Outline

- Genetic distance
- Substitution models
- Phylogeny reconstruction using ML
- Hypothesis testing (LRT, AIC, Bootstrap)

Genetic distance

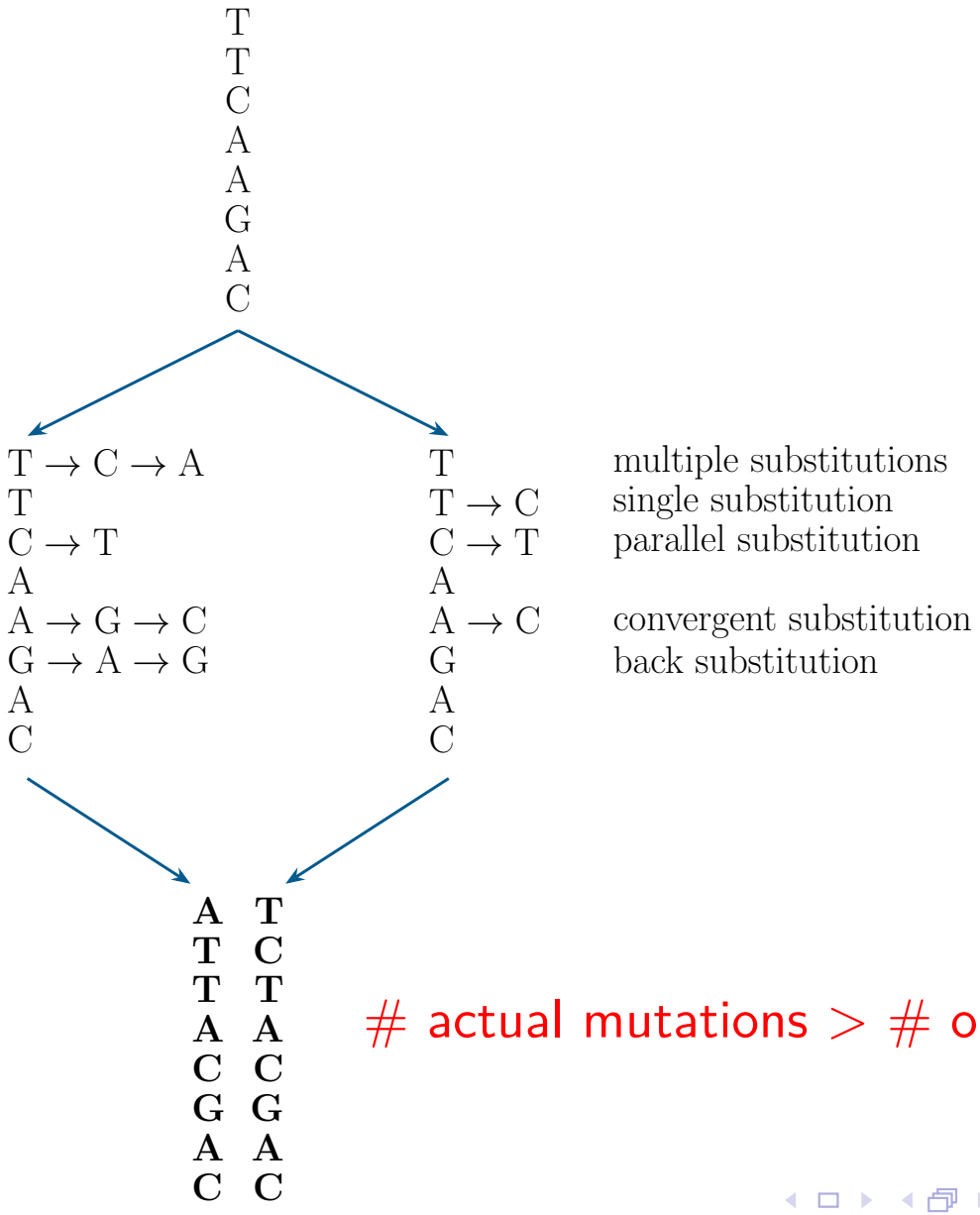
- Genetic distance is a measure of **how different two sequences are to each other.**
- In the simplest form, genetic distance can be measured by counting changes between sequences, and dividing by the number of sites (**pairwise distance**)

$$\begin{array}{c} \mathrm{ATTACGAC} \\ \mathrm{TCTACGAC} \end{array}$$

$$\text{p-distance} = 2/8 = 0.25$$

(also known as **observed** or **uncorrected** distance)

Multiple substitutions problem

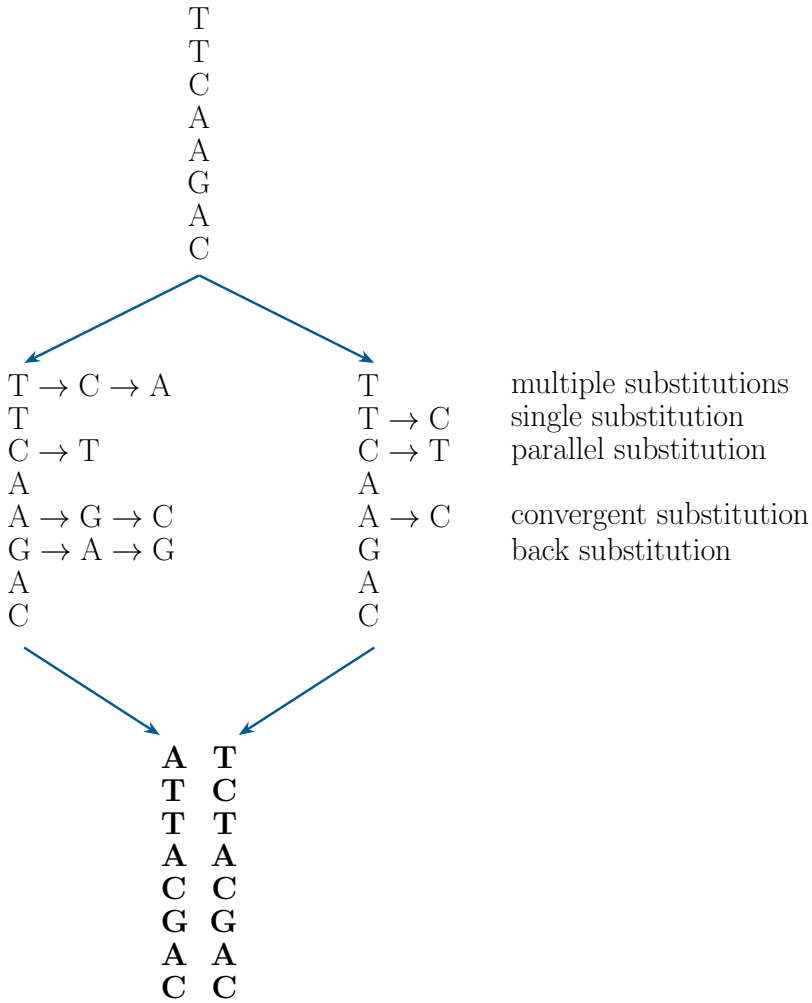


actual mutations > # observed mutations

Multiple substitutions

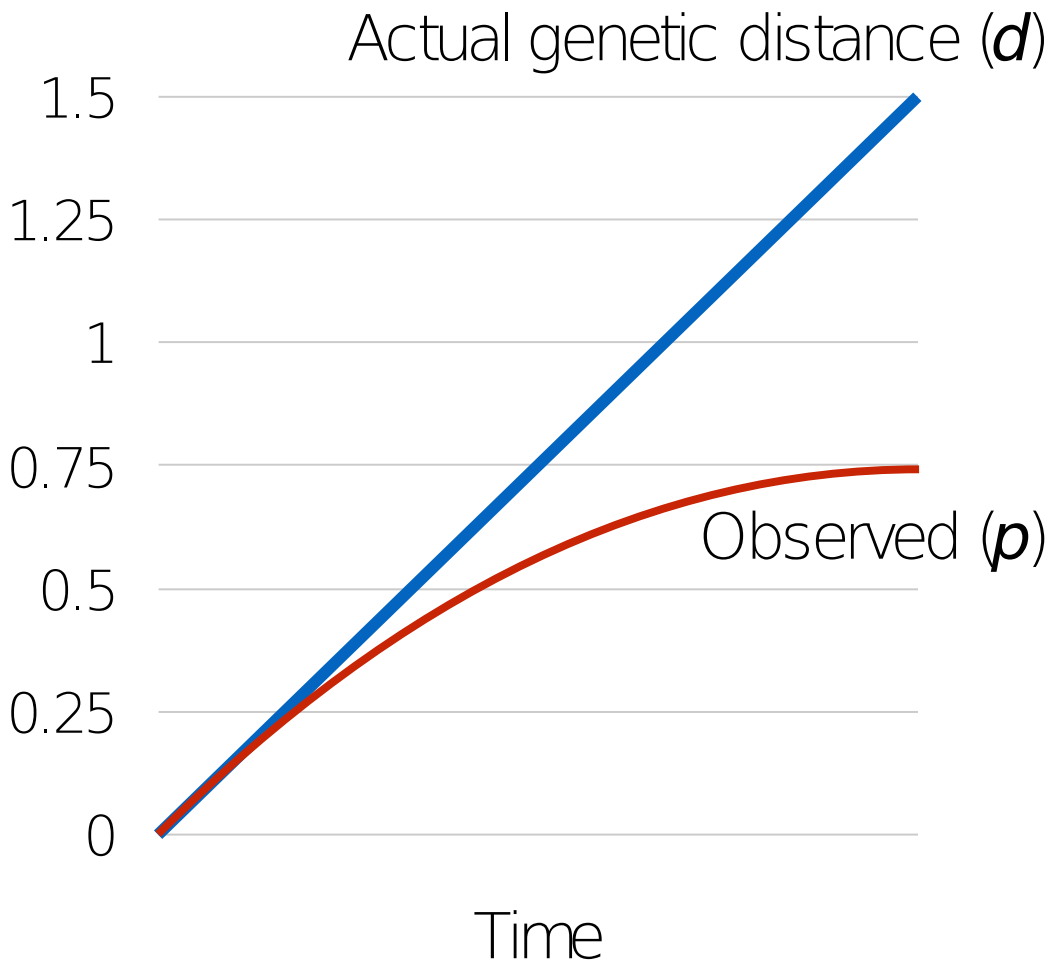
- over a long time, many substitutions will occur and the nucleotide at any particular position will be essentially random
- 25% of nucleotide sites are expected to be identical by chance

