Using phylogenetics to estimate species divergence times ...

More accurately ...

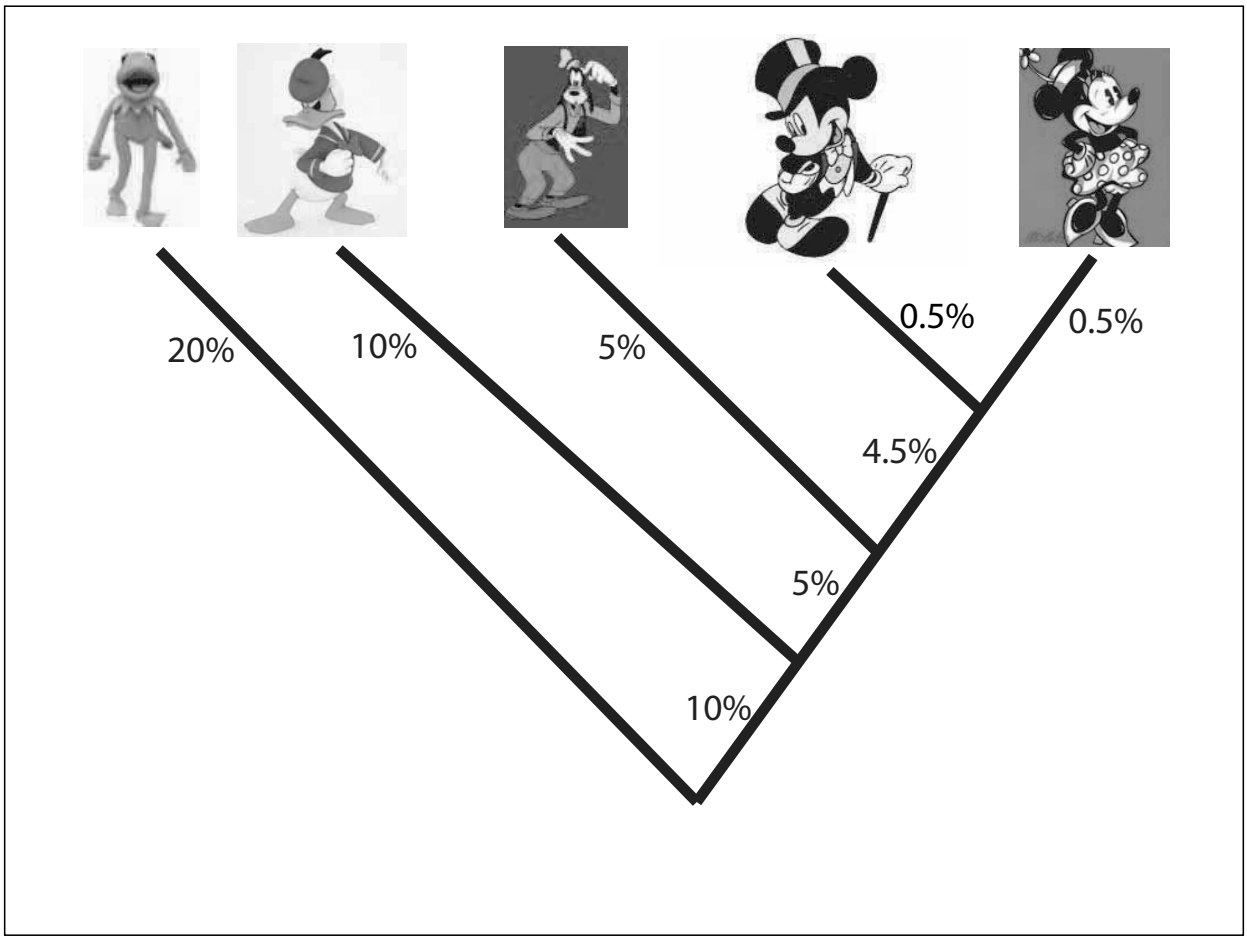
Basics and basic issues for Bayesian inference of divergence times (plus some digression)

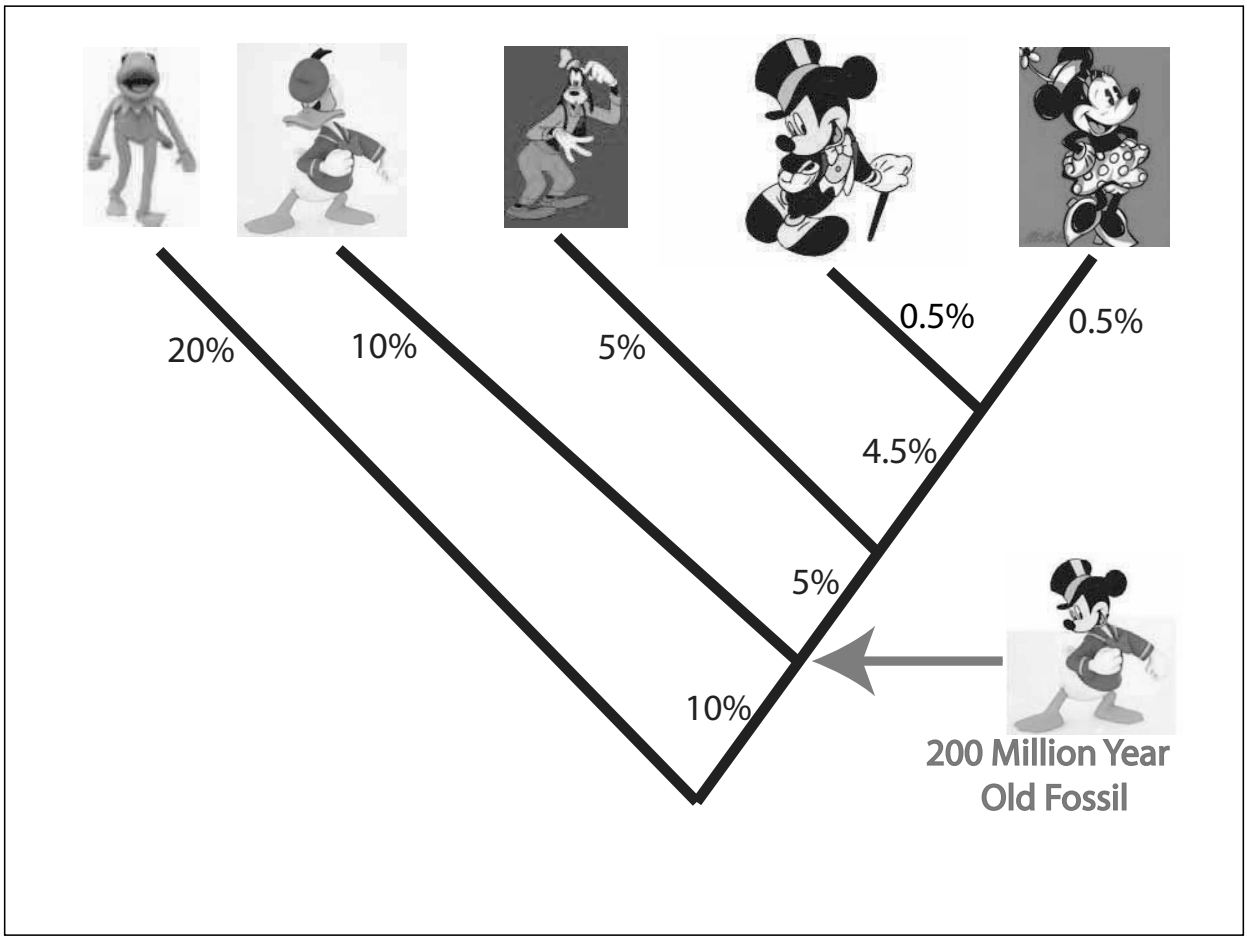
"A comparison of the structures of homologous proteins ... from different species is important, therefore, for two reasons. First, the similarities found give a measure of the minimum structure for biological function. Second, the differences found may give us important clues to the rate at which successful mutations have occurred throughout evolutionary time and may also serve as an additional basis for establishing phylogenetic relationships."

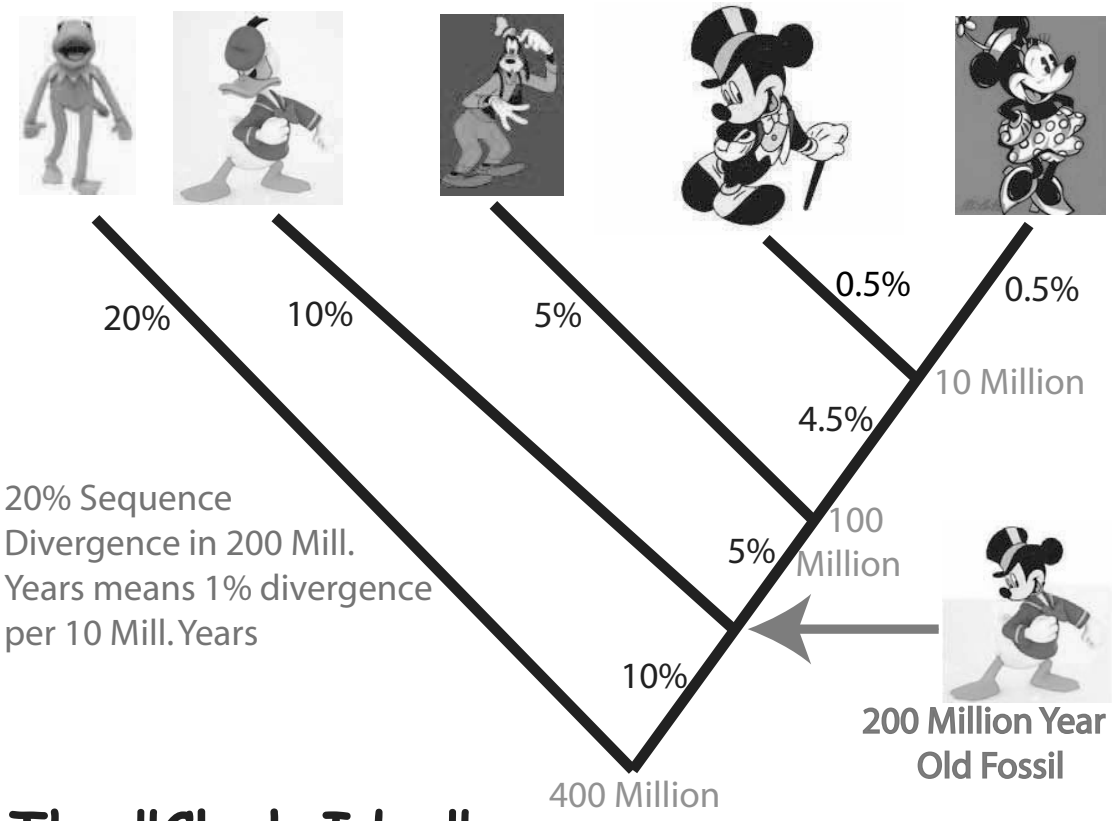
From p. 143 of

The Molecular Basis of Evolution

by Dr. Christian B. Anfinsen (Wiley, 1959)


