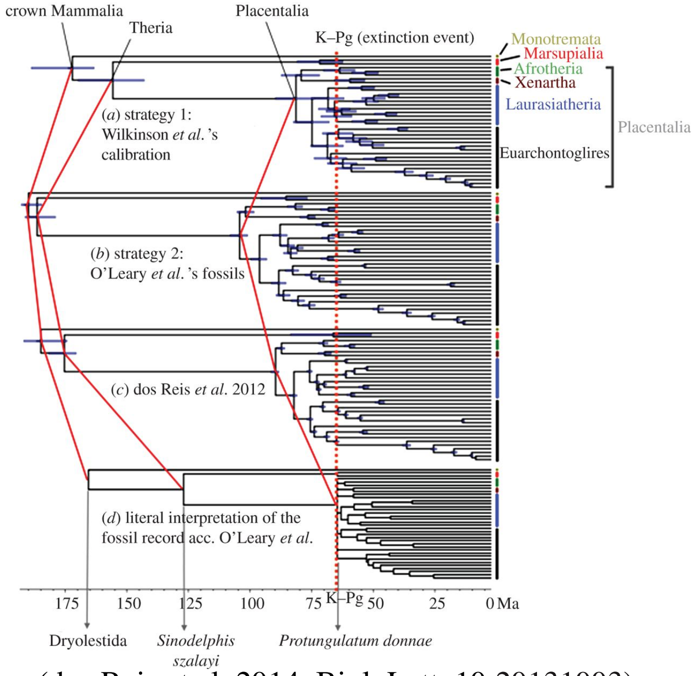
Molecular Estimates of Divergence Times

Springer et al. 2003. PNAS, 100:1056



This study indicates that most orders of mammals diverged prior to the extinction of the dinosaurs, which contradicts the classical view (although not the view held by some current paleontologists), and this has been supported by subsequent work (dos Reis et al. 2014. Biol. Lett. 10:20131003; but see – O'Leary et al. 2013. Science. 339:662.).

If you want to get a paper in a tabloid journal like *Science* or *Nature*, you can do so by estimating a date of divergence for a (charismatic) group from molecular data that differs from the fossil estimate. While there have been good studies, many high-profile papers have been shockingly bad (e.g., O'Leary et al. 2013).



(dos Reis et al. 2014. Biol. Lett. 10:20131003)

This figure illustrates the influence of different approaches to fossil calibration.

Up until fairly recently, studies such as this one relied on sequence evolution to behave in a clock-like fashion, with a uniform rate across the topology.

So, there's a really long history of inquiry into the notion of a molecular clock. Ho (2020) reviewed some of the landmarks in this history in his intro to the book *The Molecular Evolutionary Clock, Theory and Practice* (Ho, S., *ed*., Springer Nature, 249 pps.).

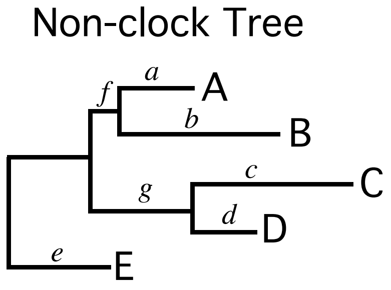
Again, the idea goes back nearly 60 years.

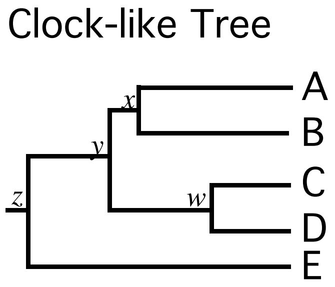
II. Tests of the Molecular Clock.

So, lots of people have proposed tests of the clock hypothesis, and there are several approaches to doing so. Let's introduce the idea of testing the clock hypothesis by looking at the following two trees.

A. LRT of the Molecular Clock

We have two identical topologies, and they differ in branch lengths.





The tree on the left has branch (2*n* - 3) lengths: *a* – *g*.

These trees converge when the following conditions are met:

1) $a = b$; $c = d$; $a + f = g + c$

2) The root occurs along branch e, such that $e'' = e' + g + c$.