  A C G T A A A C C C G G G T T T
  A C G T C G T A G T A C T A C G
- Therefore, if we take two **unrelated** sequences, they should have an **observed** genetic distance of 0.75



Multiple substitutions

- When divergence is **low**, the observed proportion of differences (p) is a **good estimator** of genetic distance (d)
- When divergence is high, p underestimates d and a correction statistic is required



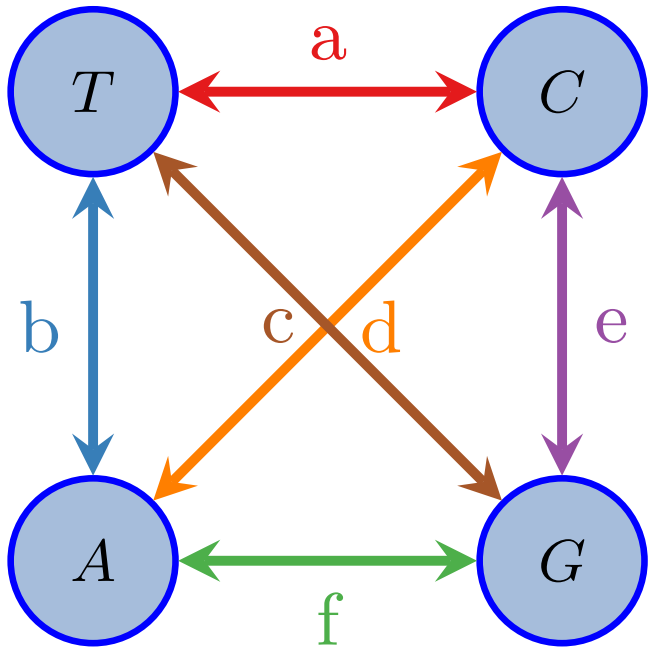
Multiple substitutions – Summary

- The **observed proportion of differences** p is the fraction of sites that differ between two sequences
- However, due to **multiple substitutions** at the same site over time (e.g., $\mathtt{A} \to \mathtt{G} \to \mathtt{T}$), p **underestimates** the actual number of substitutions that have occurred
- **Problem:** How can we estimate the true evolutionary distance d which accounts for hidden substitutions?

Substitution models

Nucleotide substitution models

Modelled as a **time-reversible** Markov process

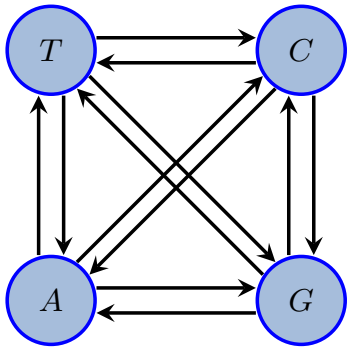


a,b,c,d,e,f = **relative** rate parameters

Jukes-Cantor model

- Simplest nucleotide substitution model (JC69)
- Corrects for multiple hits
- Q is called the **generator matrix** or **instantaneous rate matrix** of the Markov process

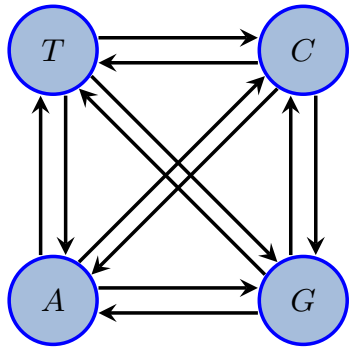
$$Q = \{q_{ij}\} = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$



Jukes-Cantor model

- Simplest nucleotide substitution model (JC69)
- Corrects for multiple hits
- Q is called the **generator matrix** or **instantaneous rate matrix** of the Markov process

$$Q = \{q_{ij}\} = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$



- Rate of change λ is the same among all nucleotides
- Total substitution rate for any nucleotide is 3λ
- Distance $d = 3\lambda t$, where t is the time that separates two sequences (**time and rate confounded**).

Transition probability (JC69)

The **transition probability matrix** $P(t)$ gives the probability of changing from one state to another over time t:

$$P(t) = e^{Qt}$$

Transition probability (JC69)

The **transition probability matrix** $P(t)$ gives the probability of changing from one state to another over time t:

$$P(t) = e^{Qt}$$

$$P(t) = \{p_{ij}(t)\} = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \end{pmatrix} \begin{array}{c} A \\ C \\ G \\ T \end{array} \end{array}$$

where $p_{ij}(t)$ is the probability a site will be in state j time t later, given it is in state i at time 0.

Transition probability (JC69)

The **transition probability matrix** $P(t)$ gives the probability of changing from one state to another over time t:

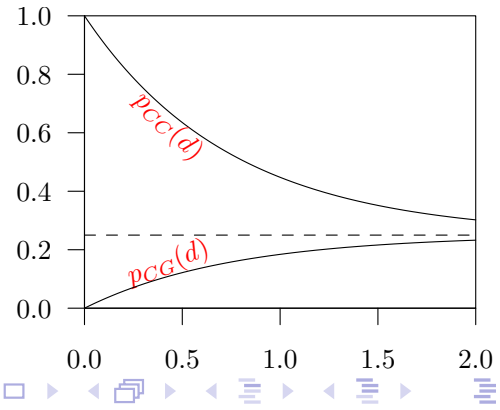
$$P(t) = e^{Qt}$$

$$P(t) = \{p_{ij}(t)\} = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \end{pmatrix} \begin{array}{c} A \\ C \\ G \\ T \end{array} \end{array}$$

where $p_{ij}(t)$ is the probability a site will be in state j time t later, given it is in state i at time 0.

Note: The matrix consists of two unique terms:

- $p_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$ (diagonal elements)
- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$ (off-diagonal elements)



Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

We assume the p-distance **approximates** the probability of change:

1. $p \approx 3p_{ij}(t)$

Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

We assume the p-distance **approximates** the probability of change:

1. $p \approx 3p_{ij}(t)$
2. $p = \frac{3}{4}(1 - e^{-4\lambda t})$

Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

We assume the p-distance **approximates** the probability of change:

1. $p \approx 3p_{ij}(t)$
2. $p = \frac{3}{4}(1 - e^{-4\lambda t})$
3. $p = \frac{3}{4}(1 - e^{-\frac{4d}{3}})$

Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

We assume the p-distance **approximates** the probability of change:

1. $p \approx 3p_{ij}(t)$
2. $p = \frac{3}{4}(1 - e^{-4\lambda t})$
3. $p = \frac{3}{4}(1 - e^{-\frac{4d}{3}})$
4. $e^{-\frac{4d}{3}} = 1 - \frac{4}{3}p$

Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

We assume the p-distance **approximates** the probability of change:

1. $p \approx 3p_{ij}(t)$
2. $p = \frac{3}{4}(1 - e^{-4\lambda t})$
3. $p = \frac{3}{4}(1 - e^{-\frac{4d}{3}})$
4. $e^{-\frac{4d}{3}} = 1 - \frac{4}{3}p$
5. $-\frac{4}{3}d = \log(1 - \frac{4}{3}p)$

Corrected distance

We can calculate the JC69 **corrected distance** between two sequences:

- $p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$: probability of change from i to j after time t
- $d = 3\lambda t$: expected number of substitutions per site (distance)
- p: observed proportion of different sites between two sequences

We assume the p-distance **approximates** the probability of change:

1. $p \approx 3p_{ij}(t)$
2. $p = \frac{3}{4}(1 - e^{-4\lambda t})$
3. $p = \frac{3}{4}(1 - e^{-\frac{4d}{3}})$
4. $e^{-\frac{4d}{3}} = 1 - \frac{4}{3}p$
5. $-\frac{4}{3}d = \log(1 - \frac{4}{3}p)$
6. $\hat{d} = -\frac{3}{4}\log(1 - \frac{4}{3}p)$ (**corrected distance**)

Distance between two sequences

Suppose two sequences consist of n sites and differ at i sites. The proportion of different sites is $p = i/n$.

Distance based on observed data (uncorrected).

$$\hat{d} = p$$

Corrected distance (Jukes-Cantor model).

$$\hat{d} = -\frac{3}{4}\log(1 - \frac{4}{3}p)$$

Example:

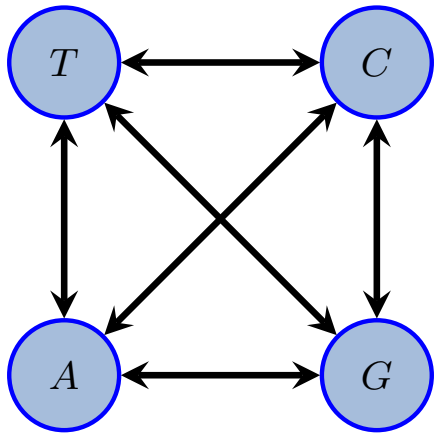
Observed p	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Actual d	0.11	0.23	0.38	0.57	0.82	1.21	2.03

p must be strictly less than 0.75, as we expect 25% of sites to be identical by chance.

Jukes Cantor (JC69)

- All bases evolve independently
- All bases are at equal frequency ($\frac{1}{4}$ each)
- Each base can change with equal probability

$$\begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix} \end{array} \begin{array}{c} T \\ C \\ A \\ G \end{array}$$

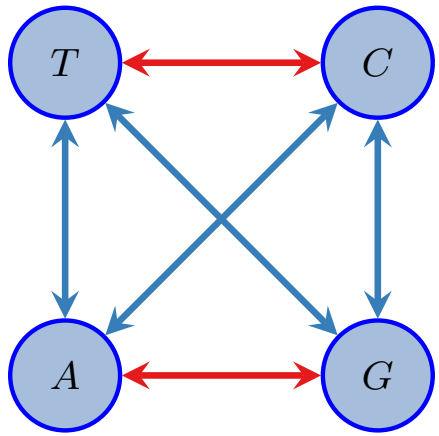


Free parameters: 0

Kimura's (1980) two-parameter model (K80)

- All bases evolve independently
- All bases are at equal frequency ($\frac{1}{4}$ each)
- **Transitions and transversions evolve at different rates**

$$\begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \begin{array}{c} T \\ C \\ A \\ G \end{array} \end{array}$$



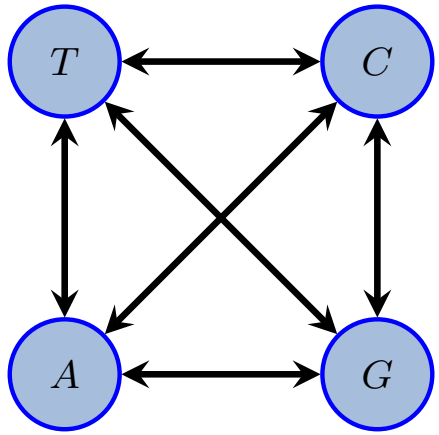
Another definition is $\kappa = \alpha/\beta$. This is 1 for no rate difference.

