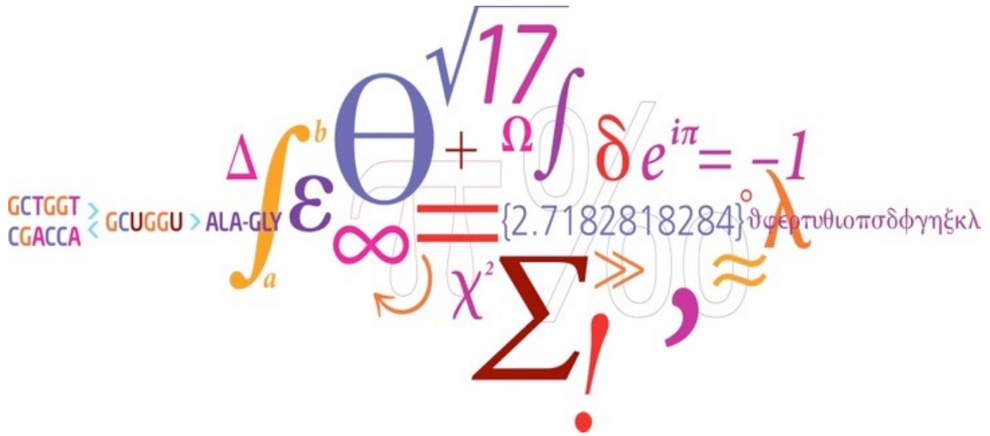
Models of Evolution



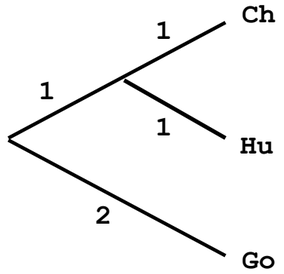
Distance Matrix Methods

```
                  ↓↓↓↓
Gorilla    :   ACGTCGTA
Human      :   ACGTTCCT
Chimpanzee:   ACGTTTCG
                    ↑ ↑
```

1. Construct multiple alignment of sequences

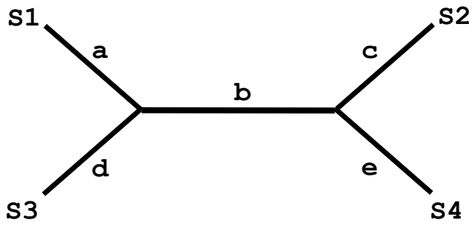
	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-

2. Construct table listing all pairwise differences (distance matrix)



3. Construct tree from pairwise distances

Optimal Branch Lengths for a Given Tree: Least Squares



Distance along tree

Goal:

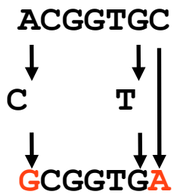
$$D_{12} \approx d_{12} = a + b + c$$
$$D_{13} \approx d_{13} = a + d$$
$$D_{14} \approx d_{14} = a + b + e$$
$$D_{23} \approx d_{23} = d + b + c$$
$$D_{24} \approx d_{24} = c + e$$
$$D_{34} \approx d_{34} = d + b + e$$

- Fit between given tree and observed distances can be expressed as “sum of squared differences”:

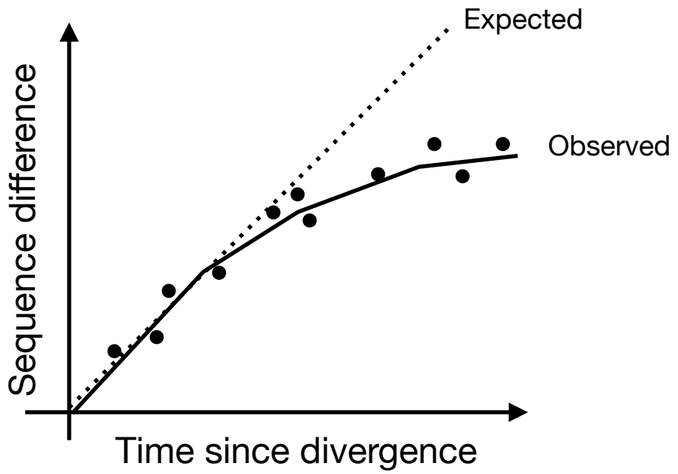
- $$Q = \sum_{j>i} (D_{ij} - d_{ij})^2$$

- Find branch lengths that minimize Q - this is the optimal set of branch lengths for this tree.

Superimposed Substitutions



- Actual number of evolutionary events:
  5
- Observed number of differences:
  2



- Distance is (almost) always underestimated

Model-based correction for superimposed substitutions

- **Goal**:
  - Try to infer the real number of evolutionary events (the real distance) based on observed data (sequence alignment)

- **This requires:**
  - Assumptions about how sequences have been changing (i.e., a hypothesis about, or model of, sequence evolution)

What is a Model?