20%
10%
5%
0.5%
0.5%
4.5%
5%
10%

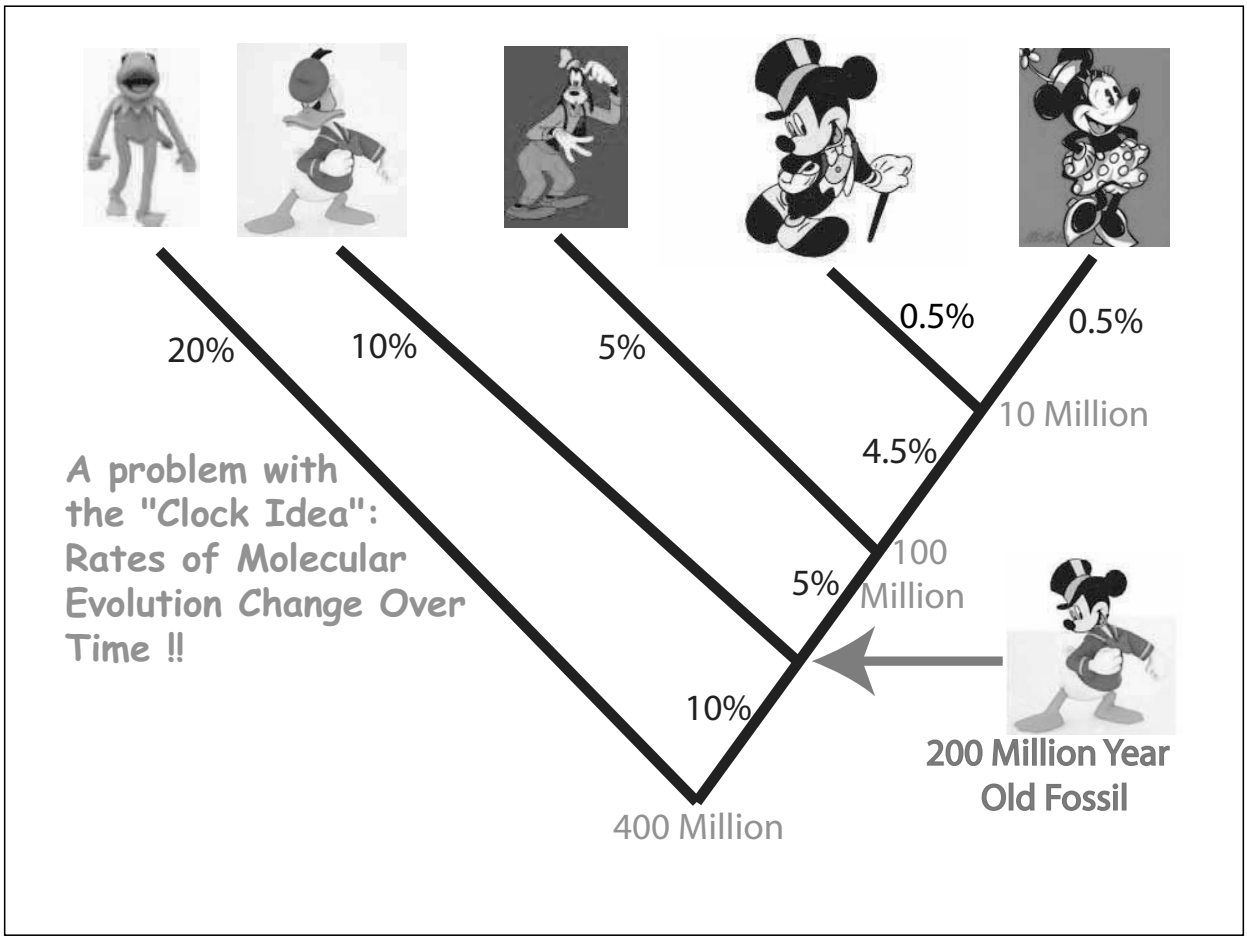

20%
10%
5%
0.5%
0.5%
4.5%
5%
10%
200 Million Year
Old Fossil

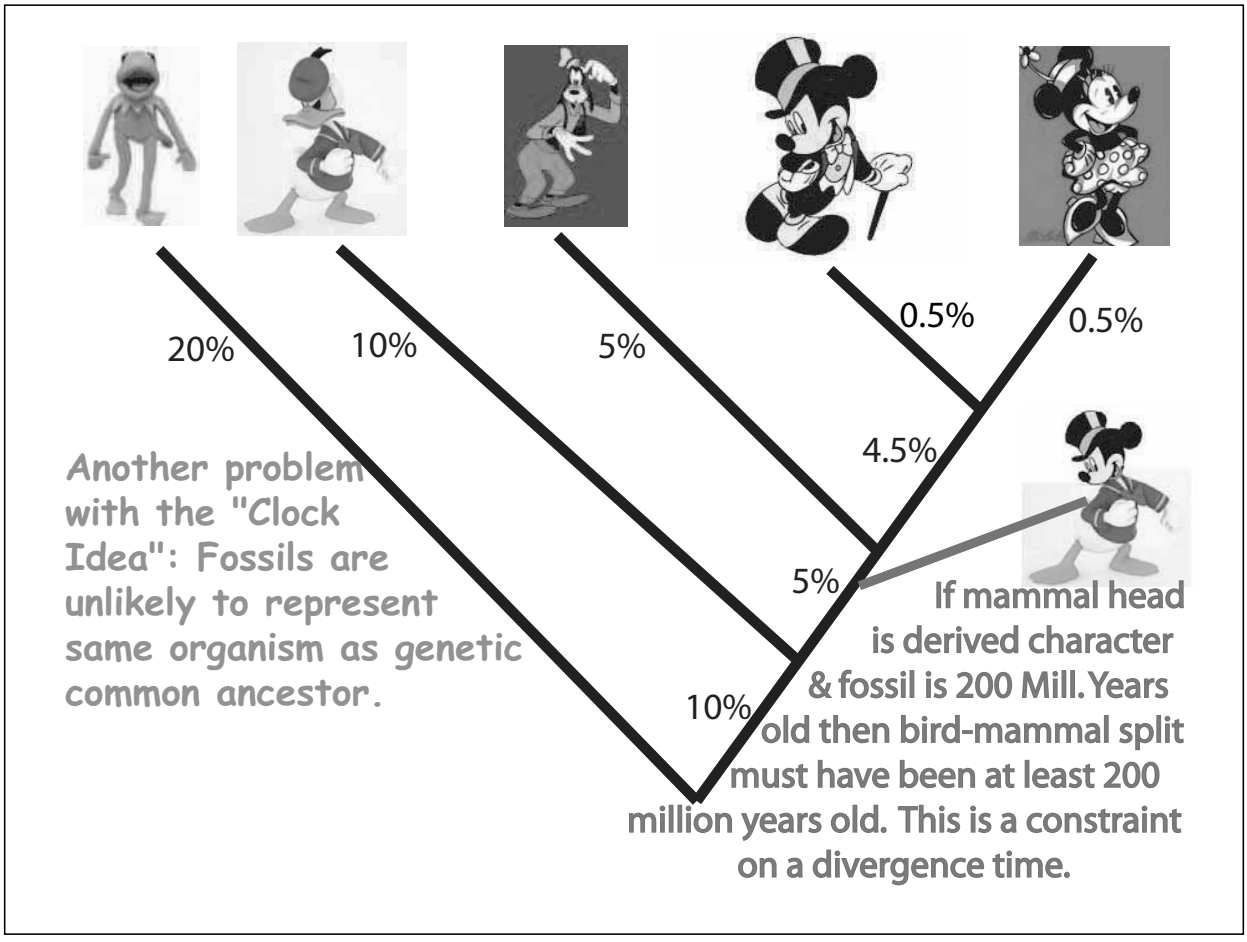


The "Clock Idea"

"Ernst Mayr recalled at this meeting that there are two distinct aspects to phylogeny: the splitting of lines, and what happens to the lines subsequently by divergence. He emphasized that, after splitting, the resulting lines may evolve at very different rates... How can one then expect a given type of protein to display constant rates of evolutionary modification along different lines of descent?"

(Evolving Genes and Proteins. Zuckerkandl and Pauling, 1965, p. 138).


20%
10%
5%
0.5%
0.5%
10 Million
4.5%
100 Million
5%
10%
400 Million
A problem with the "Clock Idea": Rates of Molecular Evolution Change Over Time !!
200 Million Year Old Fossil


20%
10%
5%
0.5%
0.5%
4.5%
5%
10%
Another problem with the "Clock Idea": Fossils are unlikely to represent same organism as genetic common ancestor.
If mammal head is derived character & fossil is 200 Mill. Years old then bird-mammal split must have been at least 200 million years old. This is a constraint on a divergence time.

Bayesian Idea:

(Prior Information)

X

(Information from data)

= Posterior Information

Basic Idea for Bayesian Divergence Time Inference

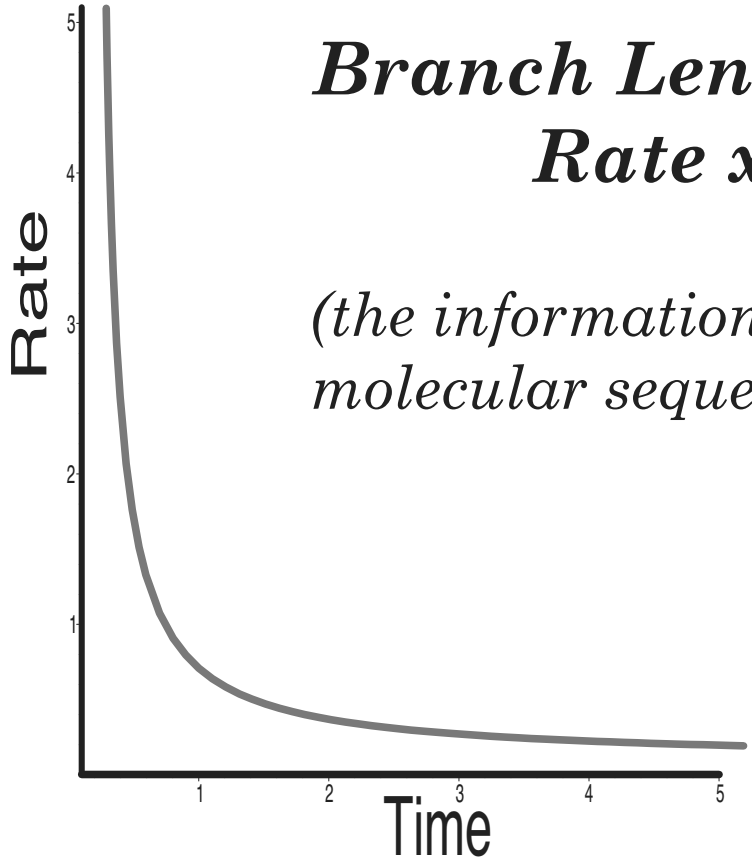
R: rates
T: node times
C: Fossil Evidence (constraints)
S: Sequence Data

$$P(R,T|S,C) = \frac{P(S,R,T|C)}{P(S|C)} = \frac{P(S|R,T,C)\,P(R|T,C)\,P(T|C)}{P(S|C)}$$

