Reconstruction of phylogenetic trees

Wessel van Wieringen
w.n.van.wieringen@vu.nl

Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

vrije Universiteit

VU medisch centrum

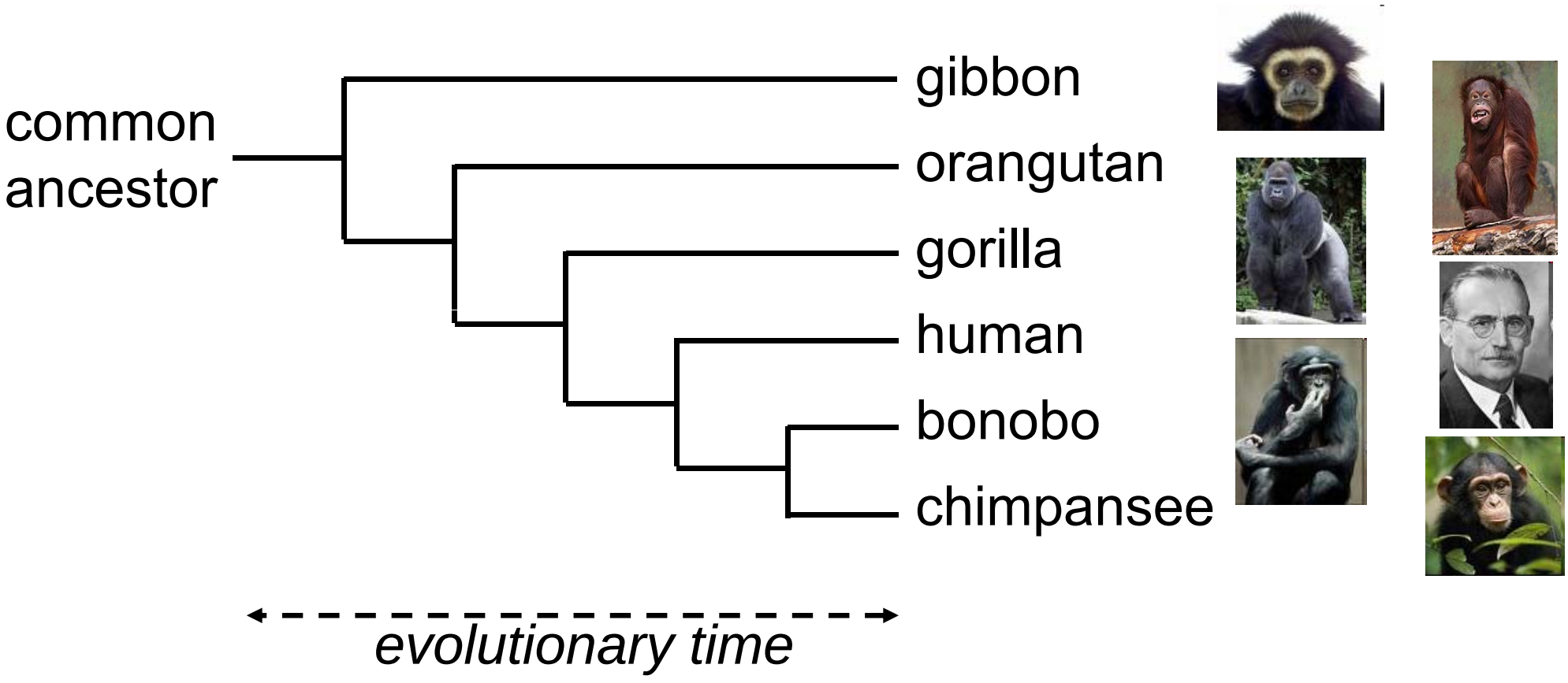
Phylogenetics

Phylogenetics

"Acceptance of the theory of evolution as the means of explaining observed similarities and differences among organisms invites the construction of trees of descent purporting to show evolutionary relationships"

-- Cavalli-Sforza, Edwards (1967)

Phylogenetics



Phylogenetics

Phylogenetics is the study of evolutionary relationships between organisms.

Goal

- Reconstruct correct genealogical ties among biogical entities.
- Estimate the time of divergence between organisms.
- Chronicle the sequence of events along evolutionary lineages.

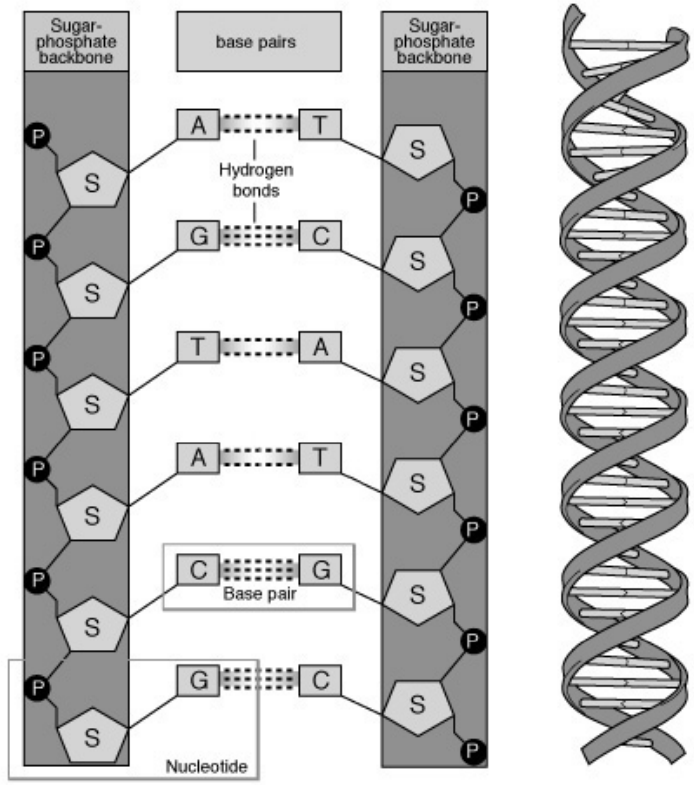
Statistical operationalization: reconstruction of phylogenetic trees on the basis of DNA sequences.

This can also be done on the basis of other characteristics.

Phylogenetics

Conceptually, **DNA** is an information-carrier.

At the molecular level DNA is a double-stranded polymer existing of four basic molecular units, called ***nucleotides***, and denoted by the letters: A, C, G and T.



… ACCCGATAGCT …

Phylogenetics

DNA of each individual is unique, but differences are small: 1 in 500 to 1000 nucleotides differ between two individuals.

Within a population each position in the DNA has a ‘pre-dominant’ nucleotide.

Over generations this ‘pre-dominant’-nucleotide of a position can change by evolution.

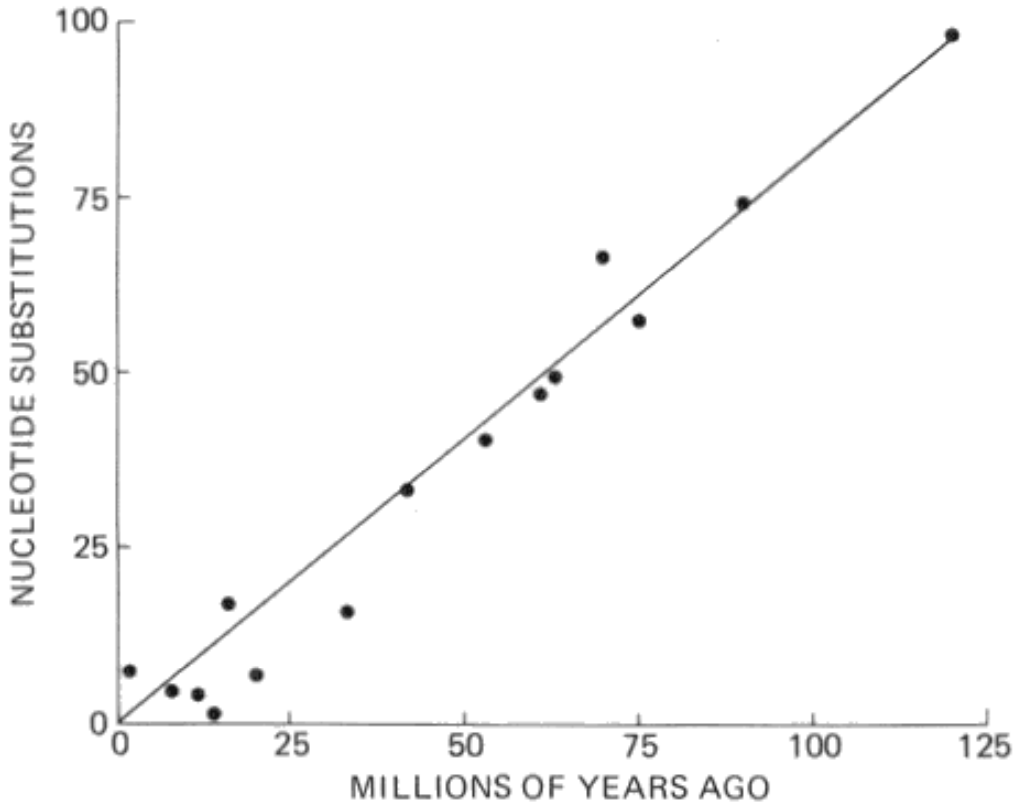
This process is called ***substitution***, and takes place over 1000s of generations.

Phylogenetics

Molecular clock-hypothesis

Pair-wise DNA differences between 17 mamal species, plotted against their ‘time-of-divergence’, determined from fossil records.

The linear relation suggests that molecular differences between pairs of species are proportional to their ‘time-of-divergence’.



(Wilson *et al*., 1977)

Phylogenetics

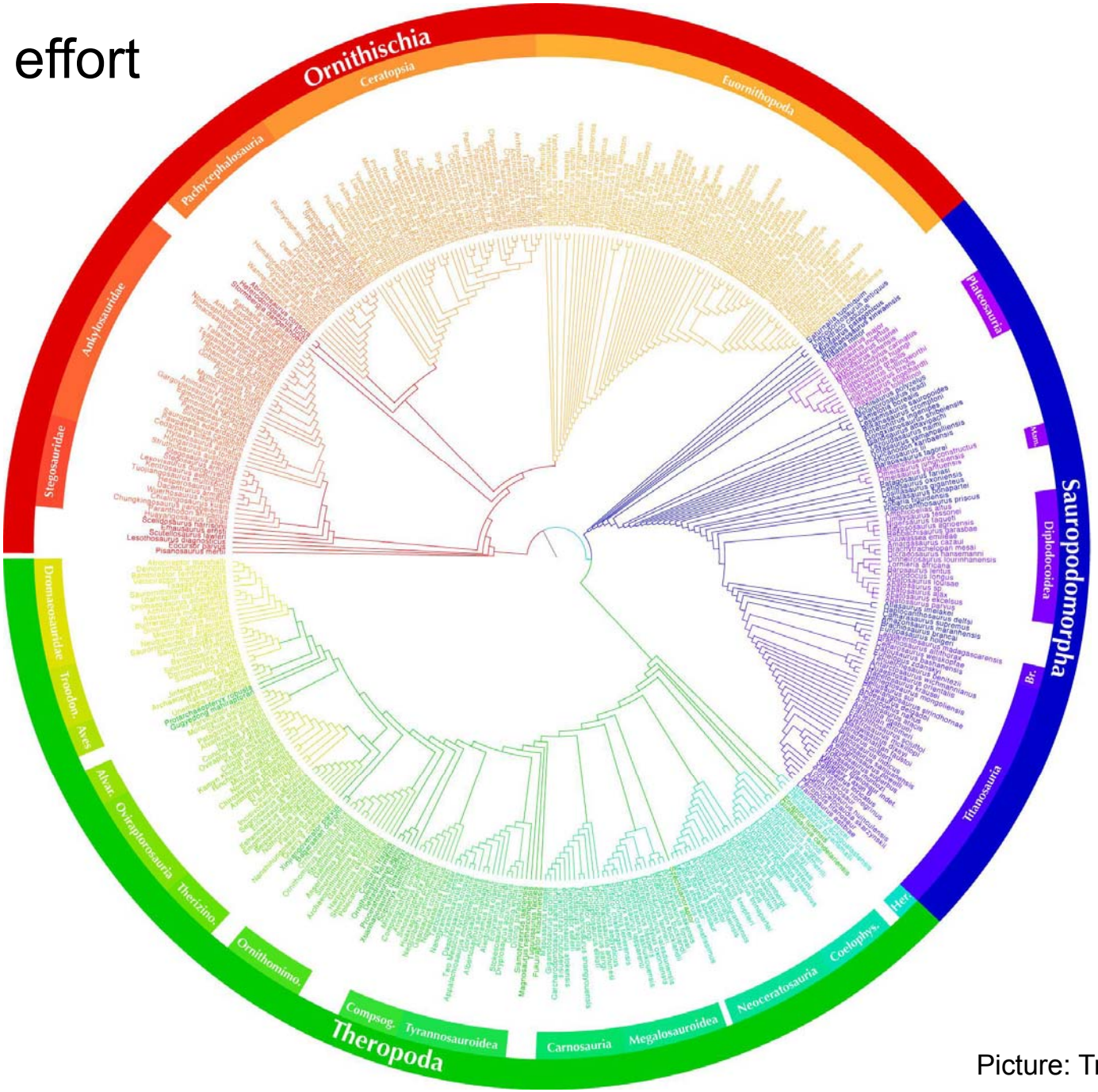
Reconstruction of molecular phylogenetic relations is a step-wise process:

1) Select sequences.
2) Build a model that describes evolution over time.
3) Find the tree that best describes the phylogenetic relations between the sequences.
4) Interpret the results.

contribution of the statistician

Phylogenetics

On-going effort



Picture: Tree of Life Web Project

Phylogenetics

The platypus: reptile or mamal?

Recently, the genome of the platypus / duck bill has been sequenced.

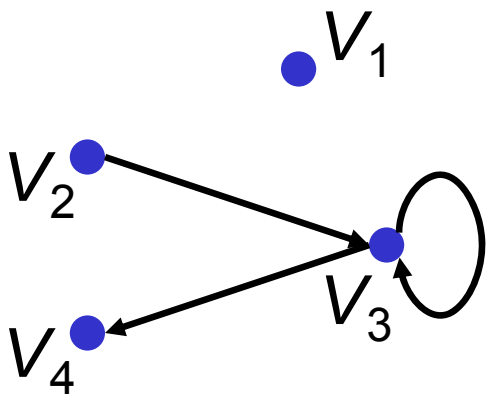
This revealed:
a) +/- 220 My ago separated from the reptiles,
b) +/- 170 My ago separated from the mamals,
and then evolved separately.

Intermezzo on graphs

Intermezzo on graphs

A ***graph*** is a system of connected components. The connections are called ***edges***, and components ***nodes***.

The ***topology*** of a graph is a pair (***V***, ***E***), where ***V*** the set of edges and ***E*** a subset of ***V*** x ***V***.

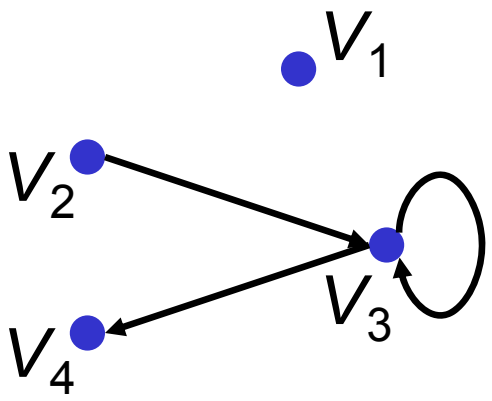


$$\boldsymbol{V} = \{ V_1, V_2, V_3, V_4 \}$$

$$\boldsymbol{E} = \{ (V_2, V_3), (V_3, V_4), (V_3, V_3) \}$$

Intermezzo on graphs

A ***path*** in a graph is a set of connected edges. In case the begin and end point of a path coincide, the path is called a ***cycle***.



Path: (V_2, V_3), (V_3, V_4)

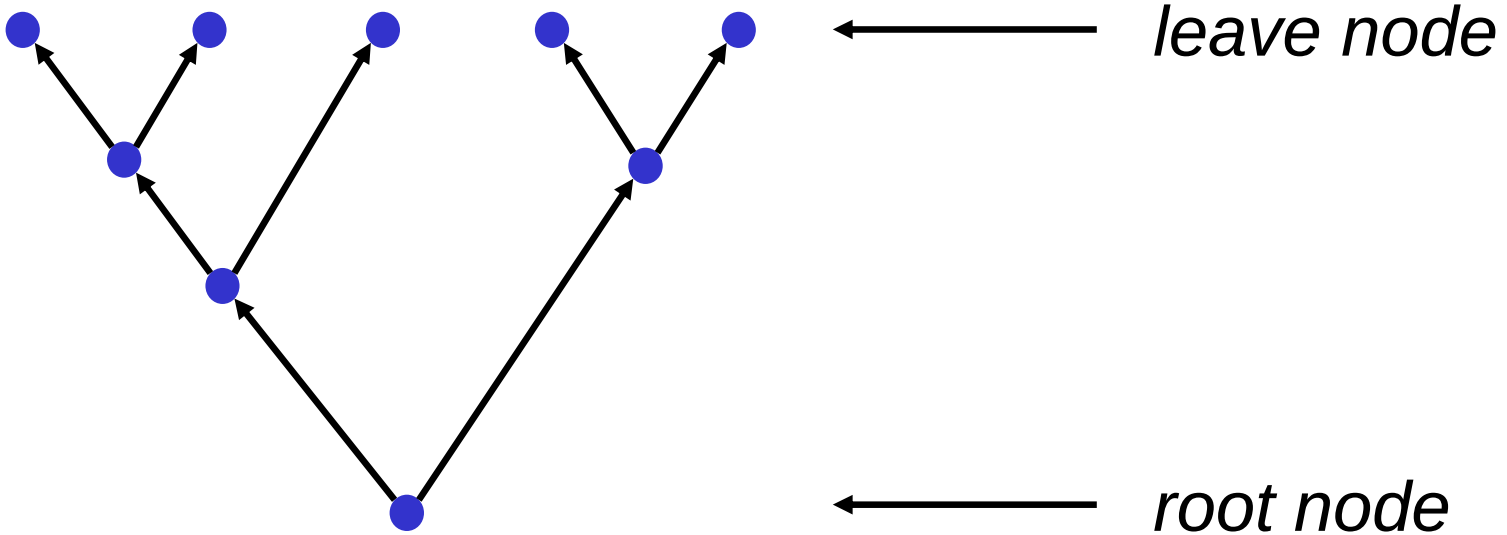
Cycle: (V_3, V_3)

Intermezzo on graphs

If a nodes of a graph are connected (i.e., there is a path between all nodes), the graph is called ***connected***.

