

Phylogenetics

Introduction to likelihood

RL-V3 MPP

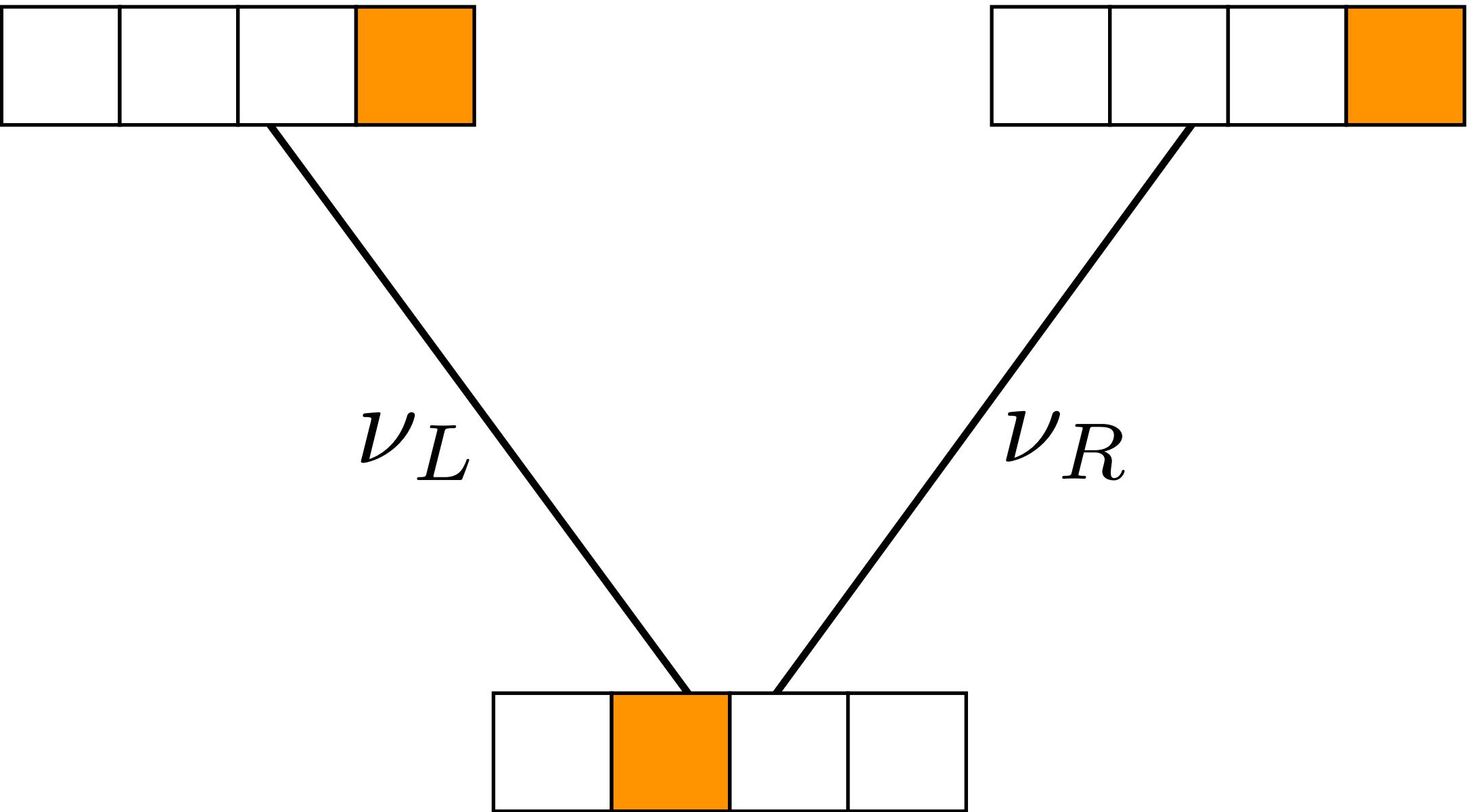
Rachel Warnock

23.04.24



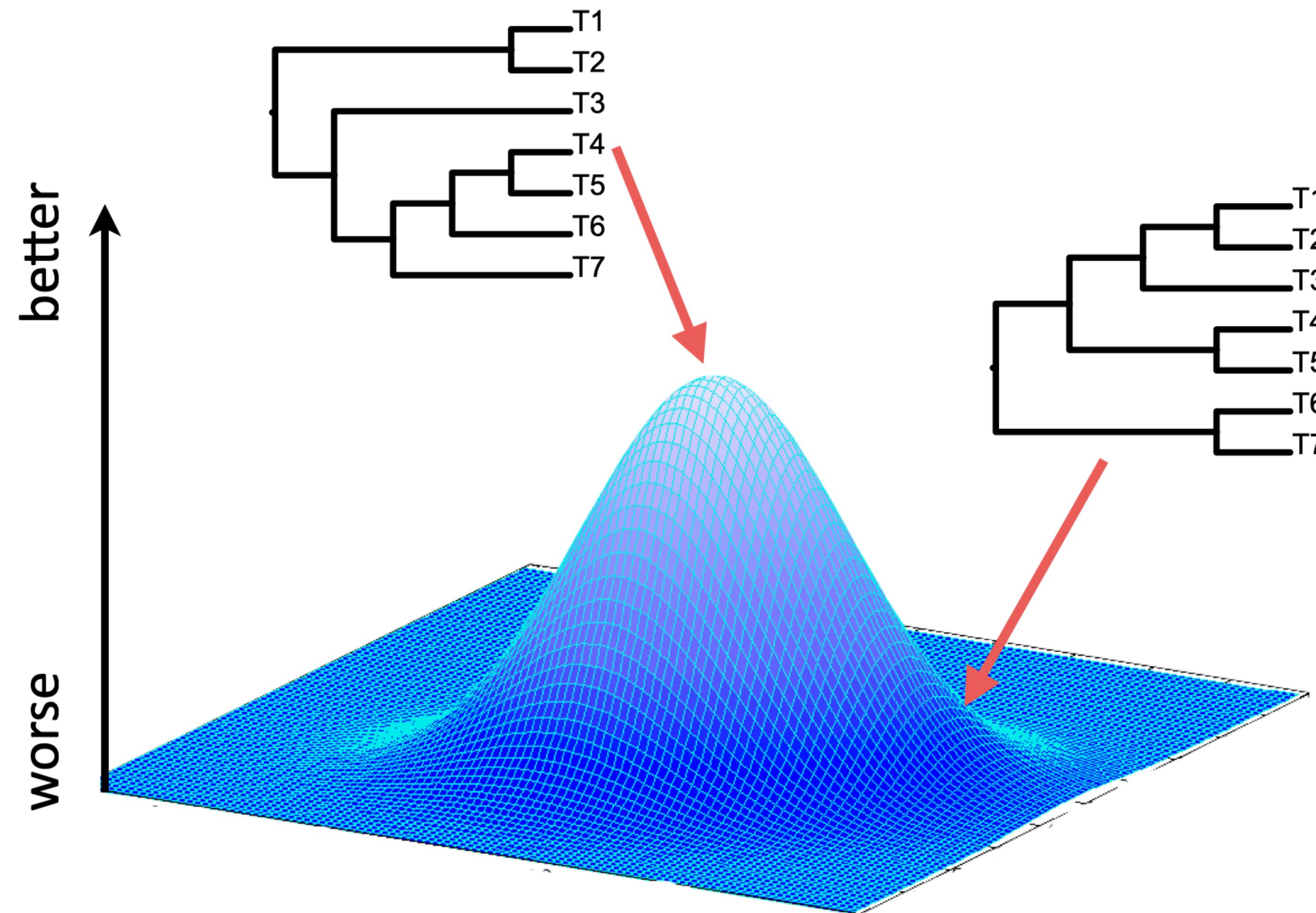
Today's objectives

- Introduction to **substitution models**
- Gain an understanding of the **likelihood** approach to tree-building



Recap

How do we find the ‘best’ tree?