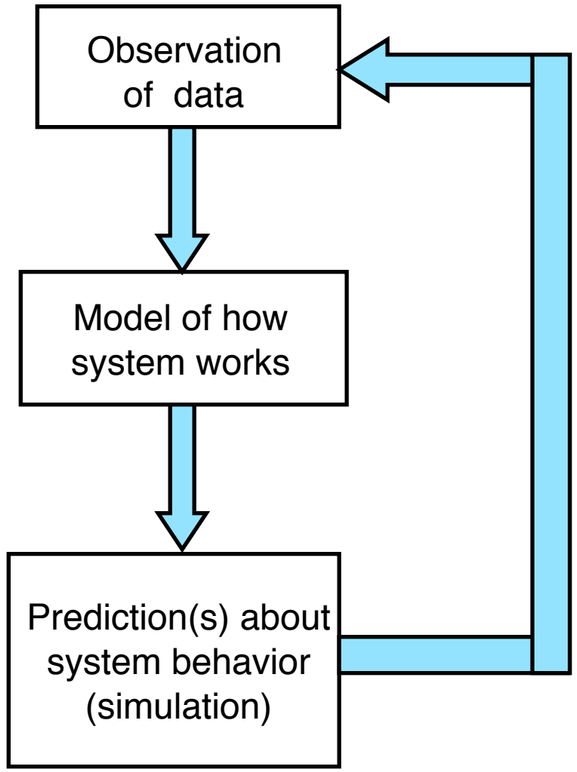
- Model = stringently phrased hypothesis !!!
- Hypothesis (as used in most biological research):
  - Precisely stated, but qualitative
  - Allows you to make qualitative predictions
  - Example: “Population size grows rapidly when there are few individuals, but growth rate declines when resources become limiting.”
- Arithmetic model:
  - Mathematically explicit (parameters)
  - Allows you to make quantitative predictions
  - Example: $$N_t = \frac{K}{1 + \left(\frac{K}{N_0} - 1\right) e^{-rt}}$$

Models do not represent full reality!

- It is typically not possible to represent full reality in a mathematical model.
- Growth model example:
  - fecundity and survival rate depend on a large number of factors
  - biological and non-biological, internal and external, some stochastic
  - for each individual in a population.
  - for each individual these are complicated functions of huge numbers of different terms.
  - it is impossible to get good estimates of this multitude of parameters from a finite data set

- One-to-one maps are difficult to read!
- Goal is instead to find good approximating model
- We assume that structure of reality has factors with “tapering effect sizes”
  - a few very important factors
  - a moderate number of moderately important factors
  - very many factors of little importance

The Scientific Method



Jukes and Cantor Model of Nucleotide Substitution

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

Relative rate matrix

$$\Rightarrow P(t) = e^{Qt} = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$$

Probability matrix
(function of time)

- Four nucleotides assumed to be equally frequent (f=0.25)
- All 12 substitution rates assumed to be equal
- Under this model the corrected distance is: $D_{JC} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D_{OBS}\right)$
- For instance: $D_{OBS} = 0.42 \implies D_{JC} = 0.62$

Other models of evolution

	A	C	G	T
A	$1-\alpha-2\beta$	β	α	β
C	β	$1-\alpha-2\beta$	β	α
G	α	β	$1-\alpha-2\beta$	β
T	β	α	β	$1-\alpha-2\beta$

	A	C	G	T
A	$1-\alpha-2\gamma$	γ	α	γ
C	δ	$1-\beta-2\delta$	δ	β
G	β	γ	$1-\beta-2\gamma$	γ
T	δ	α	δ	$1-\alpha-2\delta$

$\vdots$

	A	C	G	T
A	$1-\alpha_{12}-\alpha_{13}-\alpha_{14}$	α_{12}	α_{13}	α_{14}
A	α_{21}	$1-\alpha_{21}-\alpha_{23}-\alpha_{24}$	α_{23}	α_{24}
A	α_{31}	α_{32}	$1-\alpha_{31}-\alpha_{32}-\alpha_{34}$	α_{34}
A	α_{41}	α_{42}	α_{43}	$1-\alpha_{41}-\alpha_{42}-\alpha_{43}$

Yet more models of evolution

- Codon-codon substitution rates

  (61 x 61 matrix of codon substitution rates)