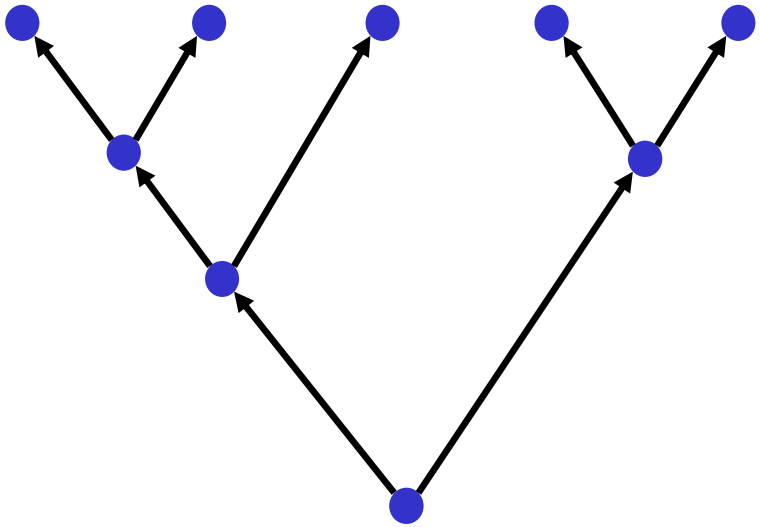
A connected graph that contains no cycles is called a ***tree***.

In a ***binary tree*** every node has either one or three edges, except for the ***root node***, if present, that has two edges.

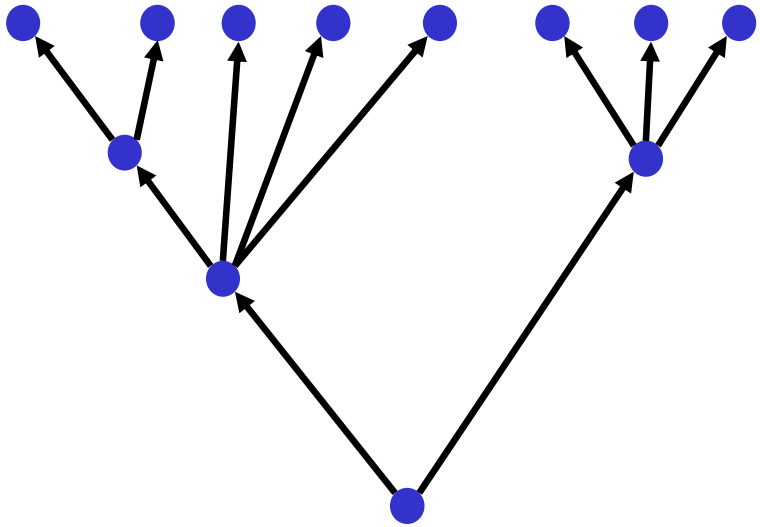


Intermezzo on graphs

Here we only consider binary trees. This rules out the possibility of one species evolving into three or more new species at a particular instance



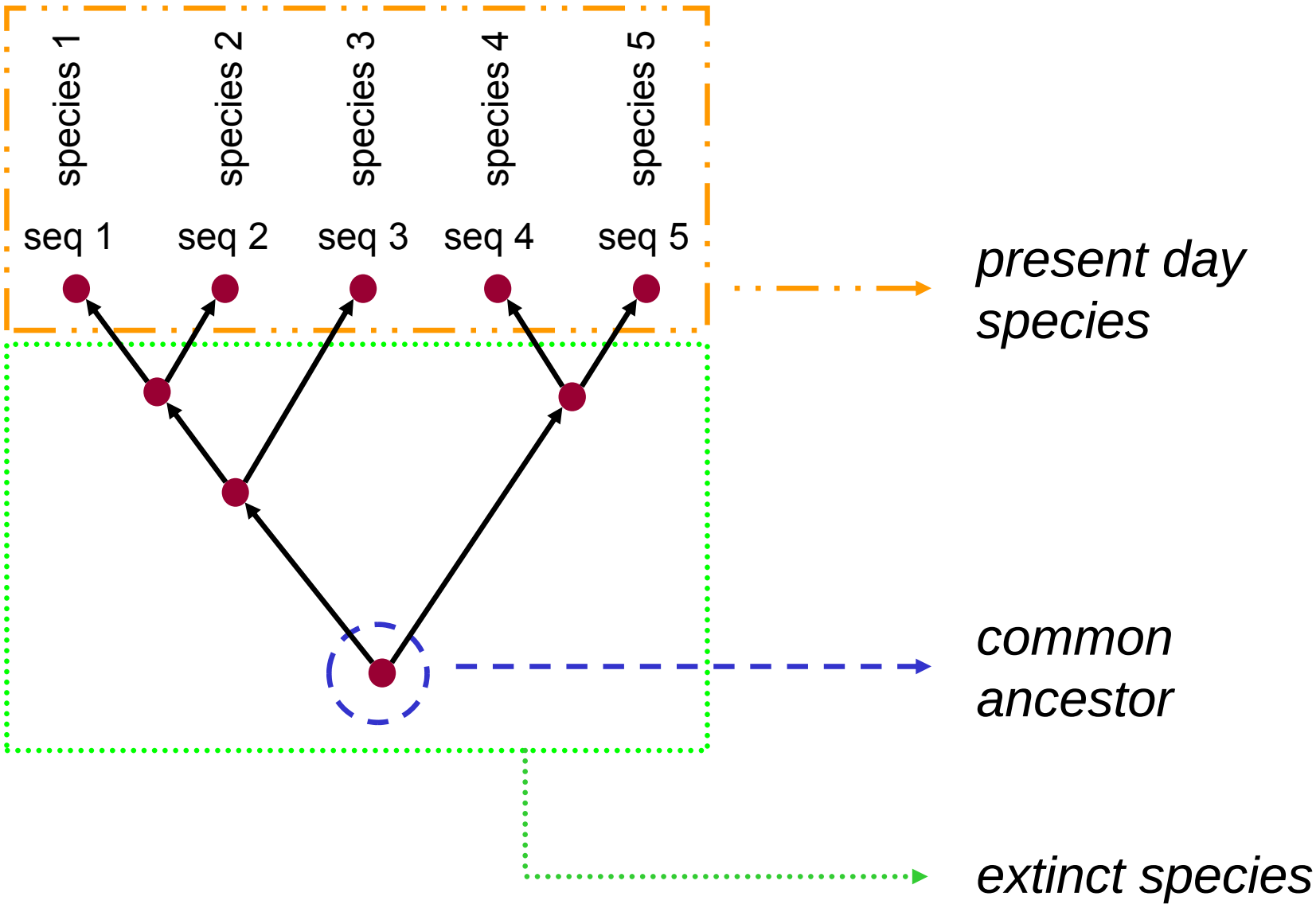
bifurcations



multifurcations

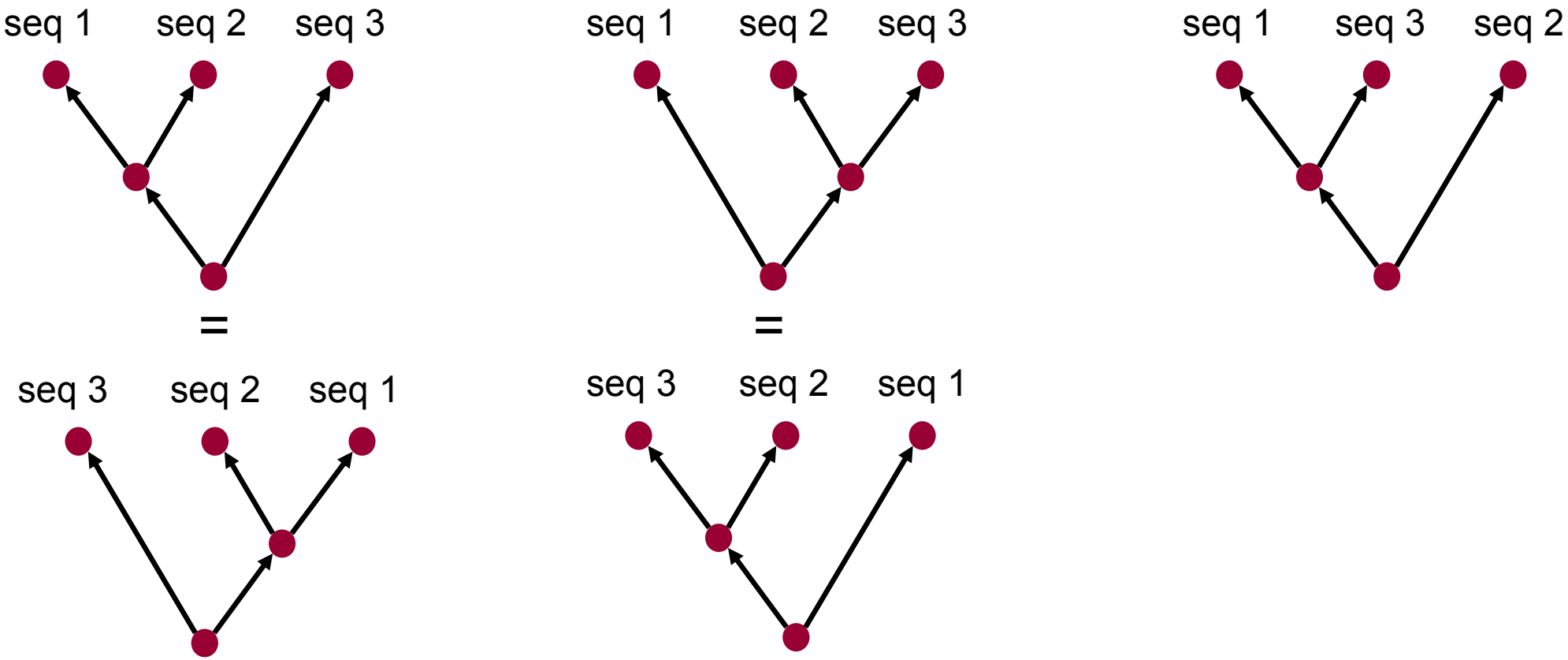
Intermezzo on graphs

In a phylogenetic tree:



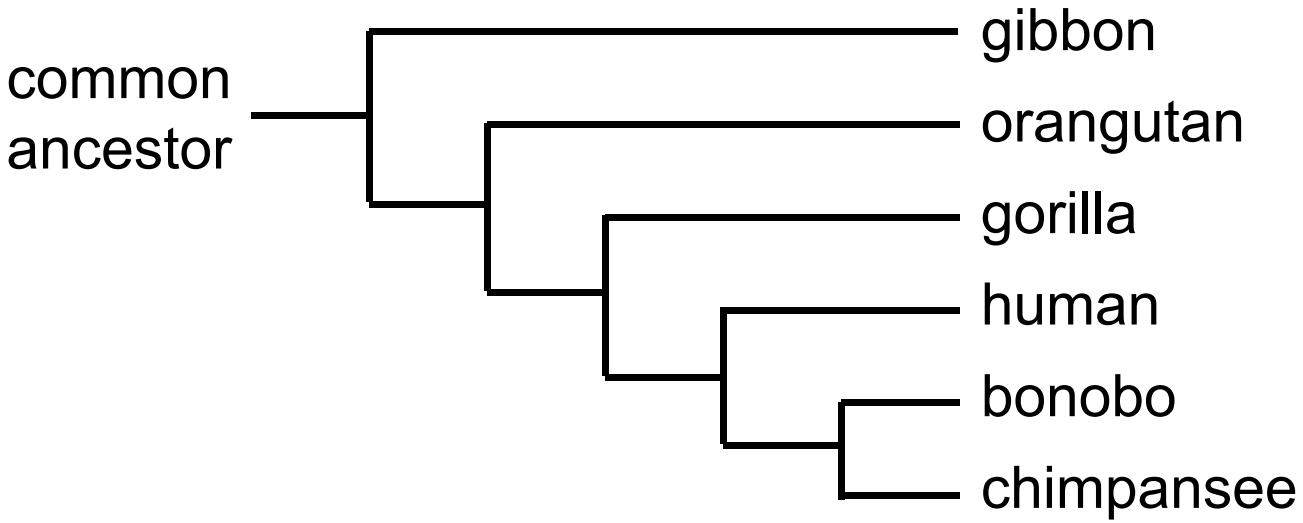
Intermezzo on graphs

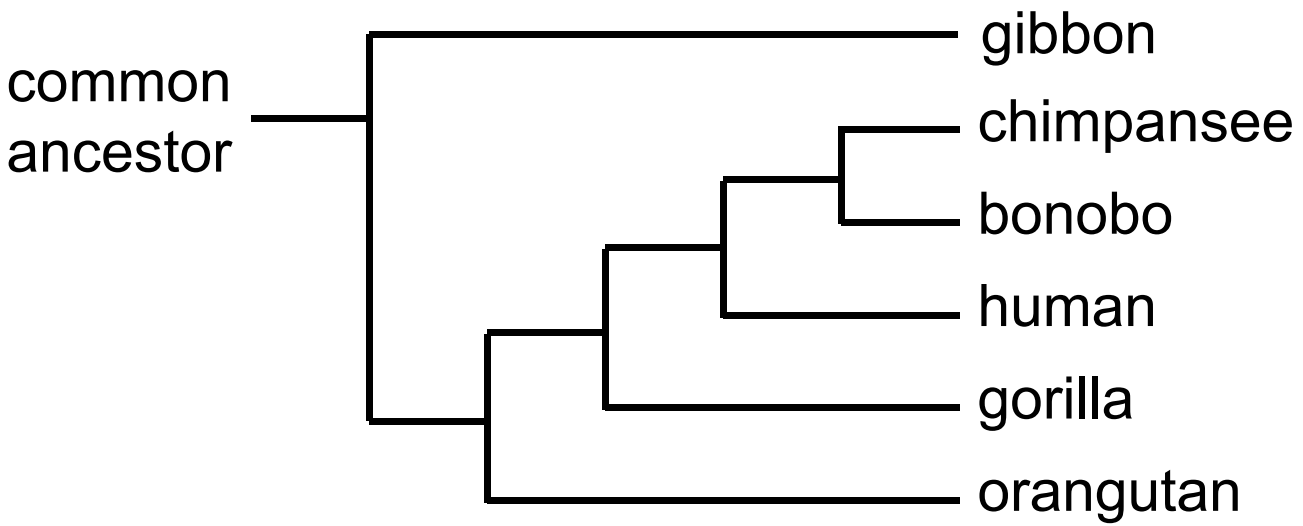
In case of three observed sequences, there are three different trees that connect the sequences:



Intermezzo on graphs

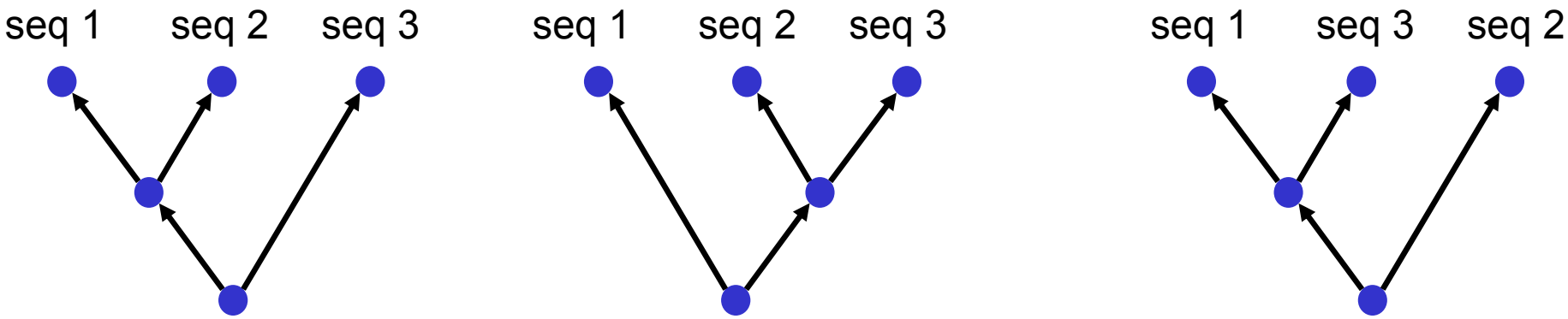
Hence, the following topologies are equivalent.





Intermezzo on graphs

If we have three observed sequences, we have three different *rooted* binary trees to connect the three sequences:



Intermezzo on graphs

The number of possible topologies is enormous. If the number of observed sequences equals n, the number of different *rooted* binary trees is:

$$(2n\text{-}3)! \,/\, 2^{n-2}\,(n\text{-}2)!$$

In case
$n = 2$: 1
$n = 3$: 3
$n = 4$: 15
$n = 5$: 105
.... : ....
$n = 10$: 34459425

And we have not even considered the branch length!

Intermezzo on graphs

The number of possible topologies is enormous. If the number of observed sequences equals n, the number of different *unrooted* binary trees is:

$$(2n\text{-}5)! \,/\, 2^{n-3}\,(n\text{-}3)!$$

In case
$n = 2$: 1
$n = 3$: 1
$n = 4$: 3
$n = 5$: 15
$n = 6$: 105
…. : ….
$n = 10$: 2027025
$n = 11$: 34459425

A model for DNA evolution

Models for DNA evolution

For an individual position the substitution process is modeled by a 1st order Markov process with the state space *S*={A, G, C, T}, now grouped by ***purines*** (A and G) and ***pyrimidines*** (C and T).