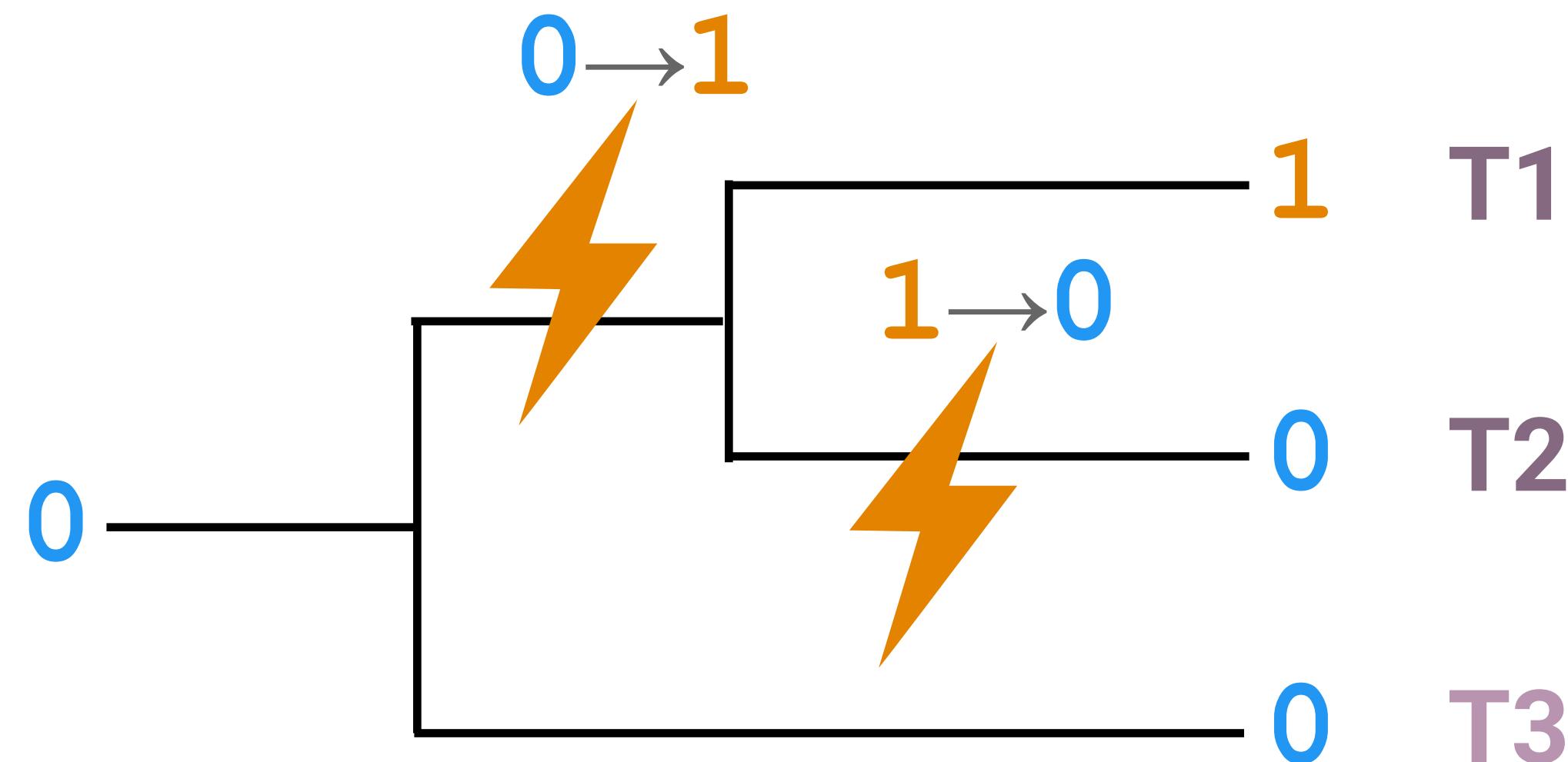
It depends how you measure ‘best’

Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Likelihood score (probability), optimised over branch lengths and model parameters
Bayesian inference	Posterior probability, integrating over branch lengths and model parameters

Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Convergence and parsimony



Hypothetical tree showing multiple transitions at the same character

Parsimony will always favour the tree with the smallest number of changes

The method does not account for multiple transitions (or “hits”), e.g.,
 $0 \rightarrow 1 \rightarrow 0$

We can only invoke convergent evolution *ad hoc* after inference

Recap: parsimony

Parsimony does not make **explicit** assumptions about the evolutionary process (although it makes **implicit** assumptions)

It has been demonstrated that in some scenarios parsimony is **statistically inconsistent**. The issue is known as **long branch attraction**

Model-based approaches on the other hand make **explicit** assumptions about evolutionary processes (+ have a wider ranger of applications in paleobio)

What do we mean by model?

(the following is my take on things – intended to be useful but not definitive)

What is a **statistical model**? When is an equation a model?

What is a **mechanistic model**?

What is the difference between an **algorithm** and a model?

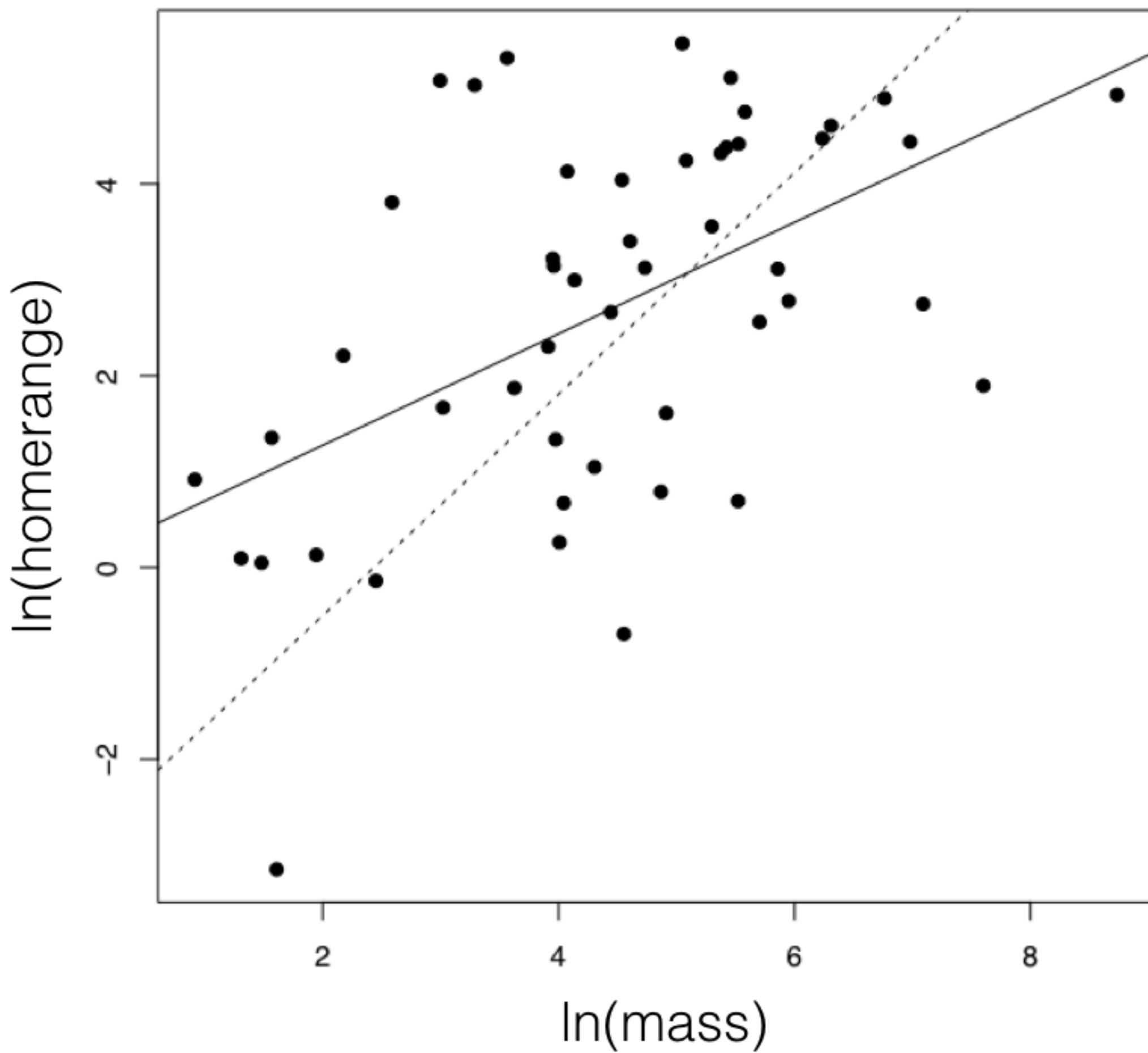
A [statistical model](#) is a type of model that includes a set of assumptions about the data-generating process