- Different mutation rates at different sites in the gene

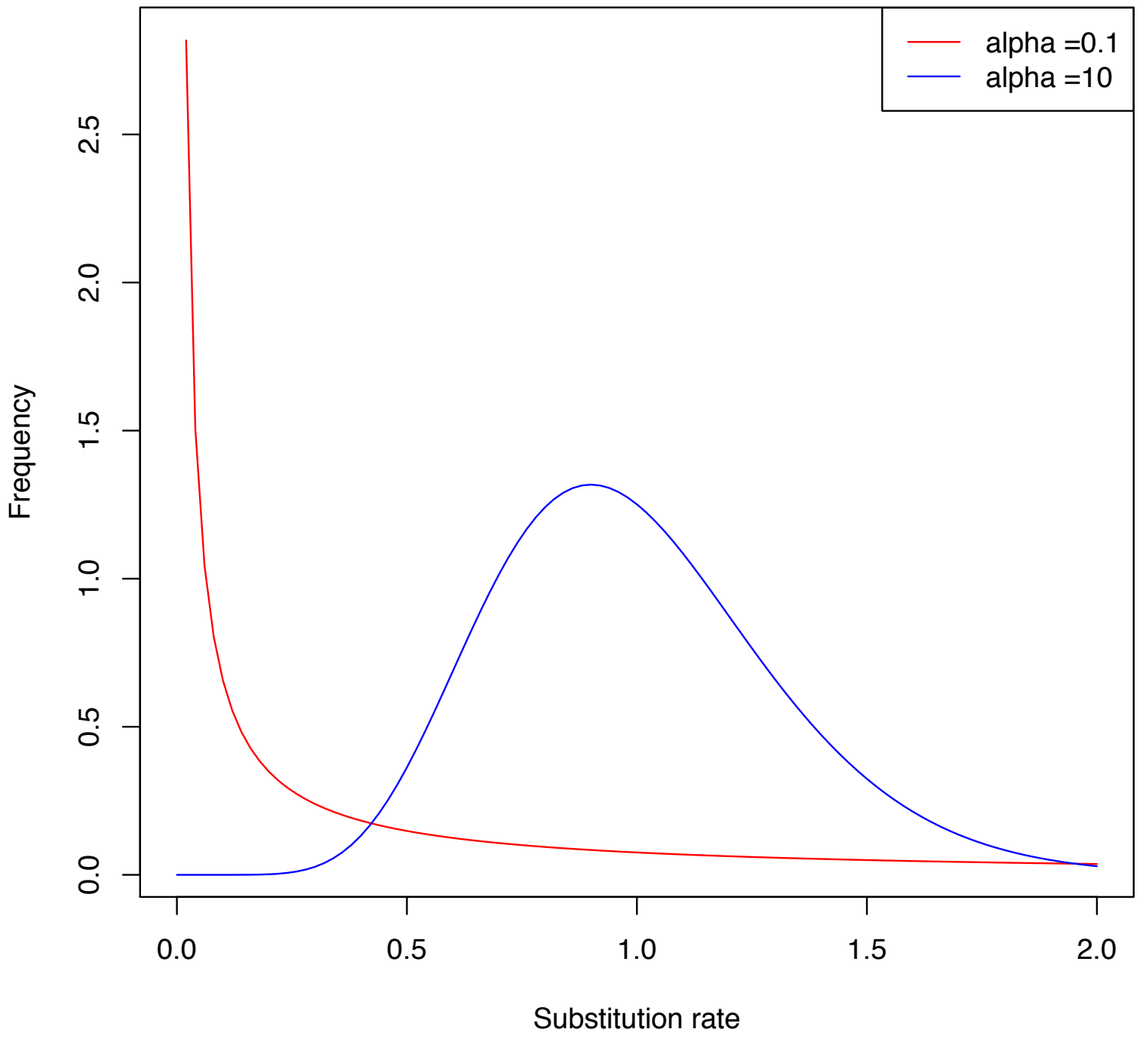
  (the “gamma-distribution” of mutation rates)

- Molecular clocks

  (same mutation rate on all branches of the tree).

- Etc., etc.

Different rates at different sites: the gamma distribution



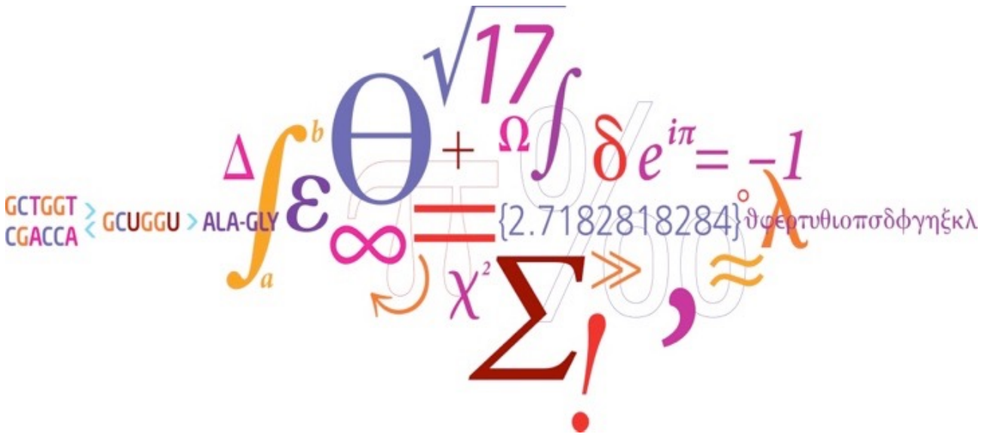
General Time Reversible Model

	A	C	G	T
A	$-$	$\pi_C\alpha$	$\pi_G\beta$	$\pi_T\gamma$
C	$\pi_A\alpha$	$-$	$\pi_G\delta$	$\pi_T\epsilon$
G	$\pi_A\beta$	$\pi_C\delta$	$-$	$\pi_T\eta$
T	$\pi_A\gamma$	$\pi_C\epsilon$	$\pi_G\eta$	$-$

