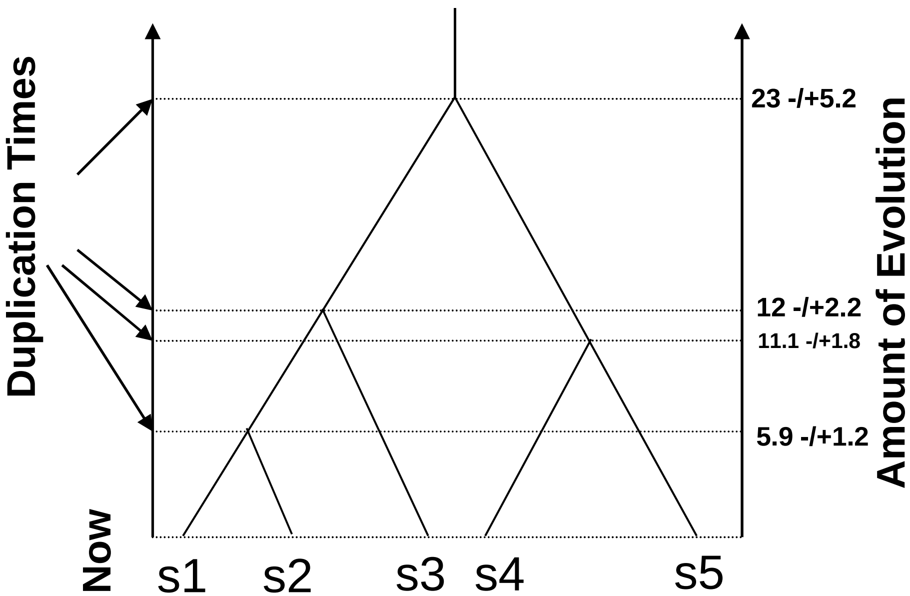
Output from Likelihood Method.

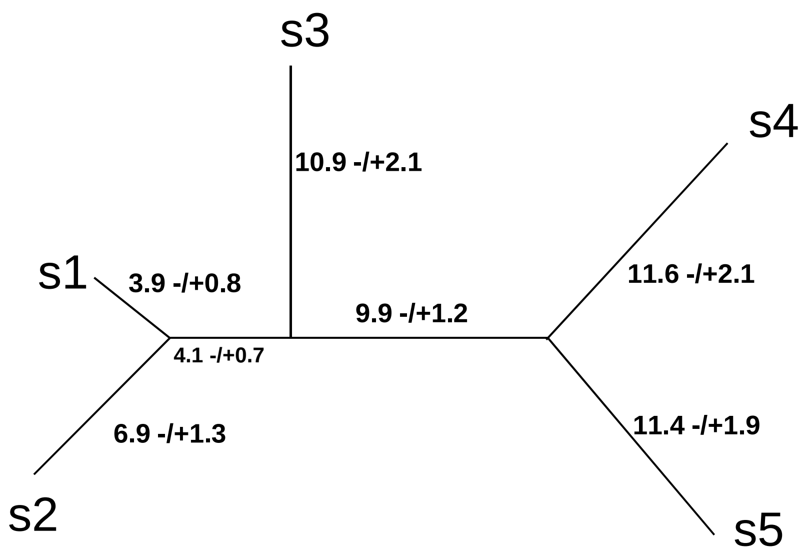
Molecular Clock



n-1 heights estimated

Likelihood: $7.9*10^{-14}$ α, β = 0.31 0.18

No Molecular Clock



2n-3 lengths estimated

Likelihood: $6.2*10^{-12}$ α, β = 0.34 0.16

$\ln(7.9*10^{-14}) - \ln(6.2*10^{-12})$ is χ^2 – distributed with (n-2) degrees of freedom

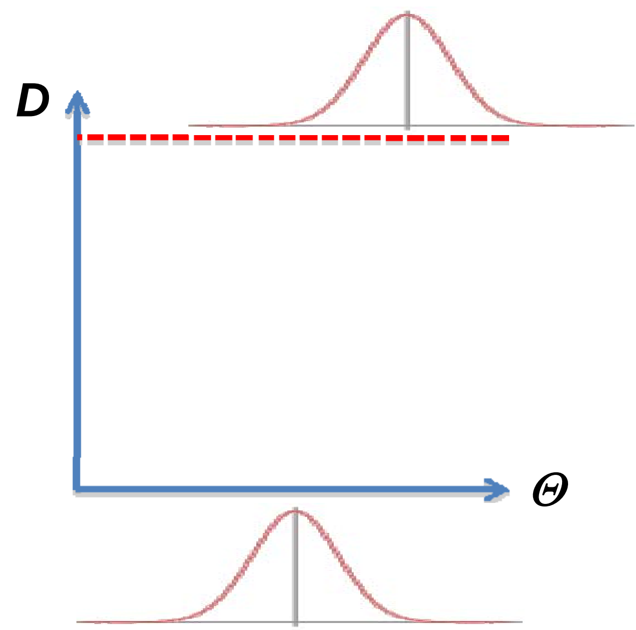
Bayesian Approach

Likelihood function L() – the probability of data as function of parameters: L(Θ,D)

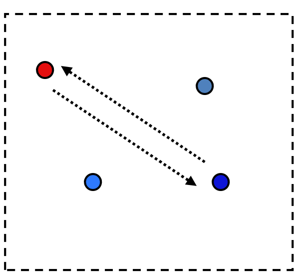
In Likelihood analysis, Θ is not stochastic variable, $\Theta_{max}(D)$ is

In Bayesian Analysis, Θ is a stochastic variable with a prior distribution before data is included in the analysis.