It should be possible to [simulate data](#) under the assumptions of the model

If we're lucky, we might also be able to [estimate parameters](#) under the model*. This isn't always possible because some models are too complex

*A fancy way of saying this is, "we can perform [inference](#) under the model"

An example



The solid black line is a linear regression line

We can estimate the parameters of the regression model

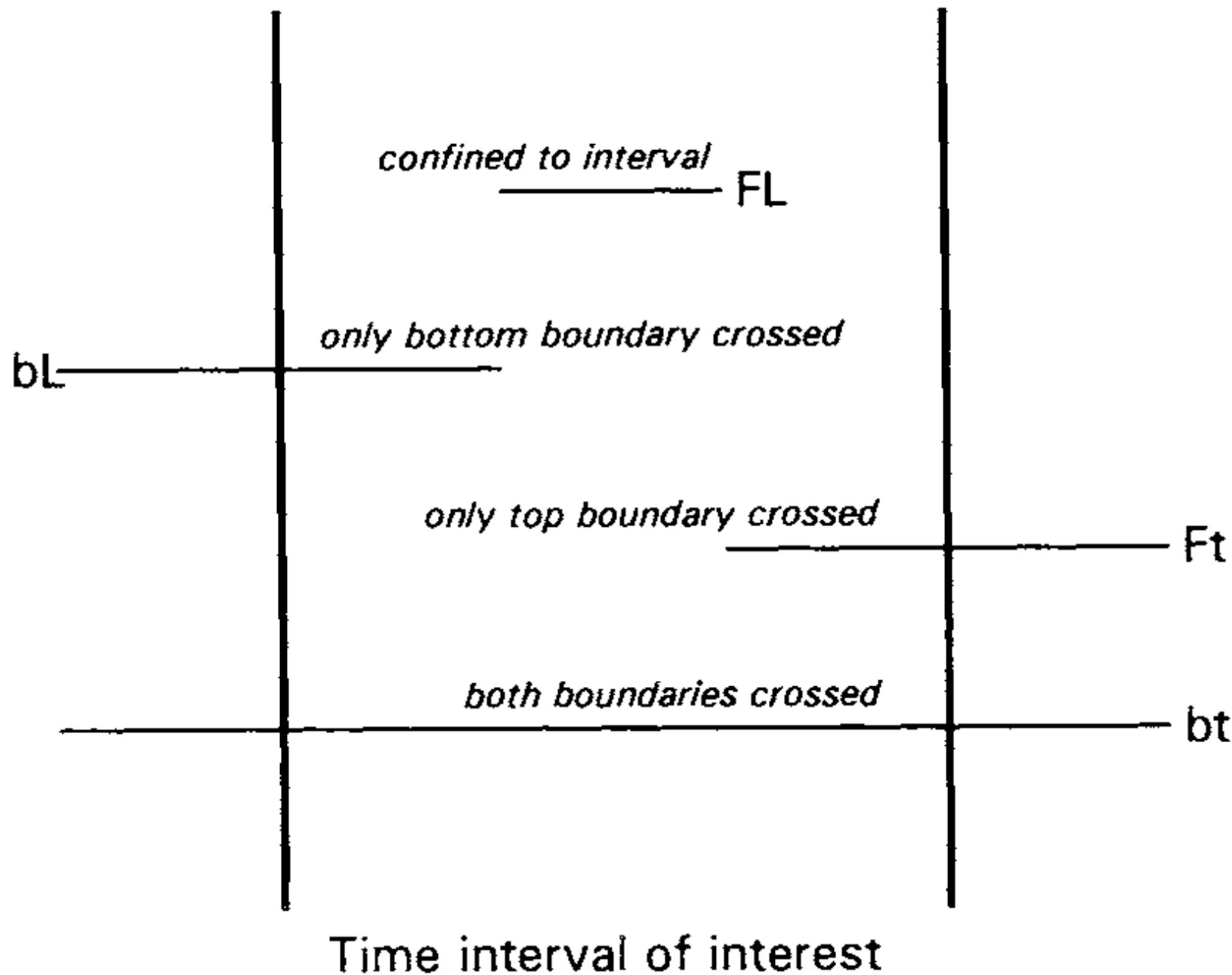
$$y = X\beta + \varepsilon$$

It's also straightforward to simulate data under this model

Image source Harmon (2019)

Non model-based approaches are still useful

Four fundamental classes of taxa



The boundary-crosser and three-timer metrics are *not* models

They provide a clever way of approximating origination and extinction rates (and often perform well), but don't describe the data generating processes

Mechanistic or process based models are based on ‘physical principles’. They describe the data as a function of a set of parameters that have a tangible biological or geological meaning

A regression model is not mechanistic – it describes the relationship between x and y but the parameters don’t have a biological meaning

Many models used in phylogenetics are mechanistic, e.g., they might include parameters for origination, extinction, or sampling

An algorithm is a precise rule (or set of rules) specifying how to solve some problem