$$= \frac{P(S|R,T)\,P(R|T)\,P(T|C)}{P(S|C)}$$

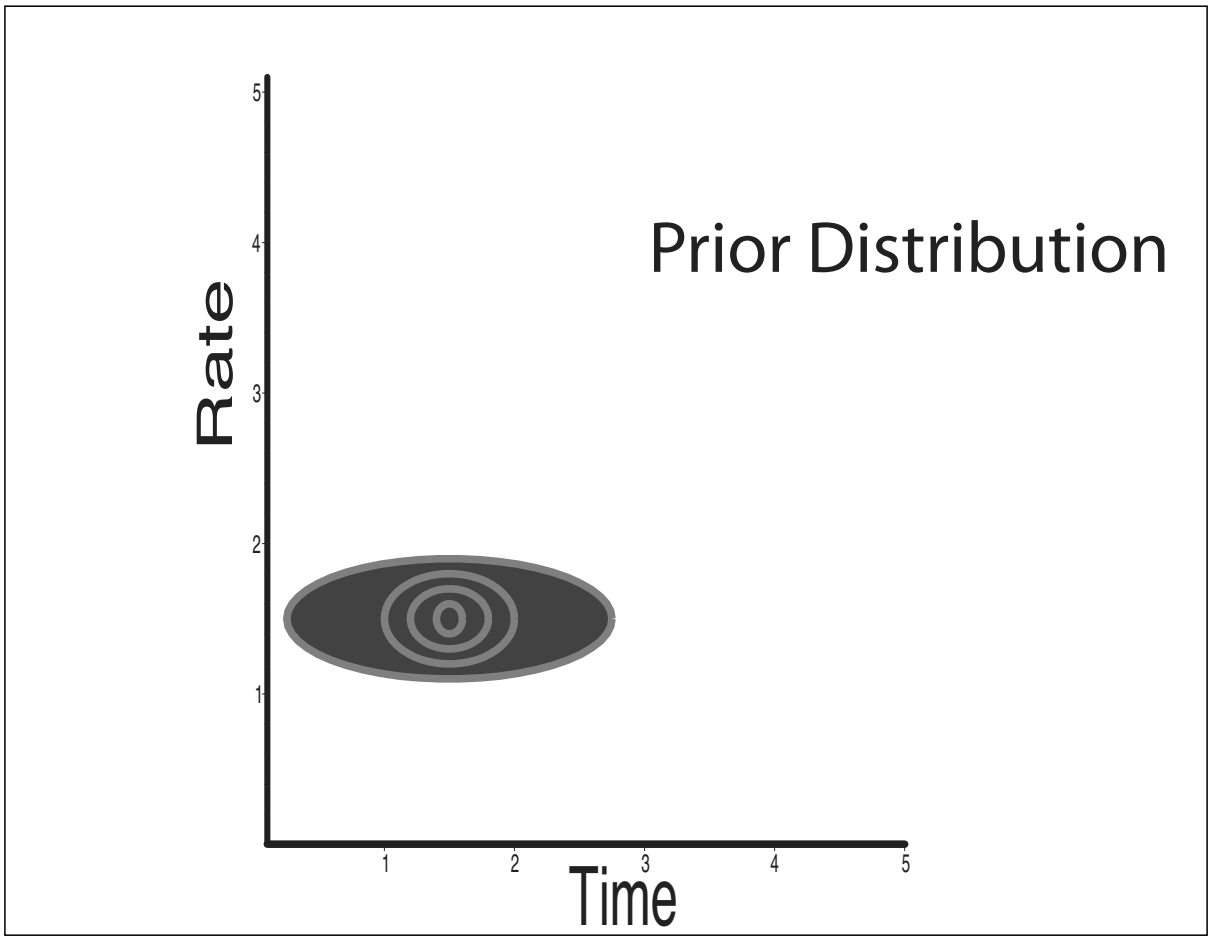
(Relaxed Clock) Bayesian Divergence Time Components

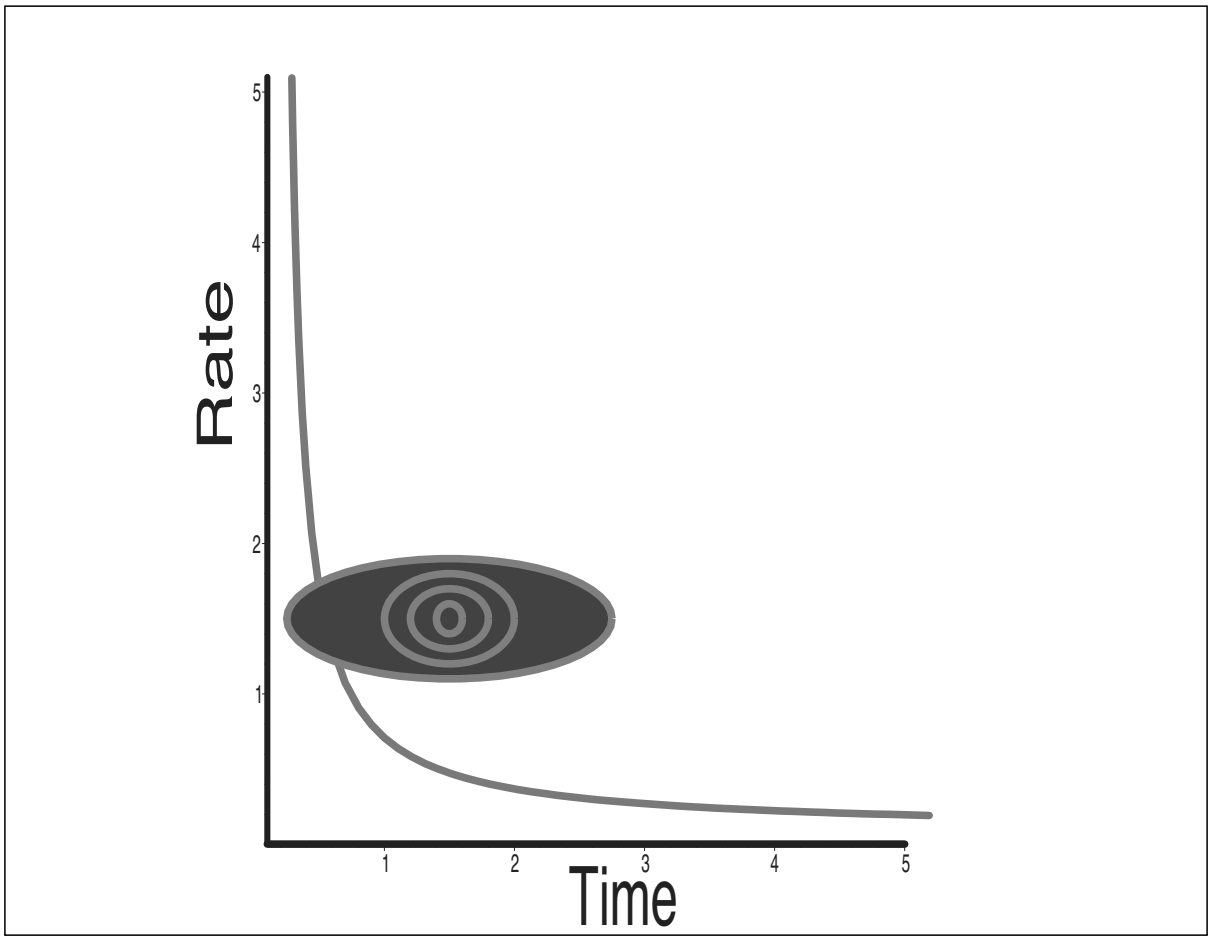
1. DNA or protein sequence data
2. Model of Sequence Change
3. Model of Rate Change
4. Prior Distributions for Rates, Times, etc.
5. Fossil or other information