Free parameters: 1

Felsenstein 1981 (F81)

- All bases evolve independently
- **Bases are at unequal frequency**
- Each base can change with equal probability

$$\begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \lambda\pi_C & \lambda\pi_A & \lambda\pi_G \\ \lambda\pi_T & \cdot & \lambda\pi_A & \lambda\pi_G \\ \lambda\pi_T & \lambda\pi_C & \cdot & \lambda\pi_G \\ \lambda\pi_T & \lambda\pi_C & \lambda\pi_A & \cdot \end{pmatrix} \begin{array}{c} T \\ C \\ A \\ G \end{array} \end{array}$$

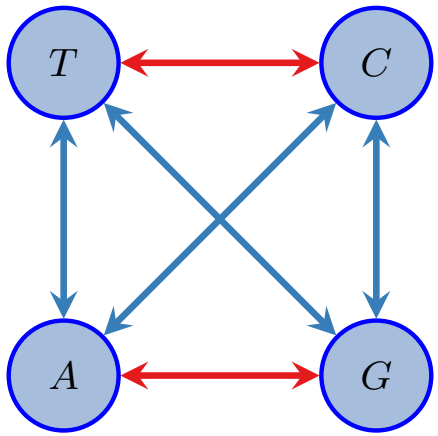


Free parameters: 3

Hasegawa, Kishino & Yano (HKY85)

- All bases evolve independently
- **Bases are at unequal frequency**
- **Transitions and transversions evolve at different rates**

$$\begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \cdot \end{pmatrix} \begin{array}{c} T \\ C \\ A \\ G \end{array} \end{array}$$

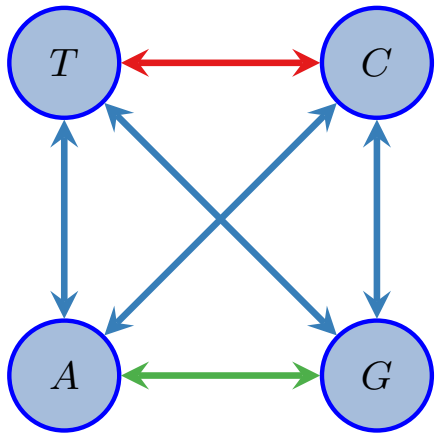


Free parameters: 4

Tamura & Nei (TN93)

- All bases evolve independently
- Bases are at unequal frequency
- **Transitions and transversions evolve at two different rates**

$$
\begin{array}{c}
\begin{array}{cccc} T & C & A & G \end{array} \\
\begin{pmatrix}
\cdot & \alpha_1\pi_C & \beta\pi_G & \beta\pi_T \\
\alpha_1\pi_T & \cdot & \beta\pi_G & \beta\pi_T \\
\beta\pi_T & \beta\pi_C & \cdot & \alpha_2\pi_T \\
\beta\pi_T & \beta\pi_C & \alpha_2\pi_G & \cdot
\end{pmatrix}
\begin{array}{c} T \\ C \\ A \\ G \end{array}
\end{array}
$$

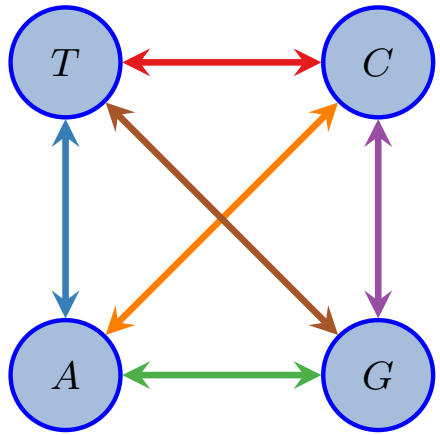


Free parameters: 5

General Time Reversible model (GTR)

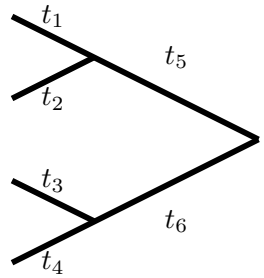
- All bases evolve independently
- Bases are at unequal frequency
- **All changes occur at different (reversible) rates**

$$\begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix} \end{array} \begin{array}{c} T \\ C \\ A \\ G \end{array}$$



Free parameters: 8

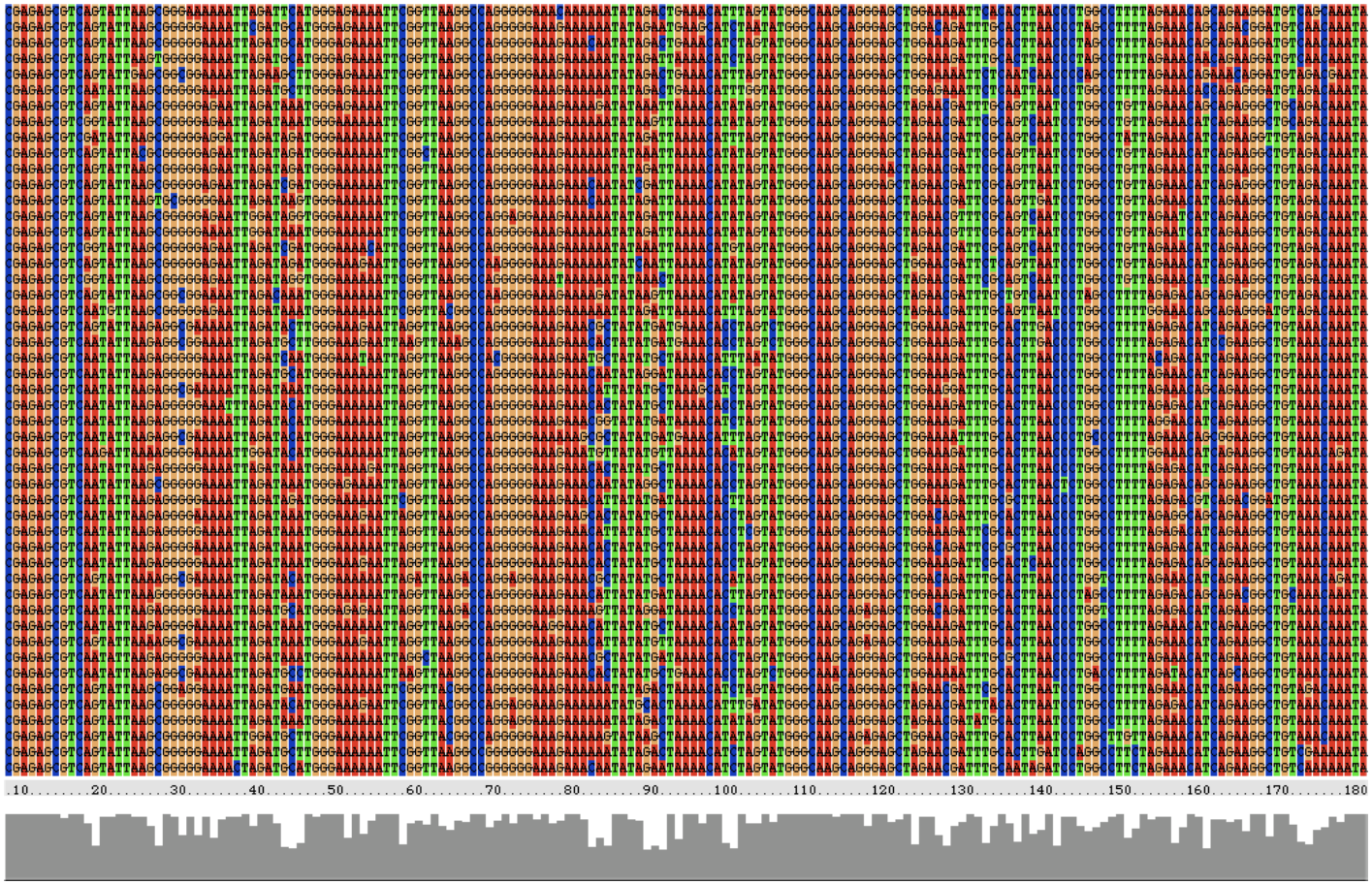
Likelihood on tree: partitioned analysis



Gene A	Gene B	Gene C	Gene D	Gene E	
$\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix}$	$\begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix}$	$\begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix}$	$\begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \cdot \end{pmatrix}$	$\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix}$	Q
$\mathrm{GTR} + \Gamma_4$	JC69	K80	HKY	$\mathrm{GTR} + \Gamma_4$	
$P_1(r_1 t) = e^{Qr_1 t}$	$P(t) = e^{Qt}$	$P(t) = e^{Qt}$	$P(t) = e^{Qt}$	$P_1(r_1 t) = e^{Qr_1 t}$	
$P_2(r_2 t) = e^{Qr_2 t}$				$P_2(r_2 t) = e^{Qr_2 t}$	
$P_3(r_3 t) = e^{Qr_3 t}$				$P_3(r_3 t) = e^{Qr_3 t}$	
$P_4(r_4 t) = e^{Qr_4 t}$				$P_4(r_4 t) = e^{Qr_4 t}$	

The rate of substitution often varies among sites

- In all models we assumed **rate homogeneity** over site (all sites evolve at the same rate)
- Rate may be heterogeneous due to different evolutionary pressures across sites or loci



Distances between primate mitochondrial genes

(a) Codon position 1 (JC69)

Human1					
Human2	**0.0053**				
Chimp1	0.0552	0.0543			
Chimp2	0.0543	0.0534	**0.0014**		
Gorilla1	0.0689	0.0692	0.0680	0.0671	
Gorilla2	0.0689	0.0698	0.0683	0.0674	**0.0025**

(b) Codon position 2 (JC69)

Human1					
Human2	**0.0011**				
Chimp1	0.0190	0.0184			
Chimp2	0.0182	0.0176	**0.0014**		
Gorilla1	0.0274	0.0268	0.0251	0.0236	
Gorilla2	0.0254	0.0248	0.0230	0.0216	**0.0025**