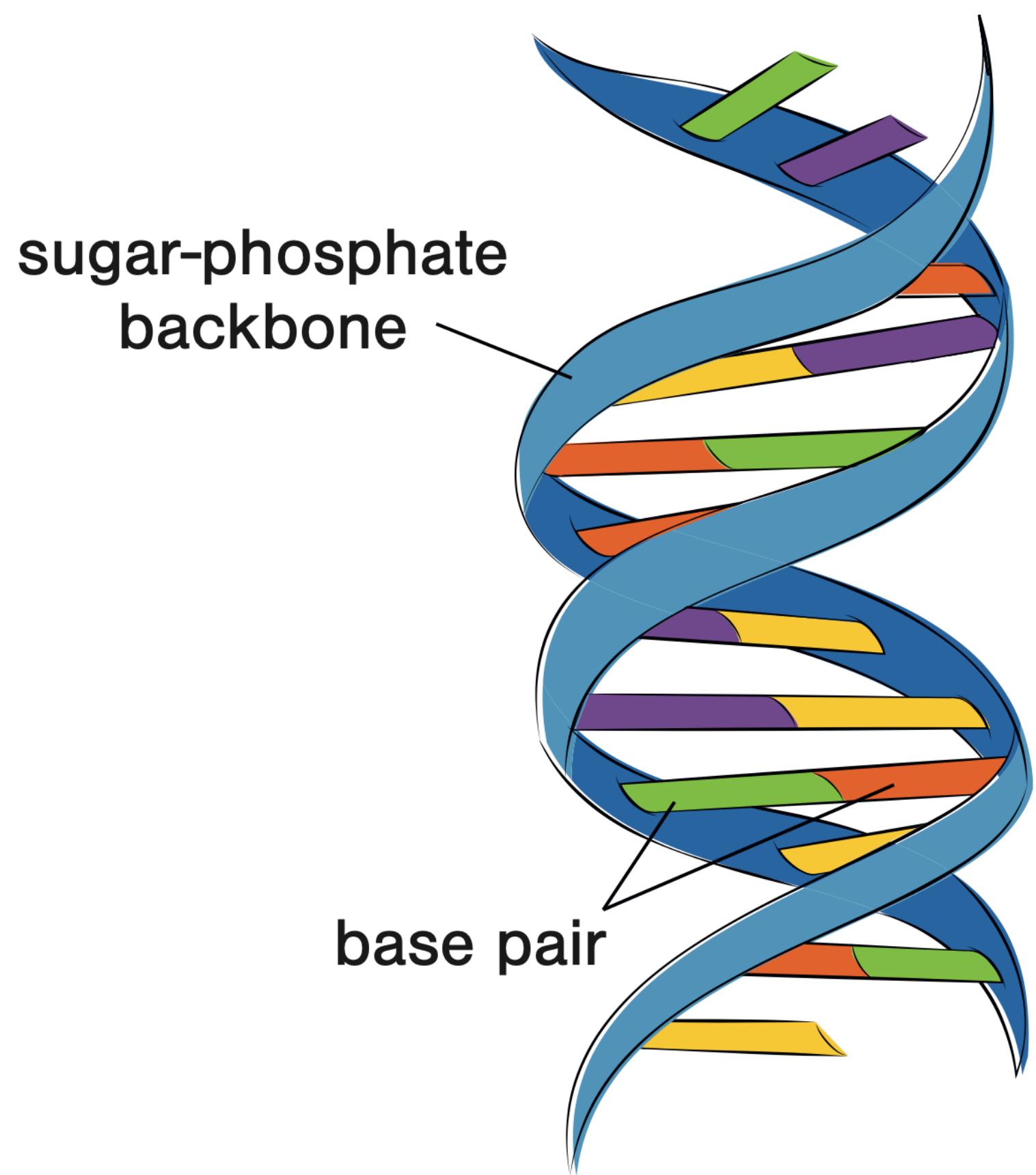
```
i = 1
while i < 11:
    print(i)
    i = i + 1
```

```
for i in range(1,11):
    print(i)
```

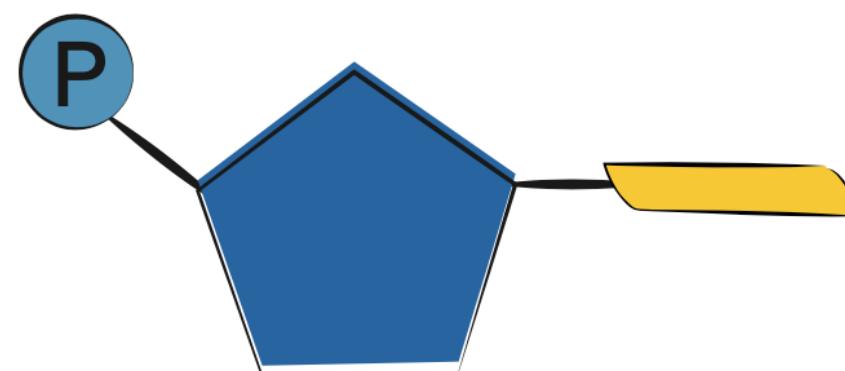
Used in phylogenetics for all sorts of tasks, e.g., traversing tree space

Molecular evolution

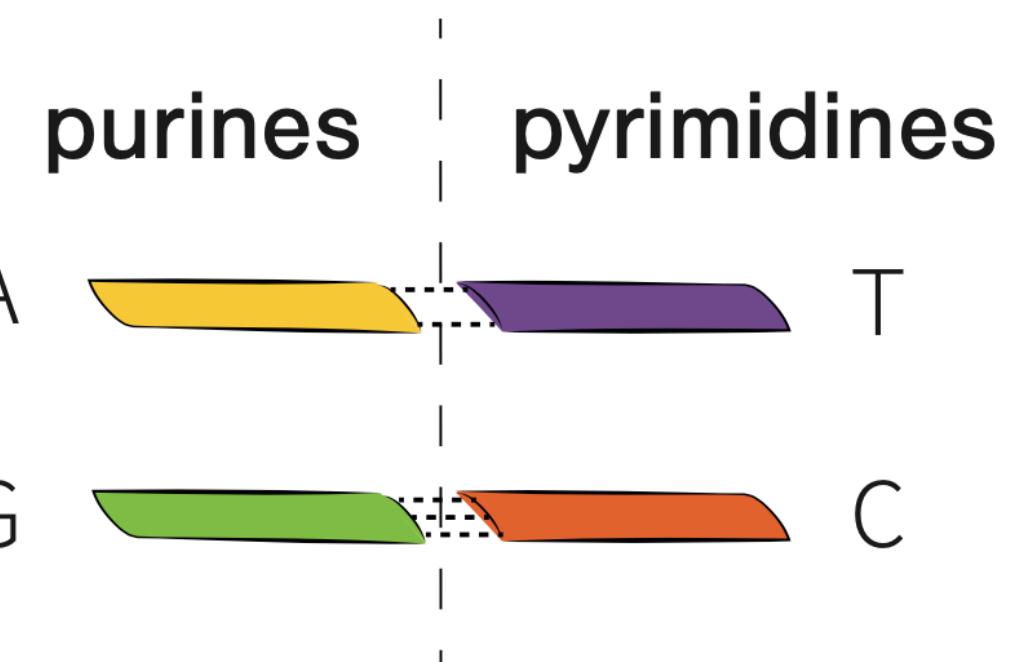
Deoxyribonucleic acid (DNA)