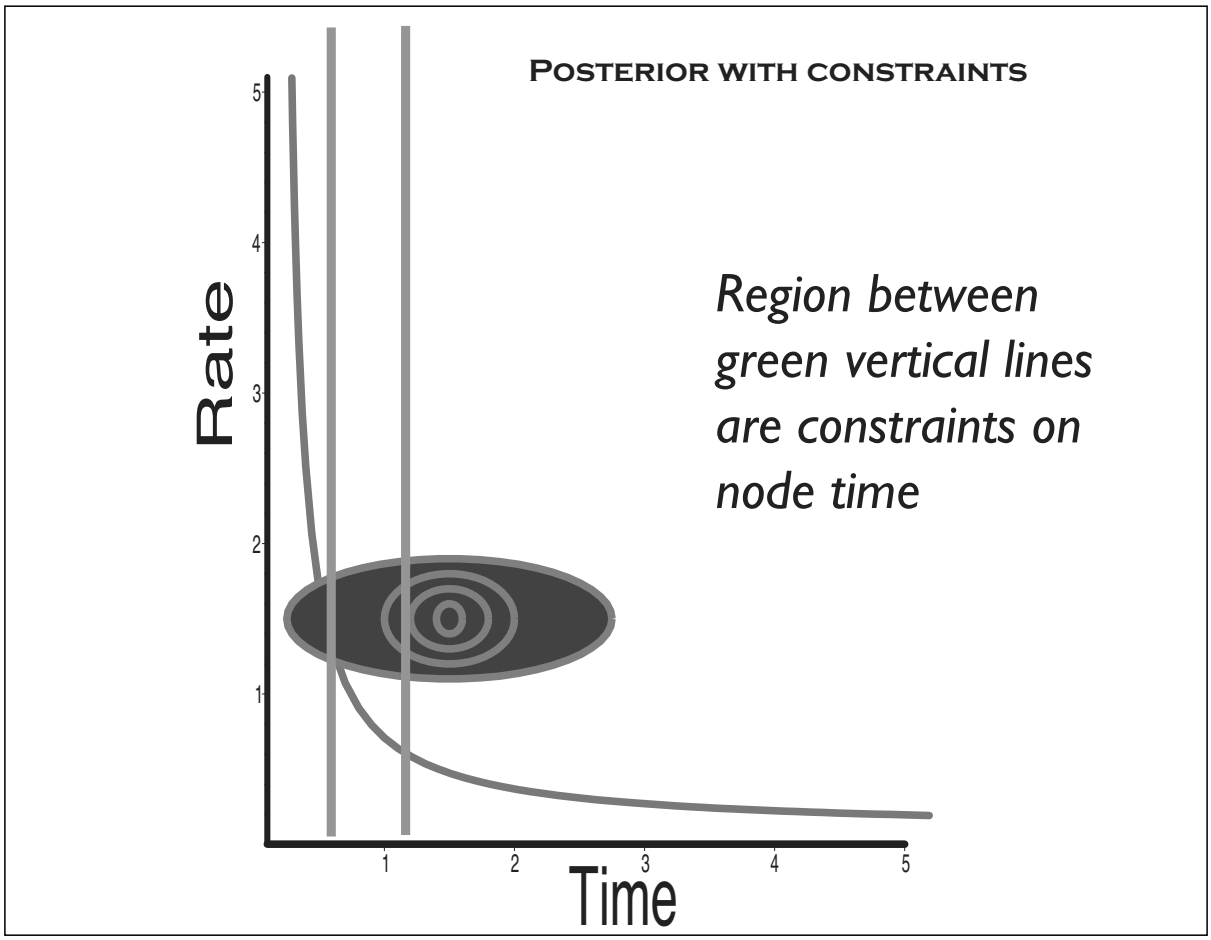


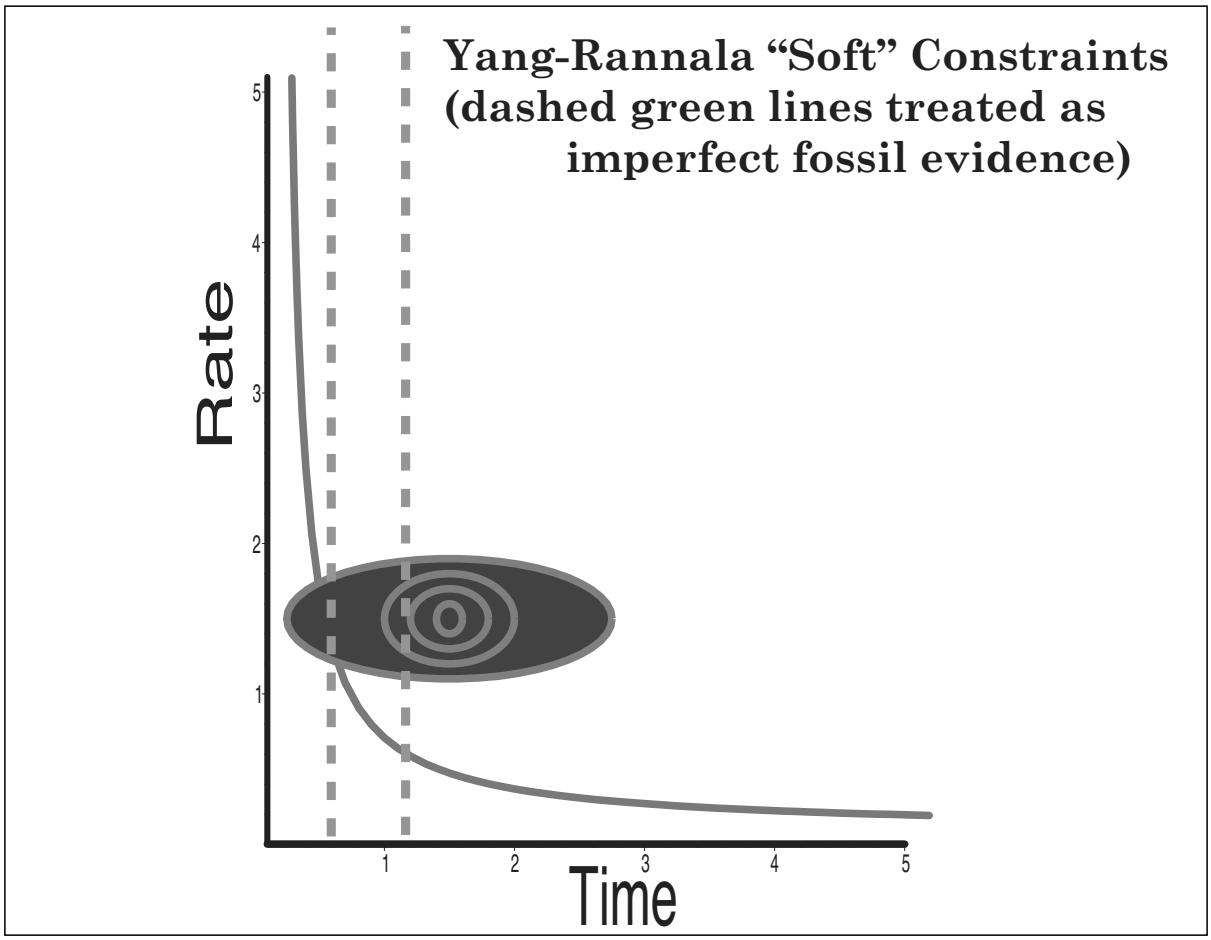
Branch Length = Rate x Time

(the information from molecular sequence data)


Prior Distribution
Rate
Time
1
2
3
4
5


Rate
Time
1
2
3
4
5

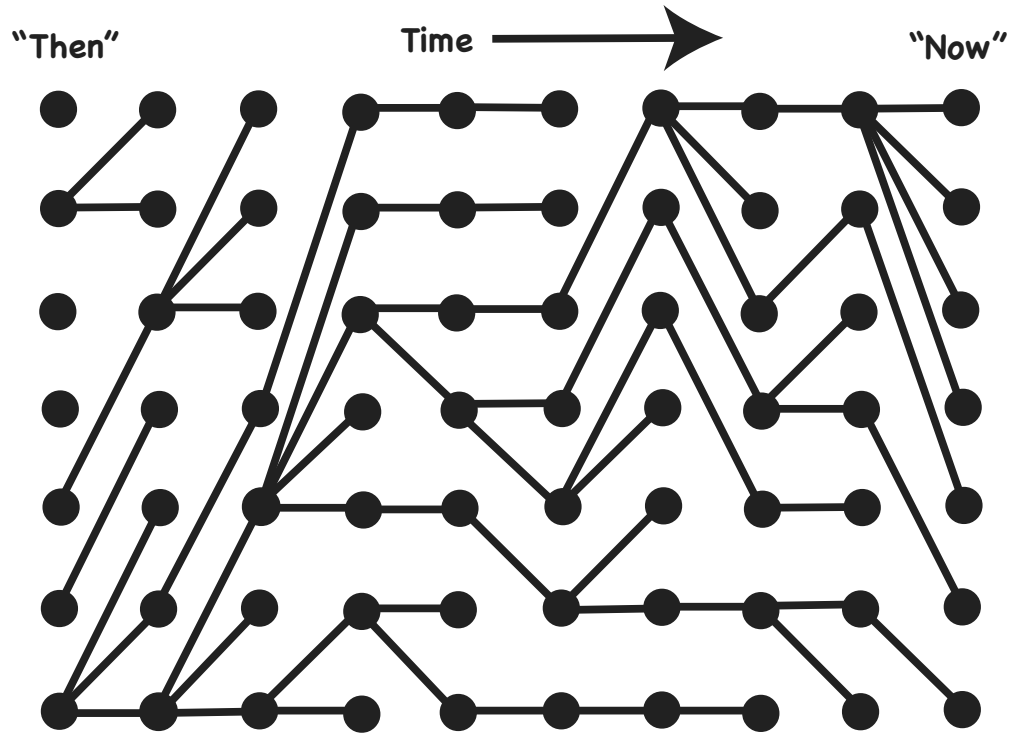

Posterior with constraints
Region between green vertical lines are constraints on node time
Rate
Time
1
2
3
4
5

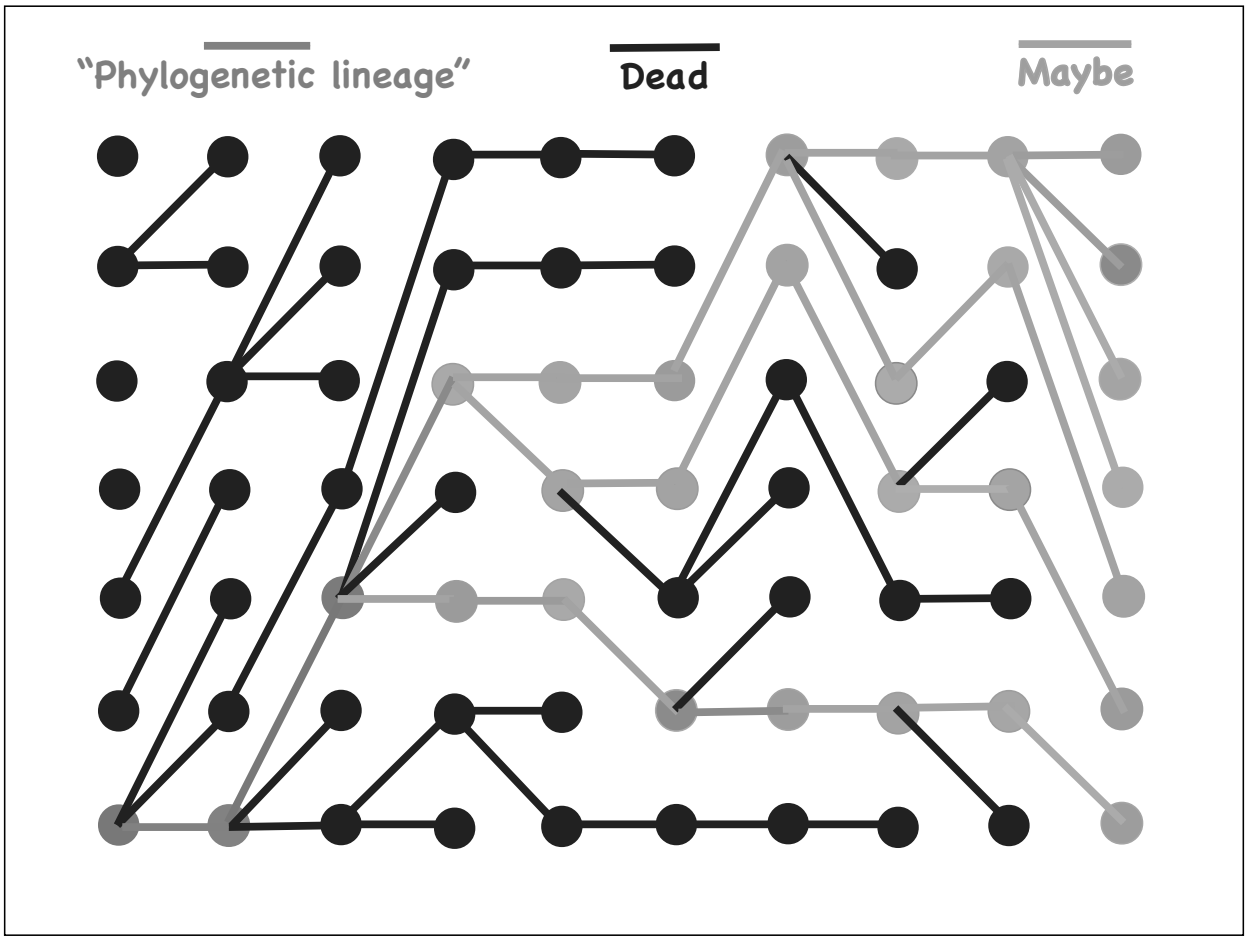

Yang-Rannala "Soft" Constraints (dashed green lines treated as imperfect fossil evidence)
Rate
Time
1
2
3
4
5

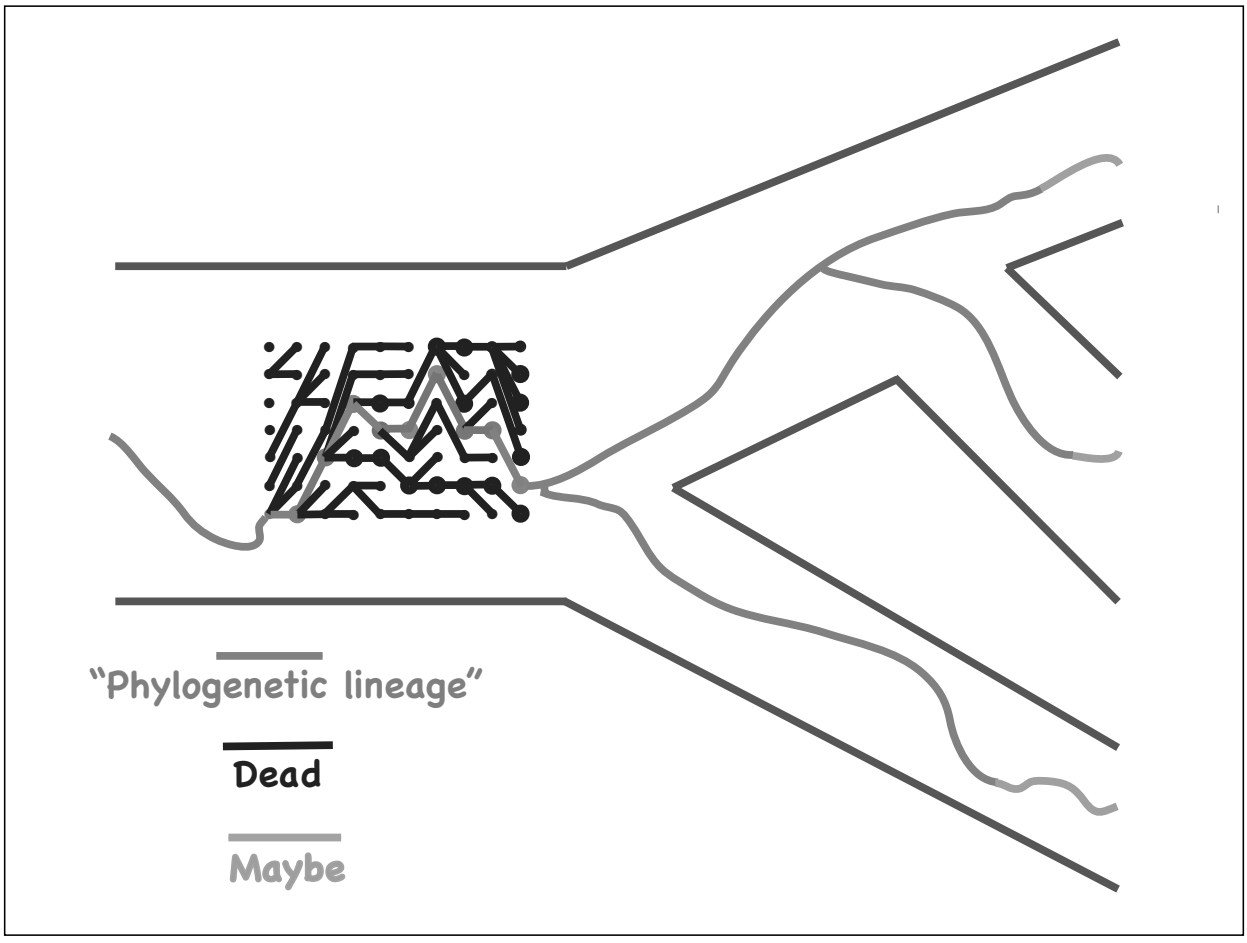
A digression:

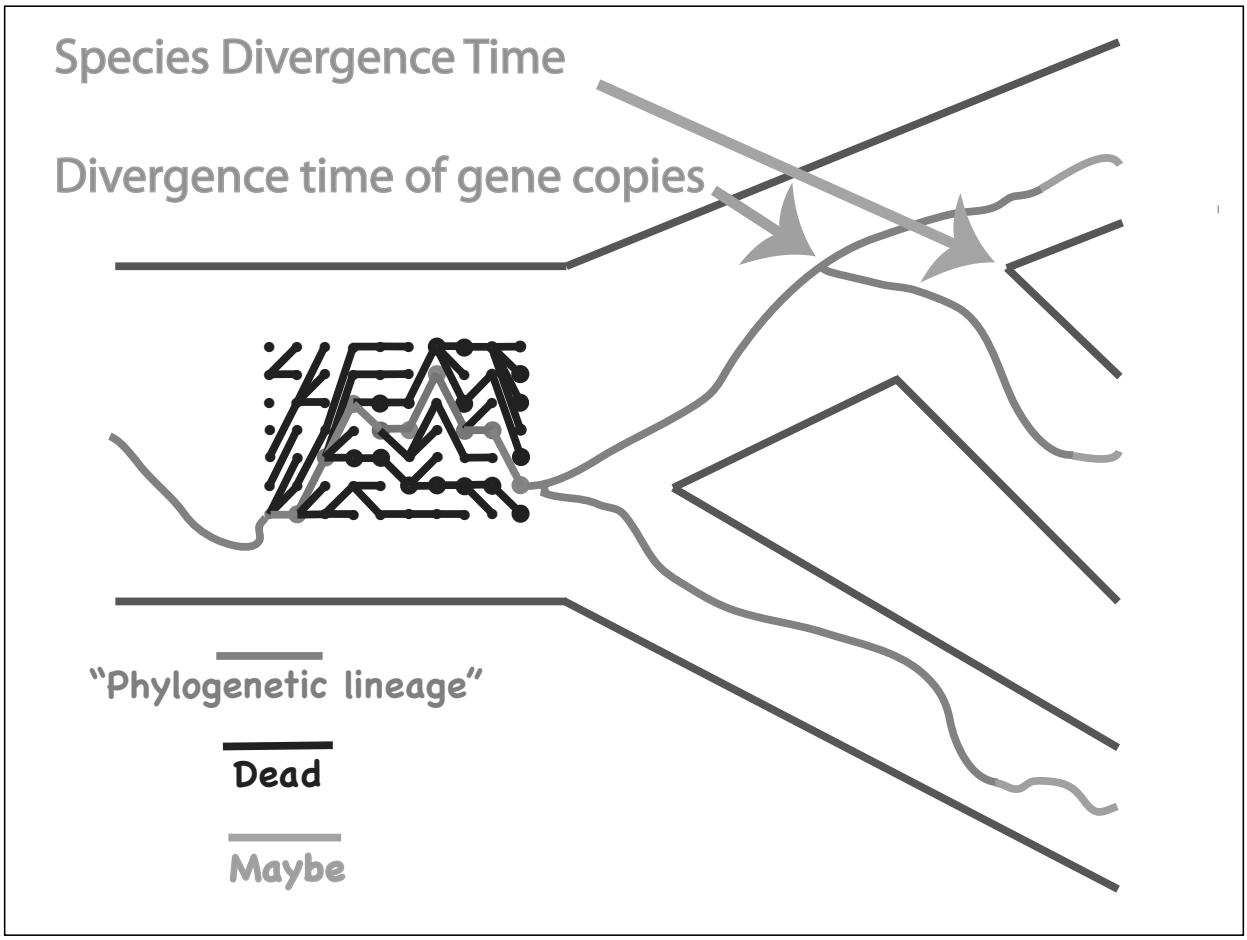
What are we really estimating when we estimate "divergence" times?

History of gene copies in a population

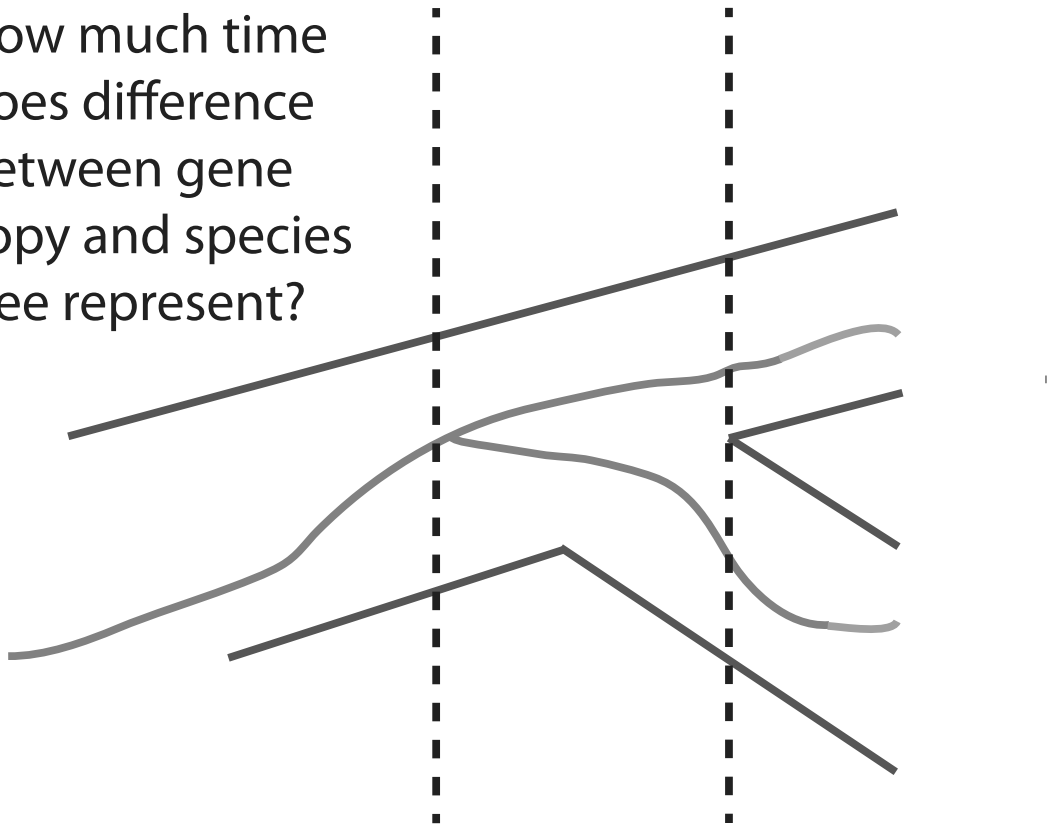



"Phylogenetic lineage"
Dead
Maybe


"Phylogenetic lineage"
Dead
Maybe

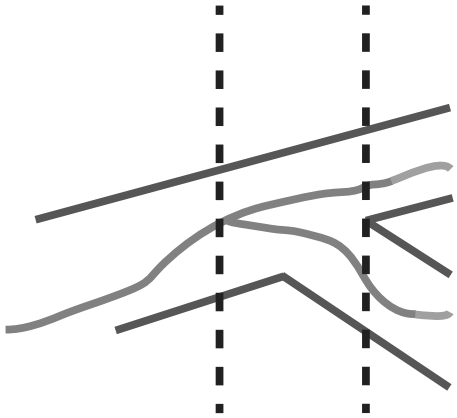


How much time does difference between gene copy and species tree represent?



How much time does difference between gene copy and species tree represent?

(N_e is effective population size)



For a coalescent process with diploid organisms, average time difference is $2N_e$ generations and standard deviation is also $2N_e$ generations ...

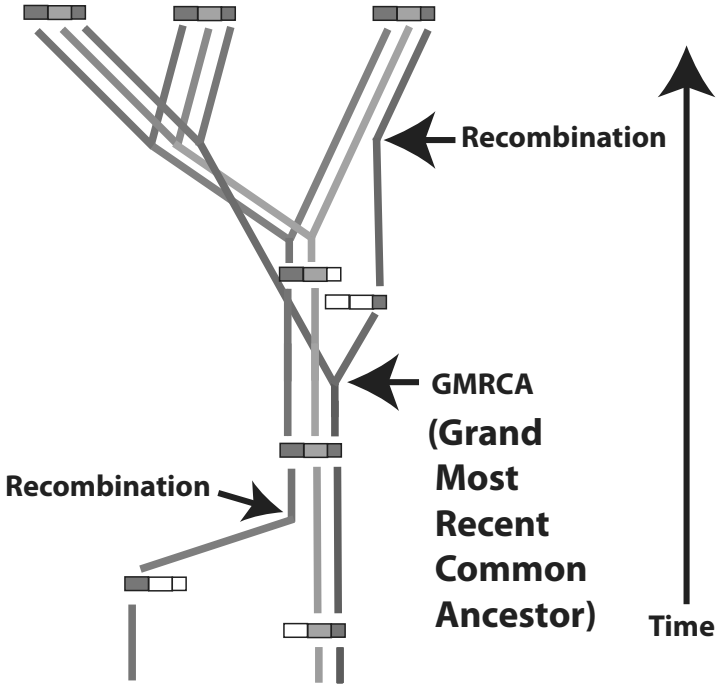
When time needed for $2N_e$ generations is large relative to species divergence times, be careful ...

and try *BEAST or BEST software?

See:

Heled & Drummond. 2012. MBE 27:570-580
Liu. 2008. Bioinformatics 24:2542-2543.

Recombination is another divergence time (and phylogenetic) challenge!



End of digression on ...

What are we really estimating when we estimate "divergence" times?

Bayesian Divergence Time Components

1. DNA or protein sequence data

Sequence data is needed for branch length (rate x time) estimation.

Sequence data does not separate rates and times.

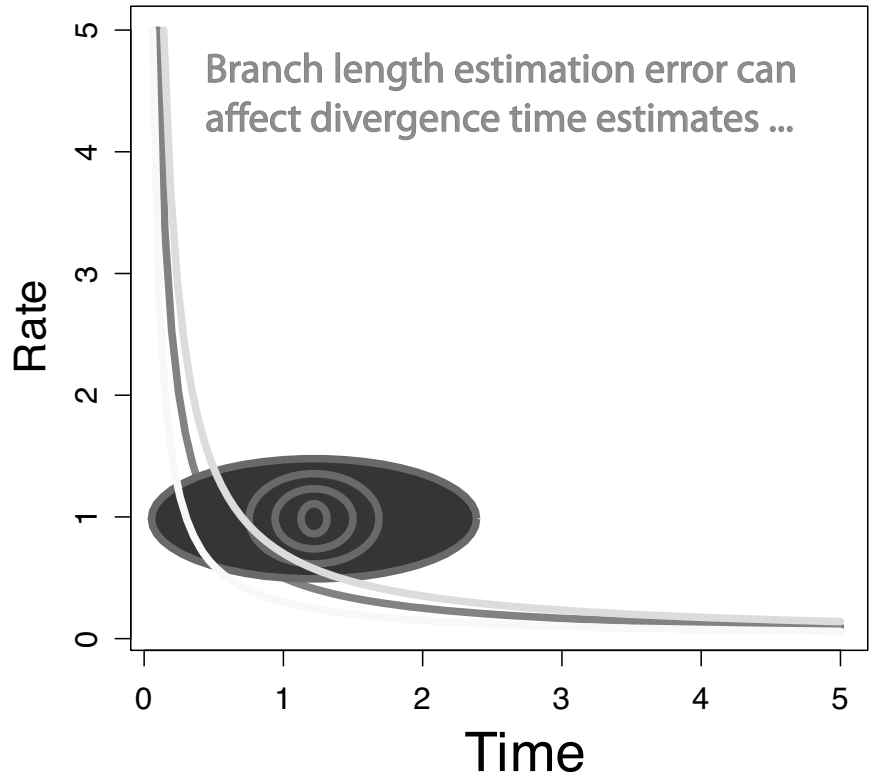
Better to invest in improving other time estimation components?