- Time-reversibility:
  The amount of change from state x to y is equal to the amount of change from y to x

$$\pi_A \times \text{rate}_{AG} = \pi_G \times \text{rate}_{GA} \Leftrightarrow \pi_A\pi_G\beta = \pi_G\pi_A\beta$$

Maximum Likelihood



The maximum likelihood approach I

- Starting point:

- You have some observed data and a probabilistic model for how the observed data was produced

- Having a probabilistic model of a process means you are able to compute the probability of any possible outcome (given a set of specific values for the model parameters).

- Example:

  - Data: result of tossing coin 10 times - 7 heads, 3 tails

  - Model: coin has probability p for heads, 1-p for tails.

  - The probability of observing h heads among n tosses is:

$$P(\text{h heads}) = \binom{n}{h} p^h (1-p)^{n-h}$$

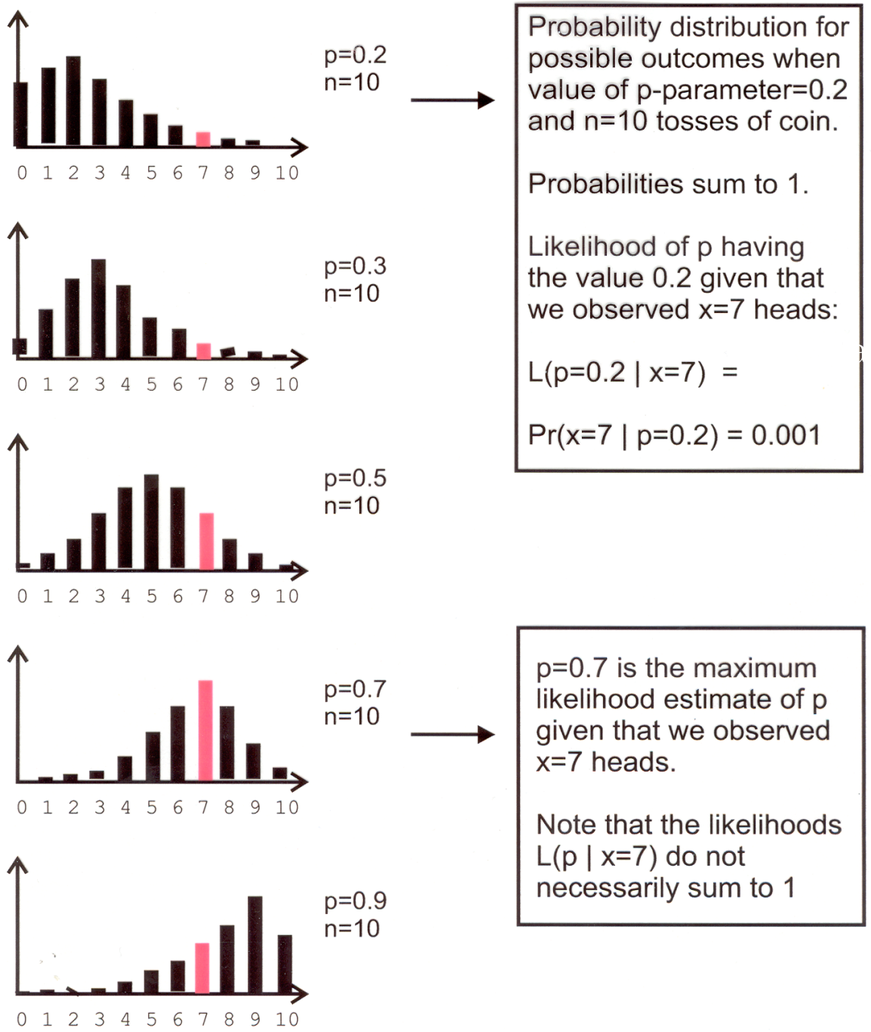
- Goal:

- You want to find the best estimate of the (unknown) parameter values based on the observations. (here the only parameter is p)

The maximum likelihood approach II

- Likelihood (Model) = Probability (Data | Model)
- Maximum likelihood: Best estimate is the set of parameter values which gives the highest possible likelihood.

Maximum likelihood: coin tossing example



Probability distribution for possible outcomes when value of p-parameter=0.2 and n=10 tosses of coin.

Probabilities sum to 1.

Likelihood of p having the value 0.2 given that we observed x=7 heads:

L(p=0.2 | x=7) =

Pr(x=7 | p=0.2) = 0.001

p=0.7 is the maximum likelihood estimate of p given that we observed x=7 heads.

Note that the likelihoods L(p | x=7) do not necessarily sum to 1

- Data: result of tossing coin 10 times - 7 heads, 3 tails
- Model: coin has probability p for heads, 1-p for tails.