Notice that, in the clock-like tree, we don't estimate the lengths of the $2n$ - 3 branches, but the times of the (n - 1) interior nodes w, x, y, & z of the rooted tree.

This suggests a very natural and intuitive test of the clock hypothesis the LRT that Felsenstein (1988. Ann. Rev. Gen., 22:521) developed 35 years ago.

Remember that LRTs are useful in testing the assumptions of restricted model relative to a more general model that relaxes those assumptions: that is, nested models.

In LRTs of the clock, the clock tree represents the special case and the non-clock tree represents the general model.

The test statistic is the standard LRT: $\delta = 2\ [\ \ln L_{(non\text{-}clock)} - \ln L_{(clock)}]$

Remember also, we can use the asymptotic χ^2-approximation, where the degrees of freedom equal to the difference in the number of parameters between the two models (note that the convergence of the test statistic with the χ^2-distribution is asymptotic).

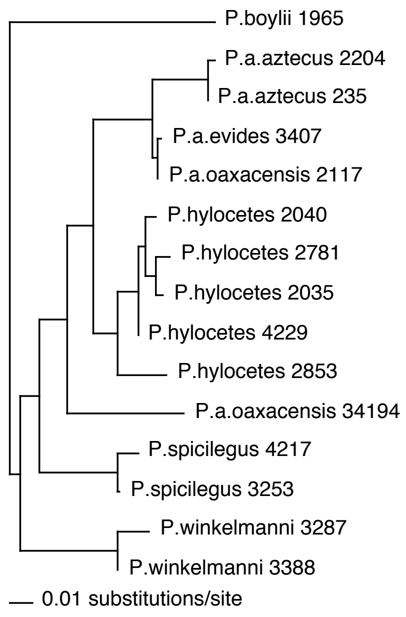
So, in the non-clock tree there are $2n - 3$ branch lengths that are estimated, whereas in the clock tree there are $n - 1$ node times to estimate.

Thus, there are $(2n - 3) - (n - 1)$ = $\boldsymbol{n - 2}$ **degrees of freedom**.

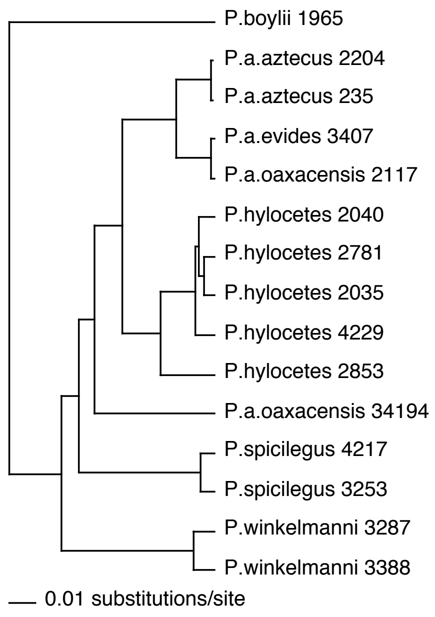
Because of the difficulties that may occur in relying on the asymptotic convergence of the test statistic to the χ^2-distribution, we may instead use a parametric bootstrap approach to generate the null distribution (like Goldman [1993] did).

Here's an example:

Non-Clock Tree: lnL = -2001.77467



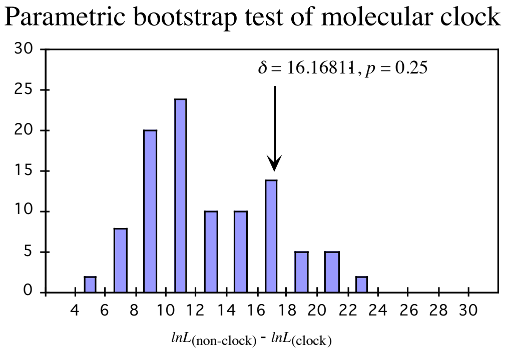
Clock Tree: lnL = -2017.94278



The test statistic is: $\delta = 32.34$. With 13 d.f., the *p*-value from the χ^2-distribution is 0.0021.

To conduct the parametric bootstrap, we would use the clock tree as the true tree and simulate data under a molecular clock. For each replicate, we find the likelihood score of the ML tree without the clock enforced (i.e., finding the optimum combination of the 28 branch lengths) and find the maximum likelihood with the clock enforced.