After the observation of Data, there will be a posterior on Θ



Bayesian Analysis have seem a major rise in use as a consequence of numerical/stochastic integration techniques such as Markov Chain Monte Carlo.



Likelihood - L($\color{red}{\bullet}$)

Probability of going from $\color{red}{\bullet}$ to $\color{blue}{\bullet}$ - q($\color{red}{\bullet}$, $\color{blue}{\bullet}$)

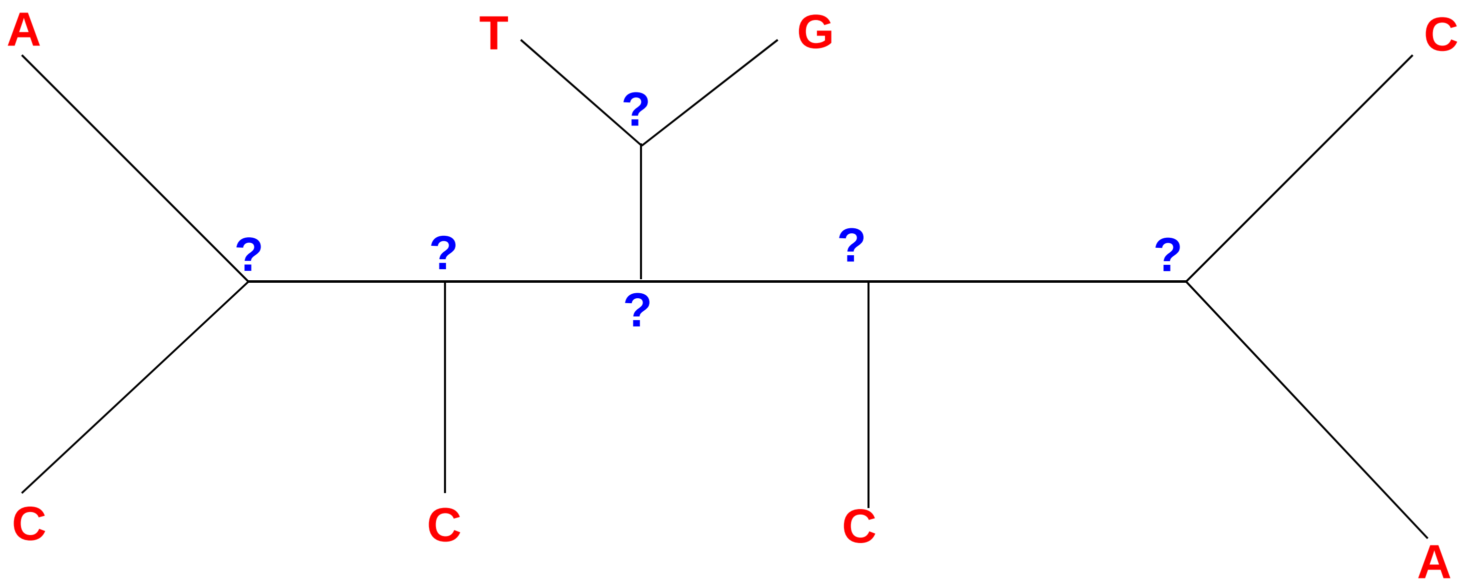
J - Jacobian

Acceptance ratio

$$\frac{L(\color{blue}{\bullet})\,q(\color{red}{\bullet},\color{blue}{\bullet})}{L(\color{red}{\bullet})\,q(\color{blue}{\bullet},\color{red}{\bullet})}J$$