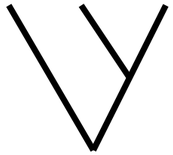
Probabilistic modeling applied to phylogeny

- **Observed data: multiple alignment of sequences**

```
H.sapiens globin      A G G G A T T C A
M.musculus globin     A C G G T T T - A
R.rattus globin       A C G G A T T - A
```

- **Probabilistic model:**
  - A model of (hypothesis about) how one ancestral sequence has evolved into the three sequences that are present in the alignment

- **Probabilistic model parameters (simplest case):**
  - Tree topology and branch lengths
  - Nucleotide frequencies: π_A, π_C, π_G, π_T
  - Nucleotide-nucleotide substitution rates (or substitution probabilities):

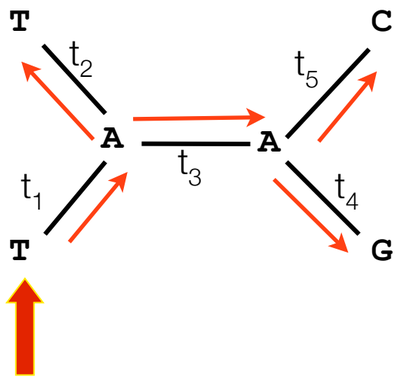


	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

$$\Rightarrow P(t) = e^{Qt} = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$$

Computing the probability of one column in an alignment given tree topology and other parameters

```
A T G G A T T C A
A T G G T T T - A
A C G G A T T - A
A G G G T T T - A
```



- Columns in alignment contain homologous nucleotides
- Assume tree topology, branch lengths, and other parameters are given. For now, assume ancestral states were A and A (we'll get to the full computation on next slide). Start computation at any internal or external node. Arrows indicate "direction" of computations ("flowing" away from the starting point).

$$Pr = \pi_T \, P_{TA}(t_1) \, P_{AT}(t_2) \, P_{AA}(t_3) \, P_{AG}(t_4) \, P_{AC}(t_5)$$

Computing the probability of an entire alignment given tree topology and other parameters

```
A T G G A T T C A
A T G G T T T - A
A C G G A T T - A
A G G G T T T - A
  j
```

$$L_{(j)} = \text{Prob}\left(\begin{array}{ccc} \text{T} & & \text{C} \\ & \text{A}\!-\!\text{A} & \\ \text{T} & & \text{G} \end{array}\right) + \text{Prob}\left(\begin{array}{ccc} \text{T} & & \text{C} \\ & \text{C}\!-\!\text{A} & \\ \text{T} & & \text{G} \end{array}\right)$$

$$+ \quad \cdots \quad + \text{Prob}\left(\begin{array}{ccc} \text{T} & & \text{C} \\ & \text{T}\!-\!\text{T} & \\ \text{T} & & \text{G} \end{array}\right)$$

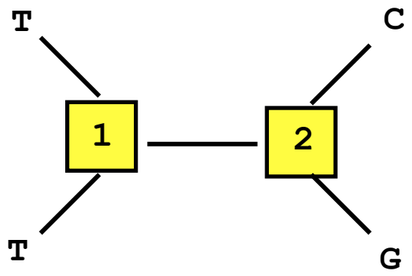
$$L = L_{(1)} \cdot L_{(2)} \cdots L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

$$\ln(L) = \ln(L_{(1)}) + \ln(L_{(2)}) + \cdots + \ln(L_{(N)}) = \sum_{j=1}^{N} \ln(L_{(j)})$$

- Probability must be summed over all possible combinations of ancestral nucleotides.
- Here we have two internal nodes giving 16 possible combinations
- Probability of individual columns are multiplied to give the overall probability of the alignment, i.e., the likelihood of the model.
- In phylogeny software these computations are done using summation of the logs of the probabilities (“log likelihoods”), because multiplication of the large number of probability terms may lead to underflow (computer problems caused by very small numbers).

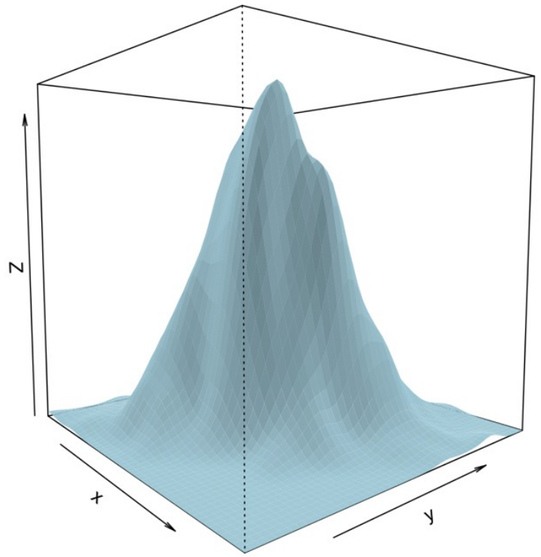
Likelihood of column in alignment: compute for each possible pair of ancestral nucleotides