Nucleotide:



phosphate + sugar + nitrogenous base



Purines

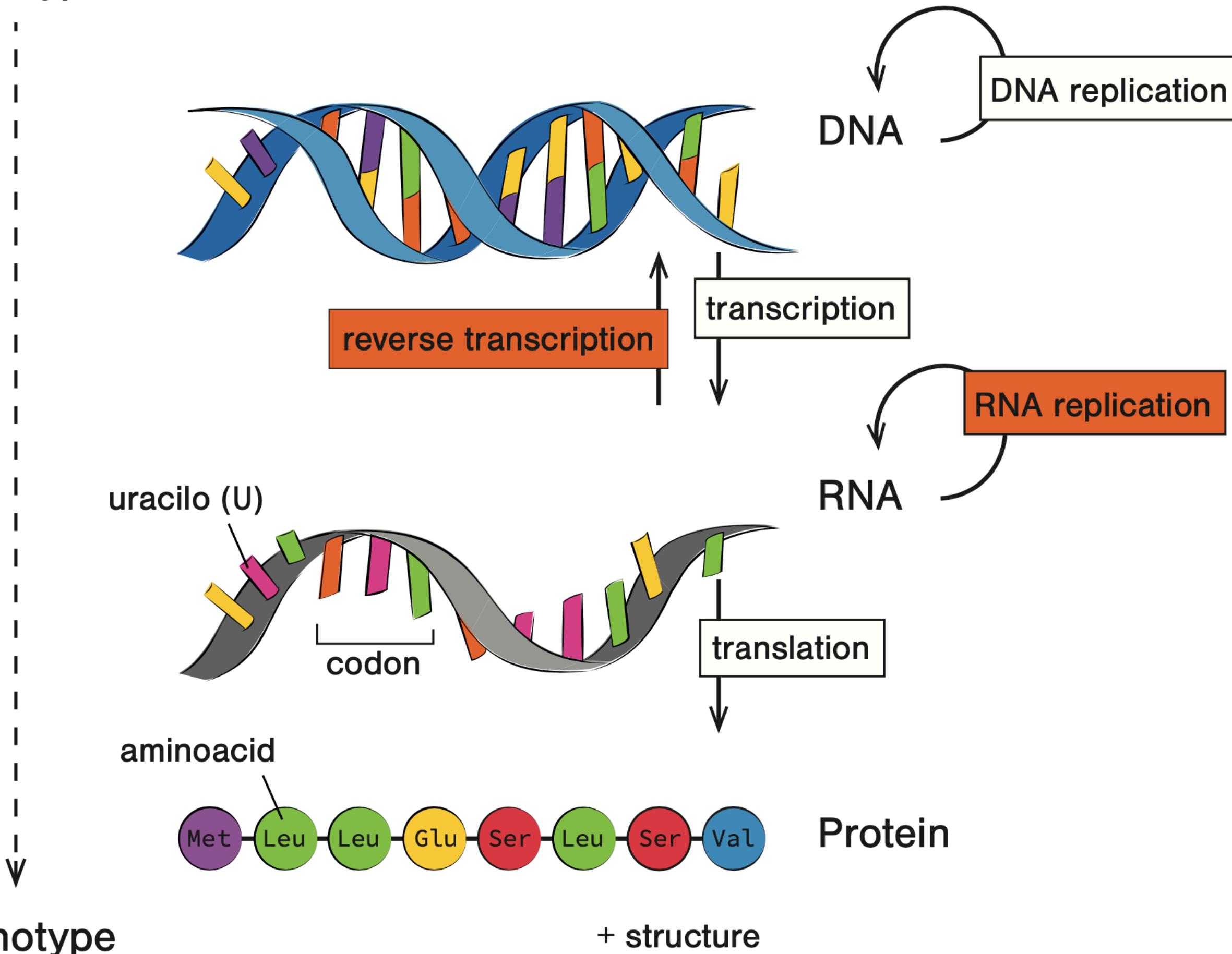
- Adenine (**A**)
- Guanine (**G**)

Pyrimidines

- Cytosine (**C**)
- Thymine (**T**)*

The central dogma of biology

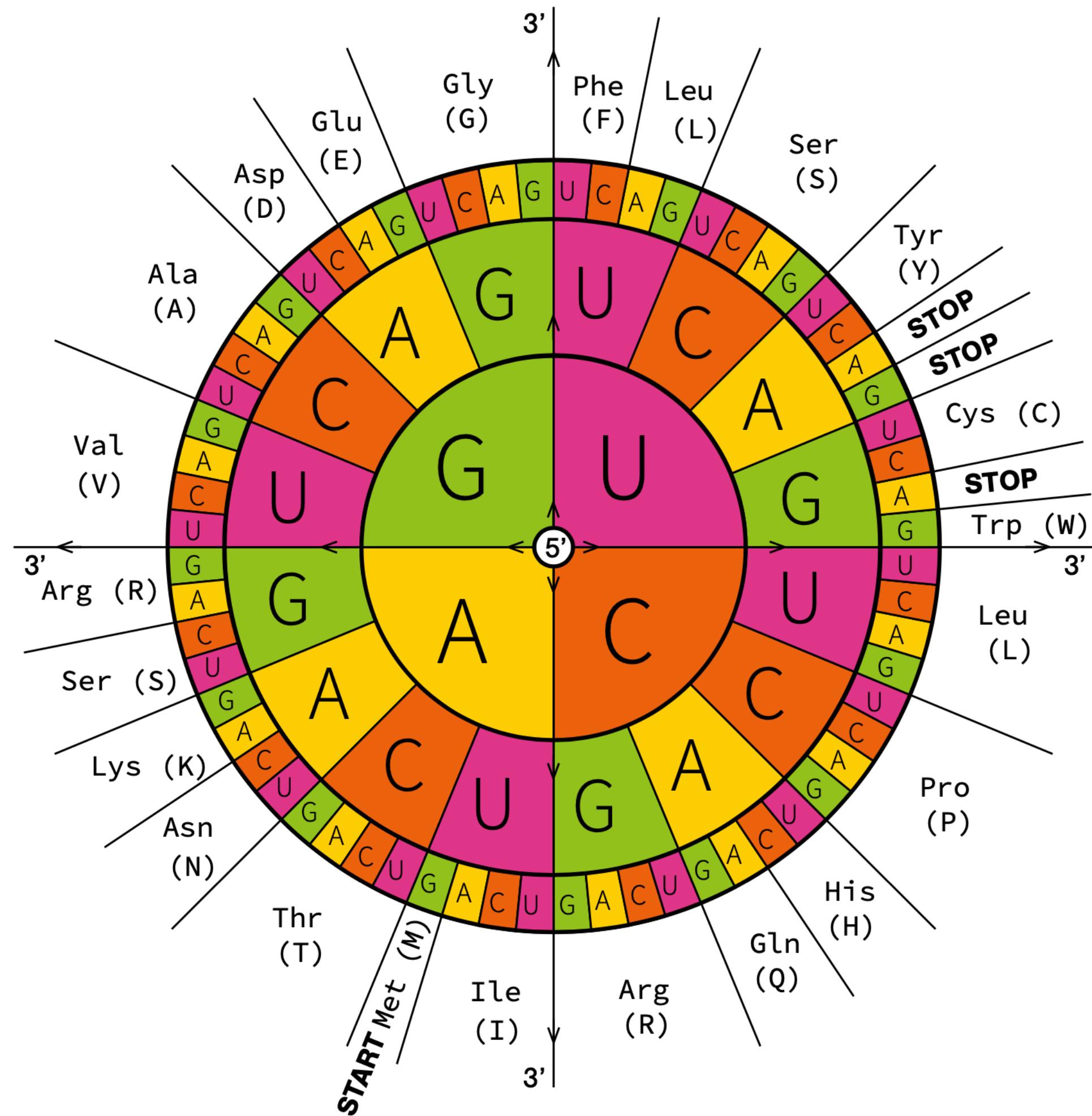
Genotype



$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$

Each group of 3 successive nucleotides in a gene is a codon that encodes an amino acid (or terminate translation)

The universal genetic code



Amino acid	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

$4^3 = 64$ combinations

3 terminate translation

21 amino acids

Mutation vs. substitution

Variation in genotypes (and in phenotypes) is due to errors that arise during DNA replication