Likelihood function L(Θ,D) is central to both approaches

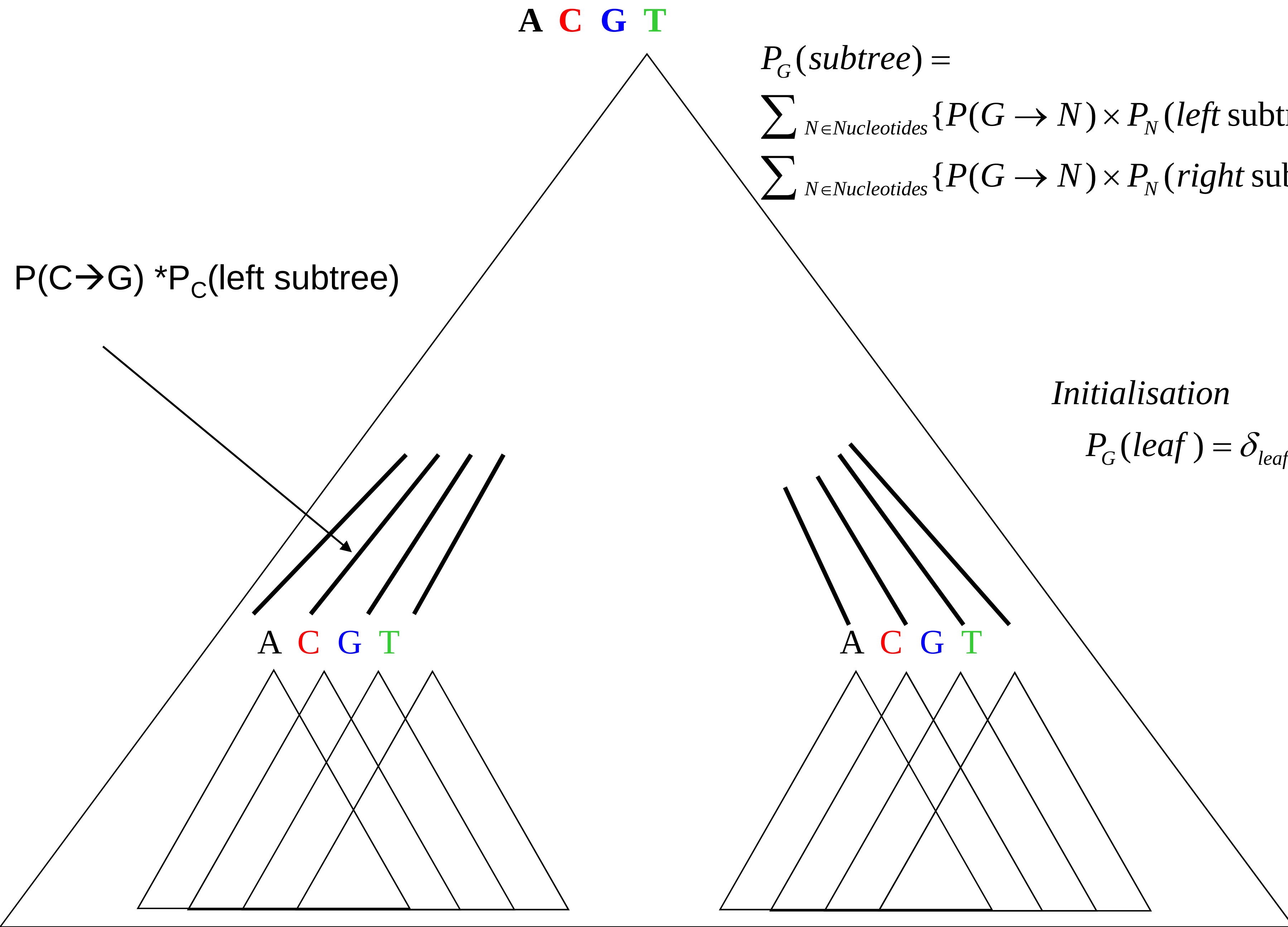
Assignment to internal nodes: The simple way.



If branch lengths and evolutionary process is known, what is the probability of nucleotides at the leaves?

```
Cctacggccatacca   a   ccctgaaagcaccccatcccgt
Cttacgaccatatca   c   cgttgaatgcacgccatcccgt
Cctacggccatagca   c   ccctgaaagcaccccatcccgt
Cccacggccatagga   c   ctctgaaagcactgcatcccgt
Tccacggccatagga   a   ctctgaaagcaccgcatcccgt
Ttccacggccatagg   c   actgtgaaagcaccgcatcccg
Tggtgcggtcatacc   g   agcgctaatgcaccggatccca
Ggtgcggtcatacca   t   gcgttaatgcaccggatcccat
```

Probability of leaf observations - summing over internal states



$$P_G(subtree) =$$

$$\sum_{N \in Nucleotides} \{P(G \rightarrow N) \times P_N(left\ \text{subtree})\} \times$$

$$\sum_{N \in Nucleotides} \{P(G \rightarrow N) \times P_N(right\ \text{subtree})\}$$

Initialisation

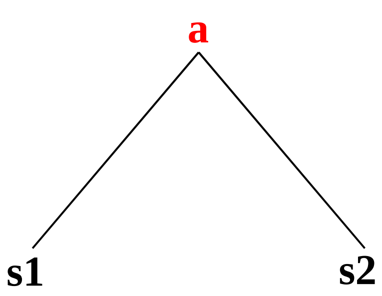
$$P_G(leaf) = \delta_{leaf,G}$$

The Molecular Clock

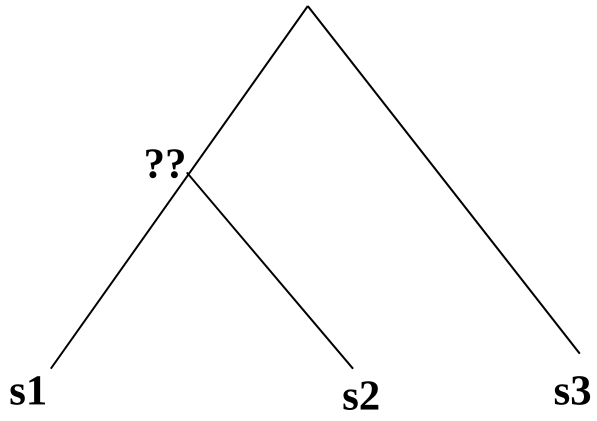
First noted by Zuckerkandl & Pauling (1964) as an empirical fact.

How can one detect it?

Known Ancestor, a, at Time t



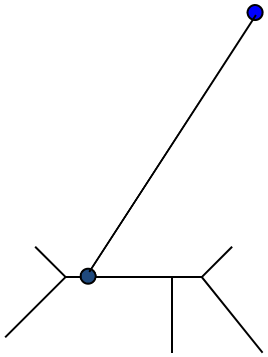
Unknown Ancestors

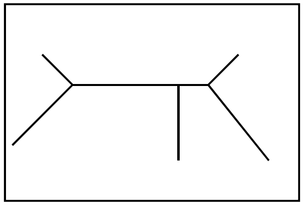


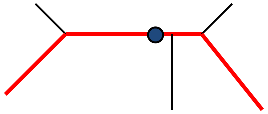
Rootings

Purpose
1) To give time direction in the phylogeny & most ancient point
2) To be able to define concepts such a monophyletic group.

