NC STATE UNIVERSITY

Bayesian phylogenetic inference

Molecular Epidemiology of Infectious Diseases

Lecture 3

January 26th, 2026

Summary of Bayesian inference

Bayes theorem tells us how to compute conditional probabilities of the form *P(A|B)* given we have information about *P(A)*. *P(A)* represents our prior beliefs about *A*.

Bayes theorem lets us compute the posterior distribution of a variable by combining prior information with new information coming from the data through the likelihood function.

Both the posterior and the prior are probability distributions over an unobserved (random) variable.

For many problems, we cannot directly compute the posterior but we can approximate it using MCMC.

Bayesian phylogenetics

“Have patience with everything that remains unsolved in your heart… live in the question.”

-Rainer Maria Rilke

From ML to Bayesian phylogenetics

Maximum likelihood (ML) methods of tree reconstruction focus on finding the tree that maximizes the likelihood of the sequence data given some model of molecular evolution.

But a single best tree can be misleading in that there may be a large number of alternative trees that can explain the data nearly equally well.

In theory then, we would like to consider a “forest” of likely trees.

The Bayesian approach

Recall the general approach to Bayesian inference:

$$p(\theta|data) = \frac{L(data|\theta)}{p(data)} p(\theta)$$

The goal of Bayesian phylogenetics is to compute the posterior distribution of trees given a sequence alignment.

$$P(\mathcal{T}|Seq)$$

Tree space

The **posterior tree distribution** is a probability distribution over the forest of all possible trees in tree space.

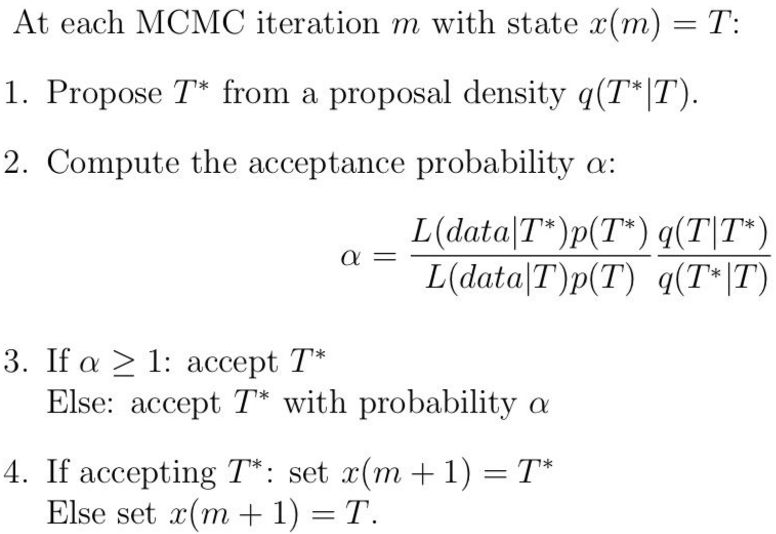
Tree space has a discrete component in that there are a finite number of possible **tree topologies** for a given number of taxa. But the number of possible topologies is typically huge.

Tree space also has a continuous component in that each branch has an associated length. There is therefore an infinite number of possible trees.

However we can still approximate the posterior tree distribution by sampling trees from the posterior using MCMC.

MCMC in tree space

We can approximate the posterior tree distribution by sampling a large number of trees from the posterior using MCMC. While tree space is complex, the basic idea is very similar to other MCMC algorithms like Metropolis-Hastings.



At each MCMC iteration m with state $x(m) = T$:

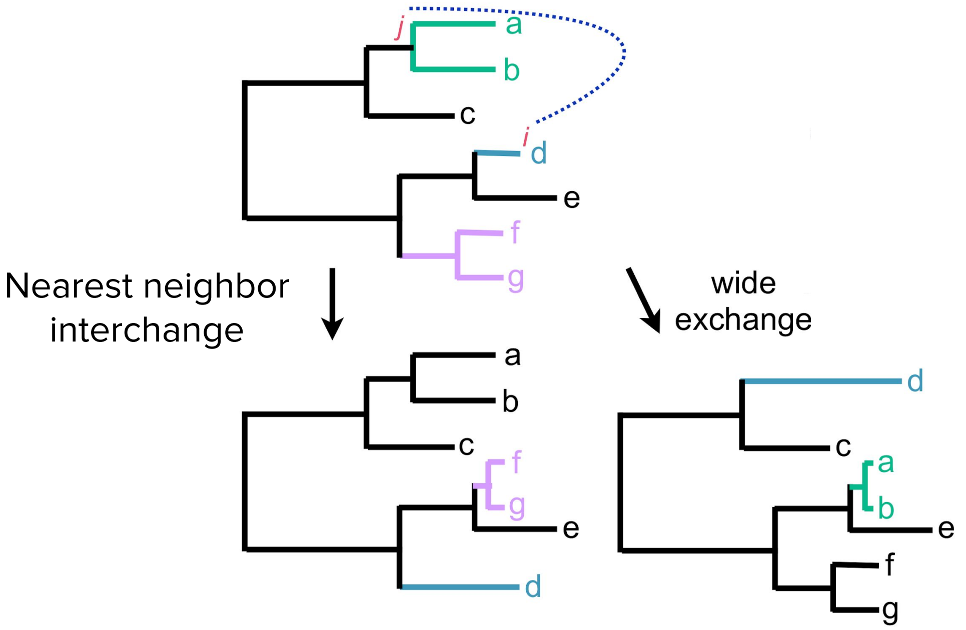
1. Propose T^* from a proposal density $q(T^*|T)$.
2. Compute the acceptance probability α:

$$\alpha = \frac{L(data|T^*)p(T^*)}{L(data|T)p(T)} \frac{q(T|T^*)}{q(T^*|T)}$$

3. If $\alpha \geq 1$: accept T^*
   Else: accept T^* with probability α
4. If accepting T^*: set $x(m+1) = T^*$
   Else set $x(m+1) = T$.

MCMC in tree space

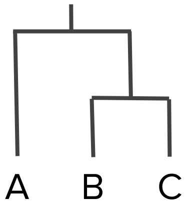
New tree topologies are proposed by rearranging the tree using subtree exchange.



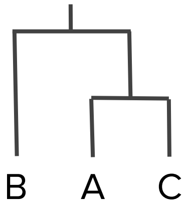
Tree support

In Bayesian phylogenetics, the support for a given tree or bipartition is given by its posterior probability. Posterior probabilities are direct measure of our (un)certainty.

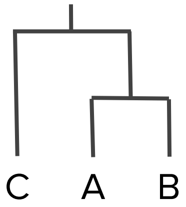
Using MCMC, the support or posterior probability for a given tree is approximated by its frequency in the posterior sample.