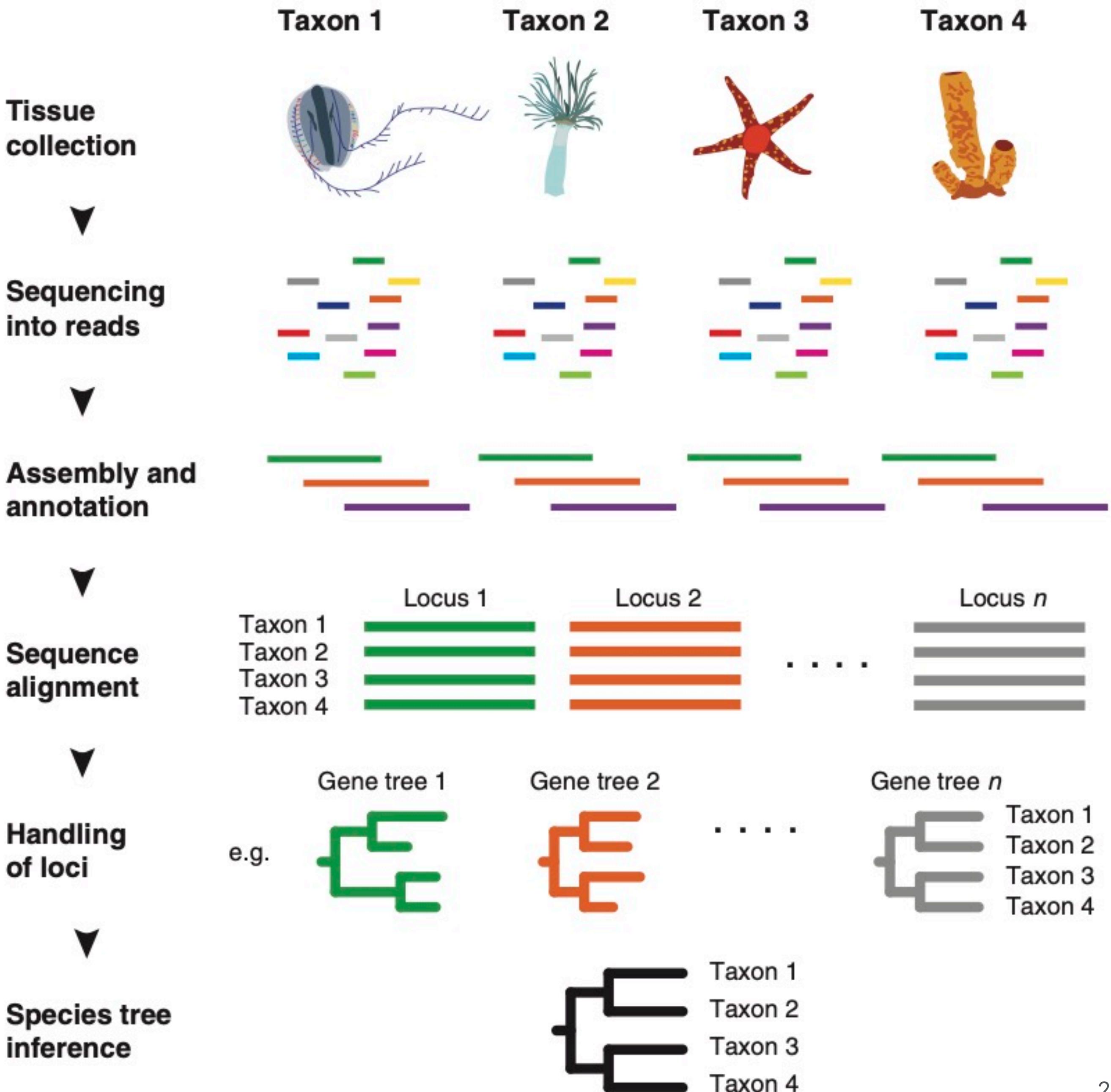
Individuals of the same species have identical characters at *most* positions in their genome (only 0.1% vary among humans)

Most mutations are repaired but can persist across generations

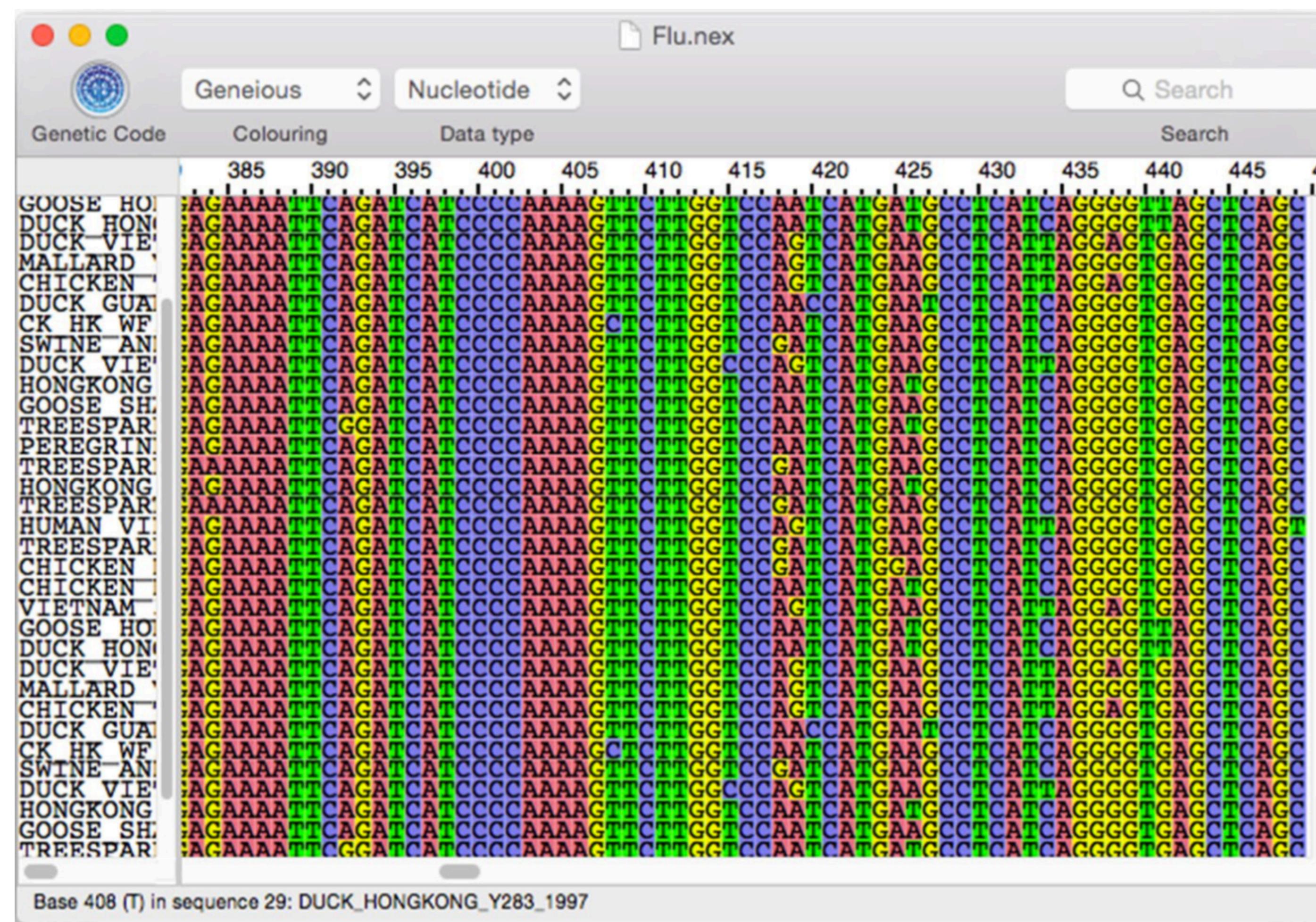
Mutations that spread throughout a population and become ‘fixed’ called **substitutions**

DNA sequencing

Multiple sequence alignment software establishes homology across sites from different species



Multiple sequence alignment



#NEXUS

[Cytochrome oxidase B genes - bears]

[Data source: <https://revbayes.github.io/tutorials/dating/>]

BEGIN DATA;

DIMENSIONS NTAX=10 NCHAR=1000;

FORMAT DATATYPE = DNA MISSING=? GAP=- ;

MATRIX

Ailuropoda_melanoleuca

Arctodus_simus

Helarctos_malayanus

Melursus_ursinus

Ursus_americanus

Ursus_arctos

Ursus_maritimus

Ursus_thibetanus

Ursus_spelaeus

Tremarctos_ornatus

ATGATCAACATCCGAAAAACTCATCCATTAGTTAAAATTATCAACAACTCATTGACCT...