The considered models differ in their parametrization of **P**:

$$\mathbf{P} = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix}$$

Models for DNA evolution (JC69)

The ***Jukes-Cantor model*** is a DNA substitution model which assumes that:

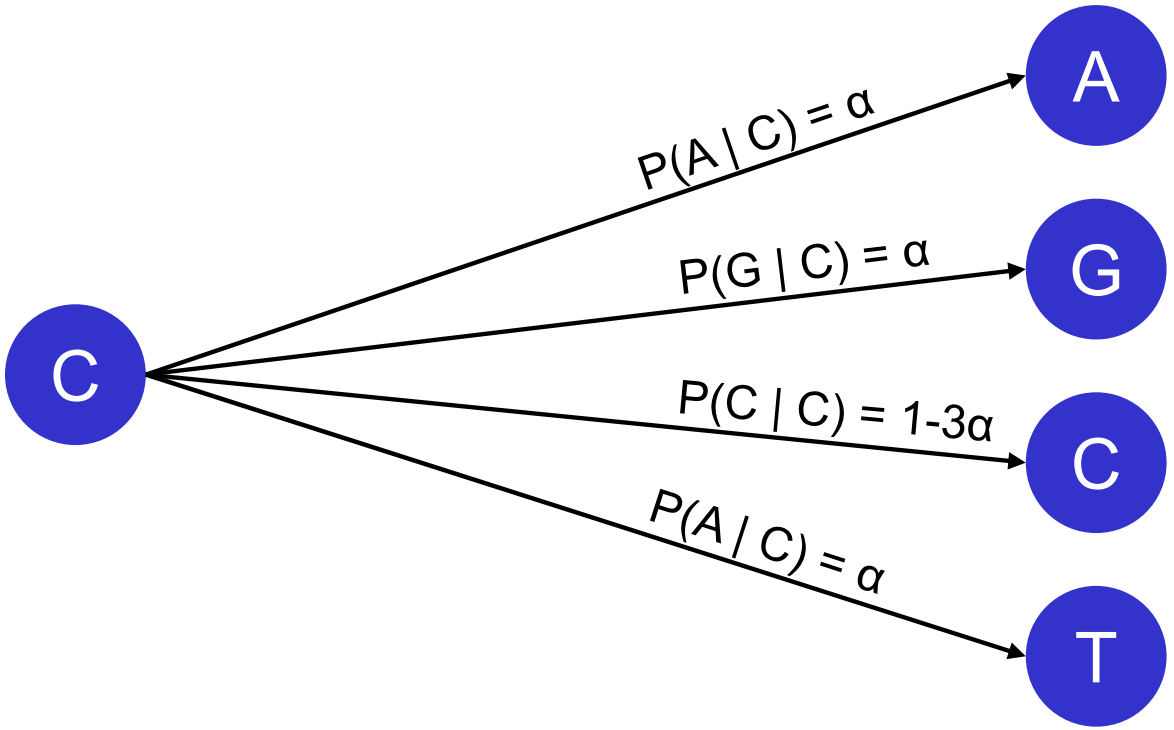
- each base in the sequence has an equal probability of being substituted.
- if a nucleotide substitution occurs, all other nucleotides have the same probability to replace it.

(As a results we expect an equal frequency of the four bases in the resulting DNA sequence.)

Models for DNA evolution (JC69)

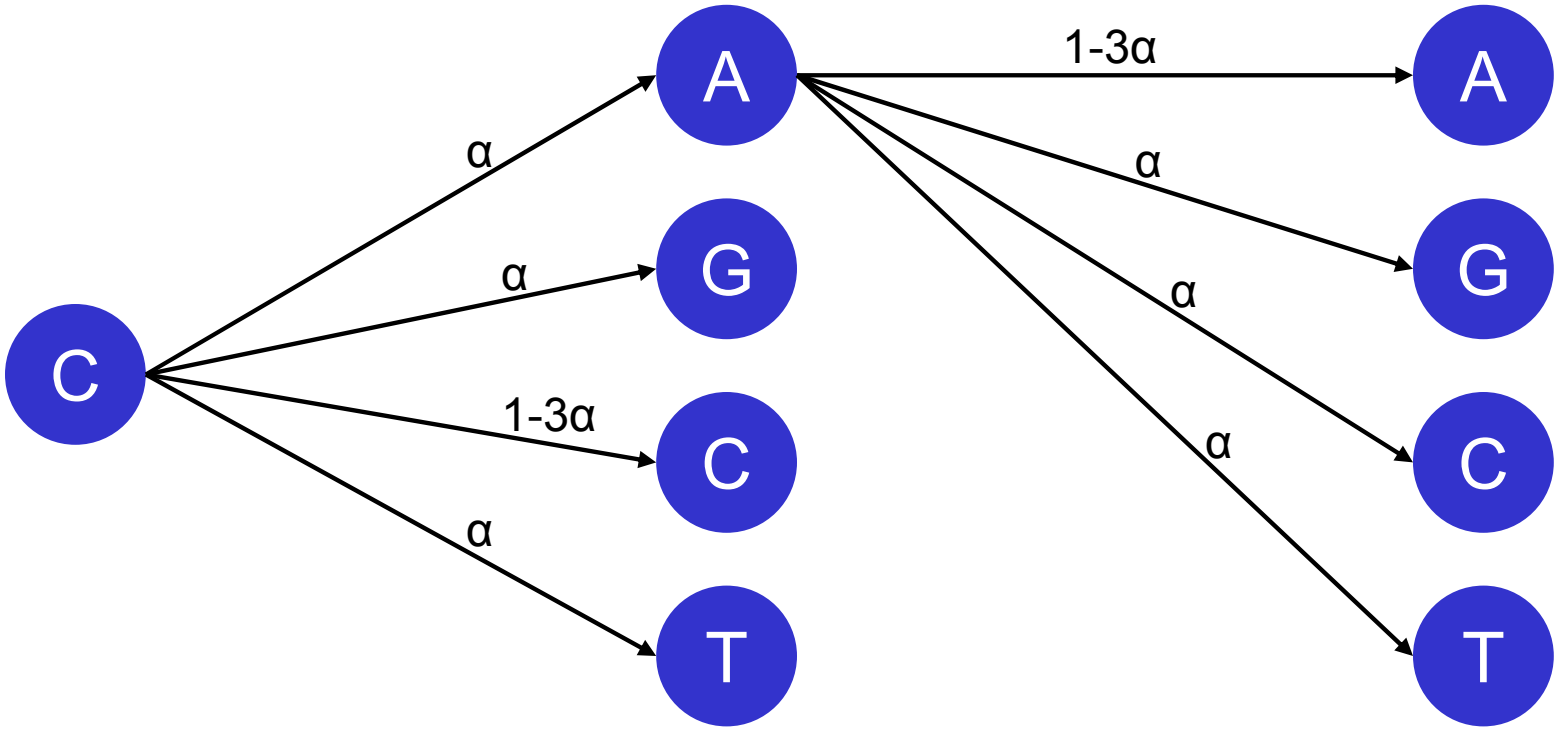
We have:

- probability α of C to substitute by A,
- probability α of C to substitute by G,
- probability α of C to substitute by T,
- probability 1-3α of C not to substitute.



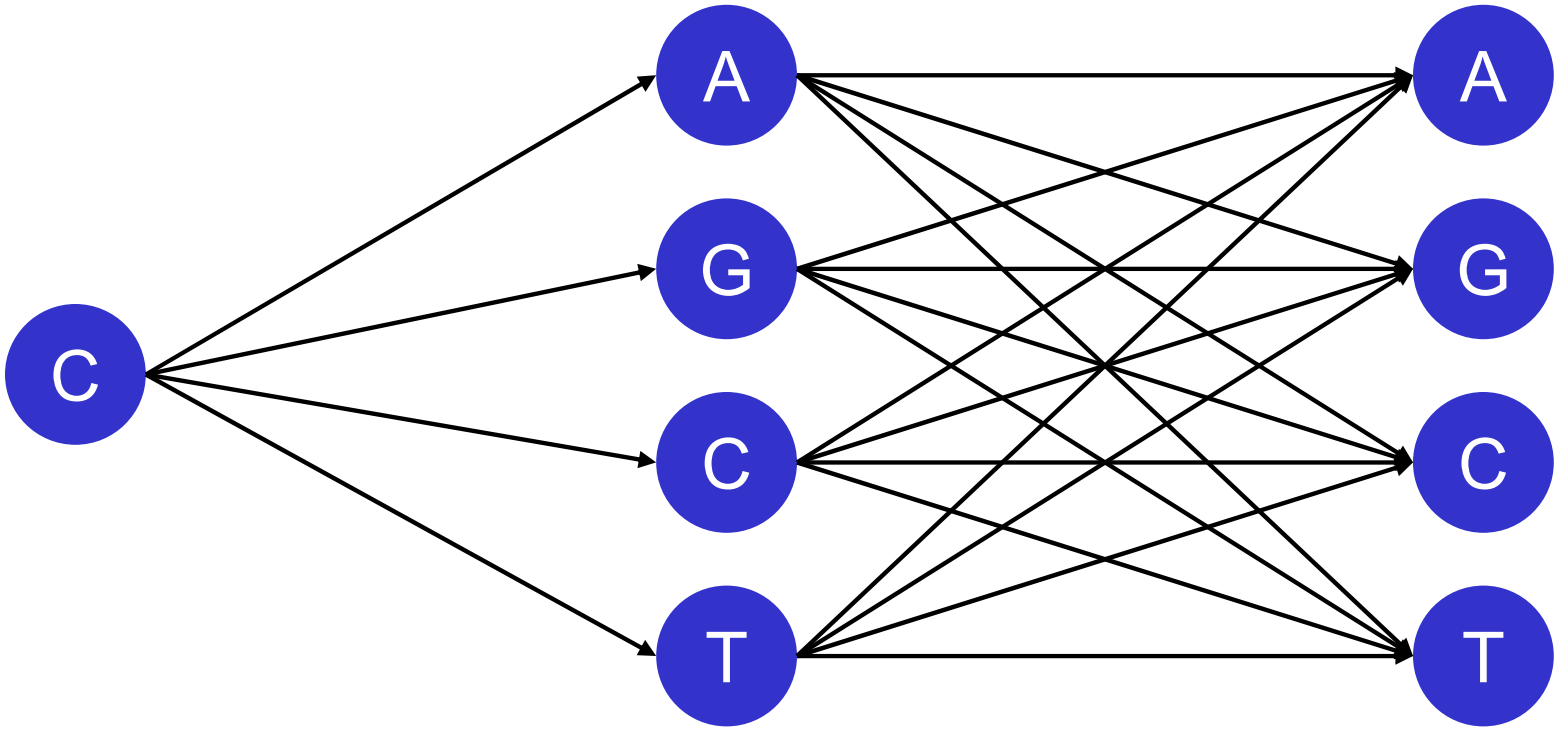
generation 0 *generation 1*

Models for DNA evolution (JC69)



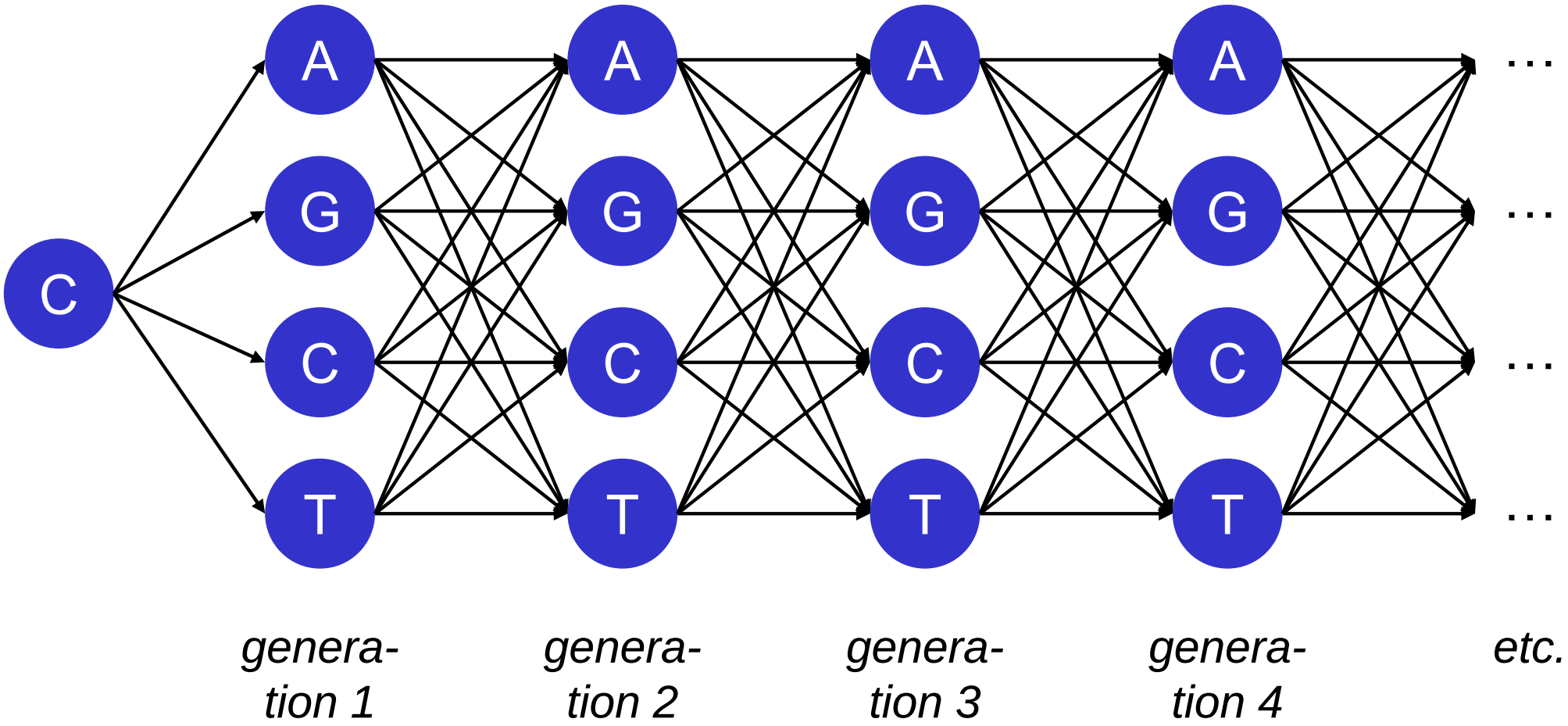
generation 0 *generation 1* *generation 2*

Models for DNA evolution (JC69)



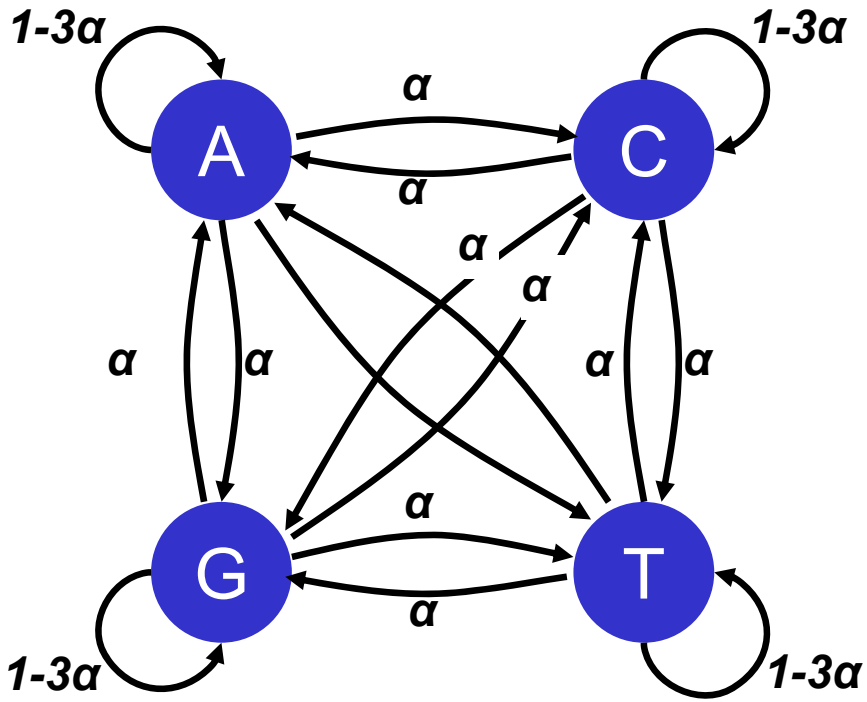
generation 0 *generation 1* *generation 2*

Models for DNA evolution (JC69)



Models for DNA evolution (JC69)

Over 1000s of generations (time homogeneity):

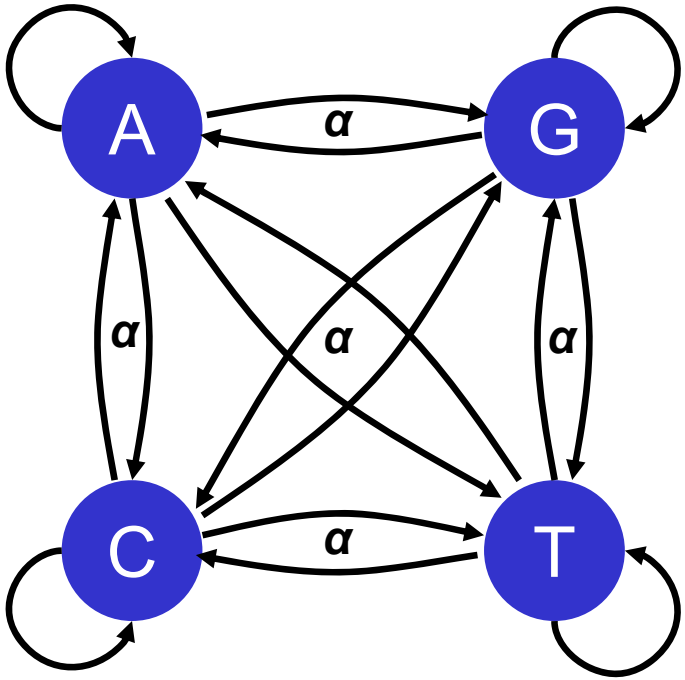


Models for DNA evolution (JC69)

The Jukes-Cantor transition matrix:

$$\mathbf{P} = \begin{pmatrix} 1-3\,\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\,\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\,\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\,\alpha \end{pmatrix}$$

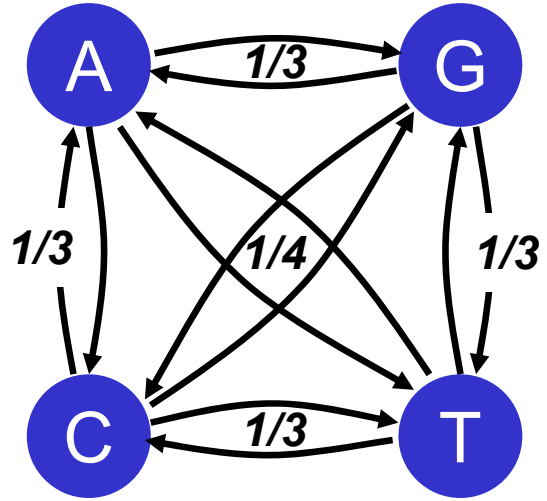
where
- $\alpha < \frac{1}{3}$,
- α depends on the step size.



Models for DNA evolution (JC69)

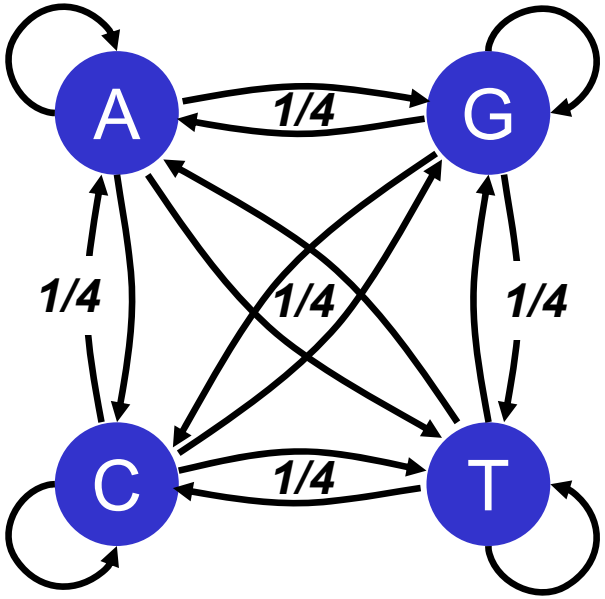
Always substitute if α=1/3:

$$\mathbf{P} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$



No Markov property if α=1/4:

$$\mathbf{P} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$



Models for DNA evolution (JC69)

Properties of the Jukes-Cantor model

The eigenvalues of **P**:

$$\lambda = 1, 1-4\alpha, 1-4\alpha, 1-4\alpha.$$

The stationary distribution corresponding to $\lambda=1$:

$$\boldsymbol{\varphi} = (¼, ¼, ¼, ¼)^T$$

Indeed, after enough generations all four states are equally likely. That is, all four nucleotides are equally likely to be the predominant nucleotide at the position under consideration.