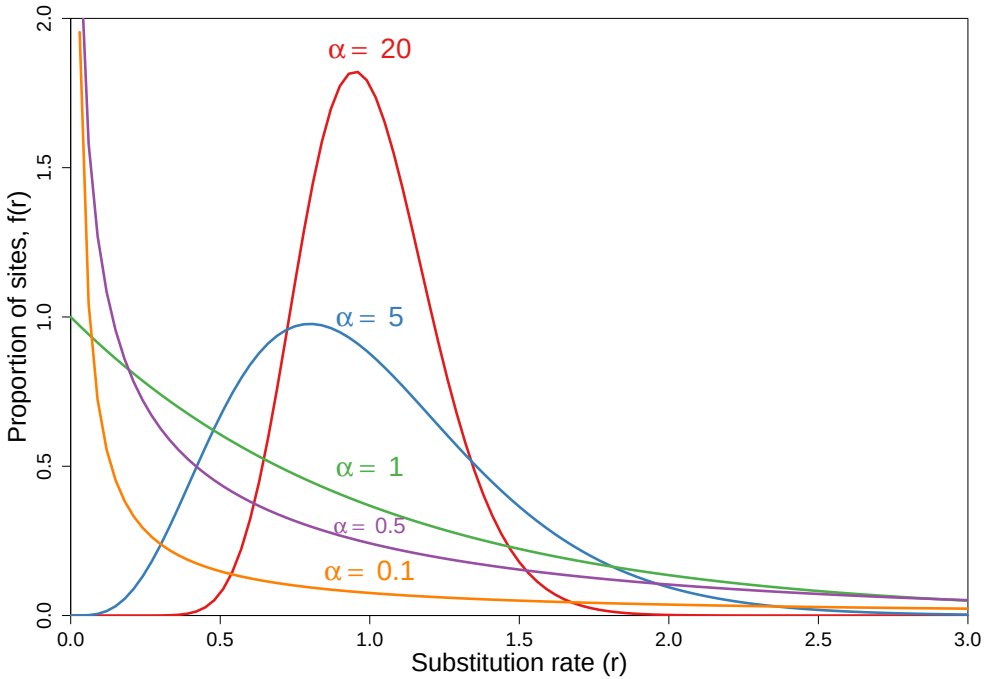
(c) Codon position 3 (JC69)

Human1					
Human2	**0.0099**				
Chimp1	0.2711	0.2695			
Chimp2	0.2711	0.2695	**0.0017**		
Gorilla1	0.3362	0.3366	0.3193	0.3184	
Gorilla2	0.3353	0.3358	0.3206	0.3197	**0.0048**

Among-site rate heterogeneity – Gamma model

- **Problem:** Some sites evolve slowly, other evolve quickly (codon positions/variable vs. conserved regions)
- **Solution:** Among-site rate heterogeneity models allow the substitution rate μ to evolve at different rates along the sequence
- One such model is the **gamma model**, which assumes that μ is distributed according to a one-parameter gamma distribution. The substitution probability is then integrated (averaged) over this distribution.

The gamma model of relative rates for sites



Gamma distribution $\Gamma(\alpha, \beta)$

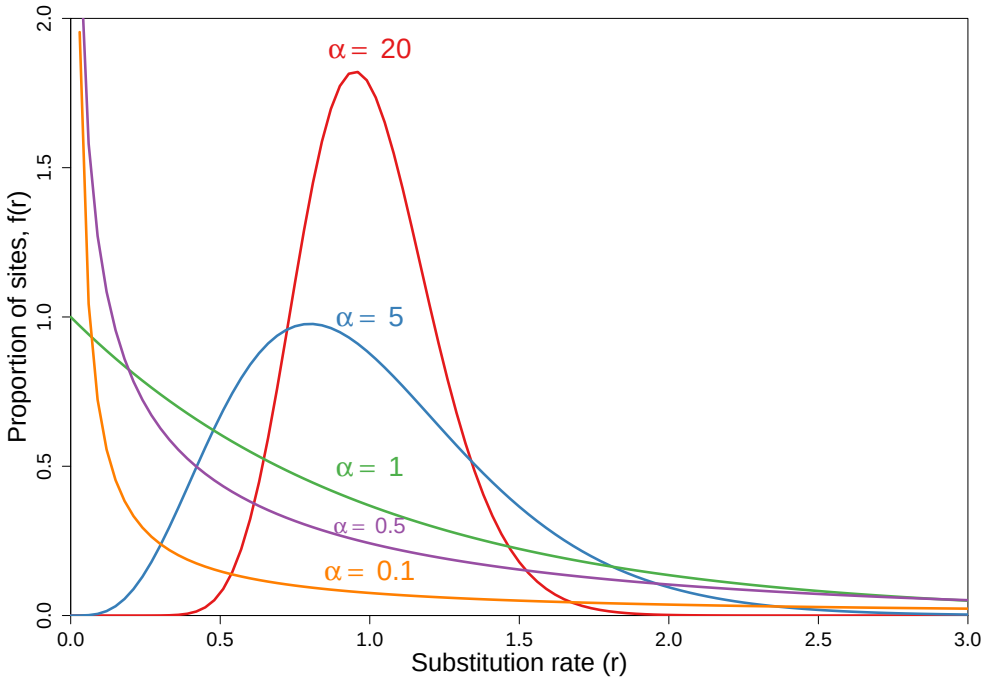
PDF $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

Mean $\frac{\alpha}{\beta}$

Variance $\frac{\alpha}{\beta^2}$

- All distributions have a mean of 1 (we set $\beta = \alpha$)
- The shape parameter α determines how variable the rates are
- Small α means more variable rates
- $\alpha = \infty$ means one rate for all sites

The gamma model of relative rates for sites



Gamma distribution $\Gamma(\alpha, \beta)$

PDF $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

Mean $\frac{\alpha}{\beta}$

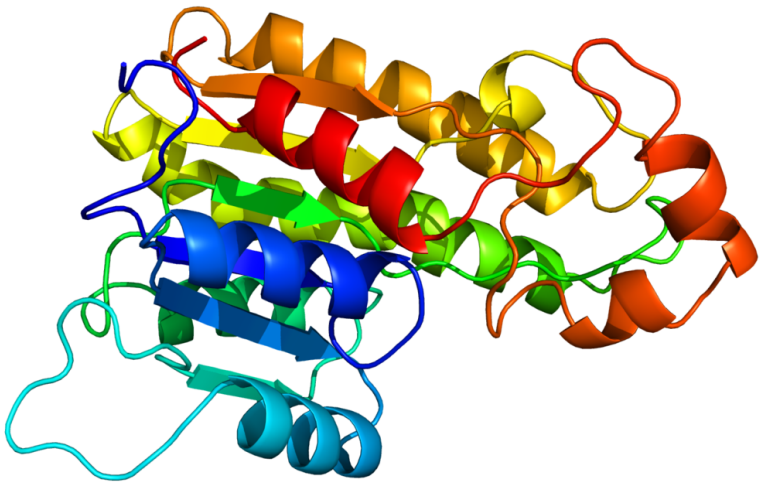
Variance $\frac{\alpha}{\beta^2}$

- All distributions have a mean of 1 (we set $\beta = \alpha$)
- The shape parameter α determines how variable the rates are
- Small α means more variable rates
- $\alpha = \infty$ means one rate for all sites

In practice..

- Alpha is usually estimated from the data using *maximum likelihood*
- Integrating the likelihood function using a continuous gamma distribution is too expensive
- Yang (1994) proposed to approximate the continuous gamma using 4 discrete categories of rates, each one with equal probability

Models of protein evolution



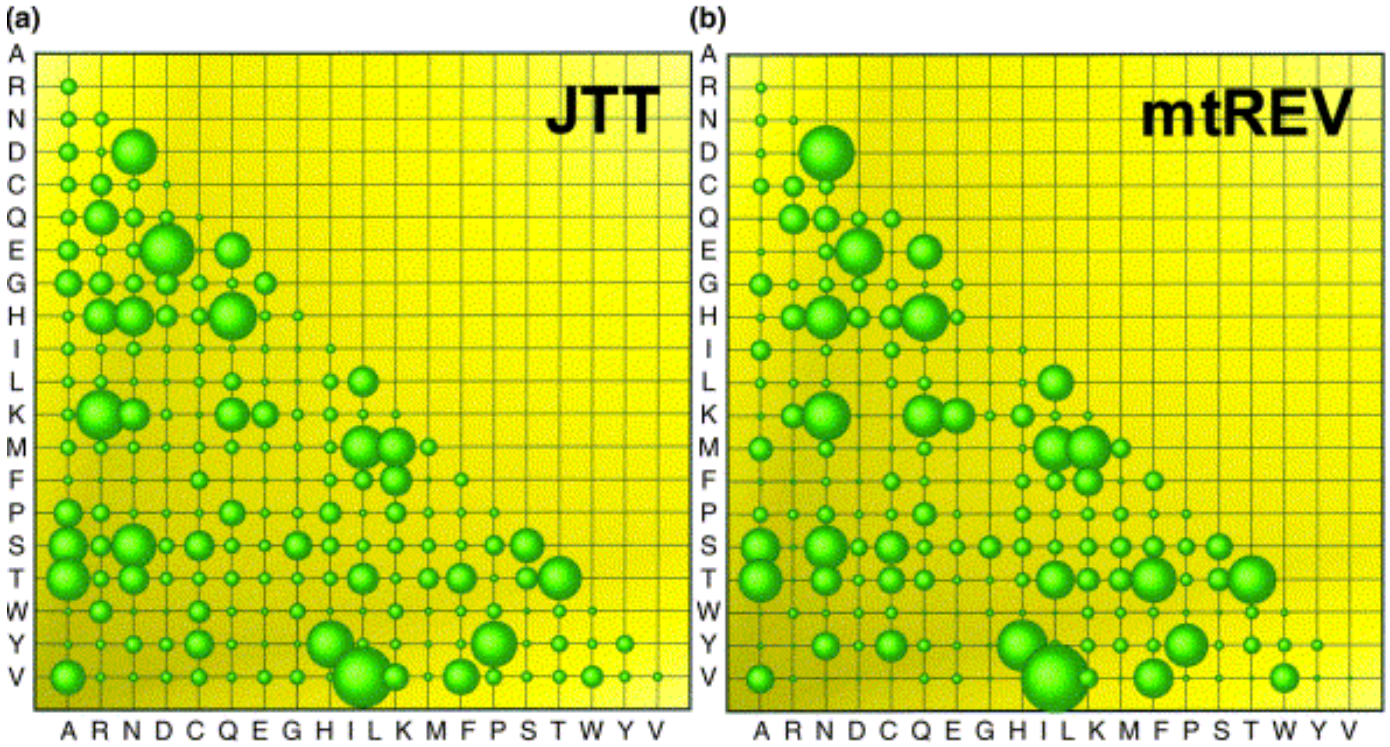
- DNA sequences have 4 states: A,C,G, and T.
- DNA models are formulated by 4×4 matrices

  $\Rightarrow$ easy to estimate params (even GTR)
- Protein sequences have 20 states: A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y.
- Protein models are formulated by 20×20 matrices

  However, parameter estimation is more complex

  $\Rightarrow$ **empirical models are often used**

Amino acid substitution matrices



- JTT derived from a range of globular proteins
- mtREV from mammalian mitochondrial genes
- No parameters in the models, empirically derived.