This is shown here:



We would fail to reject the clock.

This difference could be the result of either (or both) of two things. 1) It may be that in this case the asymptotic approximation of the χ^2-distribution doesn't work very well.

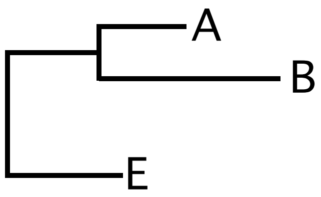
2) It also may be attributable to the fact that this data set mixes intraspecific and interspecific comparisons.

B. Relative Rates Tests

The oldest test of the clock hypothesis is that of Sarich and Wilson. It was originally a distance-based test and examines the difference in branch lengths between two ingroup taxa relative to an outgroup taxon. Therefore, it uses triplets of taxa.

If we go back to our example, the idea is to test the difference in length between branches *a* and *b*.

Relative Rates Test



Does $d_{(\text{A-E})} - d_{(\text{B-E})} = 0$?

Wu & Li (1985. PNAS, 82:1741) placed these into a statistical context. They pointed out that we can calculate the approximate variance expected under the clock and therefore assess the significance of any deviations for any set of triplets.

Tajima (1993. Genetics, 105:437) developed a test that really is based on parsimony, but works very well for closely related (i.e., intraspecific) sequences. If a clock holds, the number of sites that show the pattern *yxx* (m_1) should equal the number of sites showing the pattern *xyx* (m_2).

He calculates the statistic

$$(m_1 - m_2)^2 / (m_1 + m_2),$$

and uses a χ^2-distribution with one d.f. to assess its significance. This, for many years, was the most widely-used test.

Muse and Weir (1992) developed a RRT that uses the LRT and a χ^2-distribution with one d.f. (just as before, there are three branch lengths w/out a clock and two node times with a clock).

Whichever manifestation of the RRT one chooses, a separate test is conducted for all triplets containing the outgroup and two ingroup taxa.

So, for our initial example, one would test:

Does $d_{(A-E)} - d_{(C-E)} = 0$?	Does $d_{(B-E)} - d_{(D-E)} = 0$?
Does $d_{(A-E)} - d_{(D-E)} = 0$?	Does $d_{(C-E)} - d_{(D-E)} = 0$?
Does $d_{(B-E)} - d_{(C-E)} = 0$?	Does $d_{(A-E)} - d_{(B-E)} = 0$?

III. Estimating Divergence Times in the Absence of a Clock

It’s fairly common to be able to reject the molecular clock for large data sets. There have been a number of approaches that have been designed to estimate divergence times in the face of clock violations.

Other than ignoring the violation, there have been two approaches; identify offending lineages and eliminate them, and deal with rate variation among lineages by modeling the variation in rates.

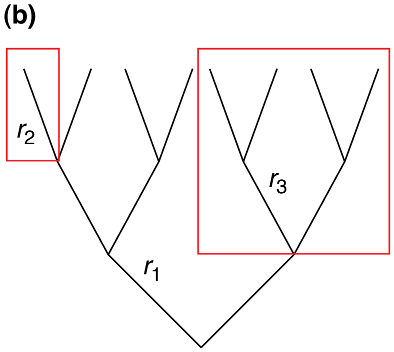
A. Linearized Trees

Takezaki et al., (1985. Mol. Biol. Evol., 12:823) proposed a series of RRT’s to identify which species or a clade that is evolving at a rate different from the rest of the group being studied.

The offending lineage can then be pruned from the data set and (if good fossil calibration is available), absolute rates can be applied to the remaining lineages to derive estimates of divergence times for them.

B. Local Clocks

A fair amount of work (e.g., Yoder & Yang, 2000. Mol. Biol. Evol. 17:1081), has focused on the notion that we might expect rate changes to be rare and that we should be able to localize them on a phylogeny and use a local clock, that is, one that only applies to a subtree of interest.



So, here we have three local clocks and we would date divergences, say within the 4-taxon tree using rate 3 (Figure from Welch & Bromham, 2005. TREE, 20:320).