Models for DNA evolution (JC69)

Properties of the Jukes-Cantor model

Its spectral decomposition:

$$\mathbf{P}^t = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \frac{1}{4}(1-4\alpha)^t \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

Models for DNA evolution (JC69)

Properties of the Jukes-Cantor model

Consider a stationary 1st order Markov chain with a Jukes-Cantor transition matrix. The probability of no substitution is given by:

$$P(X_t=A, X_{t-1}=A, \ldots, X_0=A) = P(A \mid A)^t P(A)$$
$$= (1-3\alpha)^t \varphi_A$$
$$= (1-3\alpha)^t / 4$$

Given that X_0=A, the probability that A will be the predominant nucleotide at time *t* is given by:

$$\tfrac{1}{4} + \tfrac{3}{4} (1-4\alpha)^t$$

Models for DNA evolution (JC69)

Properties of the Jukes-Cantor model

Now we know **P** and **φ**, and, hence, we can assess the reversibility of the Jukes-Cantor model by means of checking the detailed balance equations:

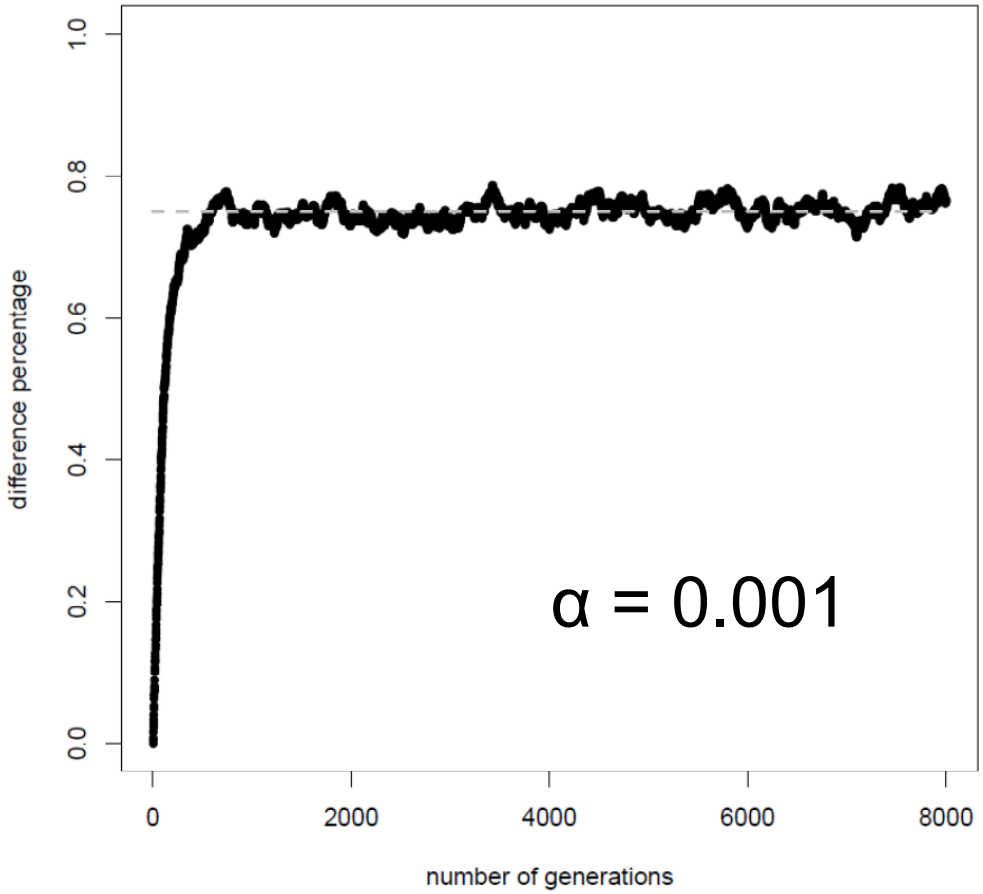
$$\varphi_i \, p_{ij} = \varphi_j \, p_{ji} \qquad \text{for all i and j.}$$

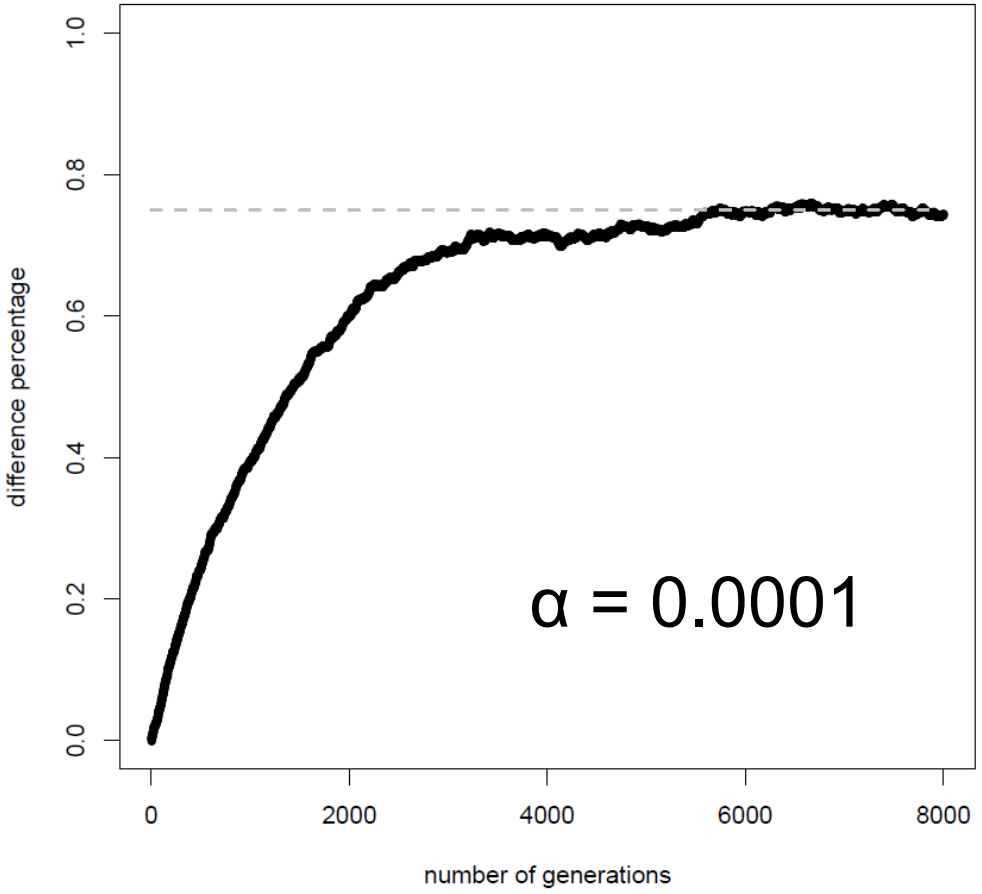
Recall

In order for the Jukes-Cantor model to link one species to another (via a common ancestor), the transition matrix **P** needs to be reversible.

Models for DNA evolution (JC69)

Proportion of site differences between two sequences in the Jukes-Canter model plotted against time (# generations), starting from the common ancestor.





Models for DNA evolution (K80)

The ***Kimura model*** is a generalization of the Jukes-Cantor model. It allows for different transition and transversion probabilities.

Similar to the Jukes-Cantor model, the Kimura is symmetrical. Therefore, after enough time it is equally likely for a base to be a purine or a pyrimidine.

Within the purine and pyrimidine categories there is complete symmetry between the nucleotides.

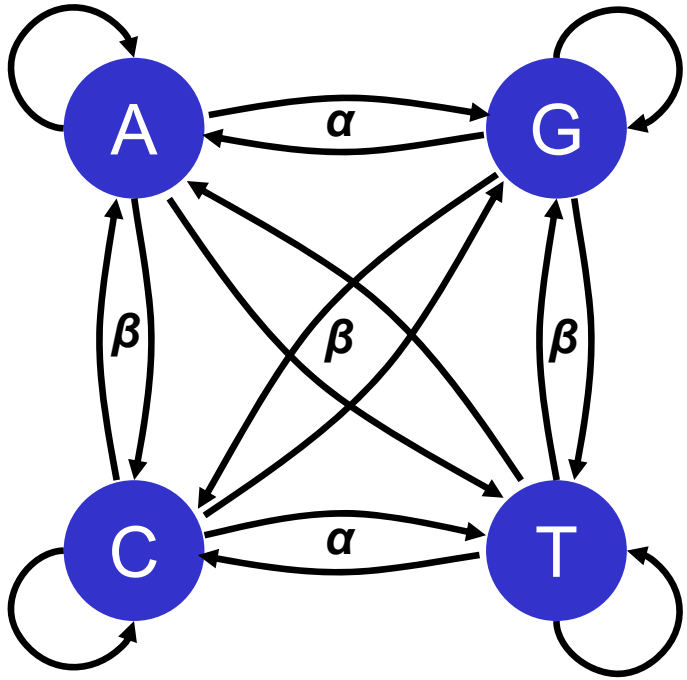
Models for DNA evolution (K80)

The Kimura transition matrix:

$$\mathbf{P} = \begin{pmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{pmatrix}$$

where
- $\alpha + 2\beta < 1$,
- α, β depend on the step size.

Take $\alpha = \beta$:
Jukes-Cantor.



Models for DNA evolution (K80)

Properties of the Kimura model

The eigenvalues of **P**:

$$\lambda = 1,\ 1-4\beta,\ 1-2(\alpha+\beta),\ 1-2(\alpha+\beta).$$

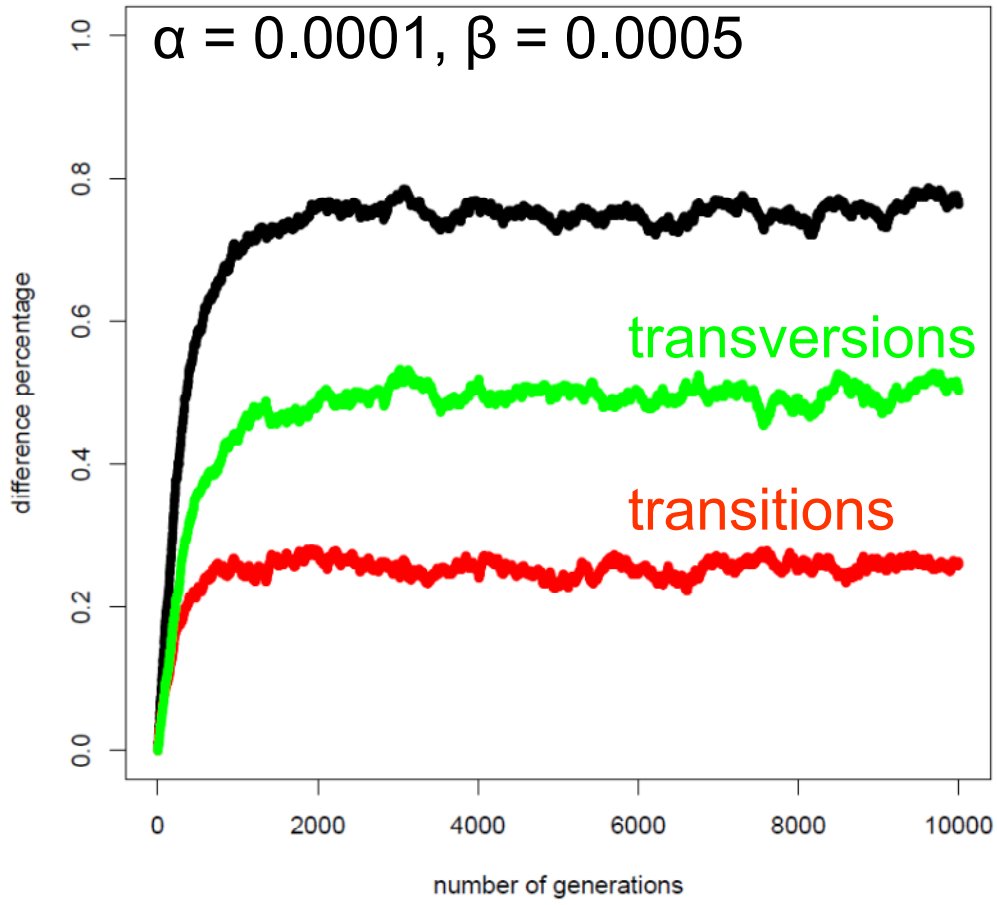
The stationary distribution corresponding to $\lambda=1$:

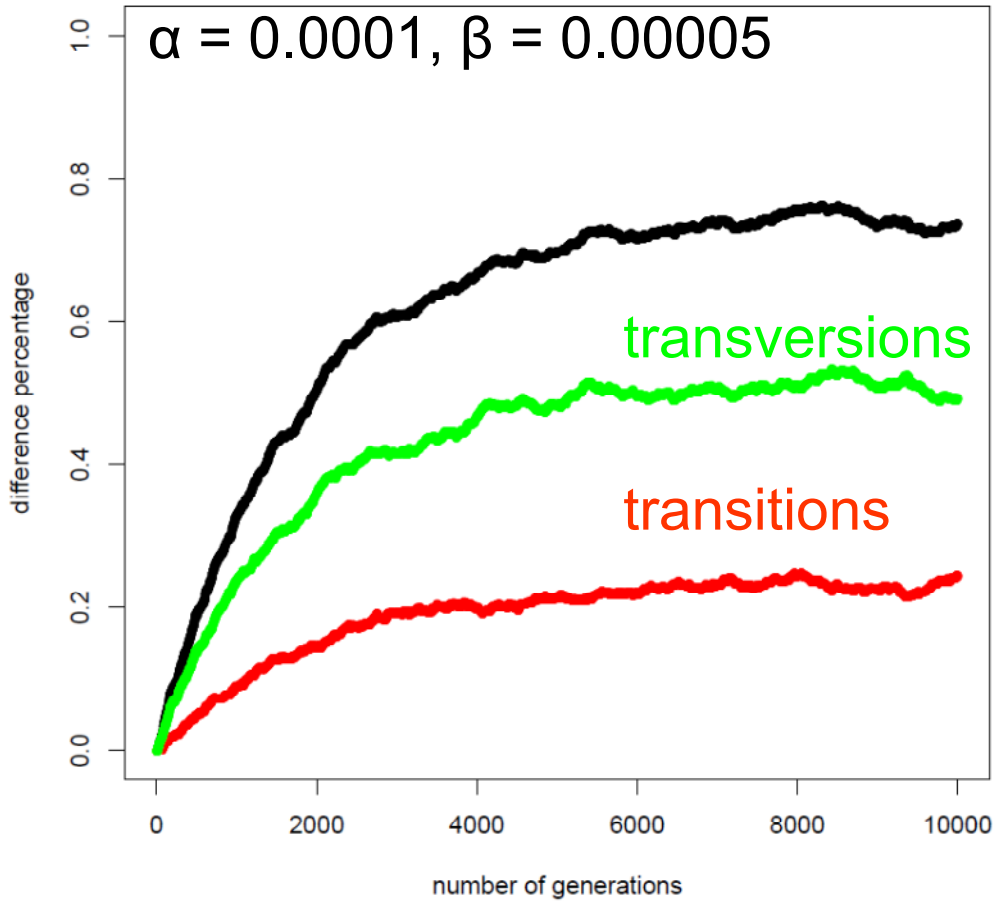
$$\boldsymbol{\varphi} = (¼, ¼, ¼, ¼)^T$$

The Kimura model is reversible (**P** is symmetric and $\boldsymbol{\varphi}$ uniform).

Models for DNA evolution (K80)

Proportion of site differences between two sequences in the Kimura model plotted against time (# generations), starting from the common ancestor.

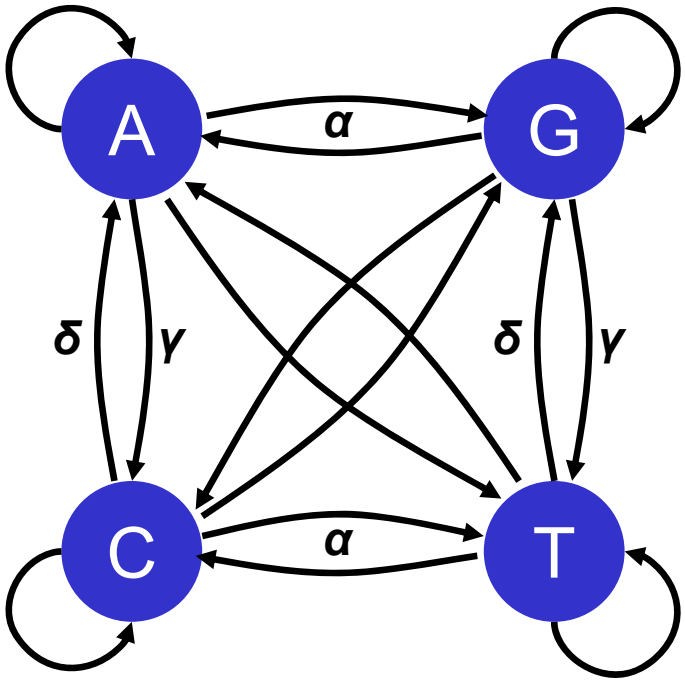


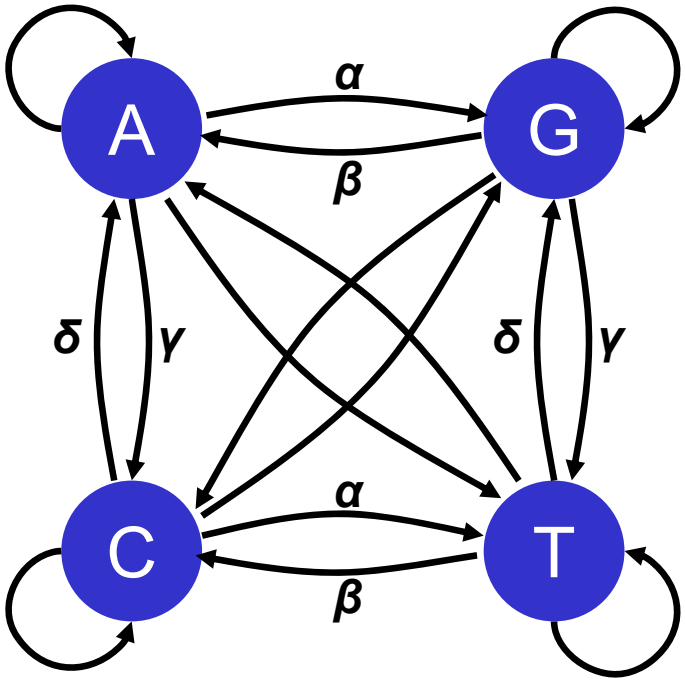


Models for DNA evolution (K80)

The Kimura model has been generalized to allow, e.g.:

- The transition probability to differ from the transversion probability.
- Different within-transition and within-transversion substitution probabilities.