Node 1	Node 2	Likelihood
A	A	0.0000009
A	C	0.0000009
A	G	0.0000009
A	T	0.0000000
C	A	0.0000001
C	C	0.0000141
C	G	0.0000014
C	T	0.0000000
G	A	0.0000001
G	C	0.0000018
G	G	0.0000150
G	T	0.0000001
T	A	0.0000248
T	C	0.0003908
T	G	0.0004028
T	T	0.0003660
Sum		0.0012198



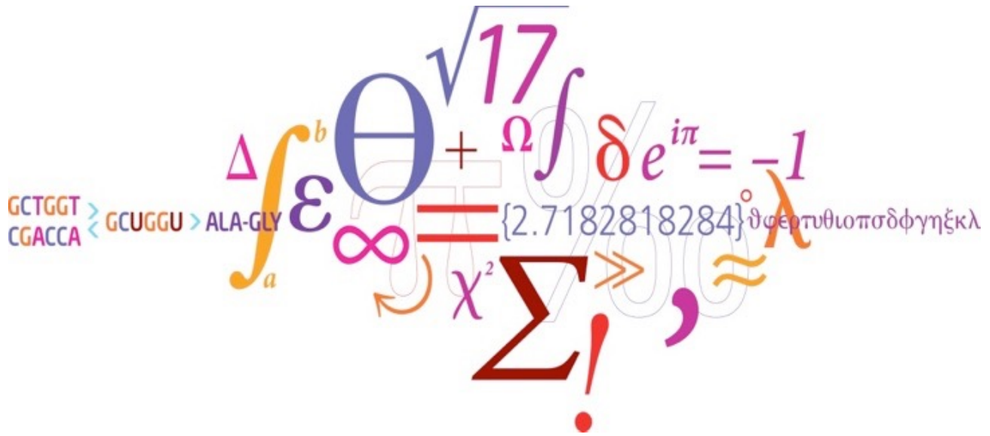
Maximum likelihood phylogeny

- **Data:**
  - sequence alignment
- **Model parameters:**
  - nucleotide frequencies, nucleotide substitution rates, tree topology, branch lengths.



- Choose random initial values for all parameters, compute likelihood
- Change parameter values slightly in a direction so likelihood improves
- Repeat until maximum found
- Results:
  - ML estimate of tree topology
  - ML estimate of branch lengths
  - ML estimate of other model parameters
  - Measure of how well model fits data (likelihood).

Ancestral Reconstruction



Likelihood of column in alignment: sum over all possible pairs of ancestral nucleotides

```
A T G G A T T C A
A T G G T T T - A
A C G G A T T - A
A G G G T T T - A
    j
```

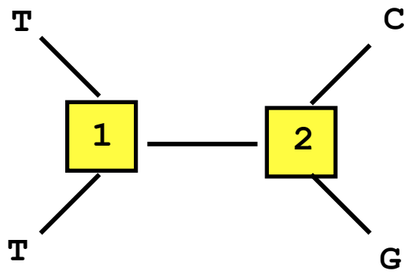
- Probability must be summed over all possible combinations of ancestral nucleotides.
- Here we have two internal nodes giving 16 possible combinations

$$L_{(j)} = \text{Prob}\left(\begin{array}{ccc} \texttt{T} & & \texttt{C} \\ & \texttt{A}\!-\!\texttt{A} & \\ \texttt{T} & & \texttt{G} \end{array}\right) + \text{Prob}\left(\begin{array}{ccc} \texttt{T} & & \texttt{C} \\ & \texttt{C}\!-\!\texttt{A} & \\ \texttt{T} & & \texttt{G} \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{ccc} \texttt{T} & & \texttt{C} \\ & \texttt{T}\!-\!\texttt{T} & \\ \texttt{T} & & \texttt{G} \end{array}\right)$$

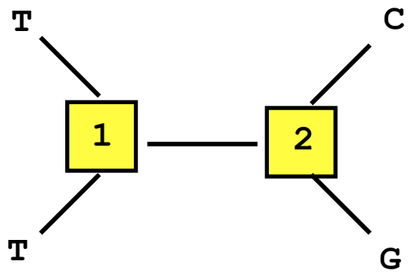
Likelihood of column in alignment: sum over all possible pairs of ancestral nucleotides

Node 1	Node 2	Likelihood
A	A	0.0000009
A	C	0.0000009
A	G	0.0000009
A	T	0.0000000
C	A	0.0000001
C	C	0.0000141
C	G	0.0000014
C	T	0.0000000
G	A	0.0000001
G	C	0.0000018
G	G	0.0000150
G	T	0.0000001
T	A	0.0000248
T	C	0.0003908
T	G	0.0004028
T	T	0.0003660
Sum		0.0012198



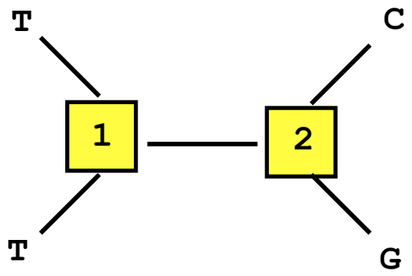
Likelihood of column in alignment: sum over all possible pairs of ancestral nucleotides