Bayesian Divergence Time Components

2. Model of Sequence Change

Branch Length (BL) Errors → **Divergence Time Errors**

Posterior distributions for times are compromise between branch length information from sequence data and prior information and fossil information.



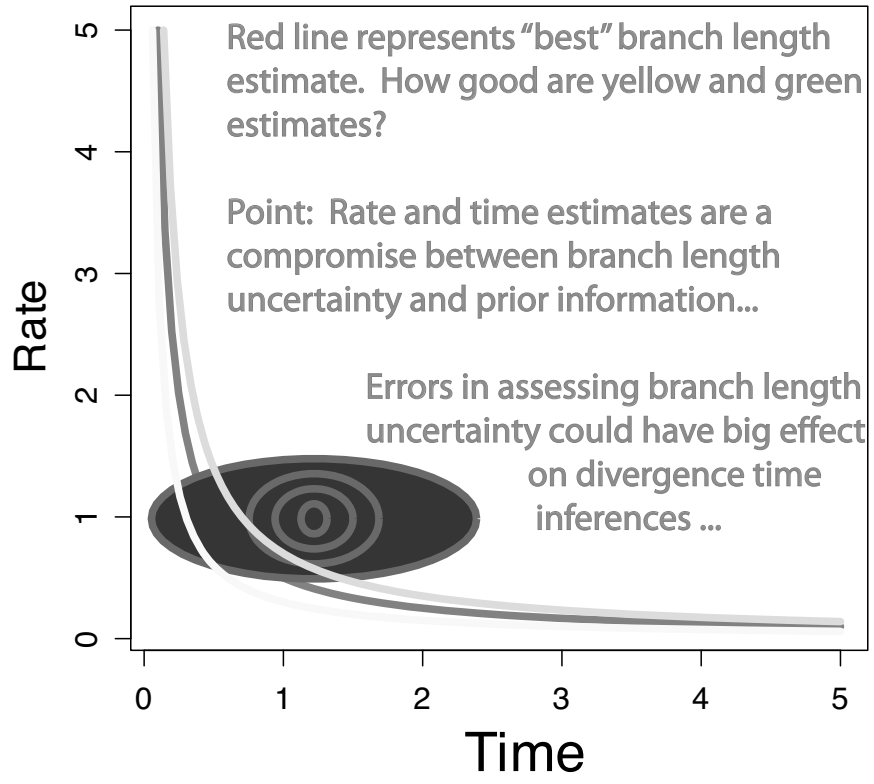
Bayesian Divergence Time Components

2. Model of Sequence Change

Branch Length (BL) Errors → **Divergence Time Errors**

Errors in BL uncertainty → **Divergence Time Errors**

Posterior distributions for times are compromise between branch length information from sequence data and prior information and fossil information.



Errors in BL uncertainty have more serious consequences for divergence time estimation than for phylogeny inference.

Sources of these errors include failure to account for dependent change among sequence positions.

Context-Dependent Mutation
Codons
Protein Tertiary Structure
RNA Secondary Structure
Other Genotype-Phenotype Connections

Bayesian Divergence Time Components

3. Model of Rate Change

How much of what appears to be rate change really is rate change?

see

Cutler, D.J. (2000) Estimating divergence times in the presence of an overdispersed molecular clock. Mol. Biol. Evol. 17:1647-1660.

A point made well by Cutler (2000)

...Rejection of constant rate hypothesis may not be due to variation of rates over time as much as being due to poor models of sequence evolution that may mislead us about how confident we can be regarding branch length estimates ...

(my viewpoint... "first principles" of evolutionary biology mean constant rate hypothesis must be formally wrong even though it may sometimes be nearly right)

Why might rates of molecular evolution change over time? Candidates include changes in ...

mutation rate per generation

generation time

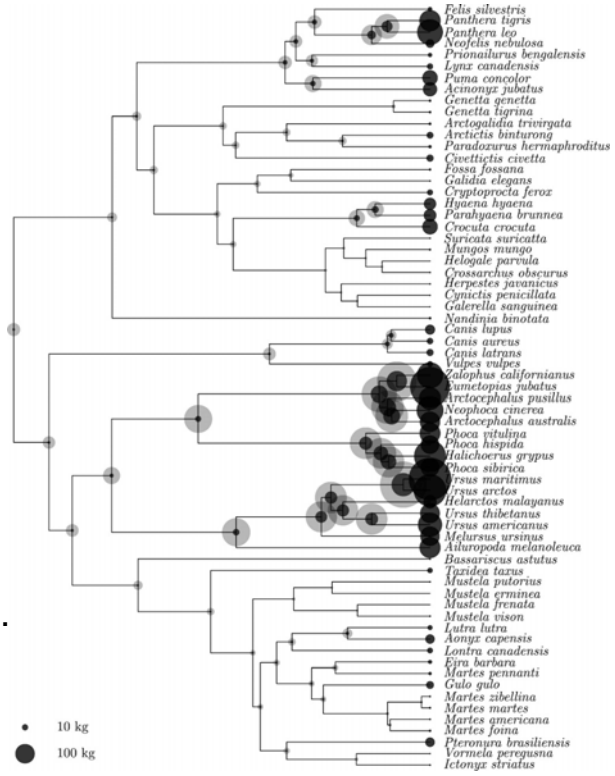
natural selection (including effects due to duplication)

population size (higher rates for small pop. size)

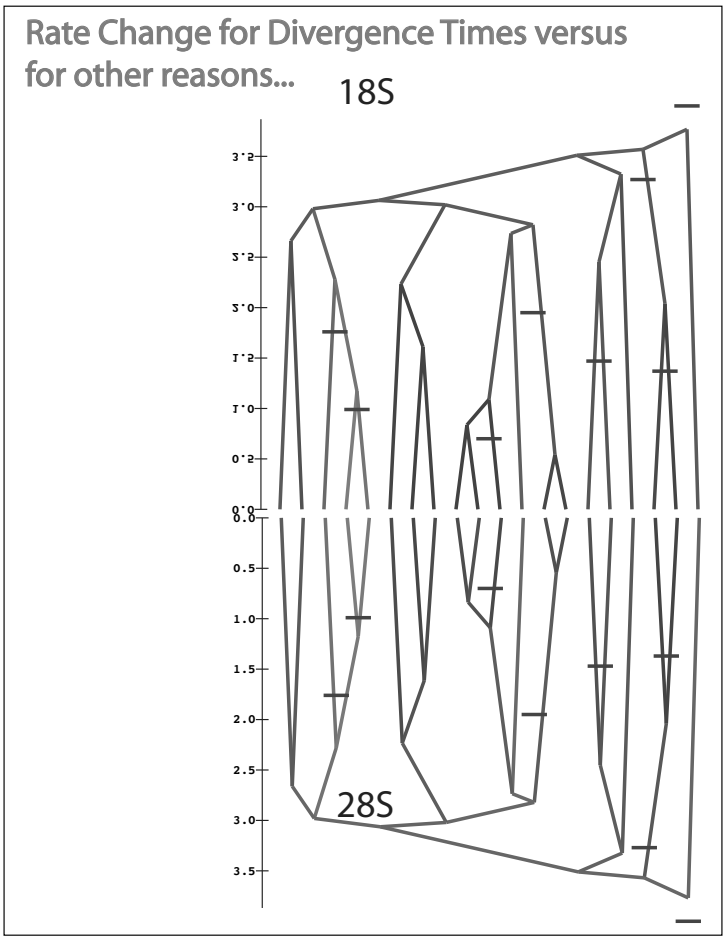
MODELING RATE VARIATION AMONG LINEAGES

- Global molecular clock (Zuckerkandl & Pauling, 1962)
- Local molecular clocks (Hasegawa, Kishino & Yano 1989; Kishino & Hasegawa 1990; Yoder & Yang 2000; Yang & Yoder 2003, Drummond & Suchard 2010)
- **Autocorrelated Rate Change (Huelsenbeck, Larget & Swofford 2000; Thorne, Kishino, & Painter 1998; Kishino, Thorne & Bruno 2001; LePage, Bryant, Philippe, & Lartillot 2007)**
- Uncorrelated/independent rates models (Drummond et al. 2006; Rannala & Yang 2007)
- Mixture models on branch rates (Heath, Holder, & Huelsenbeck 2012)

A promising idea:
By allowing them to evolve along with substitution rates, phenotypic characters that may be correlated with substitution rates can be leveraged to improved divergence time estimates



From: Lartillot N , Poujol R. 2011. Reconstruction of the evolution of body mass in carnivores. Mol Biol Evol 28:729-744