N=500
P=0.5



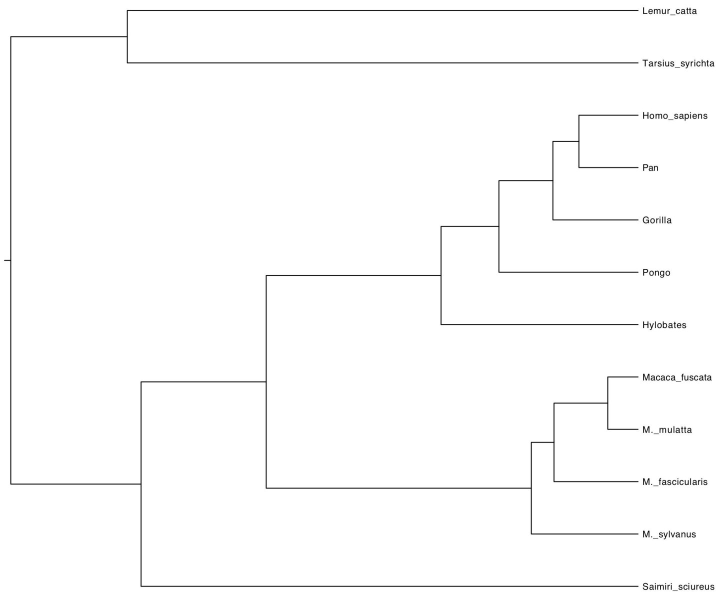
N=300
P=0.3



N=200
P=0.2

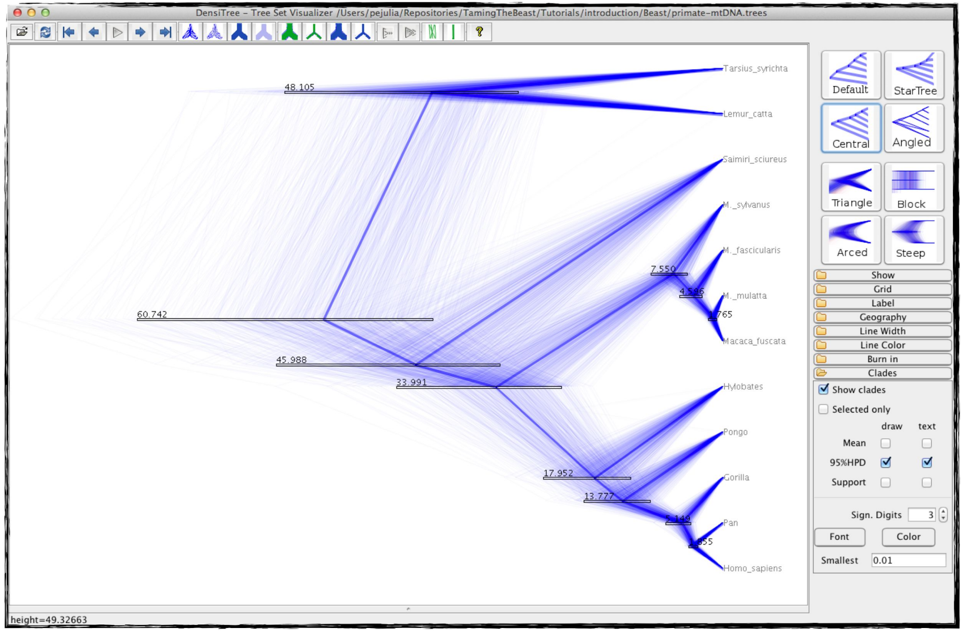
Visualizing the tree posterior

Often we summarize the tree posterior using a single consensus tree such as a Maximum Clade Credibility (MCC) tree.



Visualizing the tree posterior

DensiTree can also be used to overlay posterior tree samples.



DensiTree – Tree Set Visualizer /Users/pejulia/Repositories/TamingTheBeast/Tutorials/introduction/Beast/primate-mtDNA.trees

☑ Show clades

□ Selected only

	draw	text
Mean	□	□
95%HPD	☑	☑
Support	□	□

Sign. Digits 3

Smallest 0.01

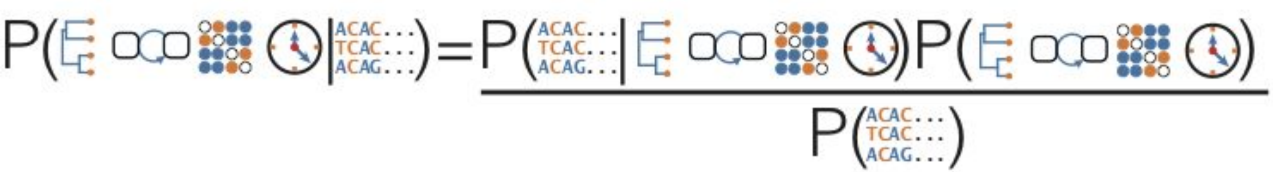
height=49.32663

Now with all the moving parts

So far we have just considered the posterior tree distribution

$$P(\mathcal{T}|Seq)$$

But typically there are several other parameters in the substitution model, molecular clock model and tree prior that must be jointly estimated together with the tree.



Now with all the moving parts

Tree

Substitution (Site) Model

Demographic Model (Tree Prior)

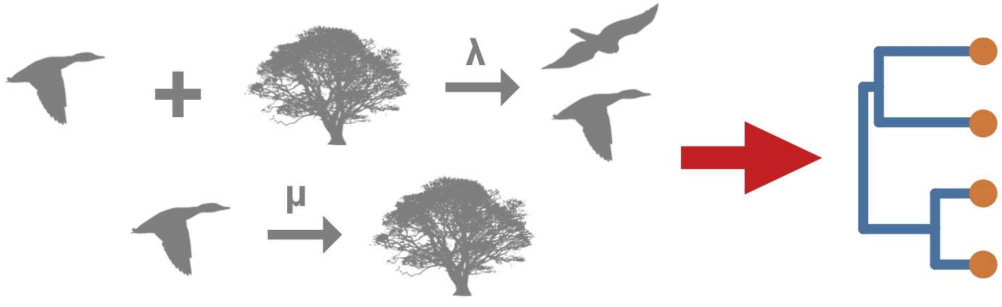
Molecular clock model

Tree priors

In Bayesian phylogenetics, we need to place a prior distribution over tree space.