Node 1	Node 2	Likelihood
A	A	0.0000009
	C	0.0000009
	G	0.0000009
	T	0.0000000
C	A	0.0000001
	C	0.0000141
	G	0.0000014
	T	0.0000000
G	A	0.0000001
	C	0.0000018
	G	0.0000150
	T	0.0000001
T	A	0.0000248
	C	0.0003908
	G	0.0004028
	T	0.0003660
Sum		0.0012198



Ancestral Reconstruction:

Node 1	Node 2	Likelihood	Sum
A	A	0.0000009	0.0000003
	C	0.0000009	
	G	0.0000009	
	T	0.0000000	
C	A	0.0000001	0.0000156
	C	0.0000141	
	G	0.0000014	
	T	0.0000000	
G	A	0.0000001	0.0000170
	C	0.0000018	
	G	0.0000150	
	T	0.0000001	
T	A	0.0000248	0.0011844
	C	0.0003908	
	G	0.0004028	
	T	0.0003660	
Sum		0.0012198	

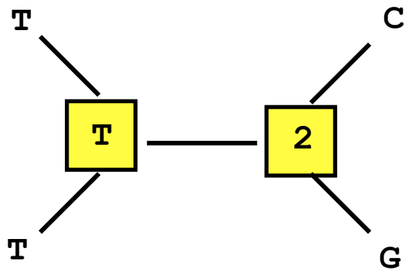


Ancestral Reconstruction:

Node 1	Node 2	Likelihood	Sum
A	A	0.0000009	0.0000003
	C	0.0000009	
	G	0.0000009	
	T	0.0000000	
C	A	0.0000001	0.0000156
	C	0.0000141	
	G	0.0000014	
	T	0.0000000	
G	A	0.0000001	0.0000170
	C	0.0000018	
	G	0.0000150	
	T	0.0000001	
T	A	0.0000248	0.0011844
	C	0.0003908	
	G	0.0004028	
	T	0.0003660	
Sum		0.0012198	

Ancestral reconstruction:

Node 1 = T



Ancestral reconstruction

- It is possible to synthesize proteins that correspond to ancestral reconstructions in the lab
- These can be investigated experimentally
- This has been done for a range of proteins including:
  - Ribonucleases
  - Chymase proteases
  - Pax transcription factors
  - Vertebrate Rhodopsins
  - Steroid receptors
  - Elongation factor EF-Tu
- Age of reconstructed ancestors: 5 million years - 1 billion years

Ancestral reconstruction: dinosaur night vision

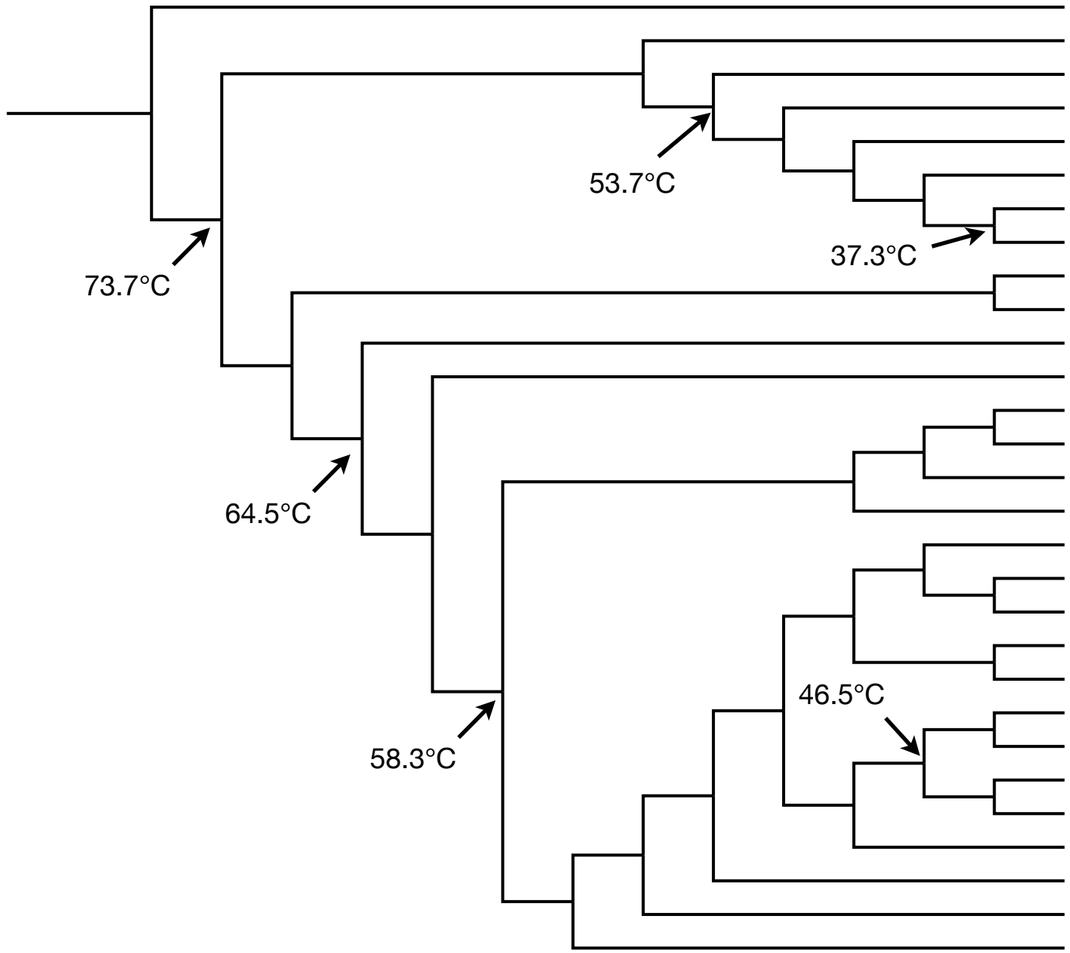
Despite its great age, the ancestral rhodopsin functioned well, carrying out all the individual steps that are required for visual function in dim light as effectively as the extant proteins in mammals, which generally have good night vision.