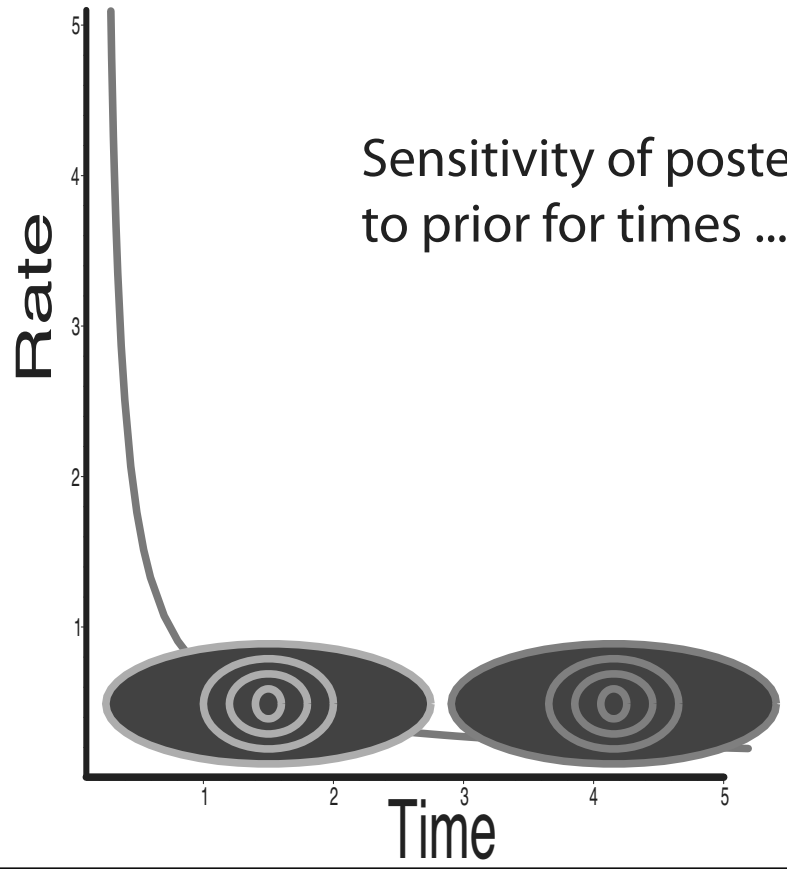


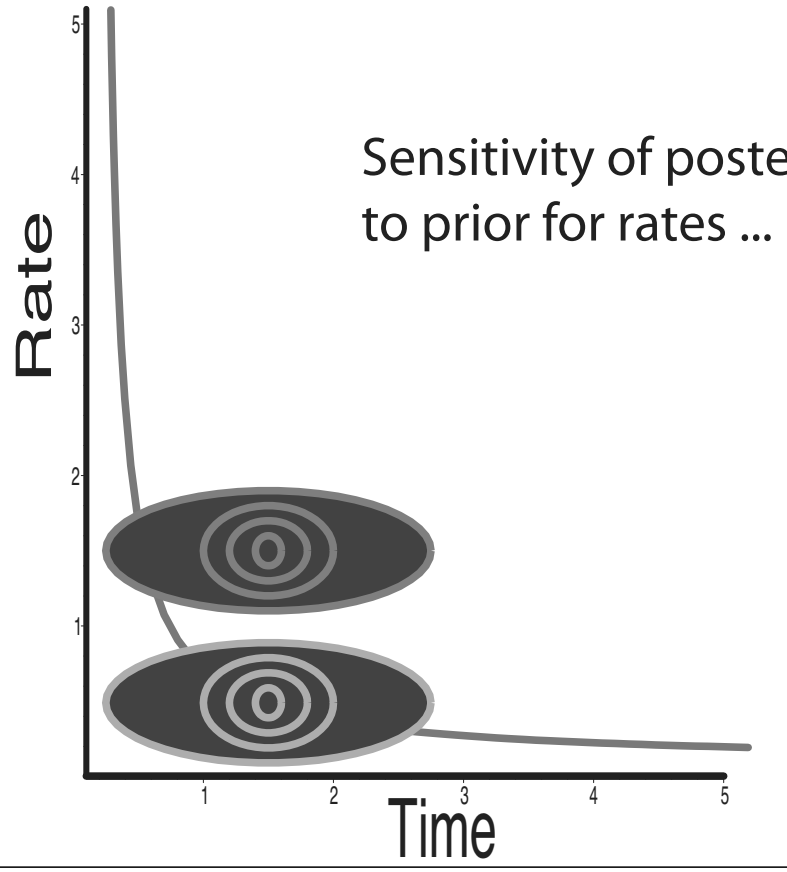
Bayesian Divergence Time Components

4. Prior Distributions for Rates, Times, etc.

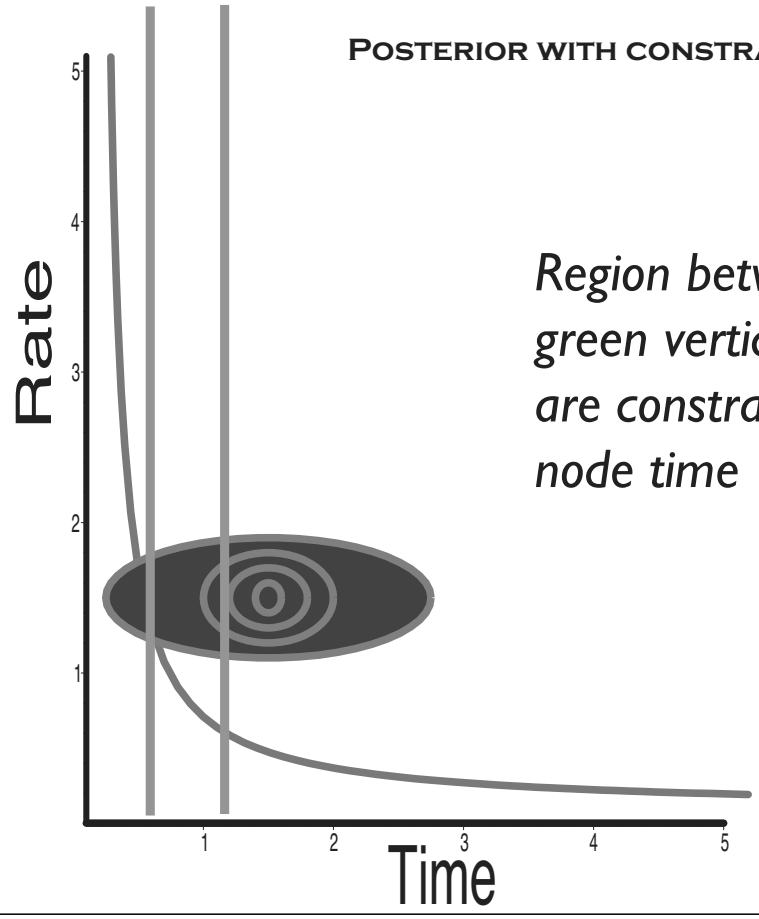
Difficulty in specifying appropriate prior distributions is arguably the biggest obstacle for Bayesian inference and this difficulty is especially great for divergence time estimation.

In many situations, prior distribution is not too important if data set is large. However, large amounts of sequence data do not overcome need for good rate and time priors here ...

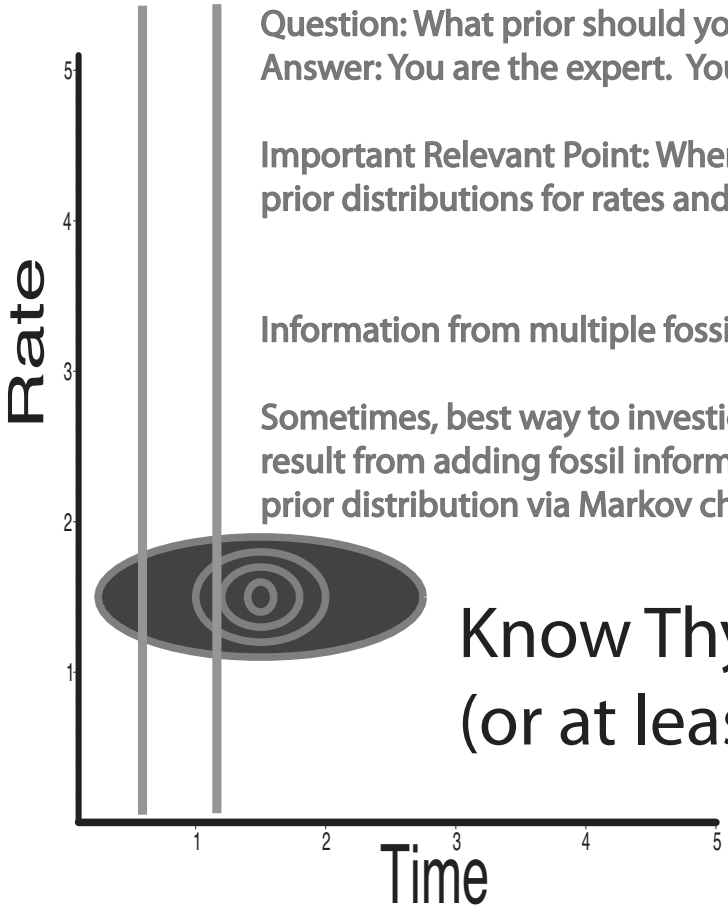

Sensitivity of posterior
to prior for times ...
Rate
Time
1
2
3
4
5


Sensitivity of posterior
to prior for rates ...
Rate
Time
1
2
3
4
5

Posterior with constraints



Region between green vertical lines are constraints on node time



Question: What prior should you use?
Answer: You are the expert. You decide.

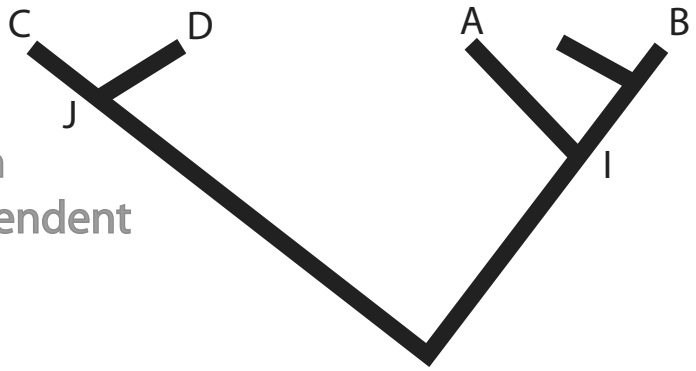
Important Relevant Point: When adding fossil information, prior distributions for rates and times can be complicated.

Information from multiple fossils can interact !

Sometimes, best way to investigate prior distributions that result from adding fossil information is to approximate prior distribution via Markov chain Monte Carlo.

Know Thy Prior!
(or at least learn it!)

Branch length between Nodes A & I and between Nodes B & I should be correlated even if rates on these branches are independent of each other.



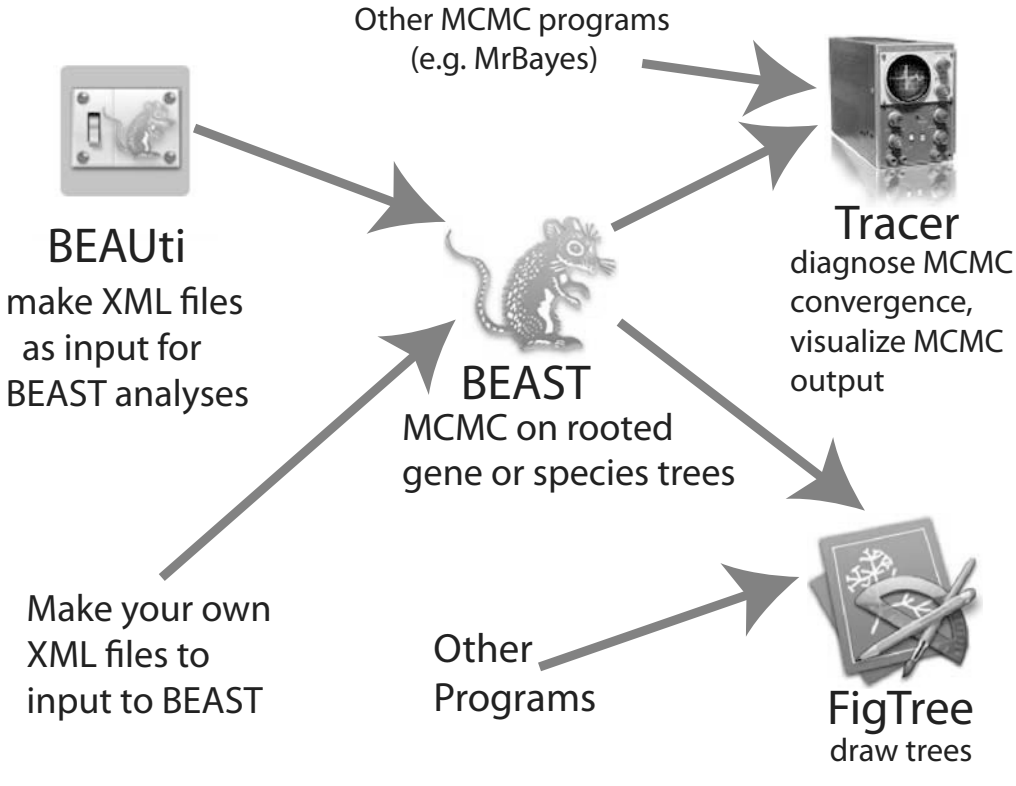
Reason: These branches represent the same amount of time.

A nice paper ...