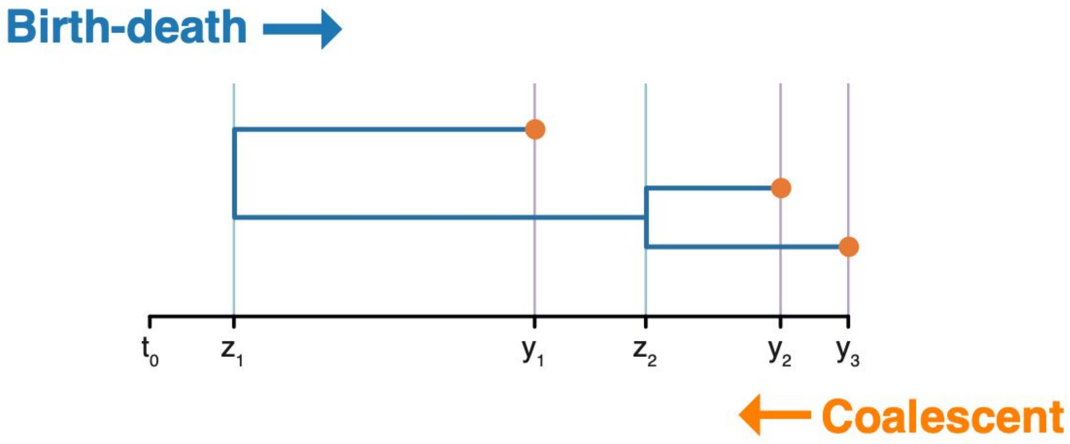
We could just assume that evolution is equally likely to produce any tree, resulting in a uniform prior over tree space.

But our tree prior should reflect our prior beliefs about the evolutionary processes that generated the tree (e.g. speciation and extinction rates).

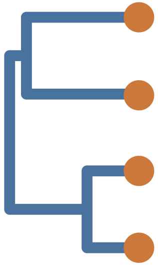


Tree priors

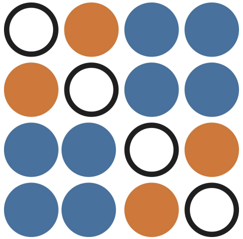
We will have several lectures on the *phylodynamic* models that are used as tree priors in molecular epidemiology, including coalescent and birth-death models.



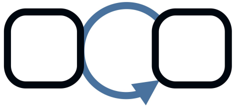
Now with all the moving parts



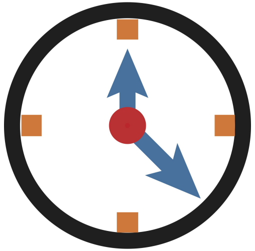
Tree



Substitution (Site) Model



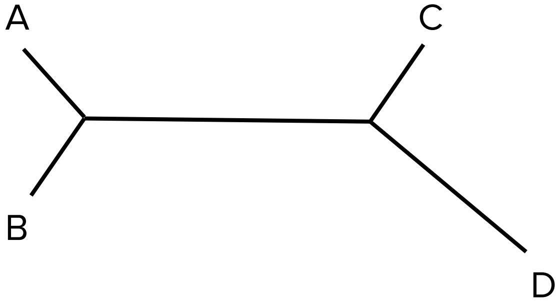
Demographic Model (Tree Prior)