Specifically, the ancestral protein bound the visual chromophore 11-cis-retinal and, when exposed to light, activated the G-protein transducin at a rate similar to that of bovine rhodopsin.

These results are consistent with the hypothesis that the ancestral archosaur possessed the ability — at the molecular level at least — to see well in dim light, and might have been active at night. This insight, of course, could never have been drawn from fossils or any other non-molecular evidence about the behaviour of ancient dinosaurs.

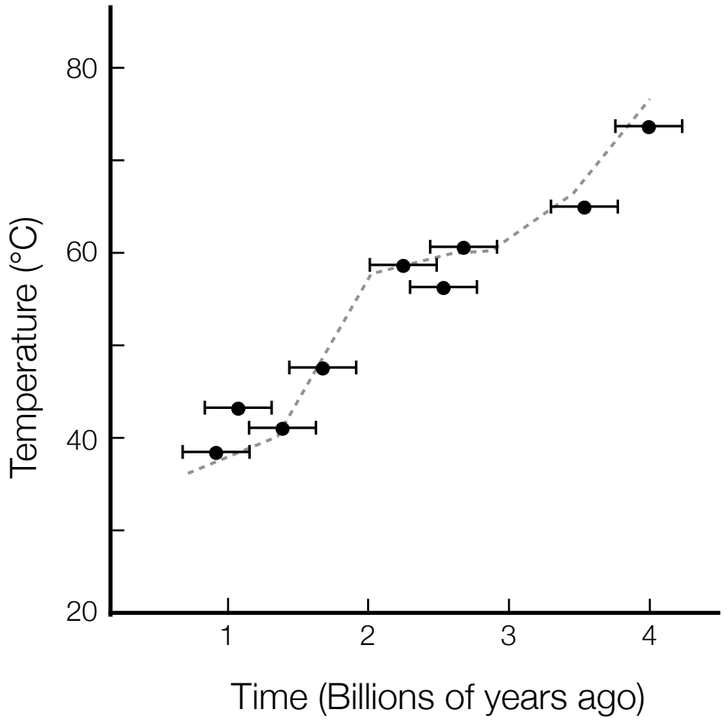
Resurrecting ancient genes: experimental analysis of extinct molecules, Nature Reviews Genetics 5, 366-375 (May 2004), Joseph W. Thornton

Ancestral reconstruction: thermostability of ancestral proteins



- **Palaeotemperature trend for Precambrian life inferred from resurrected proteins**, E. A. Gaucher, S. Govindarajan & O. K. Ganesh, Nature 451, 704-707, 2008
- Resurrection of ancient elongation factor proteins
- Melting temperatures for proteins measured in lab

Ancestral reconstruction: thermostability of ancestral proteins



Palaeotemperature trend for Precambrian life inferred from resurrected proteins, Eric A. Gaucher, Sridhar Govindarajan & Omjoy K. Ganesh, Nature 451, 704-707, 2008

Phylogeny and ancestral reconstruction for manuscripts

- Hand written manuscripts: produced by e.g., monks at convents, copying from local original (or local *copies* of original)
- Copying process resembles replication of DNA (errors introduced gradually)
- Phylogenetic methods can be used to cluster similar manuscripts: Clades typically correspond to multiple copies originating from same original
- Ancestral reconstruction can be used to make inference concerning original manuscript
- Examples:
  - **Cladistic analysis of an Old Norse manuscript tradition,** Robinson, Peter M.W., & Robert J. O'Hara. 1996, Research in Humanities Computing, 4: 115–137
  - **The phylogeny of The Canterbury Tales,** Adrian C. Barbrook, Christopher J. Howe, Norman Blake & Peter Robinson, Nature 394, 839, 1998