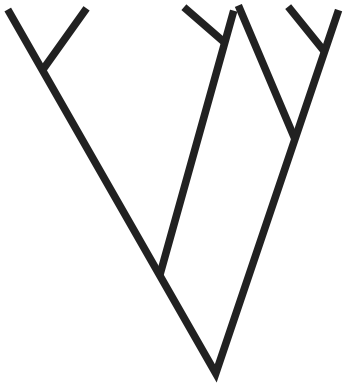
Drummond, Ho, Phillips, and Rambaut. 2006. Relaxed Phylogenetics and Dating With Confidence. PLOS Biology 4(5):e88 (see also their BEAST software)

(i) Divergence time estimation without prespecified topology
(ii) Phylogeny inference incorporating models of rate evolution

BEAST & relatives (see http://tree.bio.ed.ac.uk/software/)



Priors on node times
(and sometimes on rooted topologies):



(1) Phenomenological: Choose a hopefully flexible probability distribution (e.g., put a prior distribution on the root age and put a prior on the proportional ages of all other internal nodes relative to root age)

(2) Mechanistic: Invoke some biology to justify the prior

Yule Process (Birth process): Only speciation considered

Birth-Death Process: Speciation and Extinction considered

Taxon Sampling can also be considered (i.e., how does one decide which extant species to include in data set?)

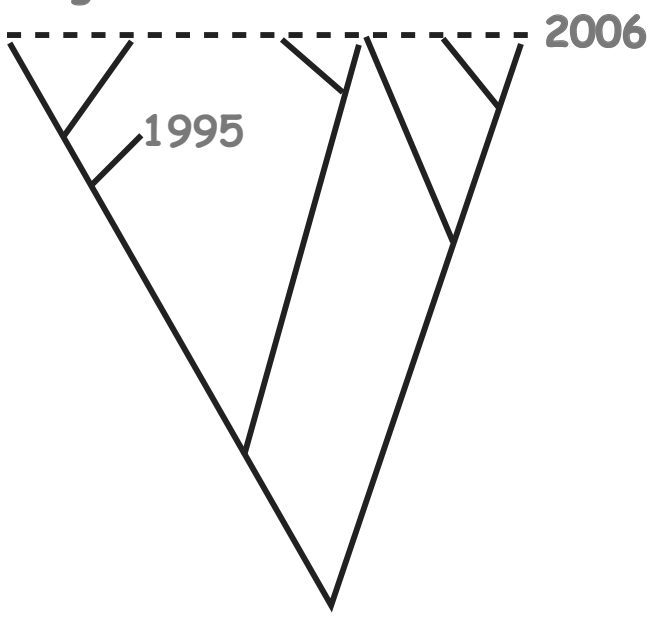
Bayesian Divergence Time Components

5. Fossil or other information

Prospects for much improved treatment of fossil evidence are good

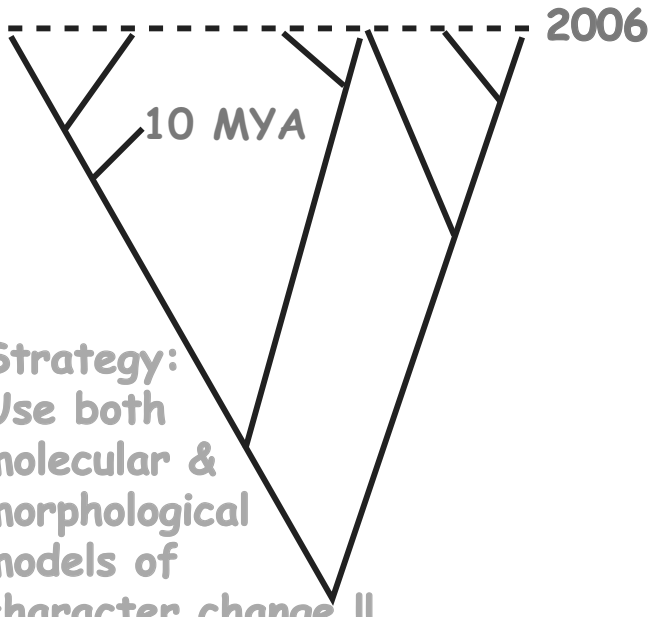
(particular progress by Ronquist et al. 2012. Syst. Biol. 61:973-999; see also Lee et al. 2009. Mol. Phylo. Evol. 50:661-666)

Can separate rates and times for quickly evolving (e.g., viral) lineages but cannot for slow lineages.



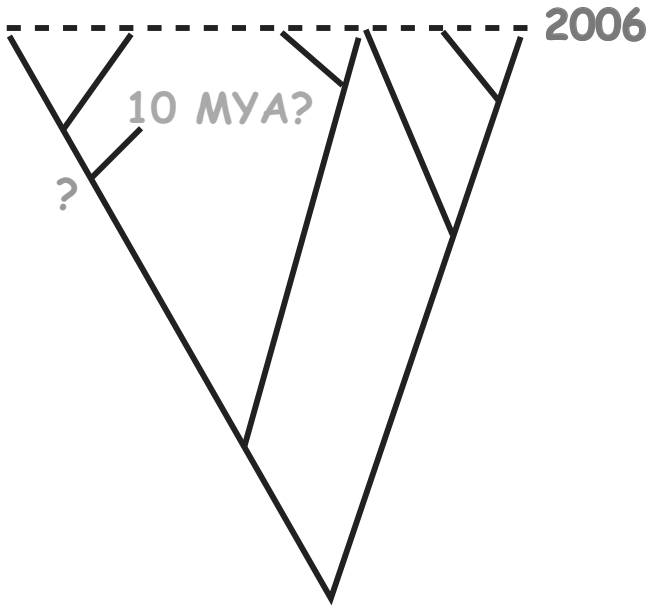
Serially Sampled Data

Can get sequence data and morphological data for 2006. Can get morphological (fossil) data for 10 million years ago!



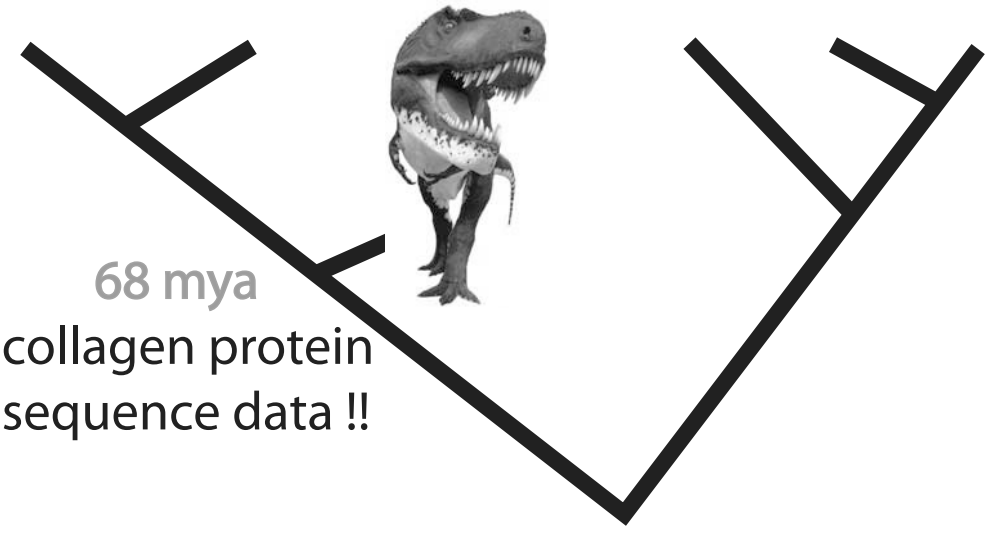
Strategy: Use both molecular & morphological models of character change !!

Bayesian techniques can (in principle) account for uncertainty in phylogenetic placement of fossils and in uncertainty of fossil dating!

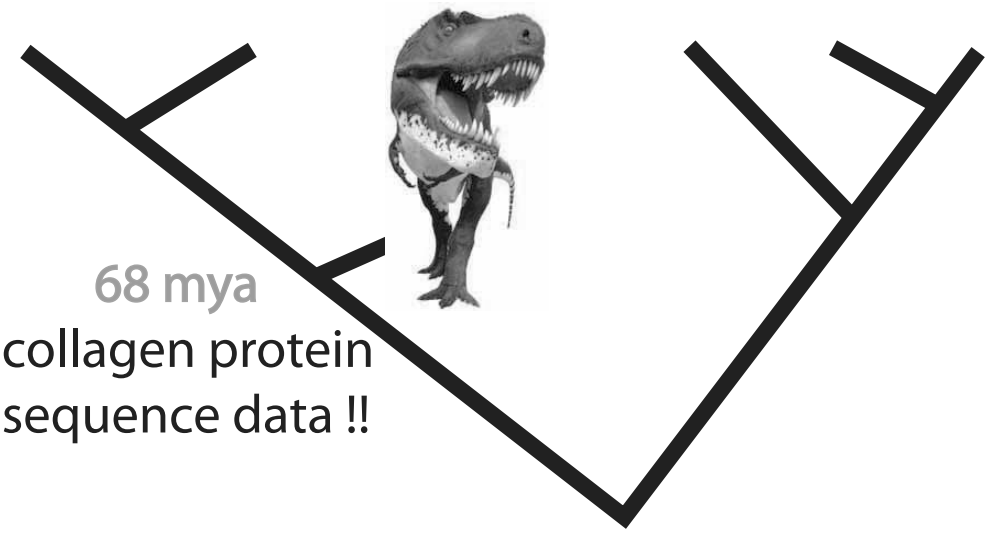


Protein Sequences from Mastodon and Tyrannosaurus Rex Revealed by Mass Spectrometry

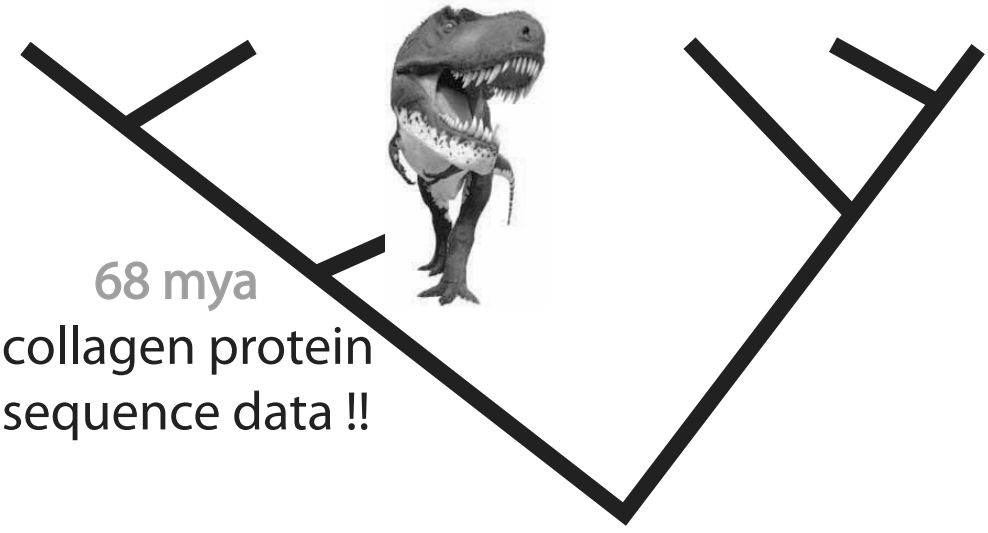
Asara et al. 2007. Science 316:280-285



Ancient protein sequences to supplement morphological fossil data (i.e., extend serially sampled techniques way way beyond HIV data) ?



With Genotype-Phenotype mapping information can we accurately predict (and validate) ancient protein/DNA sequences based on morphological evidence?



Bayesian Divergence Time Components

1. DNA or protein sequence data - Bountiful
2. Model of Sequence Change - Difficult
3. Model of Rate Change - Difficult
4. Prior Distributions for Rates, Times, etc. - ? ? ?
5. Fossil or other information - Progress !!

THE END!

Some divergence time inference software:

Beast	http://beast.bio.ed.ac.uk/
CoEvol	www.phylobayes.org/
DPPDiv	http://phylo.bio.ku.edu/content/tracy-heath-dppdiv
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html