ATGACCAACATCCGAAAGACTCACCCACTGGCCAAAATTATCAATAACTCATTGACCT...

ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATTAACAACTCACTTATTGACCT...

ATGACCAACATCCGAAAAACCCACCCACTAGCTAAAATCATTAACAACTCACTCATTGACCT...

ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCACTTATTGATCT...

ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCACTTATTGACCT...

ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCACTCATTGATCT...

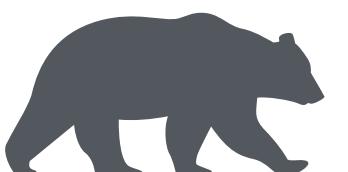
ATGACCAACATCCGAAAAACCCATCCATTAGCCAAAATCATCAACAACTCACTCATTGATCT...

ATGACCAACATCCGAAAAACCCATCCACTAGCTAAAATCATCAACAACTCACTCATTGACCT...

ATGACCAACATCCGAAAAACTCACCCACTAGCTAAAATCATCAACAACTCACTCATTGACCT...

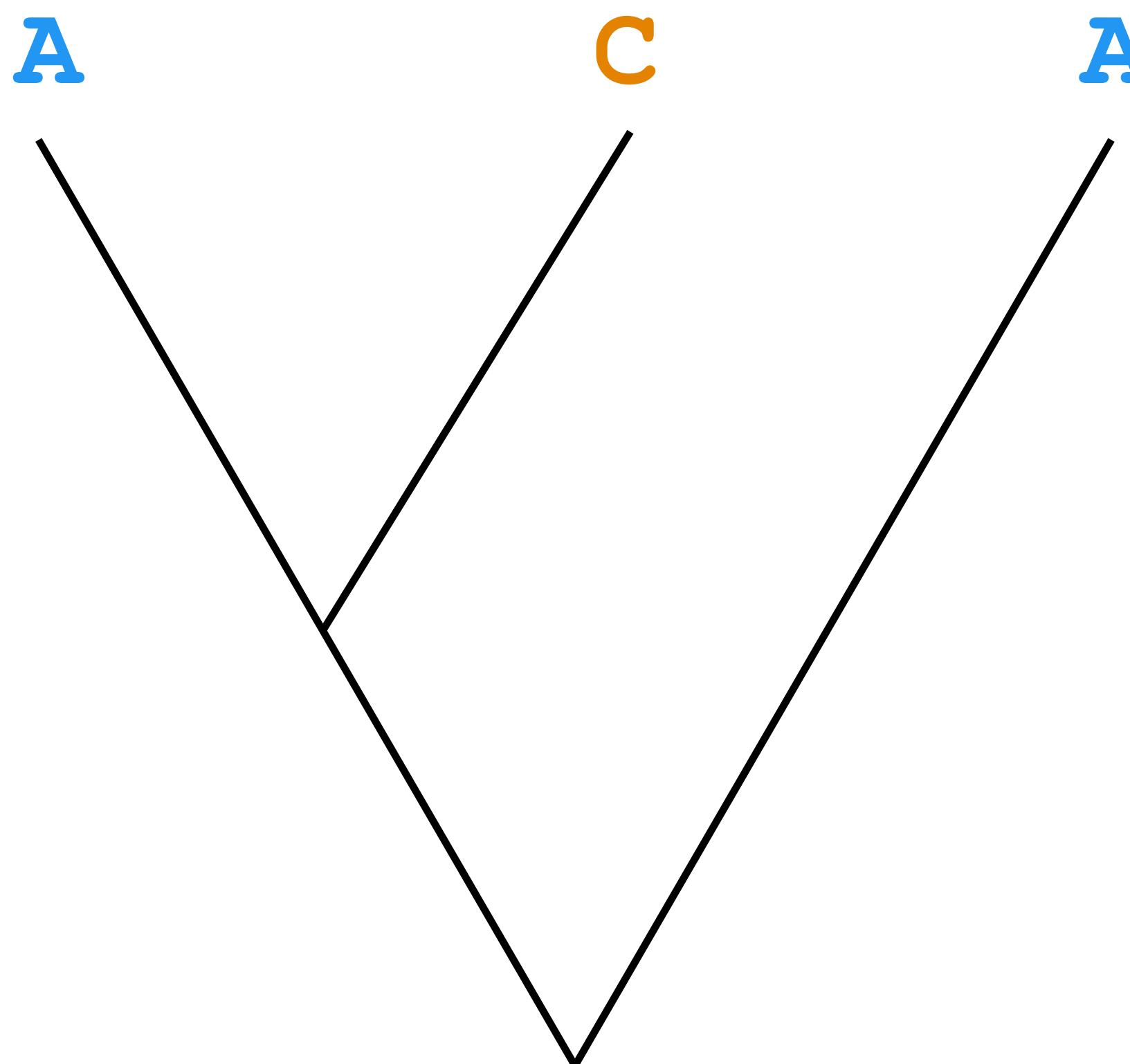
;

END;