Models for DNA evolution (F81)

The ***Felsenstein model*** is also a generalization of the Jukes-Cantor model. It relaxes the (implicit) assumption of the JC and Kimura model, both having a uniform stationary distribution.

In the Felsenstein model the probability of substitution of any nucleotide by another is proportional to the stationary probability of the substituting nucleotide.

The Felsenstein model does not distinguish between purines and pyrimidines.

Models for DNA evolution (F81)

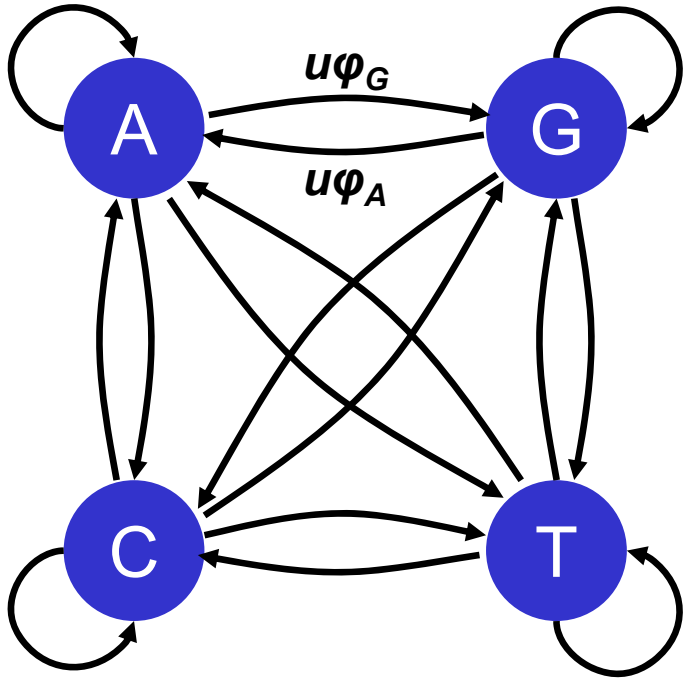
The Felsenstein transition matrix:

$$\mathbf{P} = \begin{pmatrix} 1-u+u\varphi_A & u\varphi_G & u\varphi_C & u\varphi_T \\ u\varphi_A & 1-u+u\varphi_G & u\varphi_C & u\varphi_T \\ u\varphi_A & u\varphi_G & 1-u+u\varphi_C & u\varphi_T \\ u\varphi_A & u\varphi_G & u\varphi_C & 1-u+u\varphi_T \end{pmatrix}$$

where

- $\varphi_A + \varphi_G + \varphi_C + \varphi_T = 1$.
- u a model parameter.

Take $\varphi_A=\varphi_G=\varphi_C=\varphi_T=¼$: Jukes-Cantor.



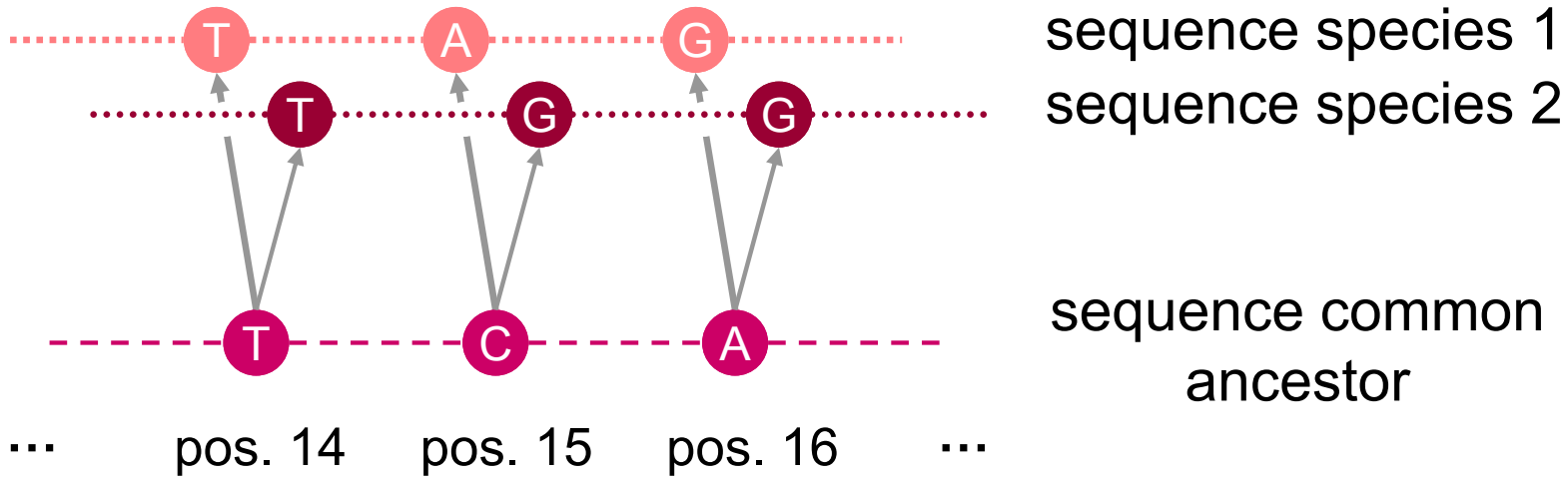
The likelihood: a simple example

The likelihood: an example

Consider two homologous sequences sampled from two different species (with a common ancestor):

```
species 1    : AATTGCGTAGCTAGATCGCTCGCTA
species 2    : AATTGCGTAGCTAGGTCGCTCGCTA
```

15th base



What is the likelihood of observing these two sequences?

The likelihood: an example

Let

X denote the sequence data of both species, and

X_{ij} denote the nucleotide at position j=1, …,25 of species i.

The likelihood for the Jukes-Cantor model is then:

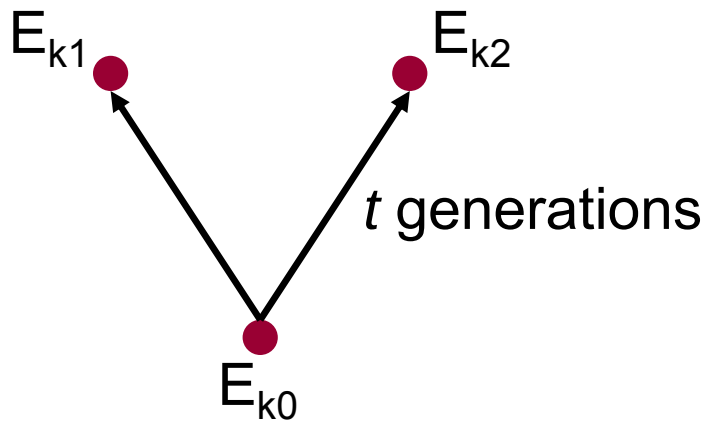
$$L(\mathbf{X}) \quad = \quad P(\mathbf{X})$$

which, assuming sites evolve independently, factorizes to

$$= \quad \prod_{j=1}^{25} P((X_{1j}, X_{2j}))$$

The likelihood: an example

Assuming $(X_{1j}, X_{2j}) = (E_{k1}, E_{k2})$ and that the species have evolved separately *t* generations since the common ancestor, then:



$$P((X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2}))$$

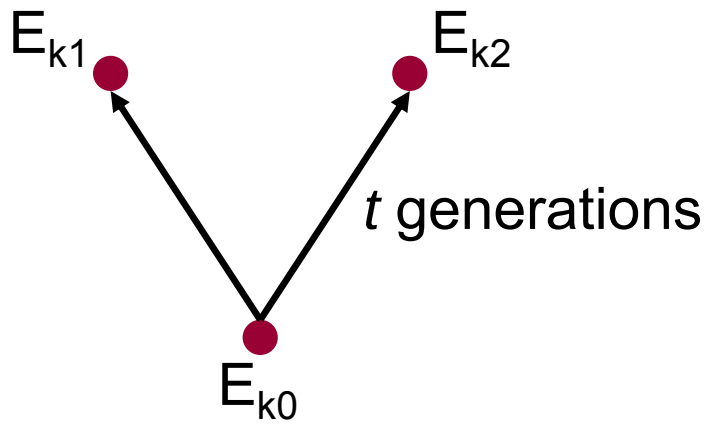
$$= \sum_{k_0 \in \mathcal{S}} \pi_{k_0} \left(\mathbf{P}^{(t)}\right)_{k_0,k_1} \left(\mathbf{P}^{(t)}\right)_{k_0,k_2}$$

$$= \sum_{k_0 \in \mathcal{S}} \pi_{k_0} \left(\mathbf{P}^{t}\right)_{k_0,k_1} \left(\mathbf{P}^{t}\right)_{k_0,k_2}$$

The likelihood: an example

Note

The life time of a generation may differ between the two present day organisms. In particular, if an evolutionary long time has passed since the common ancestor.



The solution is to use the actual time passed since the common ancestor. Modeling this requires continuous time Markov chains. Not treated here.

The likelihood: an example

In order to write down the likelihood, recall

- The Chapman-Kolmogorov equations:

$$p_{k_1 k_2}^{(t_1+t_2)} = \sum_{k_0 \in \mathcal{S}} p_{k_1 k_0}^{(t_1)} \, p_{k_0 k_2}^{(t_2)}$$

- The reversibility of the Jukes-Cantor model:

$$\pi_{k_1} \, p_{k_1 k_2} = \pi_{k_2} \, p_{k_2 k_1}$$

- The symmetry of the JC transition matrix **P**.

- Combining the last two yields:

$$\pi_{k_1} \, p_{k_1 k_2}^{(t)} = \pi_{k_2} \, p_{k_2 k_1}^{(t)}$$

The likelihood: an example

$$
\begin{aligned}
L(\mathbf{X}) &= \prod_{j=1}^{25} P((X_{1j}, X_{2j})) \\
&= \prod_{j=1}^{25} \prod_{k_1, k_2 \in \mathcal{S}} \left[P((X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2})) \right]^{I_{\{(X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2})\}}}
\end{aligned}
$$

The likelihood: an example

$$
\begin{aligned}
L(\mathbf{X}) &= \prod_{j=1}^{25} P((X_{1j}, X_{2j})) \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[P((X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2}))\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_0} \left(\mathbf{P}^t\right)_{k_0,k_1} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}
\end{aligned}
$$

1) substitute previously derived expression for probability of individual observation

2) substitution rates are the same for all sites

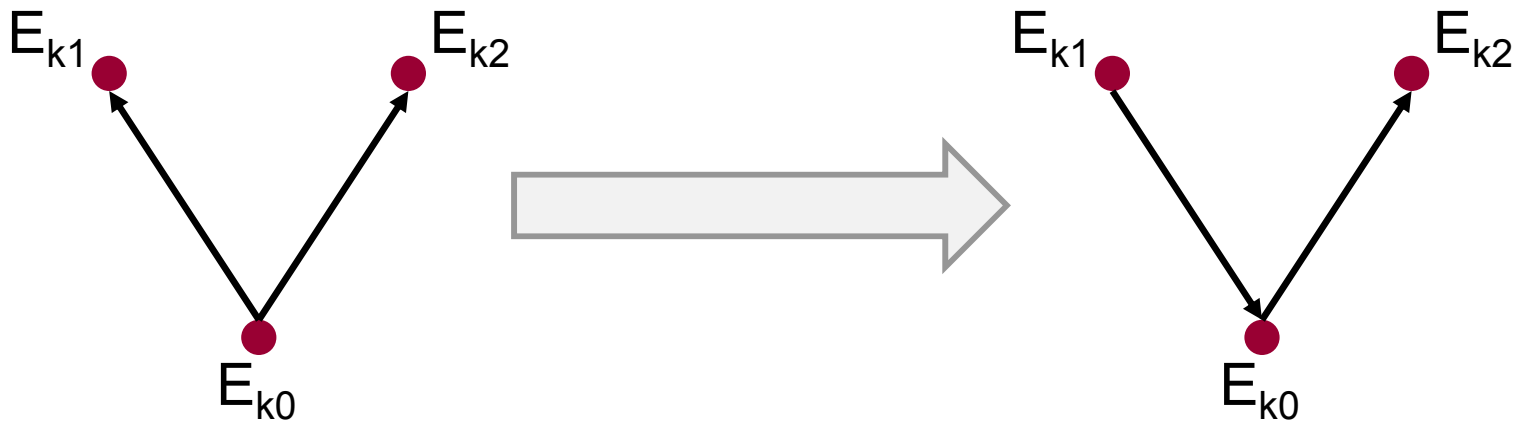
The likelihood: an example

$$
\begin{aligned}
L(\mathbf{X}) &= \prod_{j=1}^{25} P((X_{1j}, X_{2j})) \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[P((X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2}))\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_0} \left(\mathbf{P}^t\right)_{k_0,k_1} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_1} \left(\mathbf{P}^t\right)_{k_1,k_0} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}
\end{aligned}
$$

use the time reversibility of the JC model

The likelihood: an example

By using the time reversibility of the JC model, we have reversed one arrow of the phylogenetic tree:



In the formulea:

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_0} \left(\mathbf{P}^t\right)_{k_0,k_1} \left(\mathbf{P}^t\right)_{k_0,k_2} \right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_1} \left(\mathbf{P}^t\right)_{k_1,k_0} \left(\mathbf{P}^t\right)_{k_0,k_2} \right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

The likelihood: an example

$$
\begin{aligned}
L(\mathbf{X}) &= \prod_{j=1}^{25} P((X_{1j}, X_{2j})) \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[P((X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2}))\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_0} \left(\mathbf{P}^t\right)_{k_0,k_1} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\sum_{k_0 \in \mathcal{S}} \pi_{k_1} \left(\mathbf{P}^t\right)_{k_1,k_0} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\pi_{k_1} \sum_{k_0 \in \mathcal{S}} \left(\mathbf{P}^t\right)_{k_1,k_0} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}
\end{aligned}
$$

bringing π_{k1} outside the sum

The likelihood: an example

$$L(\mathbf{X}) = \prod_{j=1}^{25} P((X_{1j}, X_{2j}))$$

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[P((X_{1j}, X_{2j}) = (E_{k_1}, E_{k_2}))\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

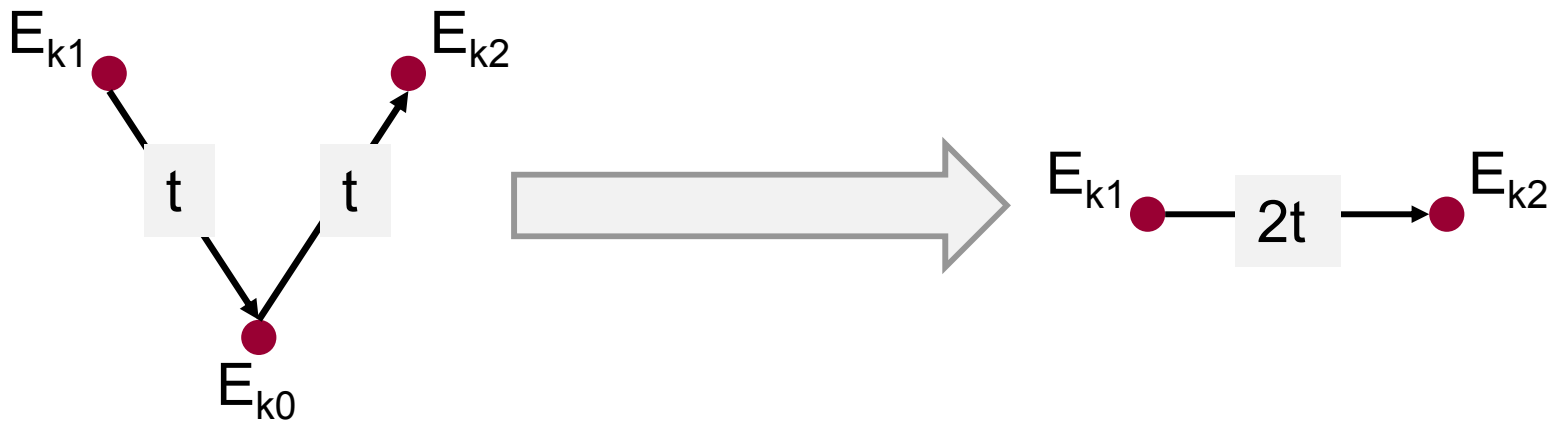
use the Chapman-Kolmogorov equations

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\pi_{k_1} \sum_{k_0 \in \mathcal{S}} \left(\mathbf{P}^t\right)_{k_1,k_0} \left(\mathbf{P}^t\right)_{k_0,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\pi_{k_1} \left(\mathbf{P}^{2t}\right)_{k_1,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

The likelihood: an example

By using the Chapman-Kolmogorov equations, we removed the common ancestor from the phylogenetic tree:



In the formulea:

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\pi_{k_1} \sum_{k_0 \in \mathcal{S}} \left(\mathbf{P}^t\right)_{k_1,k_0} \left(\mathbf{P}^t\right)_{k_0,k_2} \right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

$$= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\pi_{k_1} \left(\mathbf{P}^{2t}\right)_{k_1,k_2} \right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}}$$

The likelihood: an example

The likelihood can be further simplified, when exploiting the spectral decomposition the JC *t*-step transition matrix:

$$\mathbf{P}^t = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \frac{1}{4}(1-4\alpha)^t \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

The likelihood: an example

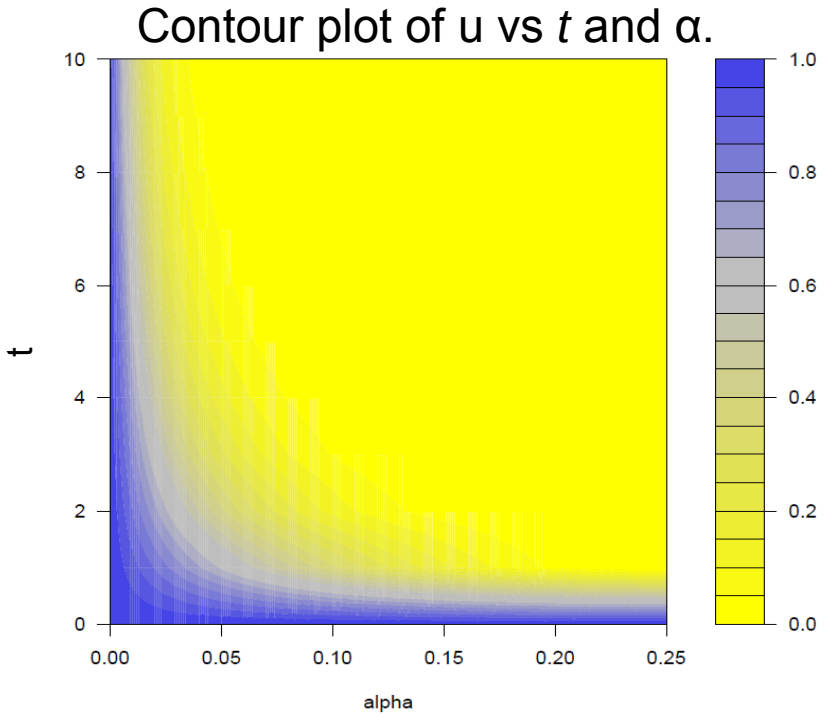
Finally, we have:

$$
\begin{aligned}
L(\mathbf{X}) &= \prod_{j=1}^{25} \prod_{k_1,k_2 \in \mathcal{S}} \left[\pi_{k_1} \left(\mathbf{P}^{2t}\right)_{k_1,k_2}\right]^{I_{\{(X_{1j},X_{2j})=(E_{k_1},E_{k_2})\}}} \\
&= \prod_{j=1}^{25} \frac{1}{4}\left[\frac{1}{4} + \frac{3}{4}(1-4\,\alpha)^{2t}\right]^{I_{\{X_{1j}=X_{2j}\}}} \\
&\qquad \times \left[\frac{1}{4} - \frac{1}{4}(1-4\,\alpha)^{2t}\right]^{I_{\{X_{1j}\neq X_{2j}\}}}
\end{aligned}
$$

where we have used that the stationary distribution of the JC model is uniform.