1) Outgrup: Enhance data set with sequence from a species definitely distant to all of them. It will be be joined at the root of the original data

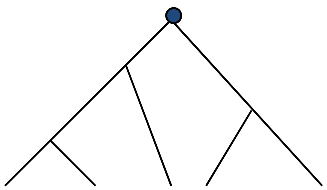






2) Midpoint: Find midpoint of longest path in tree.

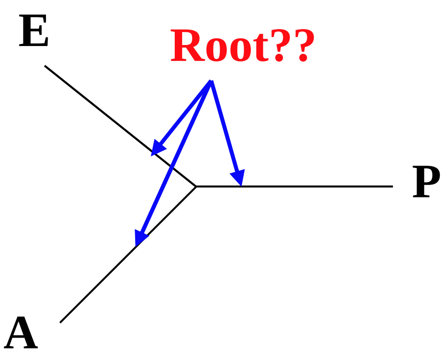
3) Assume Molecular Clock.

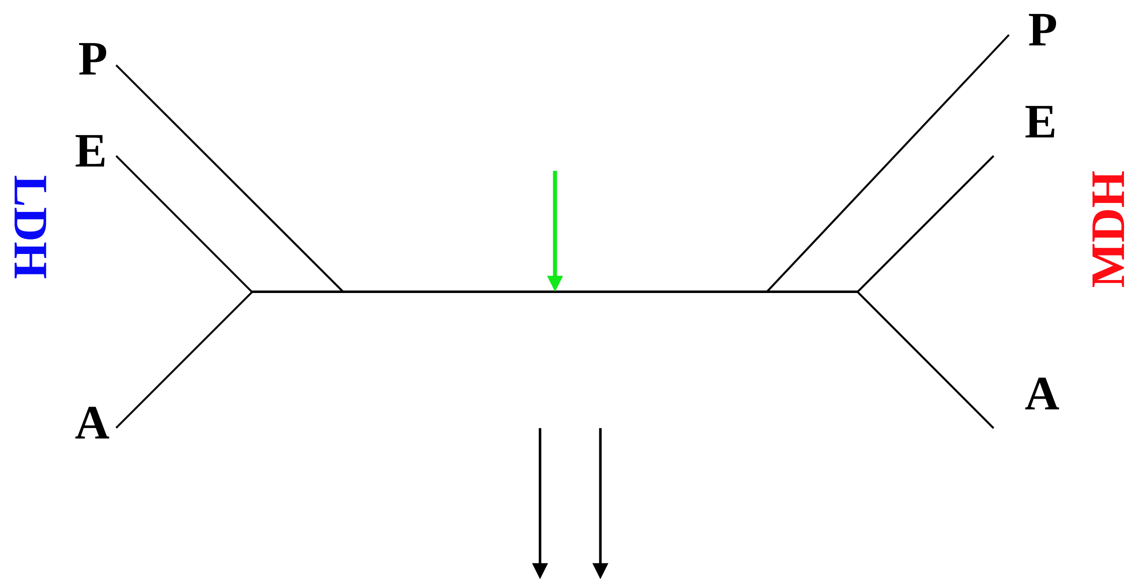


Rooting the 3 kingdoms

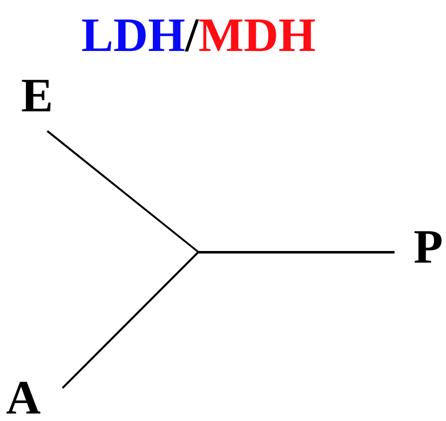
3 billion years ago: no reliable clock - no outgroup

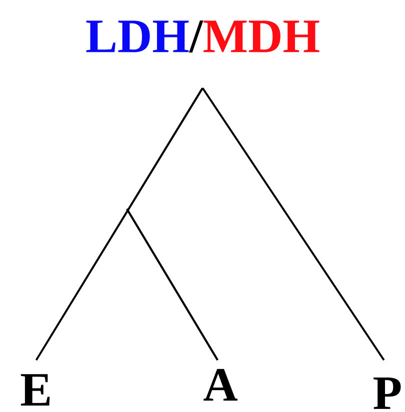
Given 2 set of homologous proteins, i.e. MDH & LDH can the archea, prokaria and eukaria be rooted?