Of course, this relies on being able to identify the points where rates change.

Drummond & Suchard (2010. BMC Biol. 8:114) developed a **Random Local Clock** (RLC) model.

The rate of branch *k*: $r_k = c\rho \text{ x } \rho_{pa(k)} \text{ x } \phi_k$,

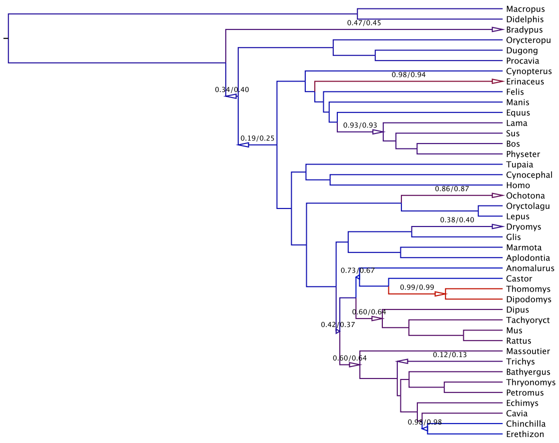
Where $c\rho$ is a scaling rate constant

$\rho_{pa(k)}$ is the rate of the parent branch of *k*

and ϕ_k is the branch-specific rate multiplier.

They use priors to restrict the number of changes (where $\phi_k \neq 1$) and calculate the probability of a new clock being set at each branch; this places a significant portion of the prior distribution on the conditions that specify a clock (i.e., rate multipliers equal 1).

With lots of data, we can identify the location of rate shifts on the tree.

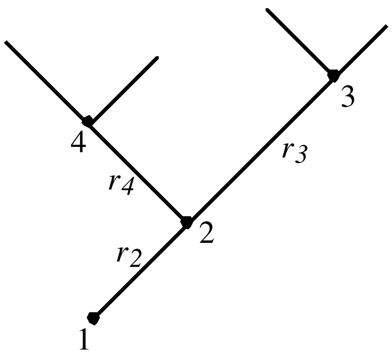


C. Autocorrelation of Rates

Other approaches have been developed where the rate of evolution is expected to change over a tree, but the rates along branches that are closely related (i.e., close in the tree) are expected to be more similar than rates along branches that are not.

Sanderson (1997. Mol. Biol. Evol., 14:1218) developed a method of called **non-parametric rate smoothing** that attempts to minimize differences in rates across branches.

It's the sum of squared differences in local rates that are minimized, as follows:



$$w_2 = (r_2 - r_4)^2 + (r_2 - r_3)^2$$

$$W = \sum w_k$$

The dependence of the rate at ancestral branch is indexed by the sum of squared differences in the rates at its daughter branches.

Since $r_k = b_k/t_k$, Sanderson originally used parsimony reconstructions as proxies for b_k's and found the combination of internal node times ($t_1, t_2, t_3, \ldots$) that minimizes W. ML branch lengths are now used.

These are relative node times, and again, it takes a fossil calibration to convert them to absolute dates. It's available in his package r8s (Sanderson. 2003. Bioinformatics, 19:301).

A weakness of NPRS (as well as Sanderson's [2002] penalized likelihood method) is that the tree and branch lengths are taken as given; uncertainty in their estimation is ignored.

$$\psi(\theta_{SAT} \mid x_1, x_2, \ldots, x_{S+M}) = \log L (\theta_{SAT} \mid x_1, x_2, \ldots, x_{S+M}) - \lambda\, \Phi\, (r1, r2, \ldots, r_{S+M})$$

$\log L (\theta_{SAT} \mid x_1, x_2, \ldots, x_{S+M})$ is likelihood of the saturated model.

λ and Φ are smoothing and roughness parameters that govern the autocorrelation of rates.