The likelihood: an example

From the likelihood, it is clear either *t* or α is identifiable not both. Many combinations (α, *t*) yield the same likelihood.



In the absence of external evidence of α, we replace:

$$u = (1 - 4\alpha)^{2t}$$

and obtain:

$$L(\mathbf{X}) = \prod_{j=1}^{25} \frac{1}{4}\left[\frac{1}{4} + \frac{3}{4}u\right]^{I_{\{X_{1j}=X_{2j}\}}} \left[\frac{1}{4} - \frac{1}{4}u\right]^{I_{\{X_{1j}\neq X_{2j}\}}}$$

The likelihood: an example

To estimate *u*, maximize the log-likehood:

$$\log(\frac{1}{4}+\frac{3}{4}u)\sum_{j=1}^{25} I_{\{X_{1j}=X_{2j}\}} + \log(\frac{1}{4}-\frac{1}{4}u)\sum_{j=1}^{25} I_{\{X_{1j}\neq X_{2j}\}}$$

This yields:

$$\hat{u} = \frac{3\sum_{j=1}^{25} I_{\{X_{1j}=X_{2j}\}} - \sum_{j=1}^{25} I_{\{X_{1j}\neq X_{2j}\}}}{3\sum_{j=1}^{25} I_{\{X_{1j}=X_{2j}\}} + 3\sum_{j=1}^{25} I_{\{X_{1j}\neq X_{2j}\}}}$$

Check that this is indeed a maximum.

The likelihood: an example

For our two-species example, with sequences:

```
species 1    : AATTGCGTAGCTAGATCGCTCGCTA
species 2    : AATTGCGTAGCTAGGTCGCTCGCTA
```

the ML estimate equals:

$$\hat{u} = \frac{3 \times 24 - 1}{3 \times 24 + 3 \times 1} = \frac{71}{75}$$

Assuming the substitution rate (α) is 1 in a million, we get:

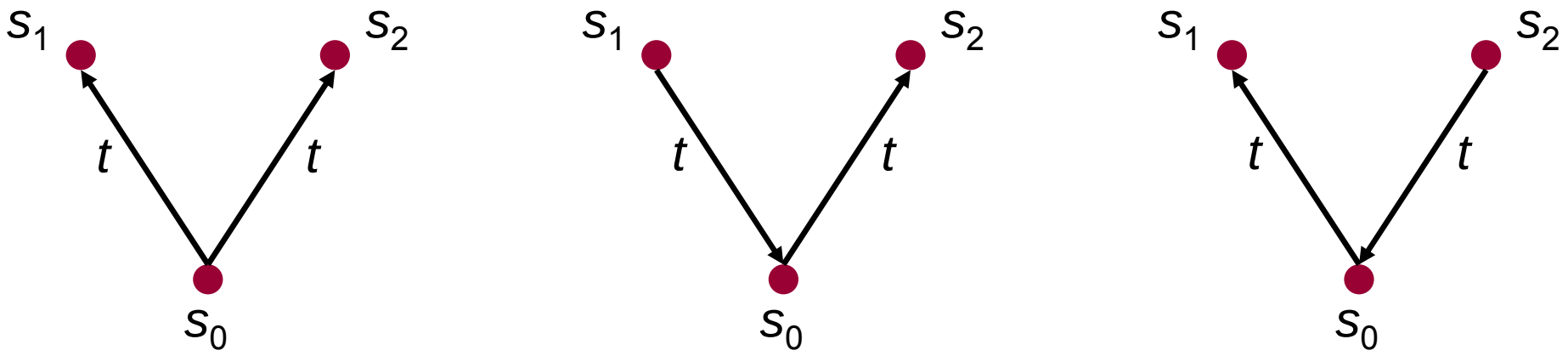
$$\hat{t} = \log(71/75)/\log(1 - \alpha/4) \approx 219233$$

This estimates suggests that the two species shared a common ancestor 219233 generations ago.

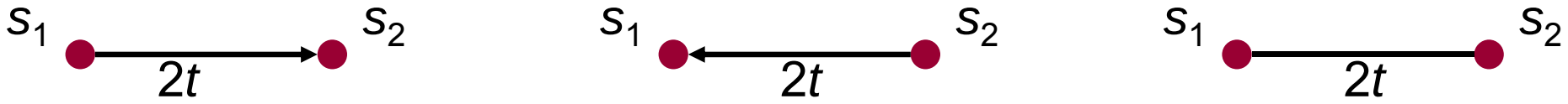
The pulley principle

The pulley principle

Due to reversibility, likelihood of trees below are equivalent:



But even to:

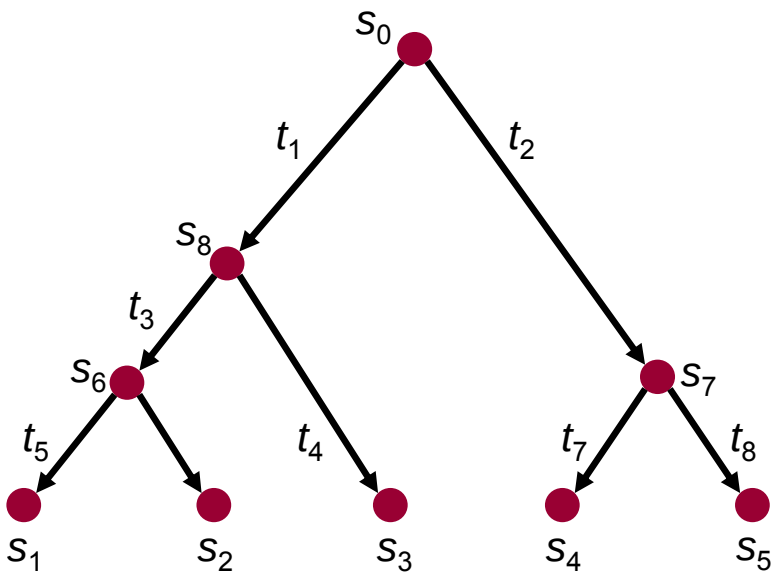


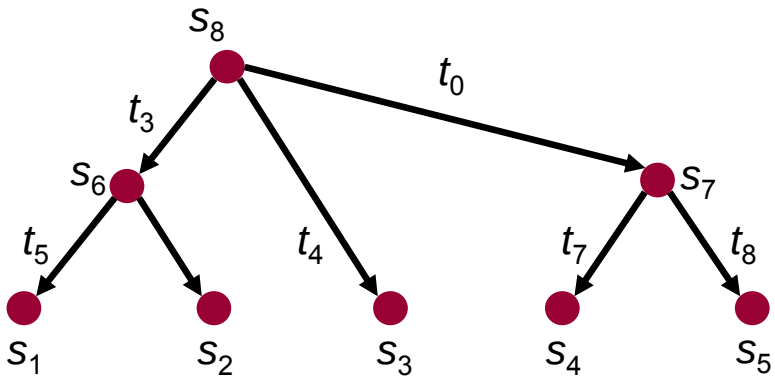
Pulley principle

The root node may be moved to any of the nodes without changing the likelihood.

The pulley principle

Due to the pulley principle, the likelihood of the following trees is equivalent:



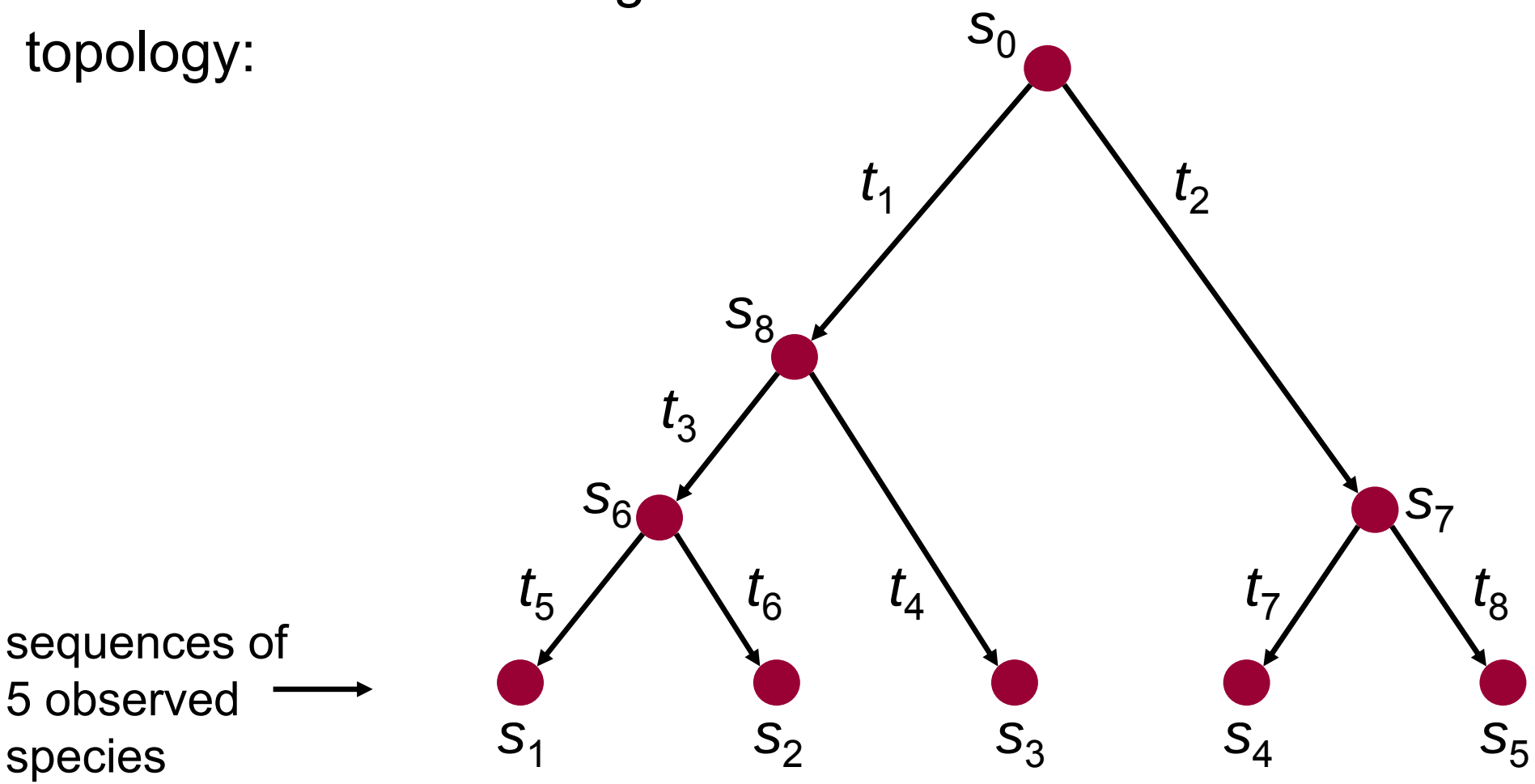


The likelihood: another example

The likelihood: another example

Consider the case where:

- DNA sequences from (say) 5 species are available.
- the sequences consist of (say) 25 bases.
- we assume the following topology:



The likelihood: another example

Step 1

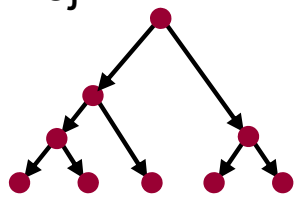
Assume the 25 sites evolve independently. The probability of evolution from (say) node / species s_7 to s_5 then becomes:

$$P(\mathbf{X}_7 \xrightarrow{t_8} \mathbf{X}_5) = \prod_{j=1}^{25} P(X_{7j} \xrightarrow{t_8} X_{5j})$$

where

$$P(X_{7j} \xrightarrow{t_8} X_{5j})$$

denotes the (conditional) probability of X_{7j} evolving to X_{5j} in t_8 generations.



The likelihood: another example

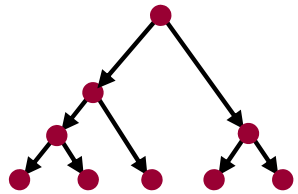
Step 1

Recall: the probability of the nucleotide at site *j* changing from X_{7j} in sequence 7 to X_{5j} in sequence 5 in t_8 generations, denoted by:

$$P(X_{7j} \xrightarrow{t_8} X_{5j})$$

is given by a multiple of the transition matrix of the evolutionary model of choice. Hence,

$$P(X_{7j} \xrightarrow{t_8} X_{5j}) \quad = \quad (\mathbf{P}^{t_8})_{X_{7j},X_{5j}}$$

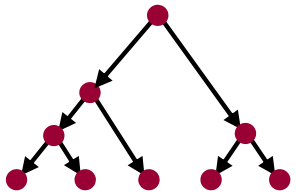


The likelihood: another example

Step 2

If the sequence of all nodes / species (s_0, …, s_8) are known, the likelihood is given by:

$$
\begin{aligned}
L \quad = \quad & P(\mathbf{X}_0) \\
& P(\mathbf{X}_0 \xrightarrow{t_2} \mathbf{X}_7)\, P(\mathbf{X}_0 \xrightarrow{t_1} \mathbf{X}_8) \\
& P(\mathbf{X}_7 \xrightarrow{t_7} \mathbf{X}_4)\, P(\mathbf{X}_7 \xrightarrow{t_8} \mathbf{X}_5) \\
& P(\mathbf{X}_8 \xrightarrow{t_4} \mathbf{X}_3)\, P(\mathbf{X}_8 \xrightarrow{t_3} \mathbf{X}_6) \\
& P(\mathbf{X}_6 \xrightarrow{t_5} \mathbf{X}_1)\, P(\mathbf{X}_6 \xrightarrow{t_6} \mathbf{X}_2)
\end{aligned}
$$

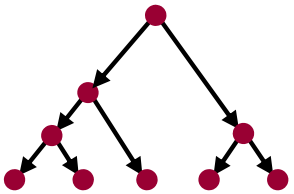


The likelihood: another example

Step 3

Since only the sequences of nodes n_1, …, n_5 are observed, the likelihood has to be summed over all possible sequences for the unobserved nodes:

$$
\begin{aligned}
L \;=\; \sum_{\mathbf{X}_0}\sum_{\mathbf{X}_6}\sum_{\mathbf{X}_7}\sum_{\mathbf{X}_8} \Big[& P(\mathbf{X}_0) \\
& P(\mathbf{X}_0 \xrightarrow{t_2} \mathbf{X}_7)\, P(\mathbf{X}_0 \xrightarrow{t_1} \mathbf{X}_8) \\
& P(\mathbf{X}_7 \xrightarrow{t_7} \mathbf{X}_4)\, P(\mathbf{X}_7 \xrightarrow{t_8} \mathbf{X}_5) \\
& P(\mathbf{X}_8 \xrightarrow{t_4} \mathbf{X}_3)\, P(\mathbf{X}_8 \xrightarrow{t_3} \mathbf{X}_6) \\
& P(\mathbf{X}_6 \xrightarrow{t_5} \mathbf{X}_1)\, P(\mathbf{X}_6 \xrightarrow{t_6} \mathbf{X}_2) \Big]
\end{aligned}
$$



The likelihood: another example

Step 3 (computational efficiency)

This likelihood can be calculated by exploiting the conditional likelihoods, e.g.:

$$L((X_{4j}, X_{5j}) \,|\, X_{7j}) \;=\; \Big[\sum_{X_{4j}} P(X_{7j} \xrightarrow{t_7} X_{4j})\, L(X_{4j})\Big] \Big[\sum_{X_{5j}} P(X_{7j} \xrightarrow{t_8} X_{5j})\, L(X_{5j})\Big]$$

which yields:

$$\begin{aligned} L \;=\; \sum_{\mathbf{X}_0} P(\mathbf{X}_0) \Big[\Big[\sum_{\mathbf{X}_7} P(\mathbf{X}_0 \xrightarrow{t_2} \mathbf{X}_7)\, P(\mathbf{X}_7 \xrightarrow{t_7} \mathbf{X}_4)\, P(\mathbf{X}_7 \xrightarrow{t_8} \mathbf{X}_5)\Big] \\ \times \Big[\sum_{\mathbf{X}_8} P(\mathbf{X}_0 \xrightarrow{t_1} \mathbf{X}_8)\, P(\mathbf{X}_8 \xrightarrow{t_4} \mathbf{X}_3)\, P(\mathbf{X}_8 \xrightarrow{t_3} \mathbf{X}_6) \\ \times \Big[\sum_{\mathbf{X}_6} P(\mathbf{X}_6 \xrightarrow{t_5} \mathbf{X}_1)\, P(\mathbf{X}_6 \xrightarrow{t_6} \mathbf{X}_2)\Big]\Big]\Big] \end{aligned}$$

The likelihood: another example

Step 3 (computational efficiency)

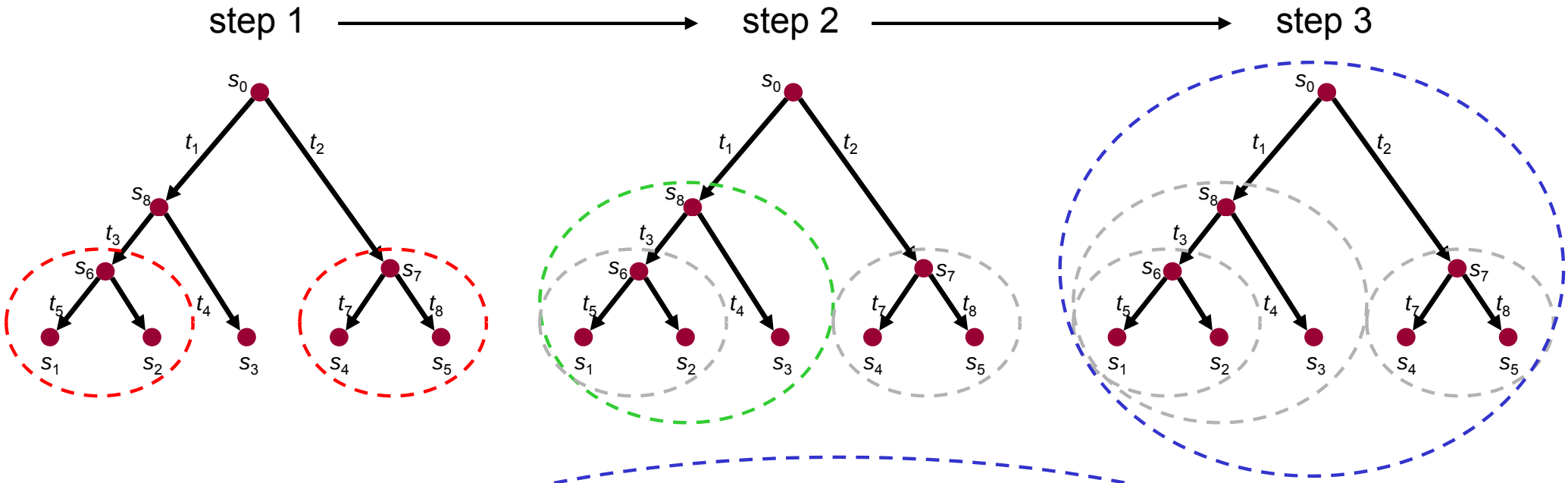
Without the exploitation of the conditional likelihood, calculation of the likelihood required the evaluation of 4^4=256 combinations (4 hidden nodes, 4 nucleotides).