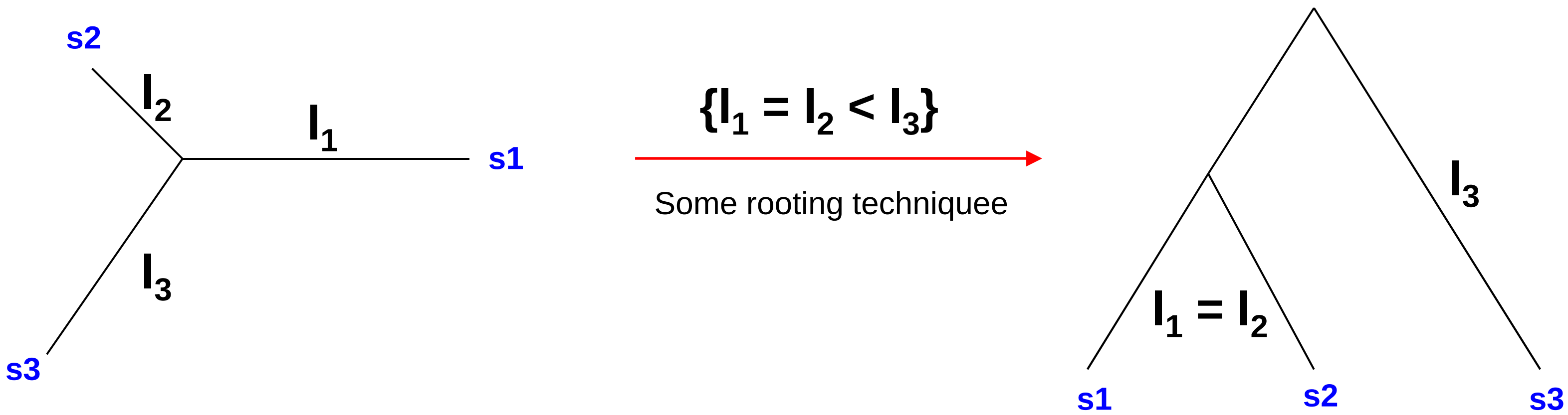
Given 2 set of homologous proteins, i.e. MDH & LDH can the archea, prokaria and eukaria be rooted?



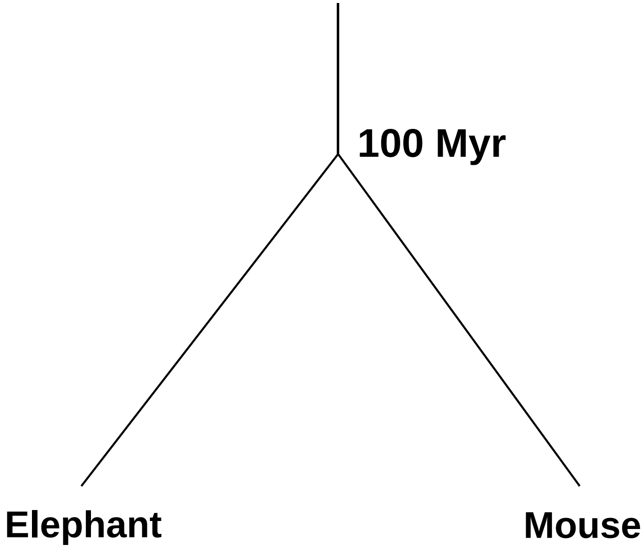


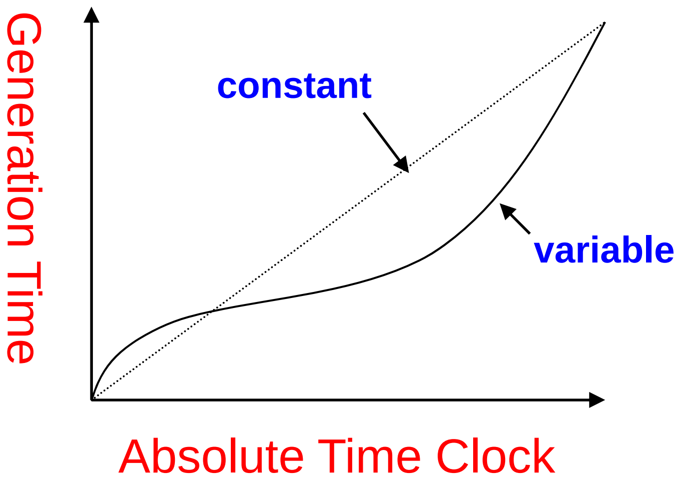
The generation/year-time clock Langley-Fitch,1973

Absolute Time Clock:

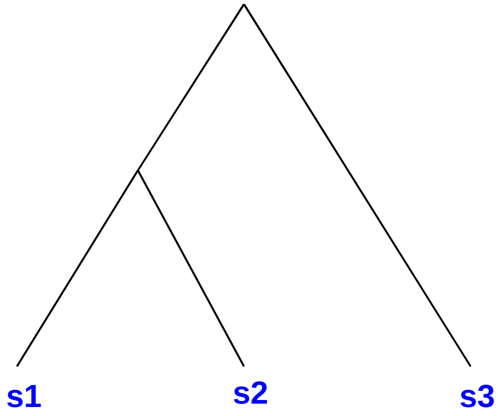


Generation Time Clock:





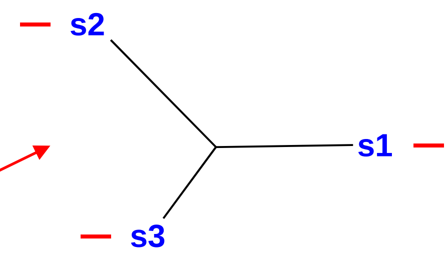
The generation/year-time clock Langley-Fitch,1973