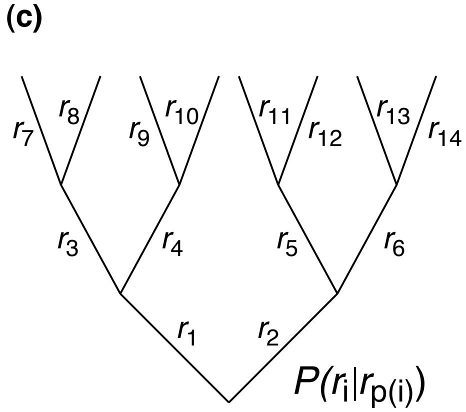
Jeff Thorne (Thorne et al., 1998. Mol. Biol. Evol., 15:1647.) has developed Bayesian methods to address autocorrelation of rates from an entirely parametric perspective using MCMC (e.g., Kishino et al., 2001. 18:352), and these approaches have been implemented in packages like BEAST (Drummond et al. 2006. PLoS Biology, 4:e88) and MULTIDIVTIME.

Here, node times and the correlated rates are treated as parameters that are estimated during the MCMC.

D. Uncorrelated Rates

A group of approaches relax the assumption that rates are correlated across the tree

Perhaps the most common approach is the Uncorrelated LogNormal (UCLN) approach of Drummond et al. (2006, PLoSBiology, 4:699.).



Here, rates are not constrained to be auto-correlated, but instead are drawn from a discretized LogNormal Distribution:

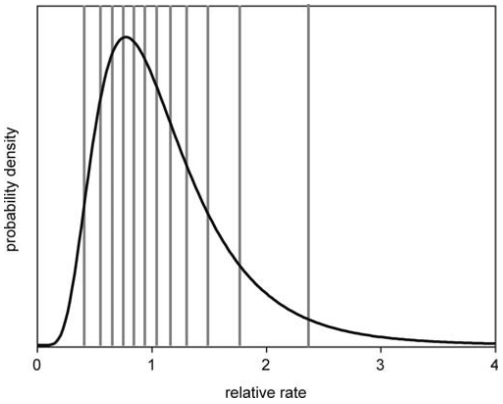
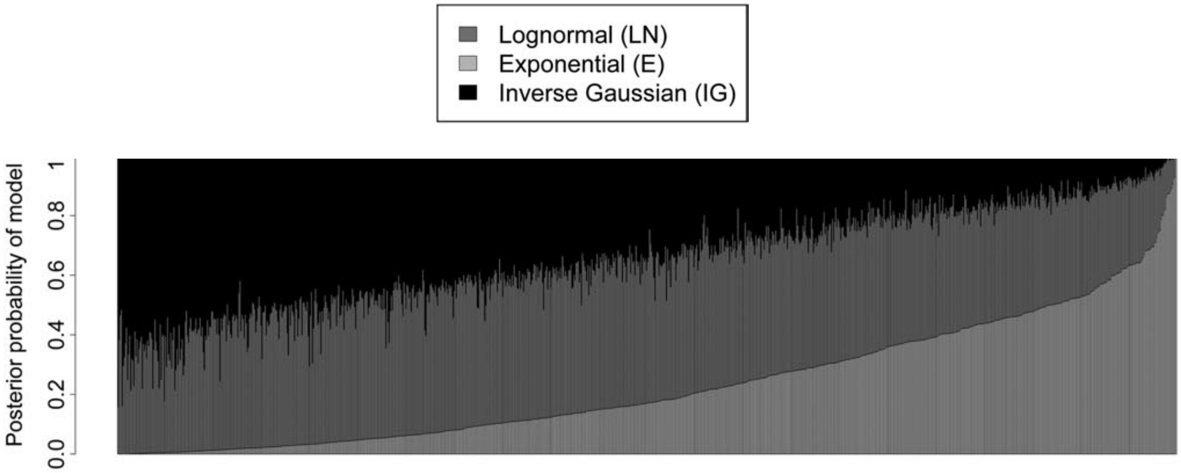


Figure 5. A Lognormal Distribution Discretized into 12 Rate Categories
Each of the 12 categories has equal probability ($p = 1/12$). The i^{th} rate category (numbered from left to right) corresponds to the (I − 0.5)/12 quantile of the lognormal distribution.
DOI: 10.1371/journal.pbio.0040088.g005

Again, node times and rates are treated as random variables that are estimated via MCMC.

Other distributions have been proposed (Inverse Gaussian, Exponential, etc.) and Li & Drummond (2012. MB&E) developed an rjMCMC to treat the relaxed clock model as a random variable and assess posterior probabilities of each of these three models for 1056 mammalian data sets.