Amino acid substitution matrices

- There are many different amino acid substitution matrices: DAYHOFF,LG,DCMUT,JTT,MTREV,WAG,RTREV,CPREV,VT,BLOSUM, MTMAM,MTART,MTZOA,PMB,HIVB,HIVW,JTTDCMUT,FLU,STMTREV
- Most are determined from empirical data, such as physiochemical properties of the amino acids
- Therefore some matrices are more appropriate for different types of data
- For example, BLOSUM, derived from pairwise comparisons of conserved amino acids, WAG from ML estimation from globular proteins

Likelihood

Maximum likelihood – a general methodology for estimating unknown parameters in the model

- Introduced by R.A. Fisher
- Closely related to the more common concept of probability
- With probability we typically refer to the probability of observing the outcome of an event
- Likelihood tries to address the probability of the nature of an event given some observations of the outcome

Probability basics

- **Probabilities of different outcomes for a certain event must always add up to 1**

  If there is a 20% chance of rain today, there must be an 80% chance of no rain.

- **If two events are independent (that is, they in no way influence each other), then the probability of a particular pair of outcomes will be the product of the two individual outcomes**

  If we toss a coin twice, the probability of getting 2 heads is $0.5 \times 0.5 = 0.25$

- **The probability of either one outcome or another is the sum of probabilities of the two individual outcomes**

  Six-sided die: probability of a 1 or a 2 is $1/6 + 1/6 = 1/3$

Models and parameters

When talking about probability, we implicitely assume some kind of **model**

(even for simple cases of the probability of observing events such as the outcome of a coin toss)

The model states that there is some certain, fixed probability for each outcome.

Models and parameters

When talking about probability, we implicitely assume some kind of **model**

(even for simple cases of the probability of observing events such as the outcome of a coin toss)

The model states that there is some certain, fixed probability for each outcome.

Example: Coin toss

The model has one parameter, p the probability of the coin landing on heads.

- If the coin is fair, then $p = 0.5$
- We can speak about the probability of observing an outcome, given specific parameter values for the model
- In this simple case, if $p = 0.5$, then the probability of the coin landing heads on any one toss is also 0.5.

Models and parameters

When talking about probability, we implicitely assume some kind of **model**

(even for simple cases of the probability of observing events such as the outcome of a coin toss)

The model states that there is some certain, fixed probability for each outcome.

Example: Coin toss

The model has one parameter, p the probability of the coin landing on heads.

- If the coin is fair, then $p = 0.5$
- We can speak about the probability of observing an outcome, given specific parameter values for the model
- In this simple case, if $p = 0.5$, then the probability of the coin landing heads on any one toss is also 0.5.

Notation:

$\Pr(X = \text{heads} \mid p = 0.5) = 0.5$ $\quad$ $\Pr(X = \text{tails} \mid p = 0.5) = 0.5$

$\Pr(X = \text{heads} \mid p) = p$ $\quad$ $\Pr(X = \text{tails} \mid p) = 1 - p$

Likelihood

The probability of an event X dependent on model parameters p is written

$$\Pr(X \mid p)$$

Now we can talk about the **likelihood function**

$$\mathcal{L}(p \mid X)$$

which gives the likelihood of the **parameters given the observed data**.

The likelihood of p is maximized by the value (of p) that gives the maximum $\Pr(X \mid p)$ — highest probability of observing the data.

Likelihood vs Probability

- For making predictions based on a set of solid assumptions we use **probabilities**
  - $\rightarrow$ the probability of certain outcomes occurring or not occurring.
  - $\rightarrow$ e.g., predicting outcomes of coin tossing

Likelihood vs Probability

- For making predictions based on a set of solid assumptions we use **probabilities**
  - $\rightarrow$ the probability of certain outcomes occurring or not occurring.
  - $\rightarrow$ e.g., predicting outcomes of coin tossing
- For data analysis, we already observed all data: once they have been observed, they are fixed.

  We are now interested in the **likelihood** of the model parameters that underlie the fixed data

Likelihood vs Probability

- For making predictions based on a set of solid assumptions we use **probabilities**

  $\rightarrow$ the probability of certain outcomes occurring or not occurring.
  $\rightarrow$ e.g., predicting outcomes of coin tossing

- For data analysis, we already observed all data: once they have been observed, they are fixed.

  We are now interested in the **likelihood** of the model parameters that underlie the fixed data

- Difference:

  probability:

  ▷ we know the parameters $\rightarrow$ prediction of outcome

  likelihood:

  ▷ observation of data $\rightarrow$ estimation of parameters

Maximum likelihood estimation – Example

Goal: Find the parameter values that make the observed data most likely.

Maximum likelihood estimation – Example

Goal: Find the parameter values that make the observed data most likely.

Coin toss example:

Rather than assume p is a certain value, we wish to find the max likelihood estimate (MLE) of $\hat{p}$, given the observed data (coin tosses).

Maximum likelihood estimation – Example

Goal: Find the parameter values that make the observed data most likely.

Coin toss example:

Rather than assume p is a certain value, we wish to find the max likelihood estimate (MLE) of $\hat{p}$, given the observed data (coin tosses).

We toss a coin 100 times and observe 56 heads and 44 tails.

Coin tosses follow a binomial distribution:

$$P(h, n|p) = \binom{n}{h} p^h (1-p)^{n-h}$$

n: number of coin tosses
h: number of heads observed
p: probability of obtaining a head on any one toss

Possible combinations

$n = 4, h = 2$

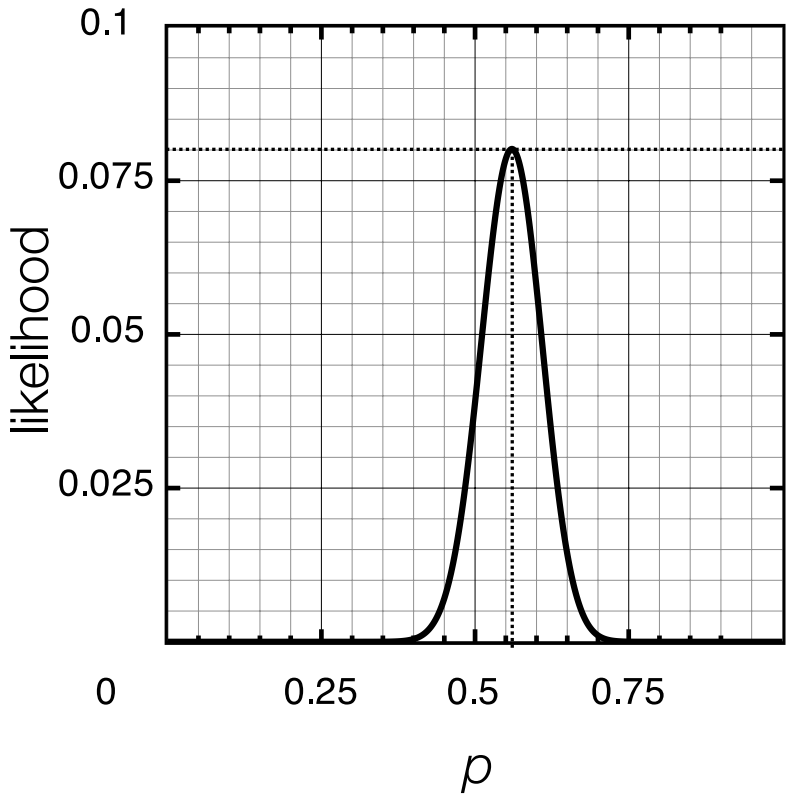
H - H - T - T
H - T - H - T
H - T - T - H
T - H - H - T
T - H - T - H
T - T - H - H

$$L(p = 0.5 \mid h = 56, n = 100) = P(h = 56, n = 100 \mid p = 0.5) = 0.0389$$

Maximum likelihood estimation

Tabulate or plot the likelihood $\mathcal{L}$ for different values of p

p	$\mathcal{L}$
0.48	0.0222
0.5	0.0389
0.52	0.0581
0.54	0.0739
0.56	**0.0801**
0.58	0.0738
0.6	0.0576
0.62	0.0378



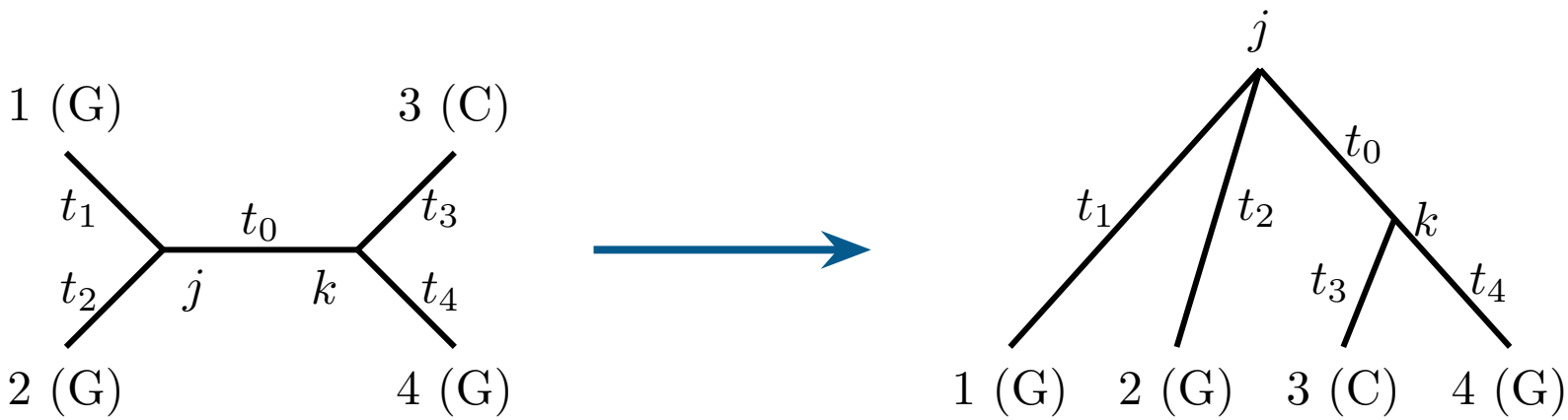
- The max likelihood estimate (MLE) of $\hat{p}$ is 0.56 with a likelihood of 0.0801
- The MLE for a binomial distribution is the observed proportion of heads