Molecular clock model

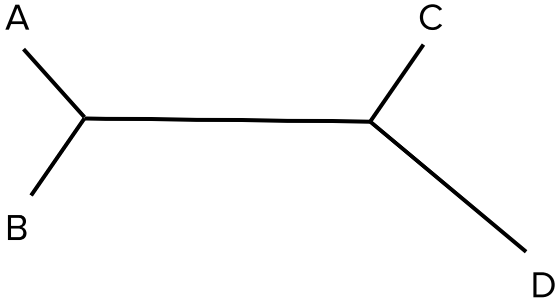
The dating problem

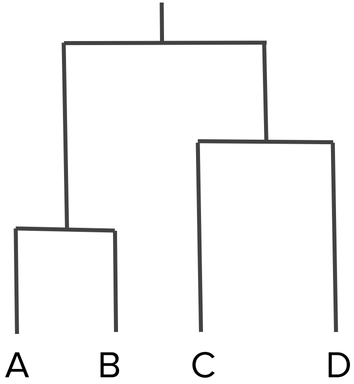
Most traditional (ML) phylogenetic methods infer unrooted trees that are not dated i.e. branch lengths are not time-calibrated in units of real time.



The dating problem

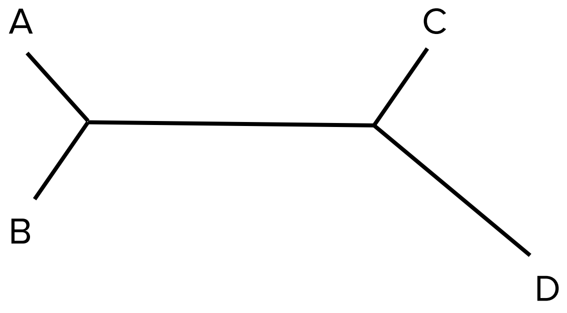
In Bayesian phylogenetics we normally work with rooted trees.



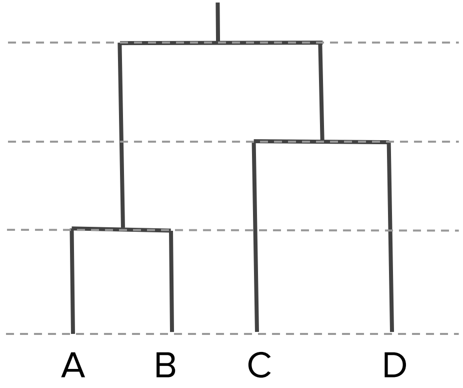


The dating problem

Rooting the tree places additional constraints on node heights and branch lengths.



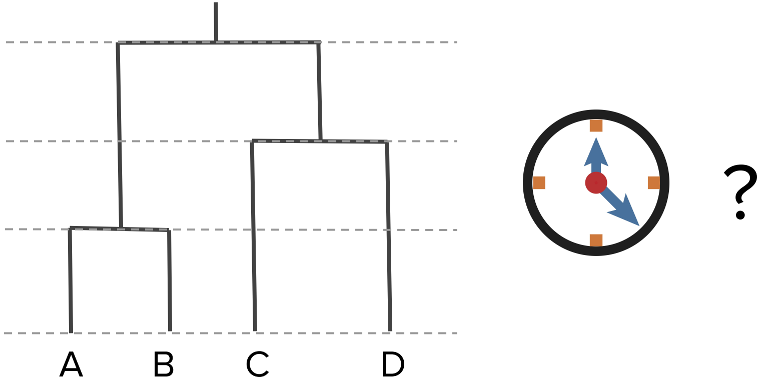
Branch lengths are unconstrained



Rooting the tree constrains branch lengths since all tips need to be equidistant to root

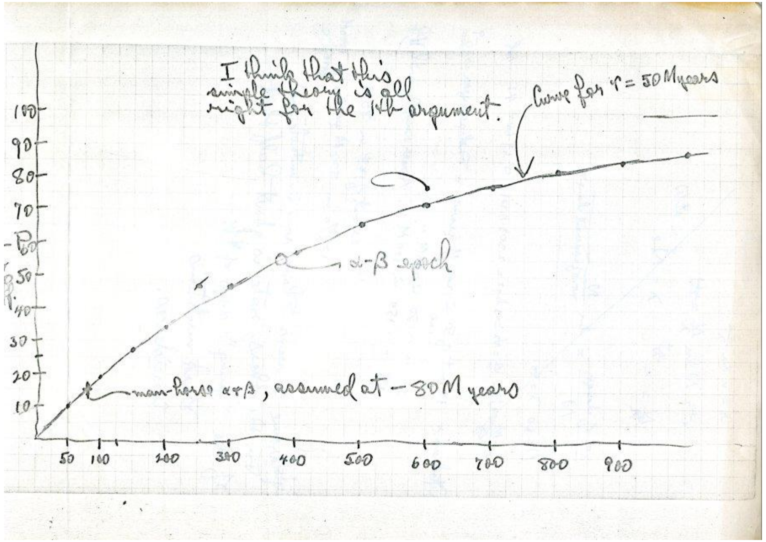
The dating problem

But how do we assign dates to the nodes in the tree?