Many data sets do not voice a preference for any of the three models, so model averaging is potentially a good way to go.

This approach has the advantage that phylogenetic uncertainty is also incorporated, because the implementation in BEAST includes the tree moves that we've discussed. In fact, Drummond et al. (2006) recommend that all phylogeny estimation be conducted this way.

It also has an enormous advantage over all other approaches with respect to how it incorporates fossil calibration data.

Node times of nodes for which there is fossil data are constrained using priors. This allows us to use a distribution to incorporate uncertainty in fossil dates in estimating divergence times.

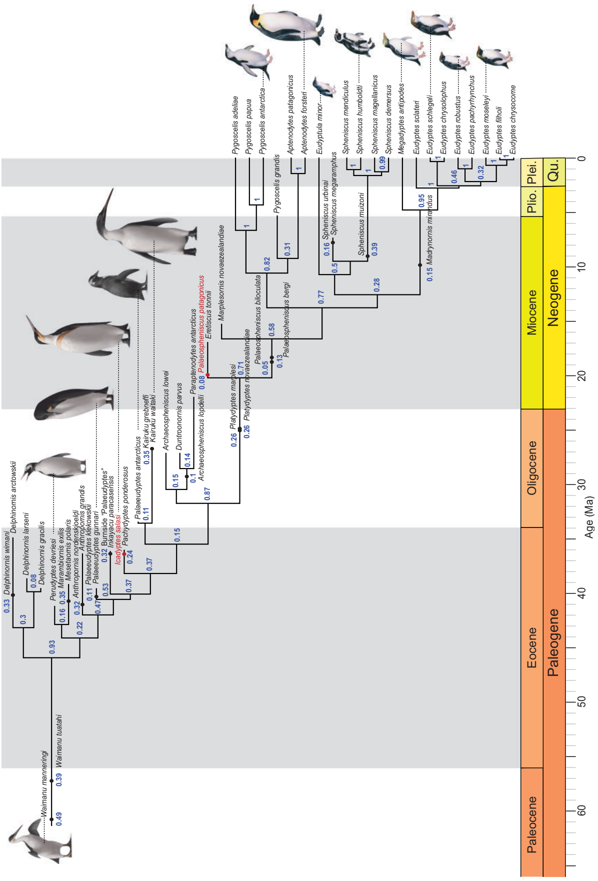
E. Inclusion of Fossils with Molecular Data

In using fossil dates to set priors on node ages, we need some idea of the topology of the tree, and we need to know with some degree of certainty where in the tree the calibration points are.

An emerging approach that circumvents this uses RevBayes (Hohna et al. 2016. Syst. Biol. 65:726) to include fossil taxa in the character by taxon matrix. The data then include both morphological and molecular characters and separate models are applied to each data type in a single analysis.

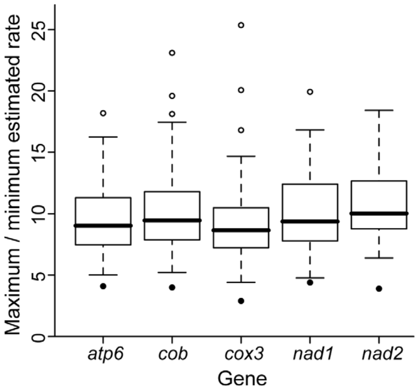
Tracey Heath gave an IBEST seminar on this approach using fossil and extant penguins last spring, and one of the huge advances is that the locations of the dated nodes in the phylogeny (i.e., the ones pertaining to the fossils) are estimated from the data, and the uncertainty in these placements is incorporated into divergence time estimates across the chronogram (i.e., dated phylogeny).

From Gavryushkina et al. (2016. Syst. Biol. 66:57).



F. A final note of caution.

A few studies (e.g., Schwartz & Mueller. 2010. PLoS One) have shown with simulated data that if data generated under a clock are subjected to these approaches, rates are usually not estimated to be equal.



To me this suggests that one needs to do one of the clock tests we discussed earlier and not apply relaxed-clock models unless a clock is (resoundingly) rejected. I think this needs much more attention.