Maximum likelihood for estimating phylogenies

Likelihood calculation on tree

Site	1	2	3	4	5	...	i	...				n
Sequence 1	C	T	C	A	T	...	**G**	...	G	T	A	A	T
Sequence 2	C	T	A	G	T	...	**G**	...	C	T	A	G	T
Sequence 3	C	T	A	G	T	...	**C**	...	G	T	A	G	T
Sequence 4	C	C	A	A	C	...	**G**	...	C	C	A	A	T
Probability	p_1	p_2				...	p_i	...					p_n

$$L = p_1 \times p_2 \times \ldots \times p_i \times \ldots \times p_n = \prod_{i=1}^{n} p_i$$

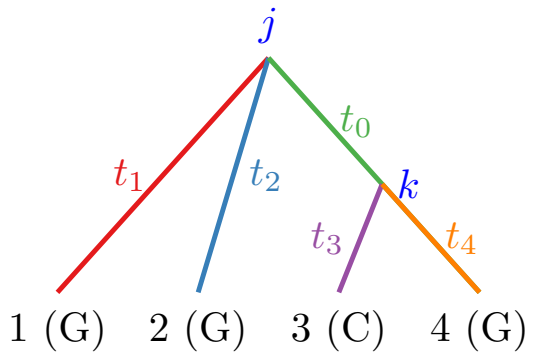


The probability of each site is a sum over all possible ancestral states

$$p_i = \Pr\left(\begin{array}{cccc} & & \text{T} & \\ & & & \diagdown \\ \text{G} & \text{G} & \text{C} & \text{T} \end{array}\right) + \Pr\left(\begin{array}{cccc} & & \text{T} & \\ & & & \diagdown \\ \text{G} & \text{G} & \text{C} & \text{C} \end{array}\right) + \Pr\left(\begin{array}{cccc} & & \text{T} & \\ & & & \diagdown \\ \text{G} & \text{G} & \text{C} & \text{A} \end{array}\right) + \ldots + \Pr\left(\begin{array}{cccc} & & \text{G} & \\ & & & \diagdown \\ \text{G} & \text{G} & \text{C} & \text{G} \end{array}\right)$$

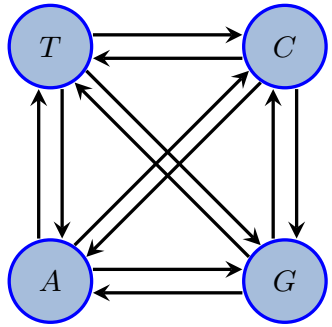
4×4 terms

$$\Pr\left(\begin{array}{cccc} & & \text{j} & \\ & & & \diagdown \\ \text{G} & \text{G} & \text{C} & \text{k} \end{array}\right) = \pi_j \, p_{j,G}(t_1) \, p_{j,G}(t_2) \, p_{j,k}(t_0) \, p_{k,C}(t_3) \, p_{k,G}(t_4)$$



JC69 model of substitution

$$Q = \{q_{ij}\} = \begin{array}{c} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$



$$P(t) = e^{Qt} = \{p_{ij}(t)\} = \begin{array}{c} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \end{pmatrix} \end{array} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

$$p_{T,T}(t) = p_{C,C}(t) = p_{A,A}(t) = p_{G,G}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$$

$$p_{T,C}(t) = p_{T,A}(t) = \ldots = p_{G,A}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$$

Summary: likelihood calculation on tree

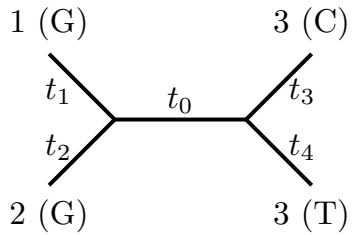
The log likelihood of a tree is the sum of log probabilities over all sites.

We assume sites evolve independently.

The probability at each site p_i is a sum over all ancestral reconstructions.

For each ancestral reconstruction, the probability is a product of transition probabilities over branches.

$$\mathcal{L}(t_0, t_1, t_2, t_3, t_4|X) = \sum_{i=1}^{n} \log(p_i)$$



Summary: likelihood calculation on tree

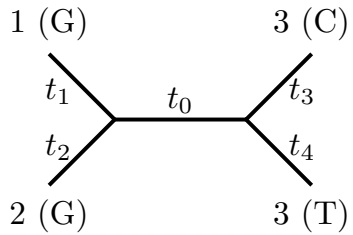
The log likelihood of a tree is the sum of log probabilities over all sites.

We assume sites evolve independently.

The probability at each site p_i is a sum over all ancestral reconstructions.

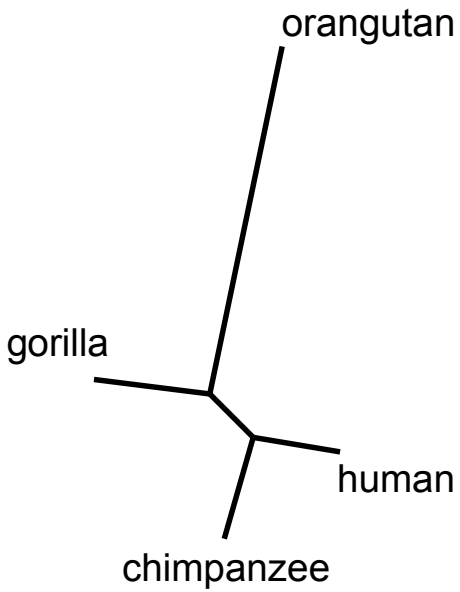
For each ancestral reconstruction, the probability is a product of transition probabilities over branches.

$$\mathcal{L}(t_0, t_1, t_2, t_3, t_4|X) = \sum_{i=1}^{n} \log(p_i)$$



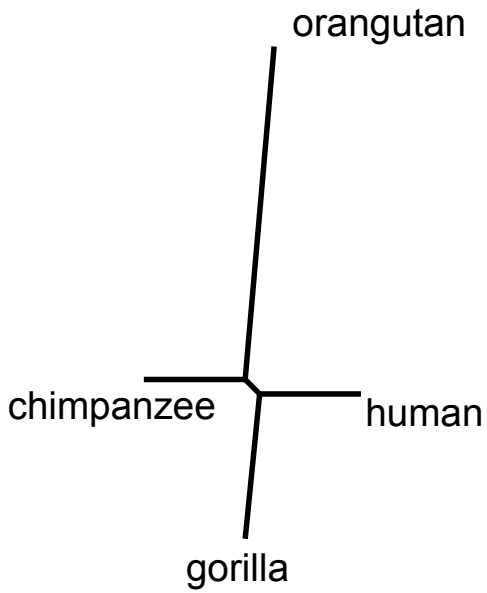
- $\mathcal{L}$ is a function of branch lengths t_0, t_1, t_2, t_3, t_4 (and any substitution parameters)
- We estimate them by maximizing $\mathcal{L}$ (via optimization).
- The optimum $\mathcal{L}$ corresponding to the MLEs of parameters is the score for the tree.
- We repeat this process for all possible trees. The maximum likelihood tree is the one with the highest score.

Ape trees for mtDNA under K80



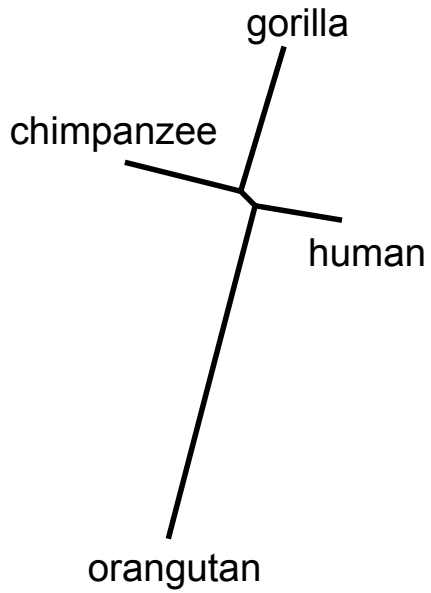
$\kappa = 11.4$

$\mathcal{L} = -2270.5$



$\kappa = 11.1$

$\mathcal{L} = -2280.6$



$\kappa = 10.7$

$\mathcal{L} = -2278.6$

Brown WM, Prager EM, Wang A, Wilson AC. **Mitochondrial DNA sequences of primates: Tempo and mode of evolution**. *J. Mol. Evol* 18:225-239, 1982.

Likelihood ratio test

Likelihood ratio test (LRT) compares a simpler (null) hypothesis H_0 against a more general (alternative) hypothesis H_1.

$$\mathrm{LR} = \frac{\mathcal{L}_1}{\mathcal{L}_0}$$

If the ratio $\mathrm{LR} > 1$ then hypothesis H_1 has a higher likelihood.

For **nested models** twice the log-likelihood difference,

$$2\Delta = \ln(\mathrm{LR}^2) = 2 \ln \mathrm{LR} = 2(\ln \mathcal{L}_1 - \ln \mathcal{L}_0)$$

is compared with the χ^2 distribution.

Degrees of freedom (df) is set to the **difference in number of parameters** between the two models.

We can perform a statistical test to detemine which hypothesis best describes the data.

Likelihood ratio test – Example

Example: LRT of JC69 against K80

H_0: JC69 ($\kappa = 1$) $\ln \mathcal{L}_0 = -1710.58$

H_1: K80 ($\kappa =$?) $\ln \mathcal{L}_1 = -1637.90$

$$2\Delta\mathcal{L} = 2[-1637.90 - (-1710.58)] = \mathbf{145.36}$$

K80 has one more parameter than JC69, d.f. $= 1$. Compare with χ^2 distribution with 1 d.f.

df	χ^2 value										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	**3.84**	**6.63**	**10.83**
2	0.1	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
***p*-value**	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Critical values are 3.84 at 5% and 6.63 at 1% and 10.83 at 0.1%. K80 fits the data significantly better

Data: human and orangutan mt 12s rRNA genes (D38112 and NC_001646), 943 sites.