The molecular clock

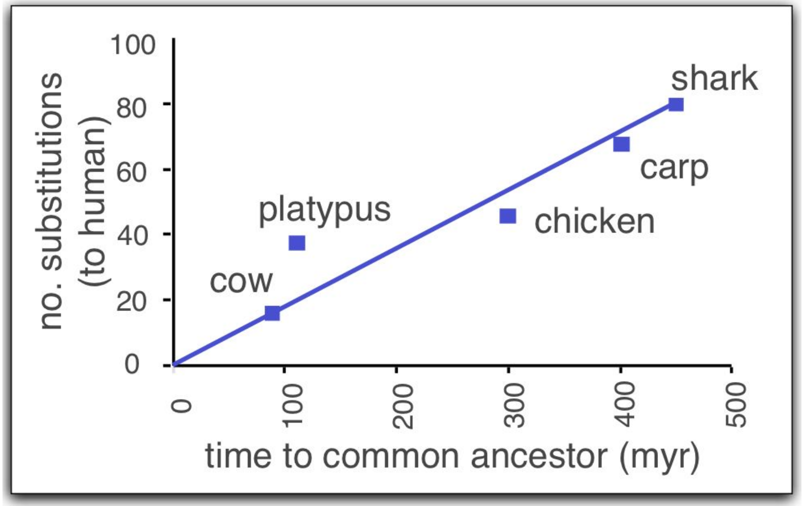
Zuckerkandl and Pauling (1962) found that the genetic distance between different animal haemoglobins increases almost linearly with the divergence time of the two species.



Linus Pauling to Emile Zuckerkandl. September 12, 1964
http://scarc.library.oregonstate.edu/coll/pauling/blood/corr/index.html

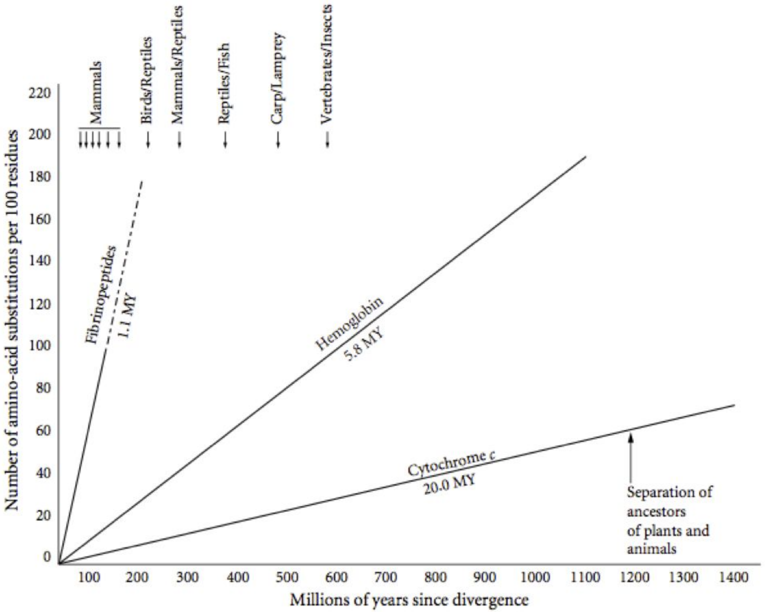
The molecular clock

Zuckerkandl and Pauling (1962) found that the genetic distance between different animal haemoglobins increases almost linearly with the divergence time of the two species.