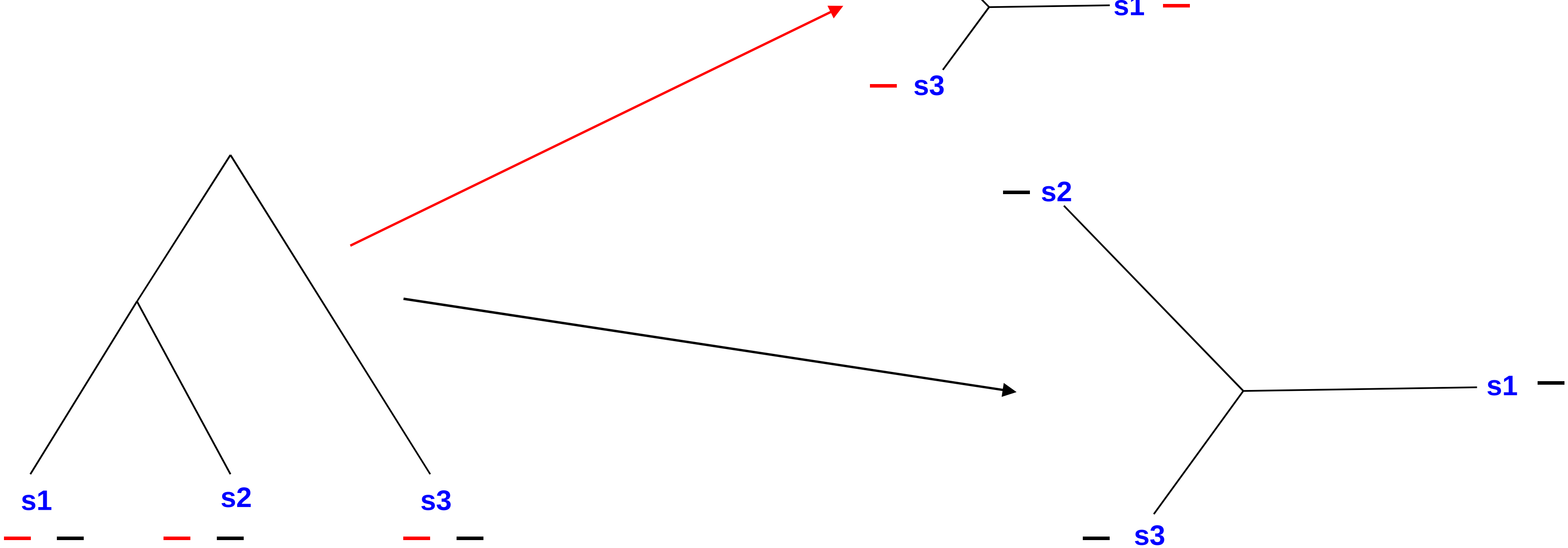


Generation Time Clock → **Any Tree**

Can the generation time clock be tested?

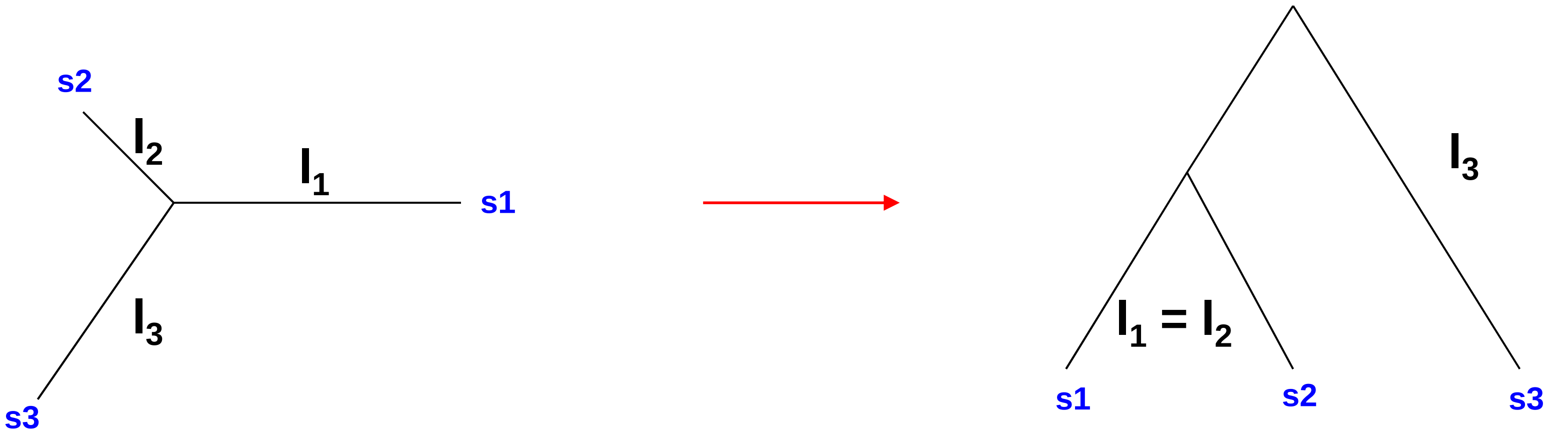
Assume, a data set: 3 species, 2 sequences each