In the reformulation on the previous slide, the likelihood is evaluated in for 4 * (4+4+4) = 48 steps.

This is (approximately) a factor 5!!!

The likelihood: another example

Pruning: calculate the likelihood by proceeding from the leaves towards the root.



$$L = \sum_{\mathbf{X}_0} P(\mathbf{X}_0) \left[\left[\sum_{\mathbf{X}_7} P(\mathbf{X}_0 \xrightarrow{t_2} \mathbf{X}_7) P(\mathbf{X}_7 \xrightarrow{t_7} \mathbf{X}_4)\, P(\mathbf{X}_7 \xrightarrow{t_8} \mathbf{X}_5) \right] \times \left[\sum_{\mathbf{X}_8} P(\mathbf{X}_0 \xrightarrow{t_1} \mathbf{X}_8) P(\mathbf{X}_8 \xrightarrow{t_4} \mathbf{X}_3)\, P(\mathbf{X}_8 \xrightarrow{t_3} \mathbf{X}_6) \times \left[\sum_{\mathbf{X}_6} P(\mathbf{X}_6 \xrightarrow{t_5} \mathbf{X}_1)\, P(\mathbf{X}_6 \xrightarrow{t_6} \mathbf{X}_2) \right] \right] \right]$$

The likelihood: another example

Step 4

As also the topology is in fact unobserved, we need to sum the likelihood from the previous step over all possible topologies.

The pulley principle comes to the rescue, partially.

- With 5 leave nodes, the number of possible rooted binary trees equals 105.
- The pulley principle tells us only to consider the unrooted binary trees, a total of 15.

Likelihood maximization

Likelihood maximization

To maximize the log-likelihood:

- ***Step 1***: Select a tree topology.
- ***Step 2***: Choose initial values for each edge.
- ***Step 3***: Maximize edges individually, given the other edges.
- ***Step 4***: Iterate step 3, until values no longer change.
- ***Step 5***: Do this for all possible topologies.

The particular form of this algorithm described below may converge to local maxima!

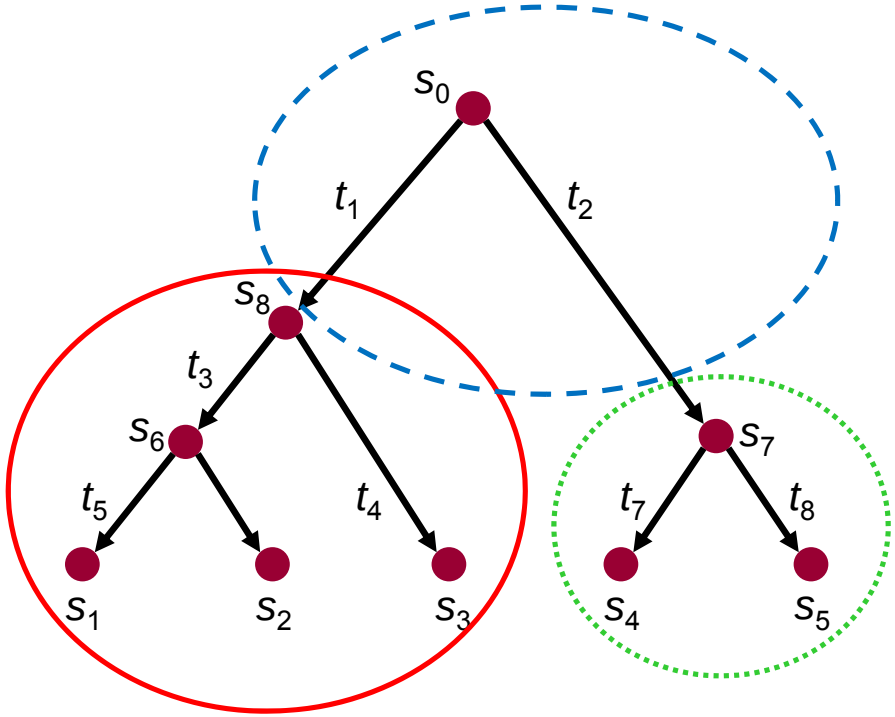
With respect to step 3:

How to maximize the log-likelihood with respect to an edge?

The likelihood

Denote the conditional likelihood of subtree rooted at node *i* with nucleotide X_{ij} by $L_{i,X_{ij}}$.

The likelihood of site *j* for our tree, now assumed to be rooted at s_8, is given by:



$$L_j = \sum_{X_{8j}} \sum_{X_{7j}} P(X_{8j}) L_{8,X_{8j}} L_{7,X_{7j}} P(X_{8j} \xrightarrow{t_1+t_2} X_{7j})$$

Likelihood maximization

Using:

$$
\begin{aligned}
P(X_{8j} \xrightarrow{t} X_{7j}) &= \frac{1}{4}[1-(1-4\,u)^t] + \delta_{X_{8j},X_{7j}}\,(1-4\,u)^t \\
&= \frac{1}{4}(1-p) + \delta_{X_{8j},X_{7j}}\,p
\end{aligned}
$$

reformulate this to:

$$
\begin{aligned}
L_j &= p \sum_{X_{8j}} \sum_{X_{7j}} \delta_{X_{8j},X_{7j}}\, P(X_{8j})\, L_{8,X_{8j}}\, L_{7,X_{7j}} \\
&\quad + (1-p) \sum_{X_{8j}} \sum_{X_{7j}} \frac{1}{4}\, P(X_{8j})\, L_{8,X_{8j}}\, L_{7,X_{7j}} \\
&= A_j\, p + B_j\, (1-p)
\end{aligned}
$$

Likelihood maximization

This holds for all sites, thus:

$$L = \prod_{j=1}^{25} [A_j\, p + B_j\, (1 - p)]$$

The log-likelihood and its derivative are given by:

$$\log(L) = \sum_{j=1}^{25} \log[A_j\, p + B_j\, (1 - p)]$$

$$\frac{\partial\, \log(L)}{\partial\, p} = \sum_{j=1}^{25} \frac{A_j - B_j}{A_j\, p + B_j\, (1 - p)} = 0$$

Likelihood maximization

The p maximizing the log-likelihood is found iteratively.

- Choose a step size $h > 0$.
- Let $p^{(k)}$ be the value of p from the k-th iteration.
- Then, define:

$$p^{(k+1)} = p^{(k)} + \frac{h}{m} \sum_{j=1}^{25} \left[\frac{A_j \, p^{(k)}}{A_j \, p^{(k)} + B_j \, (1 - p^{(k)})} - p^{(k)} \right]$$

This choice of $p^{(k+1)}$ implies the majorization:

$$\sum_{j=1}^{25} \log[A_j \, p^{(k+1)} + B_j \, (1 - p^{(k+1)})] \geq \sum_{j=1}^{25} \log[A_j \, p^{(k)} + B_j \, (1 - p^{(k)})]$$

Likelihood maximization

The majorization can be seen from:

$$
\begin{aligned}
p^{(k+1)} - p^{(k)} &= h\, p^{(k)} \left(\frac{1}{m} \sum_{j=1}^{25} \frac{A_j}{A_j\, p^{(k)} + B_j\, (1 - p^{(k)})} - 1 \right) \\
&= h\, p^{(k)} \left(1 - p^{(k)}\right) \frac{1}{m} \sum_{j=1}^{25} \frac{A_j - B_j}{A_j\, p + B_j\, (1 - p)} \\
&= h\, p^{(k)} \left(1 - p^{(k)}\right) \frac{1}{m} \left. \frac{\partial\, \log(L)}{\partial\, p} \right|_{p=p^{(k)}}
\end{aligned}
$$

which has the same sign at the derivative of the log-likelihood, evaluated in the current estimate of *p*!

Why the likelihood approach?

Why the likelihood approach?

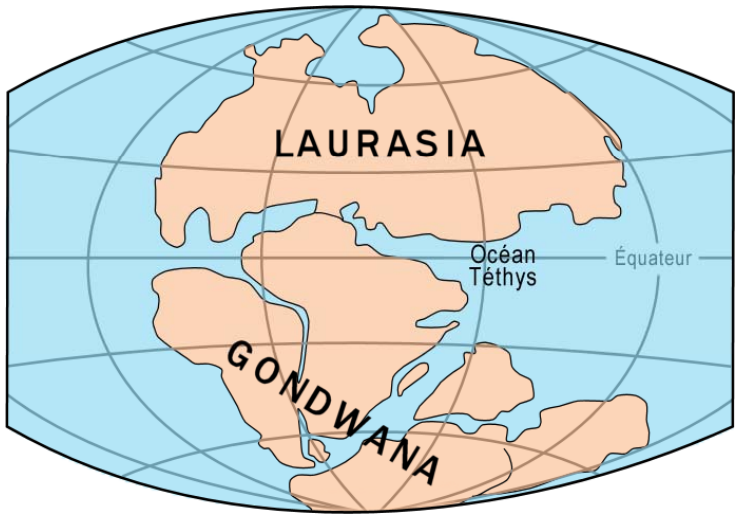
Why use the likelihood approach when also the methodologically simpler ***distance matrix*** and ***maximum parsimony methods*** are available?

- The likelihood approach makes assumptions explicit. This enables us to assess their validity.
- Within the likelihood framework we may compare nested models using a likelihood ratio test.

Example

Example

Laurasiatheria is a group of mammals originating from the former continent Laurasia.



The phylogenetic relationships between the Laurasiatherians are still uncertain.

Available:

- RNA sequence data of 47 Laurasiatherians.
- Sequence is 3179 bases long.

Reconstruct their phylogenetic tree.

Picture: Wikipedia

Example

In R:

```
> # activate library
> library(phangorn)

> # load data
> data(Laurasiatherian)
```

```
Platypus  ttaaaggtttggtcctagccttactgttagatttgattagatttatacatgcagtatcc...
Wallaroo  ccaaaggtttggtcctggccttactgttaattgtagttagacctacacatgcagtttcc...
Possum    ccaaaggtttggtcctagccttactgttaattataattaaacctacacatgcagtttcc...
Bandicoot ccaaaggtttggtcctagcctttctattaattttaattaaacctacacatgcagtctcc...
Opposum   ccataggtttggtcctagccttattattagttctaattagacctacacatgcagtttcc...
Armadillo ccacaggtctggtcctagccttactattaattcataacaaaattacacatgcagtatca...
Elephant  ccaaaggtttggtcccggccttcttattggttactaggaaacttatacatgcagtatcc...
Aardvark  ttaaaggtttggtcctagcctttctattagttgacagtaaatttatacatgcagtatct...
Tenrec    ttaaaggtttggttctagccttttttattagttcttaataaaattatacatgcagtatcc...
Hedghog   aataaggtctggtcccagccttcctattttctattagtagaattacacatgcagtatca...
...       ...
```

Example

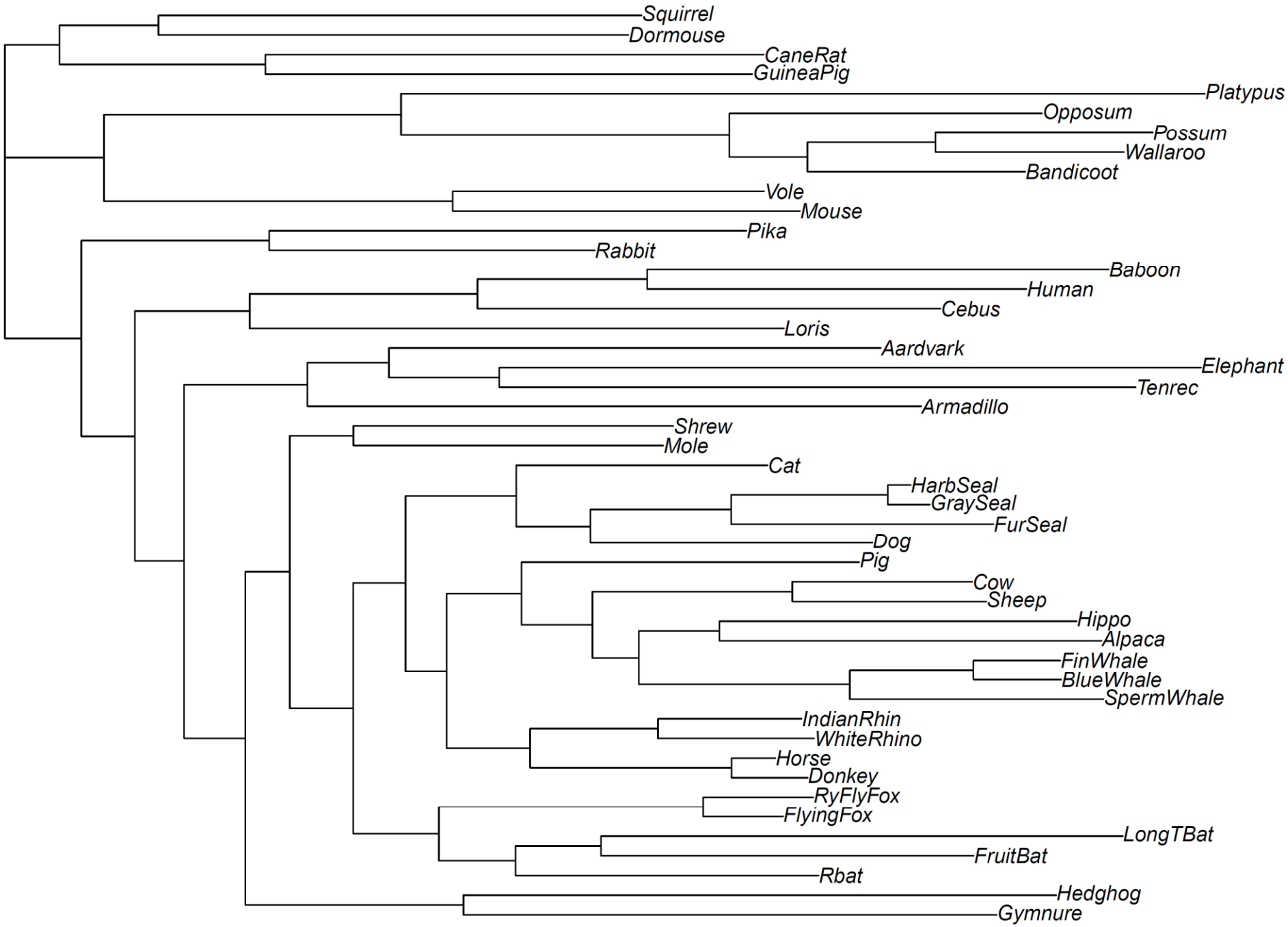
Now fit the model:

```
> # construct a starting tree
> distMat <- dist.logDet(Laurasiatherian)
> tree <- NJ(distMat)

> # fit Jukes-Cantor model
> fitJC <- pml(tree, Laurasiatherian, model="JC")
> fitJC <- optim.pml(fitJC, optNni=TRUE,
                     optEdge=TRUE, model="JC")
> plot(fitJC$tree)
```

Note: this fits a model with continuous time, instead of discrete time as treated in the lecture.