Likelihood ratio test - models must be nested

If the simpler model is not a special case of the more complex model, we cannot use this test statistic

i.e., must be able to set the parameters of the complex model to specific values to obtain the simpler model.

Examples:

Rate heterogeneity can be set to 0

$\Rightarrow$ HKY is a special case of HKY+G

Substitution rates can be set equal to each other

$\Rightarrow$ JC69, F81, K80 are all special cases of HKY

Likelihood ratio test

The LRT is designed to compare two **nested models**

- Tell us whether the more complex model provides a significantly better fit to the data than the simpler model, given the extra parameters.
- It cannot tell us the best model overall, i.e., it does not provide an **absolute measure** of model quality.

Model selection

How can we compare **non-nested models**?

Model selection

How can we compare **non-nested models**?

The **Kullback-Leibler** (K-L) divergence (or **relative entropy**) is a measure of the divergence between two probability distributions:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \qquad D_{KL}(P||Q) = \int P(x) \log \left(\frac{P(x)}{Q(x)} \right) dx$$

Model selection

How can we compare **non-nested models**?

The **Kullback-Leibler** (K-L) divergence (or **relative entropy**) is a measure of the divergence between two probability distributions:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \qquad D_{KL}(P||Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

Suppose our data is generated by some **unknown process** $\boldsymbol{P}$.

We consider two **candidate models** $\boldsymbol{Q_1}$ **and** $\boldsymbol{Q_2}$.

If we knew P, we would **calculate** $\boldsymbol{D_{KL}(P||Q_1)}$ **and** $\boldsymbol{D_{KL}(P||Q_2)}$ and **pick** the candidate that minimizes divergence.

Model selection

How can we compare **non-nested models**?

The **Kullback-Leibler** (K-L) divergence (or **relative entropy**) is a measure of the divergence between two probability distributions:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \qquad D_{KL}(P||Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

Suppose our data is generated by some **unknown process** $\boldsymbol{P}$.

We consider two **candidate models** $\boldsymbol{Q_1}$ **and** $\boldsymbol{Q_2}$.

If we knew P, we would **calculate** $\boldsymbol{D_{KL}(P||Q_1)}$ **and** $\boldsymbol{D_{KL}(P||Q_2)}$ and **pick** the candidate that minimizes divergence.

Unfortunately, we do not know P

Akaike Information Criterion

Akaike derived an approximation of KL using the MLE of the model parameters.

AIC estimates the expected, **relative** Kullback-Leibler divergence (information loss) between the true model P and candidate Q.

If $\mathcal{L}$ is the maximum likelihood value for some model Q with K (free) parameters, then:

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2K$$

In other words, we can estimate how much more (or less) information is lost by Q_1 compared to Q_2. **The preferred model is the one with the minimum AIC value.**

Akaike Information Criterion

Akaike derived an approximation of KL using the MLE of the model parameters.

AIC estimates the expected, **relative** Kullback-Leibler divergence (information loss) between the true model P and candidate Q.

If $\mathcal{L}$ is the maximum likelihood value for some model Q with K (free) parameters, then:

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2K$$

In other words, we can estimate how much more (or less) information is lost by Q_1 compared to Q_2. **The preferred model is the one with the minimum AIC value.**

Example: JC69 against K80

H_0: JC69 ($\kappa = 1$) $\quad \ln \mathcal{L}_0 = -1710.58$
H_1: K80 ($\kappa =$?) $\quad \ln \mathcal{L}_1 = -1637.90$

H_0: $\mathrm{AIC} = 2 \times 1710.58 = 3421.16$
H_1: $\mathrm{AIC} = 2 \times 1637.90 + 2 = 3277.80$ **Preferred model**

Bayesian Information Criterion

Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = -2\ln(\mathcal{L}) + K\ln(n)$$

Akaike Information Criterion (AIC)

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2K$$

Bayesian Information Criterion (BIC) is similar to AIC, but penalizes models with more parameters more heavily in particular for large sample sizes.

The penalty term for AIC is $2K$, while for BIC it is $K\ln(n)$ where n is the sample size (sites) and K the number of parameters in the model.

Statistics for phylogenies

- Measures of phylogenetic support

  Bootstrap (non-parametric)

  Parametric bootstrap

  Bayesian posterior probabilities

  Jackknifing

  others....

Note: different tree topologies are **not** nested, and therefore χ^2 approximating to the likelihood ratio is not valid.

Bootstrap

What is the bootstrap?

- A measure of confidence of our phylogenetic tree
- Originally introduced for phylogenetics in 1985 by Felsenstein.
- Can be applied to any method that starts from a sequence alignment, e.g., parsimony, likelihood, distance methods
- For each boostrap sample:
  - Create a new pseudo-replicate alignment by sampling the columns of the original alignment
  - Construct a tree for the pseudo-replicate alignment
  - Count the frequency of the nodes of the original tree in the bootstrap trees

1

Bootstrap

Original Alignment

Site:	1	2	3	4	5	6	7	8	9	10
Human	N	E	N	L	F	A	S	F	I	A
Chimpanzee	N	E	N	L	F	A	S	F	A	A
Bonobo	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap Alignment

Site:
Human
Chimpanzee
Bonobo
Gorilla
Orangutan
Sumatran
Gibbon

Task: Create a new alignment of equal size to the original by sampling sites from the original alignment with replacement.

Bootstrap

Original Alignment

Site:	1	2	3	4	5	6	7	8	9	10
Human	N	E	N	L	F	A	S	F	I	A
Chimpanzee	N	E	N	L	F	A	S	F	A	A
Bonobo	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap Alignment

Site:	2
Human	E
Chimpanzee	E
Bonobo	E
Gorilla	E
Orangutan	E
Sumatran	E
Gibbon	E

Task: Create a new alignment of equal size to the original by sampling sites from the original alignment with replacement.

Bootstrap

Original Alignment

Site:	1	2	3	4	5	6	7	8	9	10
Human	N	E	N	L	F	A	S	F	I	A
Chimpanzee	N	E	N	L	F	A	S	F	A	A
Bonobo	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap Alignment

Site:	2	1
Human	E	N
Chimpanzee	E	N
Bonobo	E	N
Gorilla	E	N
Orangutan	E	N
Sumatran	E	N
Gibbon	E	N

Task: Create a new alignment of equal size to the original by sampling sites from the original alignment with replacement.

Bootstrap

Original Alignment

Site:	1	2	3	4	5	6	7	8	9	10
Human	N	E	N	L	F	A	S	F	I	A
Chimpanzee	N	E	N	L	F	A	S	F	A	A
Bonobo	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap Alignment

Site:	2	1	4	9
Human	E	N	L	I
Chimpanzee	E	N	L	A
Bonobo	E	N	L	A
Gorilla	E	N	L	I
Orangutan	E	N	L	T
Sumatran	E	N	L	I
Gibbon	E	N	L	A

Task: Create a new alignment of equal size to the original by sampling sites from the original alignment with replacement.

Bootstrap

Original Alignment

Site:	1	2	3	4	5	6	7	8	9	10
Human	N	E	N	L	F	A	S	F	I	A
Chimpanzee	N	E	N	L	F	A	S	F	A	A
Bonobo	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap Alignment

Site:	2	1	4	9	1
Human	E	N	L	I	N
Chimpanzee	E	N	L	A	N
Bonobo	E	N	L	A	N
Gorilla	E	N	L	I	N
Orangutan	E	N	L	T	N
Sumatran	E	N	L	I	N
Gibbon	E	N	L	A	N

Task: Create a new alignment of equal size to the original by sampling sites from the original alignment with replacement.

Bootstrap

Original Alignment

Site:	1	2	3	4	5	6	7	8	9	10
Human	N	E	N	L	F	A	S	F	I	A
Chimpanzee	N	E	N	L	F	A	S	F	A	A
Bonobo	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap Alignment

Site:	2	1	4	9	1	8	9	5	3	7
Human	E	N	L	I	N	F	I	F	N	S
Chimpanzee	E	N	L	A	N	F	A	F	N	S
Bonobo	E	N	L	A	N	F	A	F	N	S
Gorilla	E	N	L	I	N	F	I	F	N	S
Orangutan	E	N	L	T	N	F	T	F	D	P
Sumatran	E	N	L	I	N	F	I	F	S	P
Gibbon	E	N	L	A	N	F	A	F	N	S

Task: Create a new alignment of equal size to the original by sampling sites from the original alignment with replacement.

Bootstrap pipeline

Original Alignment

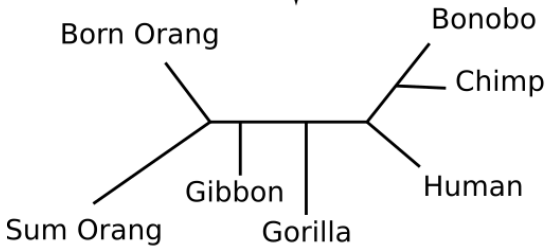
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
N	E	N	L	F	A	S	F	I	A
N	E	N	L	F	A	S	F	A	A
N	E	N	L	F	A	S	F	A	A
N	E	N	L	F	A	S	F	I	A
N	E	D	L	F	T	P	F	T	T
N	E	S	L	F	T	P	F	I	T
N	E	N	L	F	T	S	F	A	T