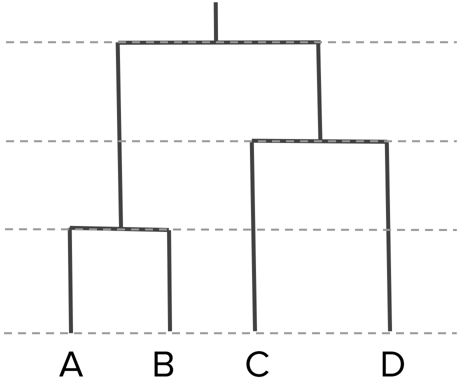
The molecular clock

This pattern holds for many different proteins, implying that substitutions accumulate at a constant rate over time but at different rates in different proteins.



The molecular clock model

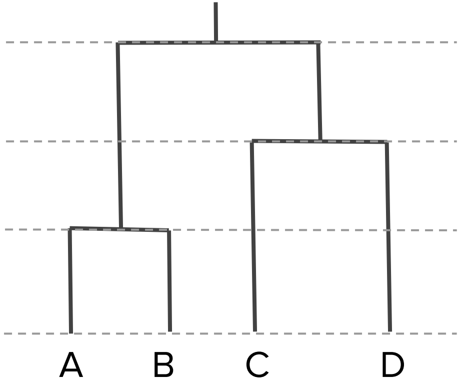
Genetic distance = clock rate X time



$$\text{Time} = \frac{\text{Genetic distance}}{\text{clock rate}}$$

The molecular clock model

But we need additional information/constraints in order to estimate the clock rate in terms of absolute time..



$$\text{Time} = \frac{\text{Genetic distance}}{\text{clock rate}}$$

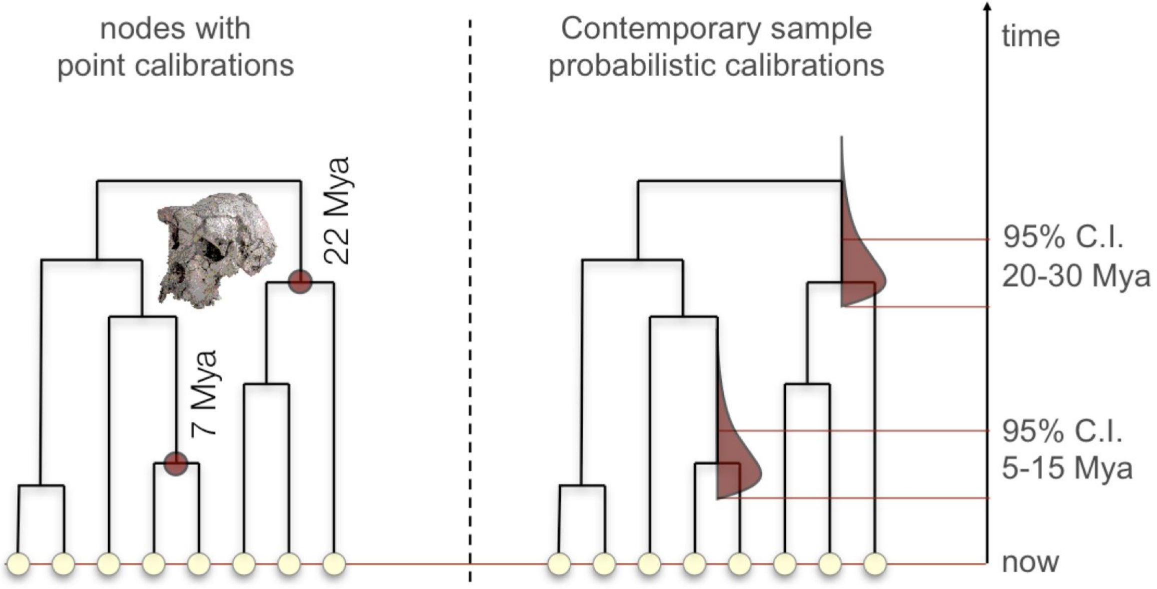
Two ways to calibrate the molecular clock