The generation/year-time clock

Langley-Fitch,1973

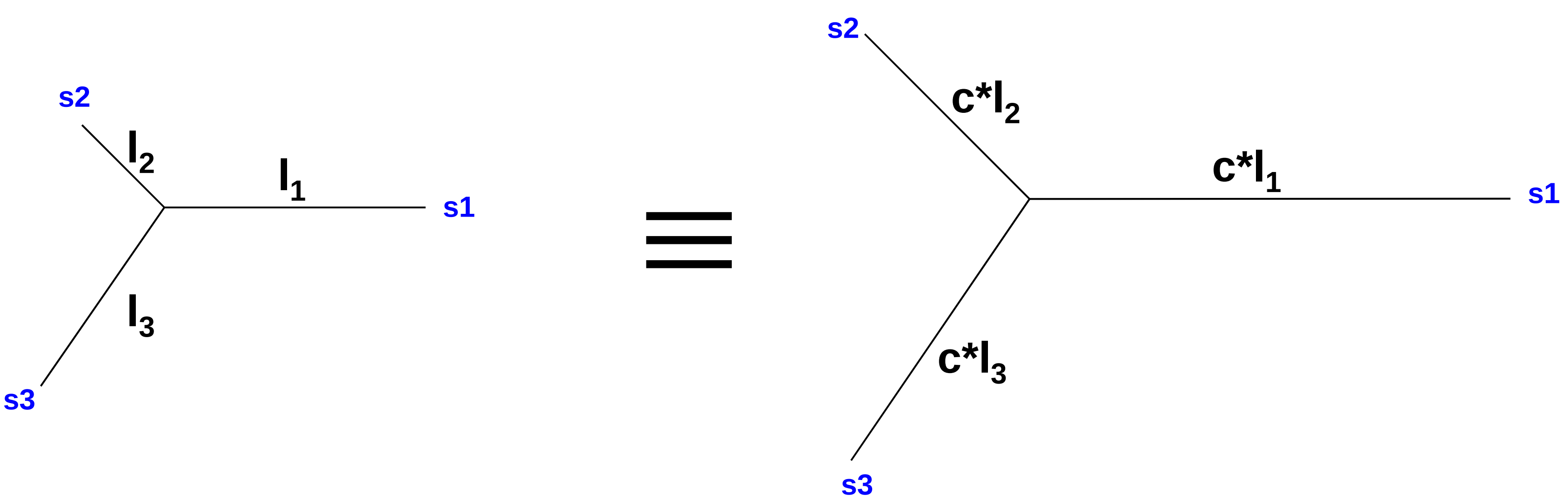


k=3: degrees of freedom: 3

k: dg: 2k-3

dg: 2

dg: k-1



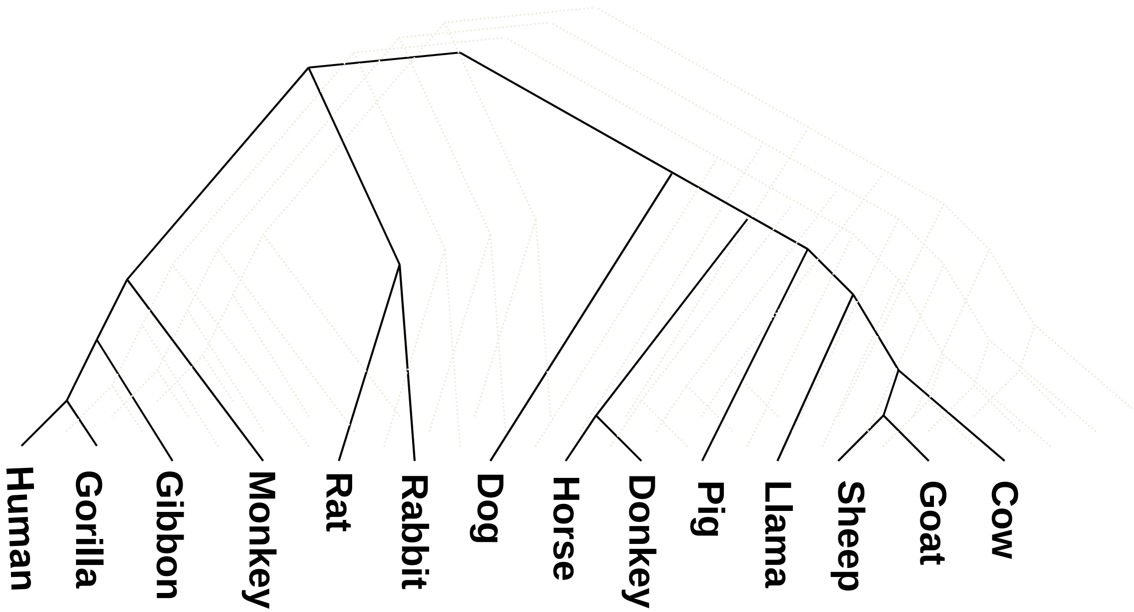
k=3, t=2: dg=4

k, t: dg =(2k-3)-(t-1)

α & β – globin, cytochrome c, fibrinopeptide A & generation time clock

Langley-Fitch,1973

Fibrinopeptide A phylogeny:



Relative rates

α-globin	0.342
β– globin	0.452
cytochrome c	0.069
fibrinopeptide A	0.137

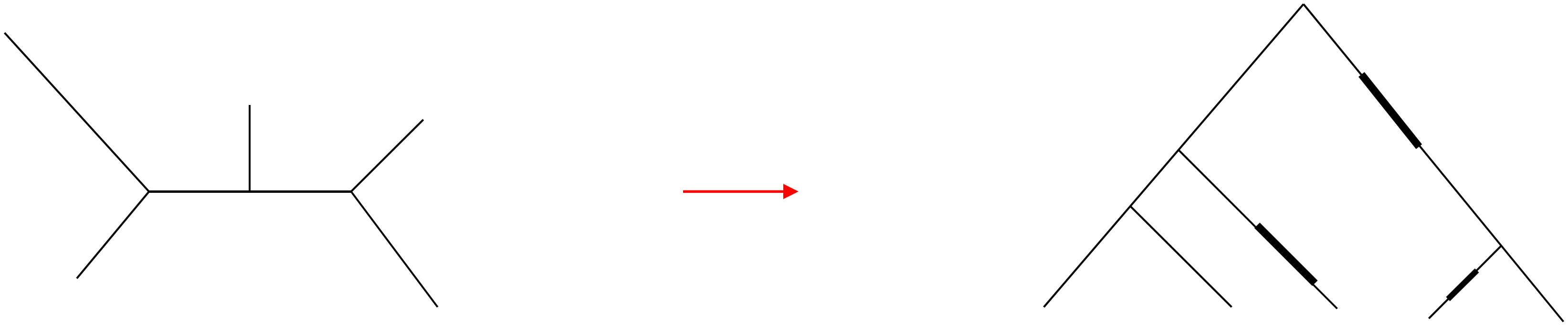
Almost Clocks

(MJ Sanderson (1997) "A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy" Mol.Biol.Evol.14.12.1218-31), J.L.Thorne et al. (1998): "Estimating the Rate of Evolution of the Rate of Evolution." Mol.Biol.Evol. 15(12).1647-57, JP Huelsenbeck et al. (2000) "A compound Poisson Process for Relaxing the Molecular Clock" Genetics 154.1879-92.)

I Smoothing a non-clock tree onto a clock tree (Sanderson)

II Rate of Evolution of the rate of Evolution (Thorne et al.).
The rate of evolution can change at each bifurcation

III Relaxed Molecular Clock (Huelsenbeck et al.).
At random points in time, the rate changes by multiplying with random variable (gamma distributed)

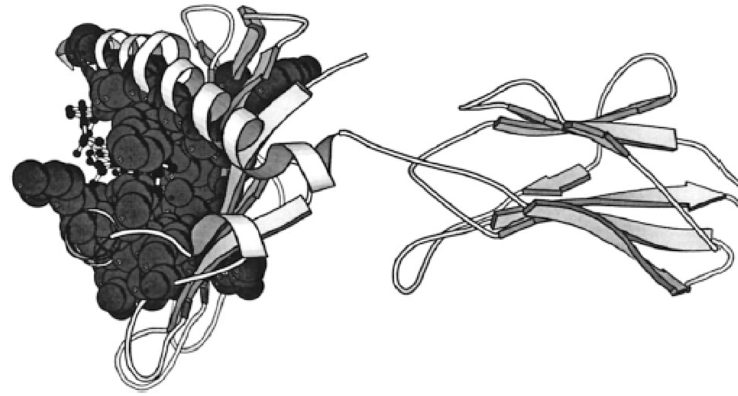


Comment: Makes perfect sense. Testing no clock versus perfect is choosing between two unrealistic extremes.

Adaptive Evolution

Yang, Swanson, Nielsen,..

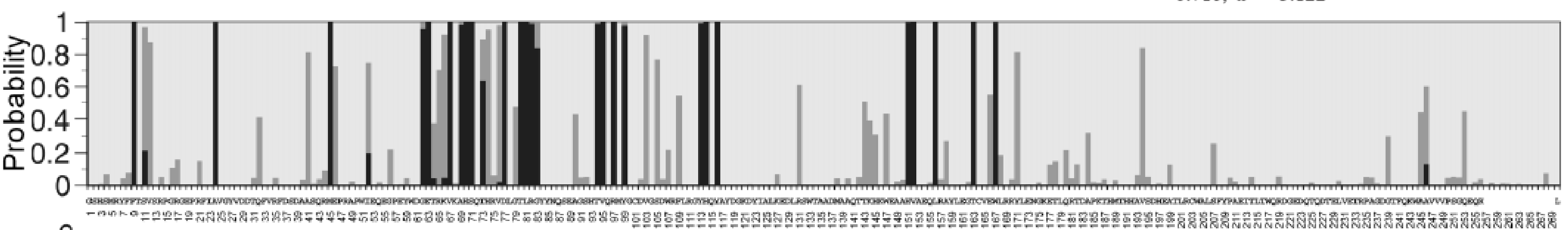
- ***Models with positive selection.***



- ***Positive Selection is interesting as it is as functional change and could at times be correlated with change between species.***

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

Model Code	p	ℓ	Estimates of Parameters
M0 (one-ratio).....	392	−8225.16	$\hat{\omega} = 0.612$
M1 (neutral)	392	−7719.46	$\hat{p}_0 = 0.585$ ($\hat{p}_1 = 0.415$)
M2 (selection).....	394	−7296.69	$\hat{p}_0 = 0.566$, $\hat{p}_1 = 0.332$, ($\hat{p}_2 = 0.102$), $\hat{\omega}_2 = 8.092$
M3 (discrete)......	396	−7226.51	$\hat{p}_0 = 0.780$, $\hat{p}_1 = 0.133$ ($\hat{p}_2 = 0.086$) $\hat{\omega}_0 = 0.069$, $\hat{\omega}_1 = 1.328$, $\hat{\omega}_2 = 6.048$
M7 (beta).........	393	−7498.97	$\hat{p} = 0.103$, $\hat{q} = 0.354$
M8 (beta & ω)	395	−7232.68	$\hat{p}_0 = 0.900$, ($\hat{p}_1 = 0.100$), $\hat{p} = 0.168$, $\hat{q} = 0.710$, $\hat{\omega} = 5.122$



Summary

Combinatorics of Trees

Principles of Phylogeny Inference

- *Distance*
- *Parsimony*
- *Probablistic Methods*

Applications

- *Clocks*
- *Selection*