Bootstrap pipeline

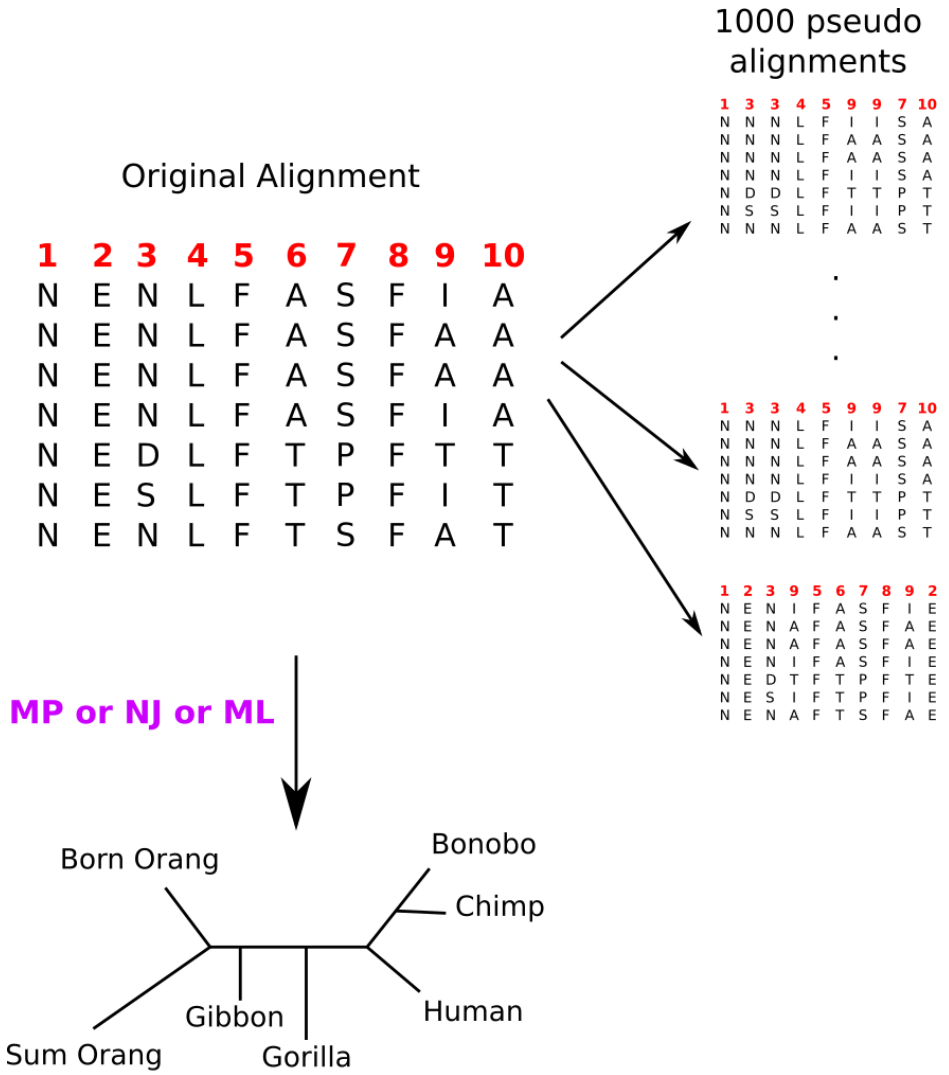
Original Alignment

1	2	3	4	5	6	7	8	9	10
N	E	N	L	F	A	S	F	I	A
N	E	N	L	F	A	S	F	A	A
N	E	N	L	F	A	S	F	A	A
N	E	N	L	F	A	S	F	I	A
N	E	D	L	F	T	P	F	T	T
N	E	S	L	F	T	P	F	I	T
N	E	N	L	F	T	S	F	A	T

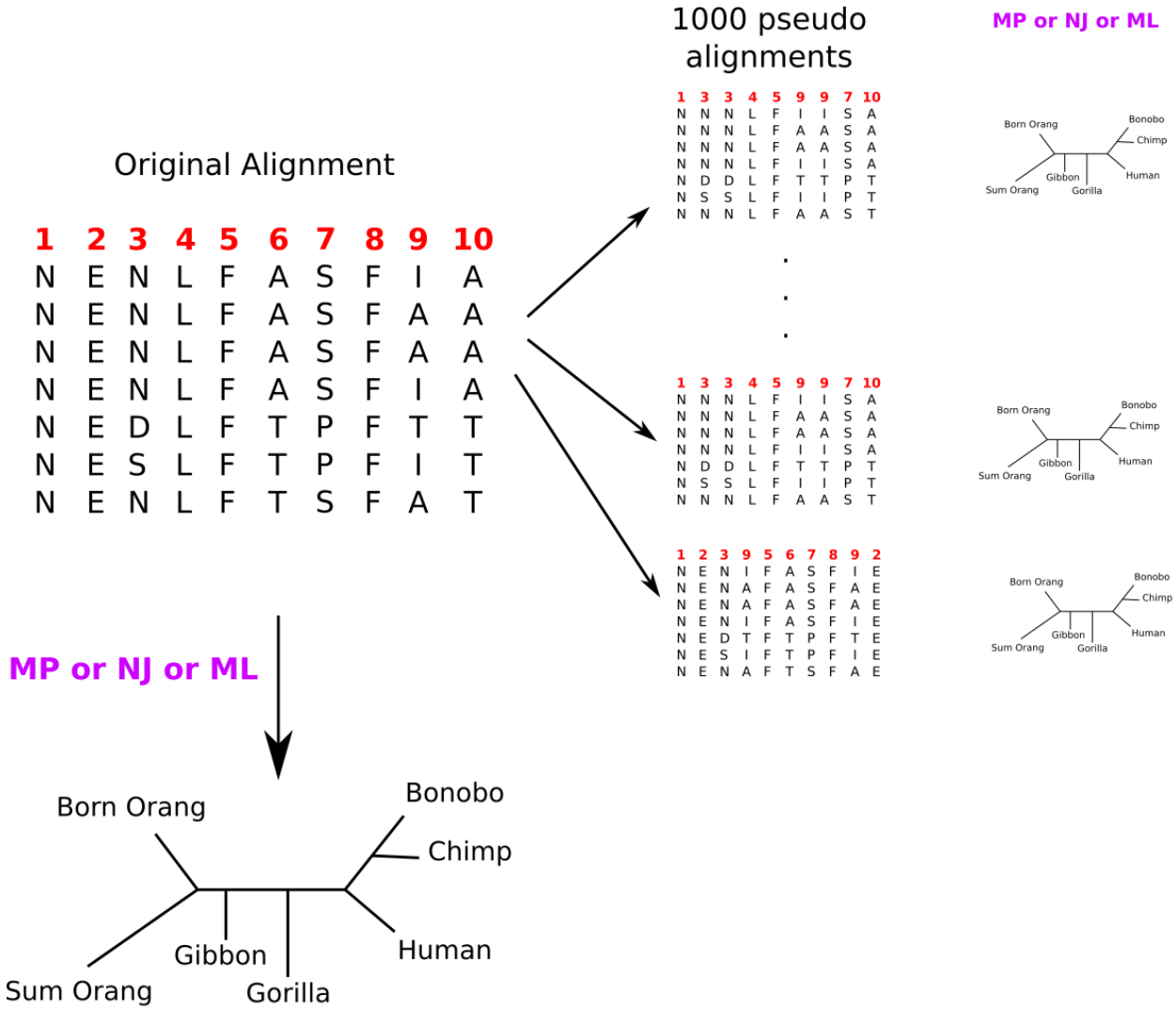
MP or NJ or ML



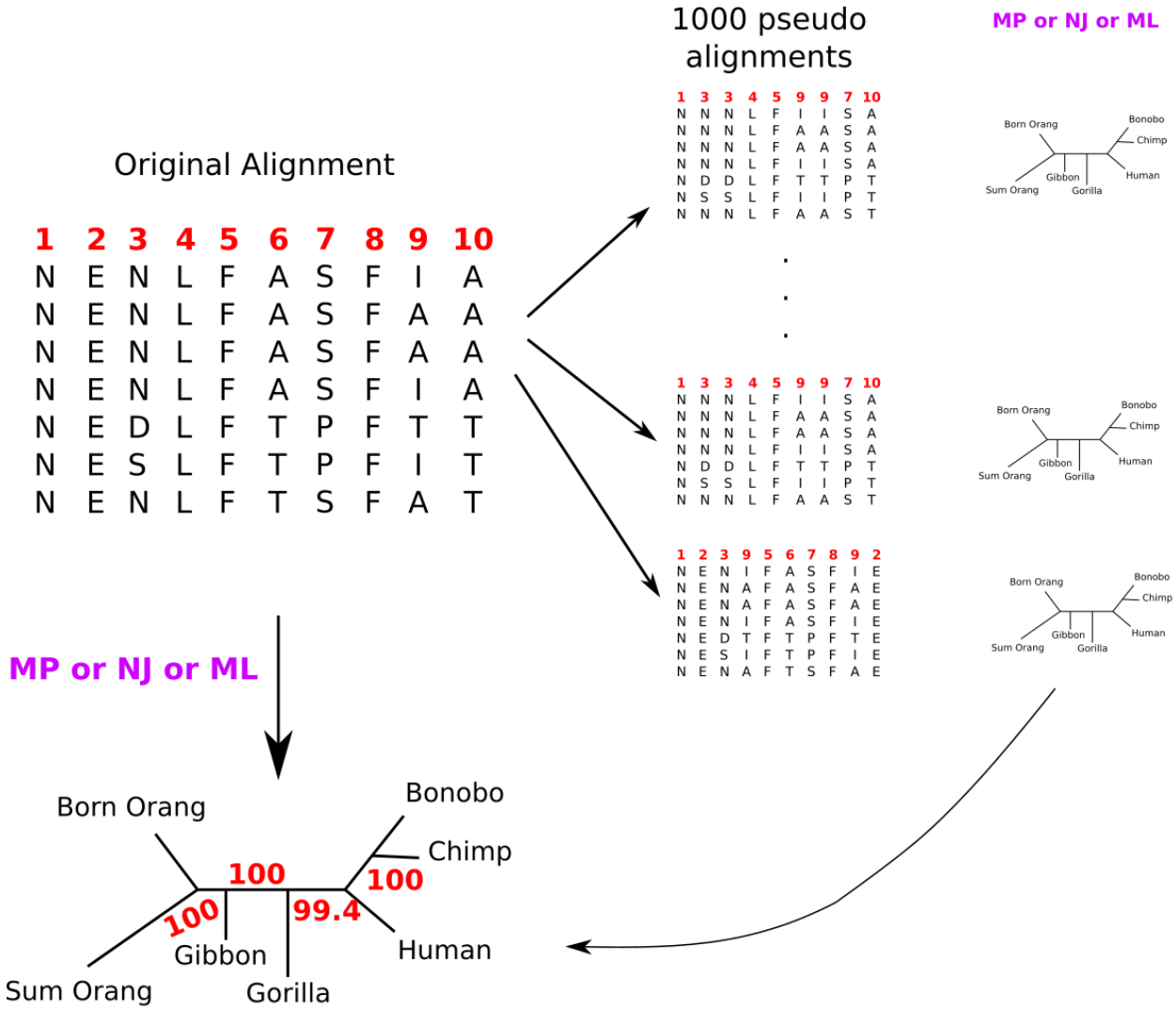
Bootstrap pipeline



Bootstrap pipeline



Bootstrap pipeline



Bootstrap pipeline

- Bootstrap replicates help us assess whether all sites support the same topology or if there is conflicting signal
- Bootstrap values vary from 0 to 100 (higher values stronger support)
- Bootstrap values help us evaluate potential stochastic errors but can be misleading under model violations