1. Time point calibrations on divergence times

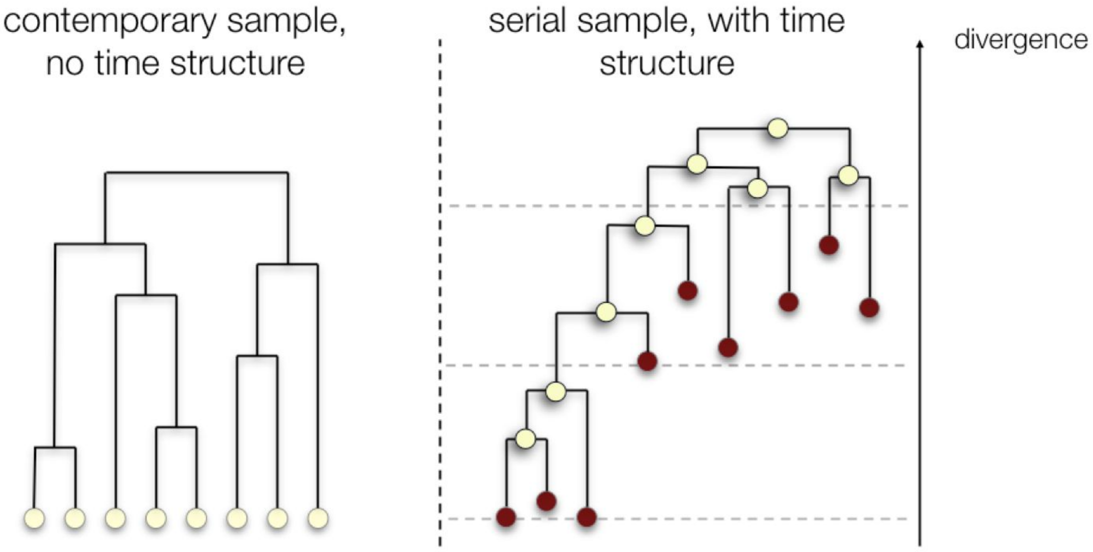
2. Serially sampled (heterochronous) sequence data

Calibrating from divergence times

Specifying divergence times on particular nodes.



Calibrating from serial tip times



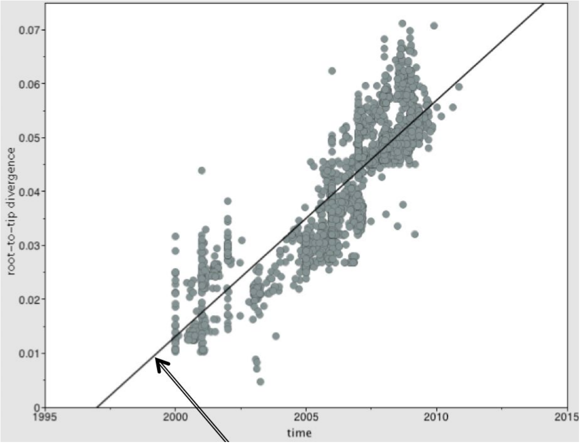
Rambaut A. (2000) *Bioinformatics*, **16**, 395-399.

Calibrating from serial tip times

In Bayesian phylogenetics, we jointly infer the clock rate and all node heights/branch lengths using MCMC. But the idea is conceptually similar to regressing root-to-tip genetic divergence against sampling times in terms of where the information is coming from. The slope is an estimate of the clock rate..

Influenza A H1N1
2000-2011

Clock rate = 4.38 x 10^{-3}



Calibrating from serial tip times

There must be sufficient time for mutations to occur between sampling events in order for the clock rate to be estimated.

The slower the mutation rate or the more closely together sequences are sampled in time the less information there will be about the clock rate.

Now with all the moving parts

Tree

Substitution (Site) Model

Demographic Model (Tree Prior)

Molecular clock model

Conclusion

Bayesian phylogenetic analysis can be complex with many interacting models each with their own set of parameters/priors.

Bayesian phylogenetics lets us quantify uncertainty about these parameters and the phylogeny by inferring the posterior tree distribution using MCMC.

We may only be interested in the tree or specific parameters, but performing joint inference in a Bayesian setting allows us to take into account uncertainty about all parameters including the tree.

Why Beast2?

Beast2

Bayesian evolutionary analysis by sampling trees

Implements many popular evolutionary and phylodynamic models. Plus many add-on packages.

Very efficient MCMC due to optimized proposals

Written in Java, runs everywhere.

Well-documented with lots of online community support. Check out https://taming-the-beast.org/tutorials/