Example


Squirrel
Dormouse
CaneRat
GuineaPig
Platypus
Opposum
Possum
Wallaroo
Bandicoot
Vole
Mouse
Pika
Rabbit
Baboon
Human
Cebus
Loris
Aardvark
Elephant
Tenrec
Armadillo
Shrew
Mole
Cat
HarbSeal
GraySeal
FurSeal
Dog
Pig
Cow
Sheep
Hippo
Alpaca
FinWhale
BlueWhale
SpermWhale
IndianRhin
WhiteRhino
Horse
Donkey
RyFlyFox
FlyingFox
LongTBat
FruitBat
Rbat
Hedghog
Gymnure

Example

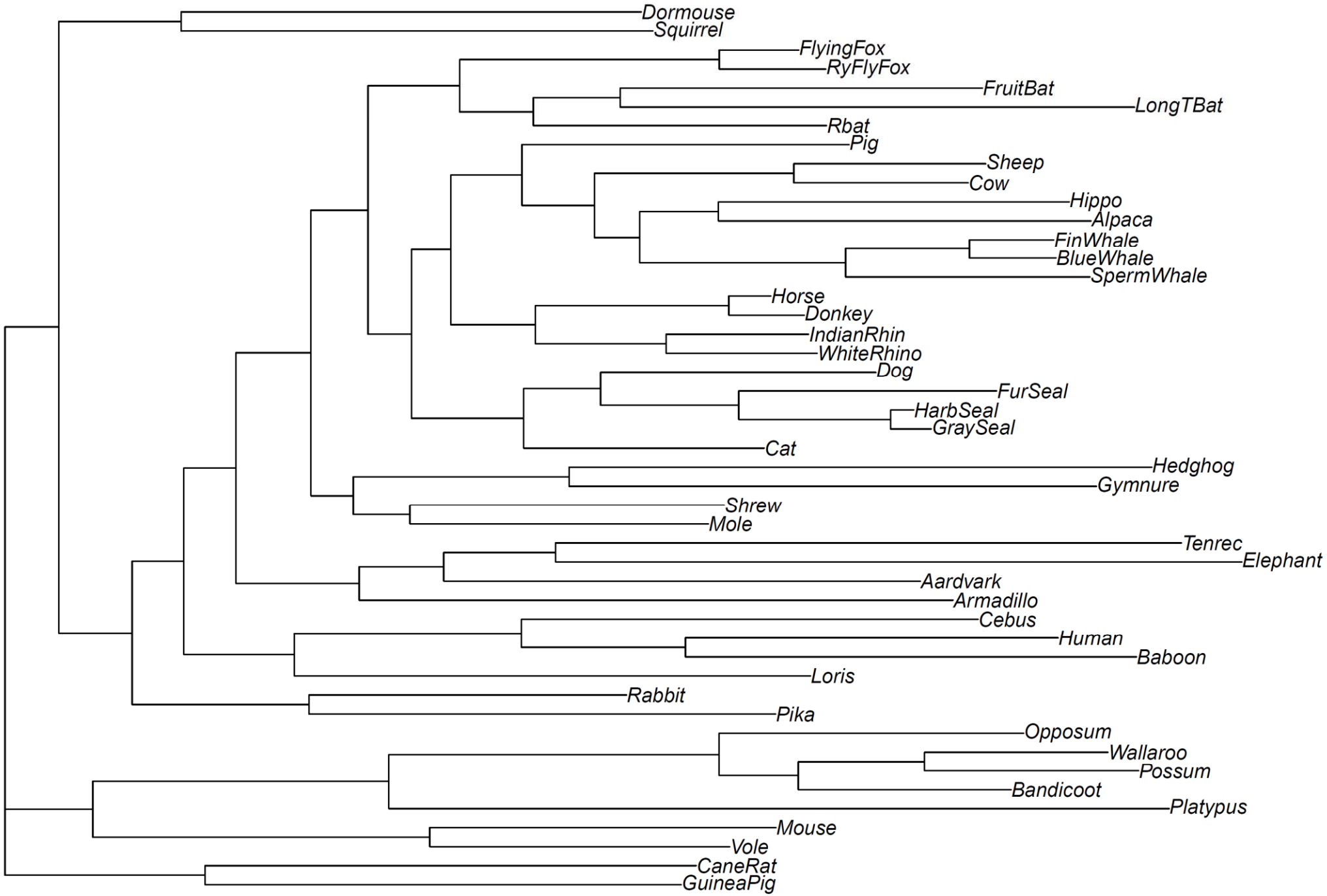
The Jukes-Cantor model is just one evolutionary model. Many more exist.

Fit different model:

```
> # construct a starting tree
> distMat <- dist.logDet(Laurasiatherian)
> tree <- NJ(distMat)

> # fit Felsenstein model
> fitF81 <- pml(tree, Laurasiatherian, model="F81")
> fitF81 <- optim.pml(fitF81, optNni=TRUE,
                      optBf=TRUE, optEdge=TRUE,
                      model="F81")
> plot(fitF81$tree)
```

Example


Dormouse
Squirrel
FlyingFox
RyFlyFox
FruitBat
LongTBat
Rbat
Pig
Sheep
Cow
Hippo
Alpaca
FinWhale
BlueWhale
SpermWhale
Horse
Donkey
IndianRhin
WhiteRhino
Dog
FurSeal
HarbSeal
GraySeal
Cat
Hedghog
Gymnure
Shrew
Mole
Tenrec
Elephant
Aardvark
Armadillo
Cebus
Human
Baboon
Loris
Rabbit
Pika
Opposum
Wallaroo
Possum
Bandicoot
Platypus
Mouse
Vole
CaneRat
GuineaPig

Assumptions

Assumptions

Transition-substitution:

A, G → A, G

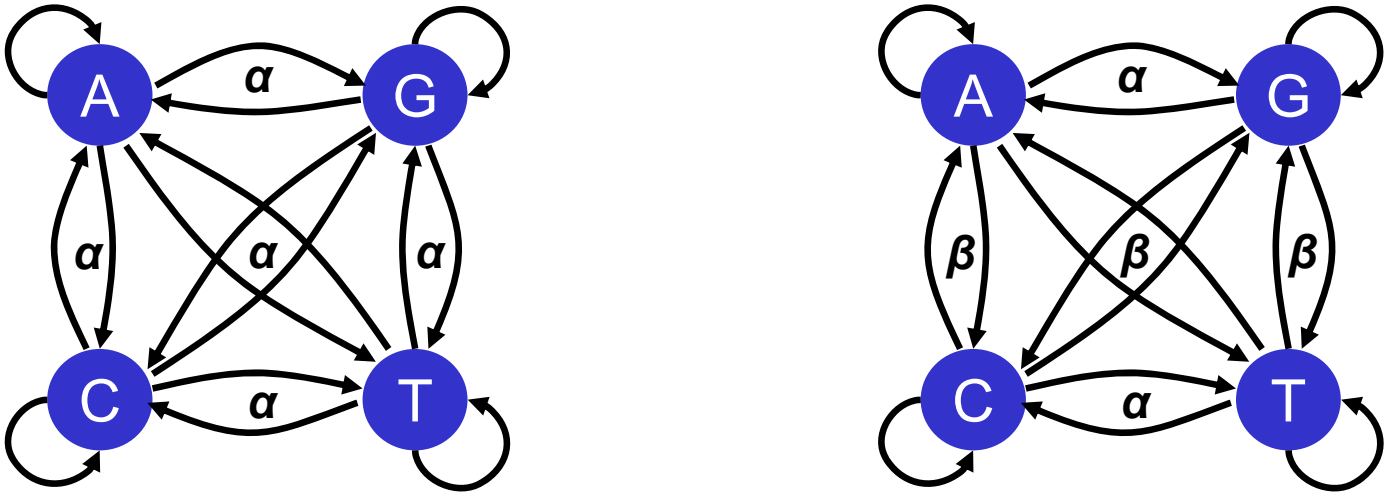
C, T → C, T

Transversion-substitution:

A, G → C, T

C, T → A, G

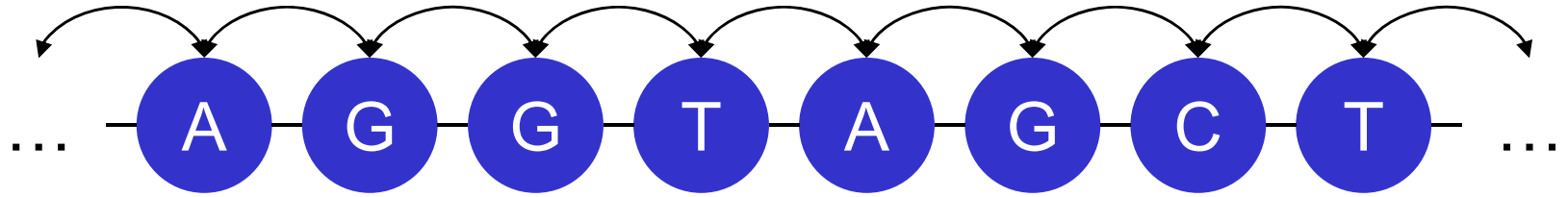
Transition- and transversion probabilities differ:



More complicated models than Jukes-Cantor available.

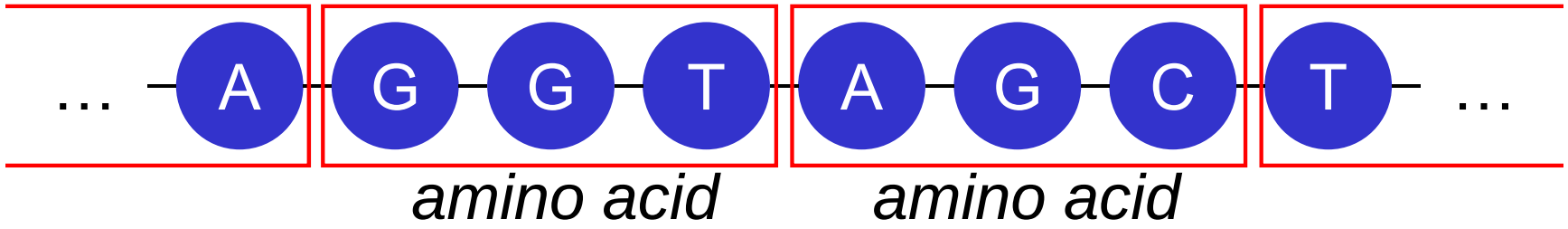
Assumptions

Positions do not evolve independently (***covarion***):



But also …

… three contiguous bases code for one amino acid:



Assumptions

Heterotachy is a general term for within-site rate variation over time. Under heterotachy, evolutionary rates at different sites may vary in different ways over subtrees.

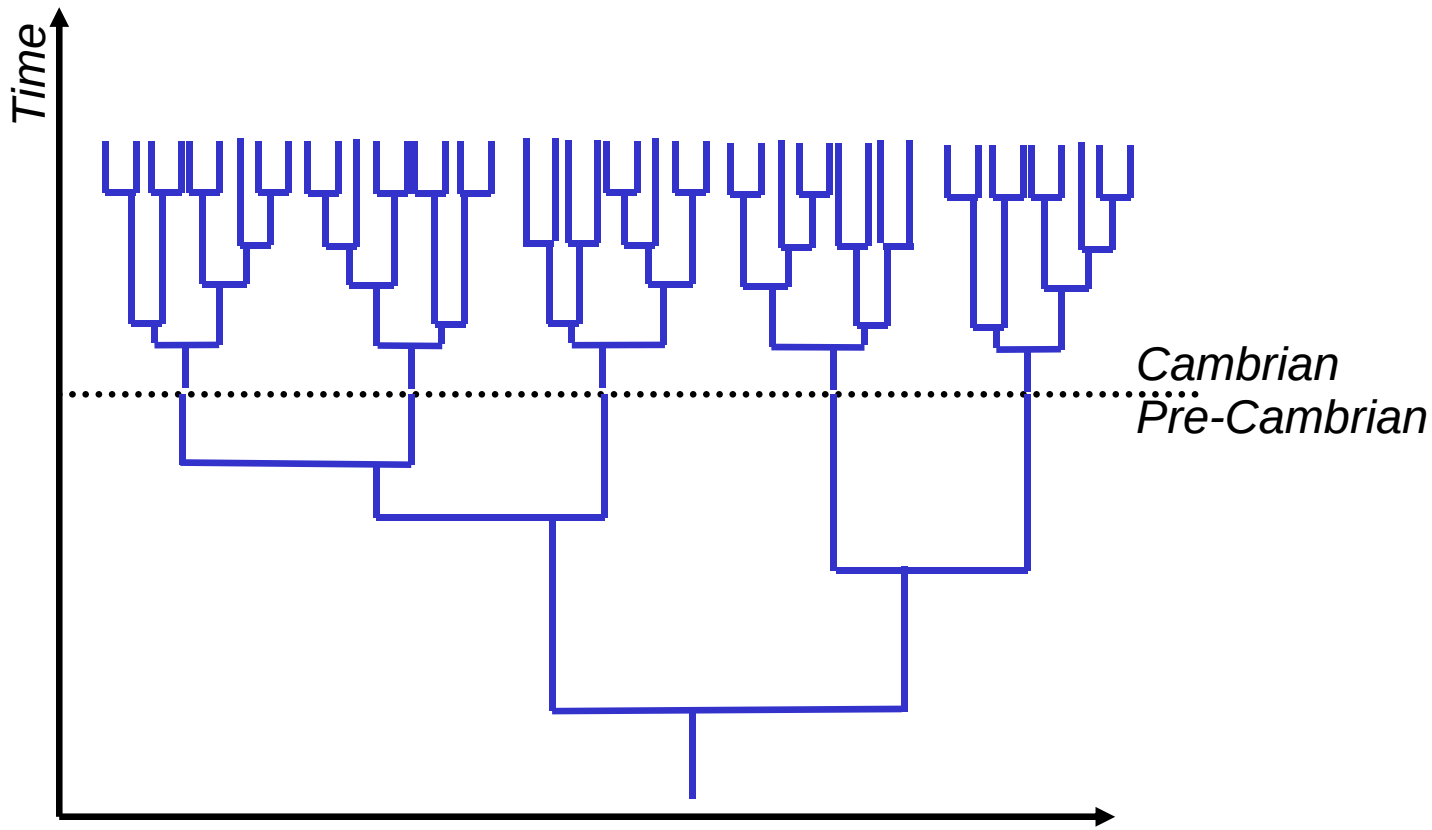
Hence, under heterotachy, the time-homogeneity assumption may be invalid. That is, the rate of nucleotide substution (the transition probability) may not be constant of time.

The molecular hypothesis should be applied with care.

Assumptions

The ***Cambrian explosion*** refers to the period around 530 My ago in which the evolutionary pace seems accelarated.

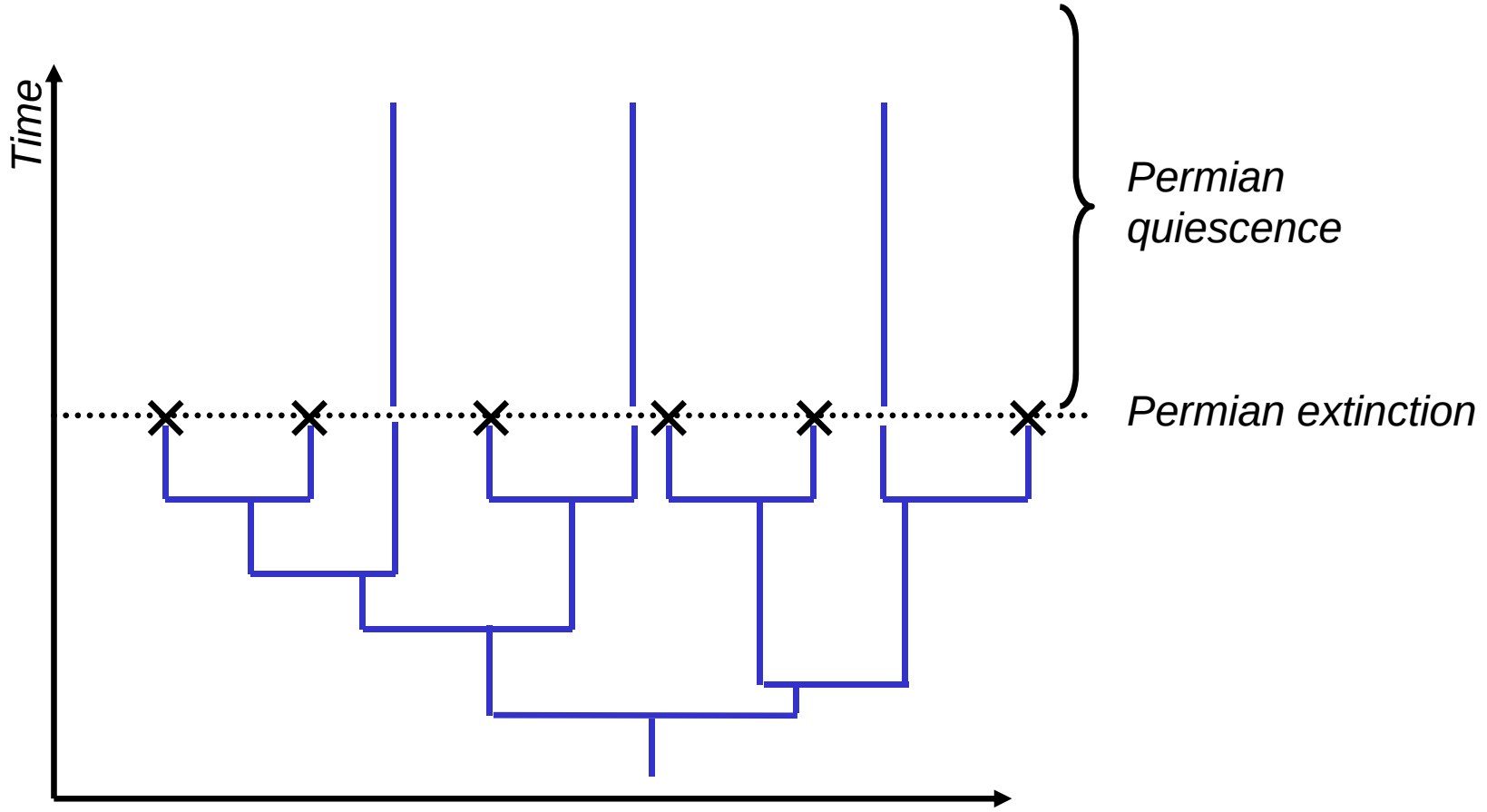
→ substitution-rate varies over time.



Assumptions

The ***Permian quiescence*** refers to the period after the Permian extinction (250 My ago), where the evolutionary pace seemed to have slowed down.

→ substitution-rate varies over time.



Assumptions

Implicitly, it has been assumed that organism evolve independently.

However, often there is ***co-evolution***:

References & further reading

References and further reading

Cavalli-Sforza, L., Edwards, A.W.F. (1967), “Phylogenetic analysis. Models and estimation procedures”, *Evolution*, **21**, 550-570.

Clote, P., Backoefen, R. (2000), *Computational Molecular Biology: An Introduction*, John Wiley, New York.

Ewens, W.J, Grant, G. (2006), *Statistical Methods for Bioinformatics*, Springer, New York.

Felsenstein, J. (1981), “Evolutionary trees from DNA sequences: a maximum likelihood approach”, *Journal of Molecular Evolution*, **17**, 368-376.

Felsenstein, J. (2004), *Inferring Phylogenies*, Sinauer Associates, Sunderland, Massachusetts.

Graur, D., Li, W.-H. (2000), *Fundamentals of Molecular Evolution*, 2nd Ed., Sinauer: Sunderland, Massachusetts.

Schliep, K.P. (2010), “phangorn: Phylogenetic analysis in R”, *Bioinformatics*, …

Wilson, A.C., Carlson, S.S., White, T.J. (1977), “Biochemical Evolution”, *Annual Review of Biochemistry*, **46**, 573-639.