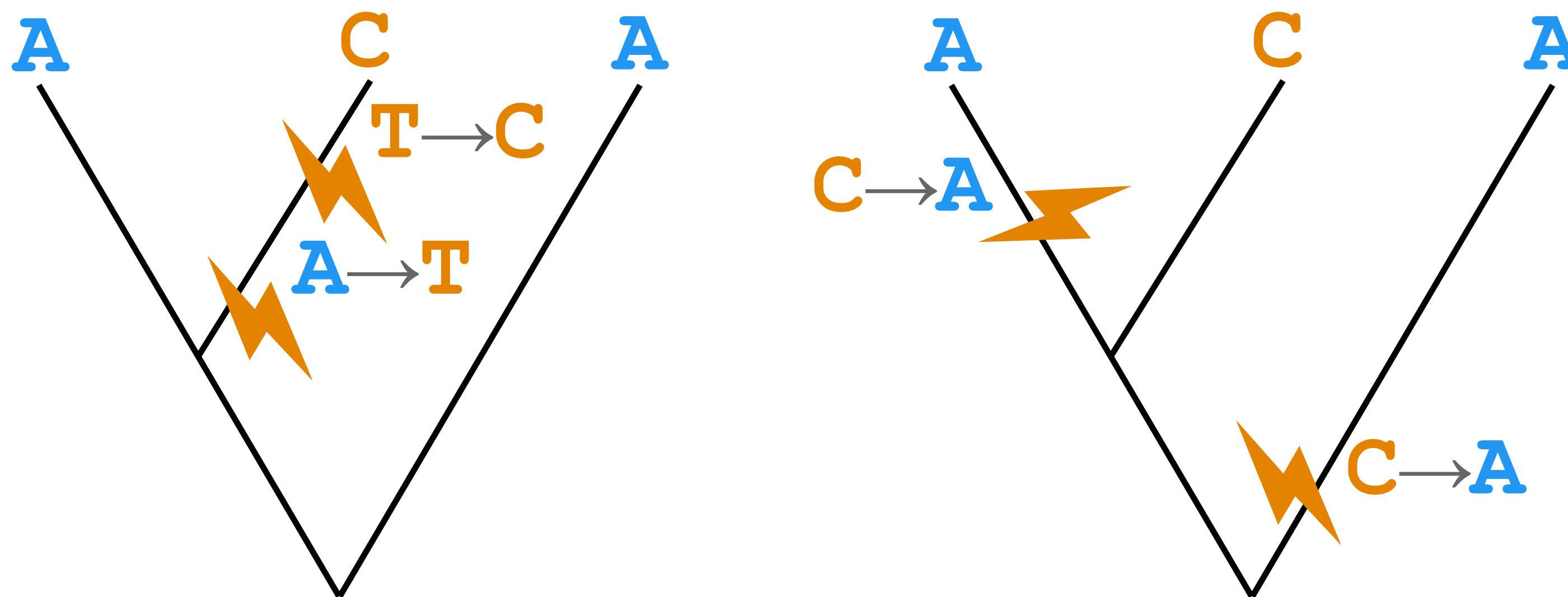
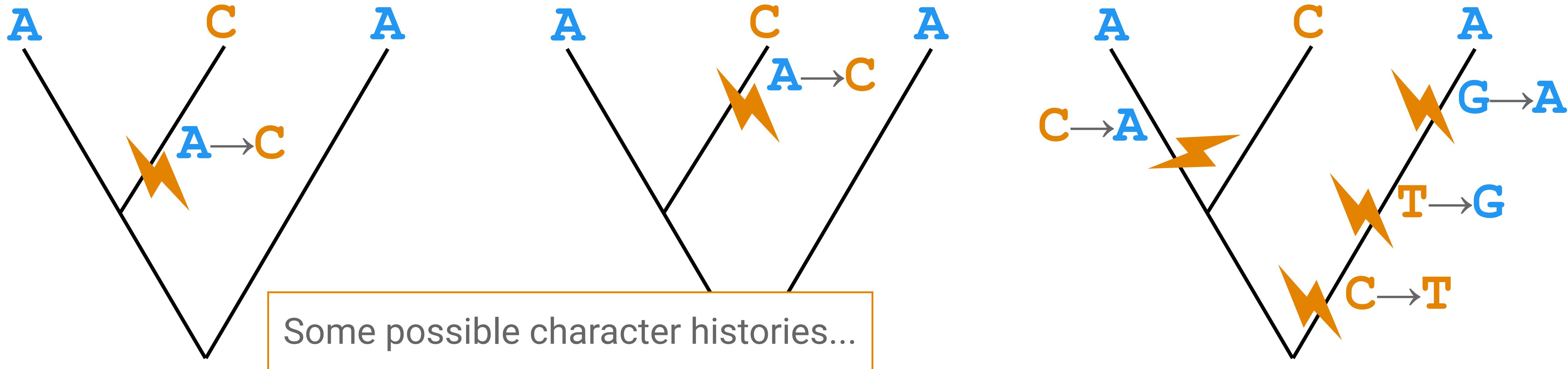


The data are the observed states at the tips

How probable is our data, given my tree?



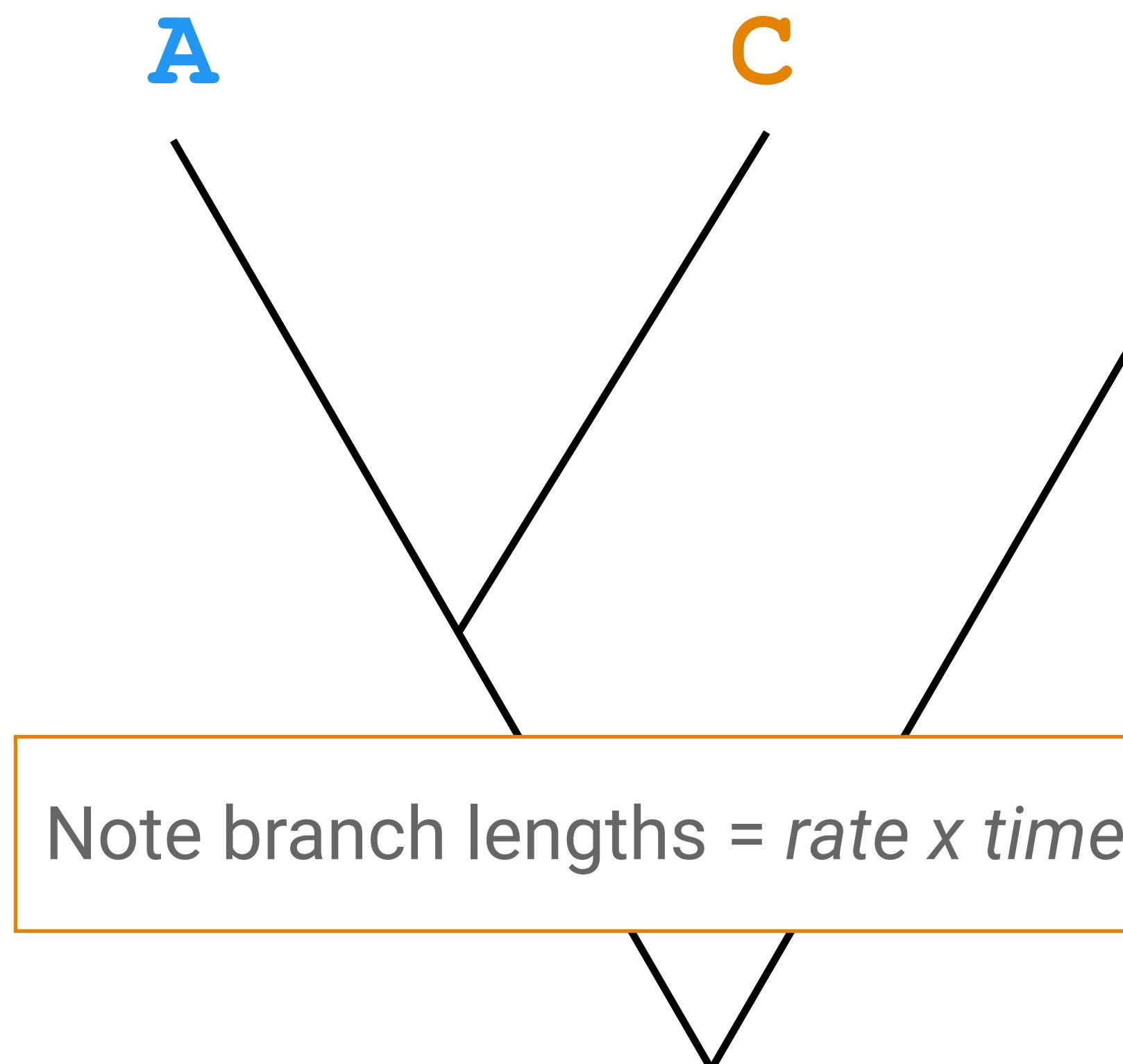
To apply a model based approach
we need to be able to compute
the probability of our sequence
alignment (or character matrix)



There are **infinitely** many histories leading to the same observed data

The data are the observed states at the tips

How probable is our data, given my tree?



To compute P , we need:

- A model of sequence (or character) evolution
- A way of calculating the probability for given a phylogeny (tree topology + branch lengths)

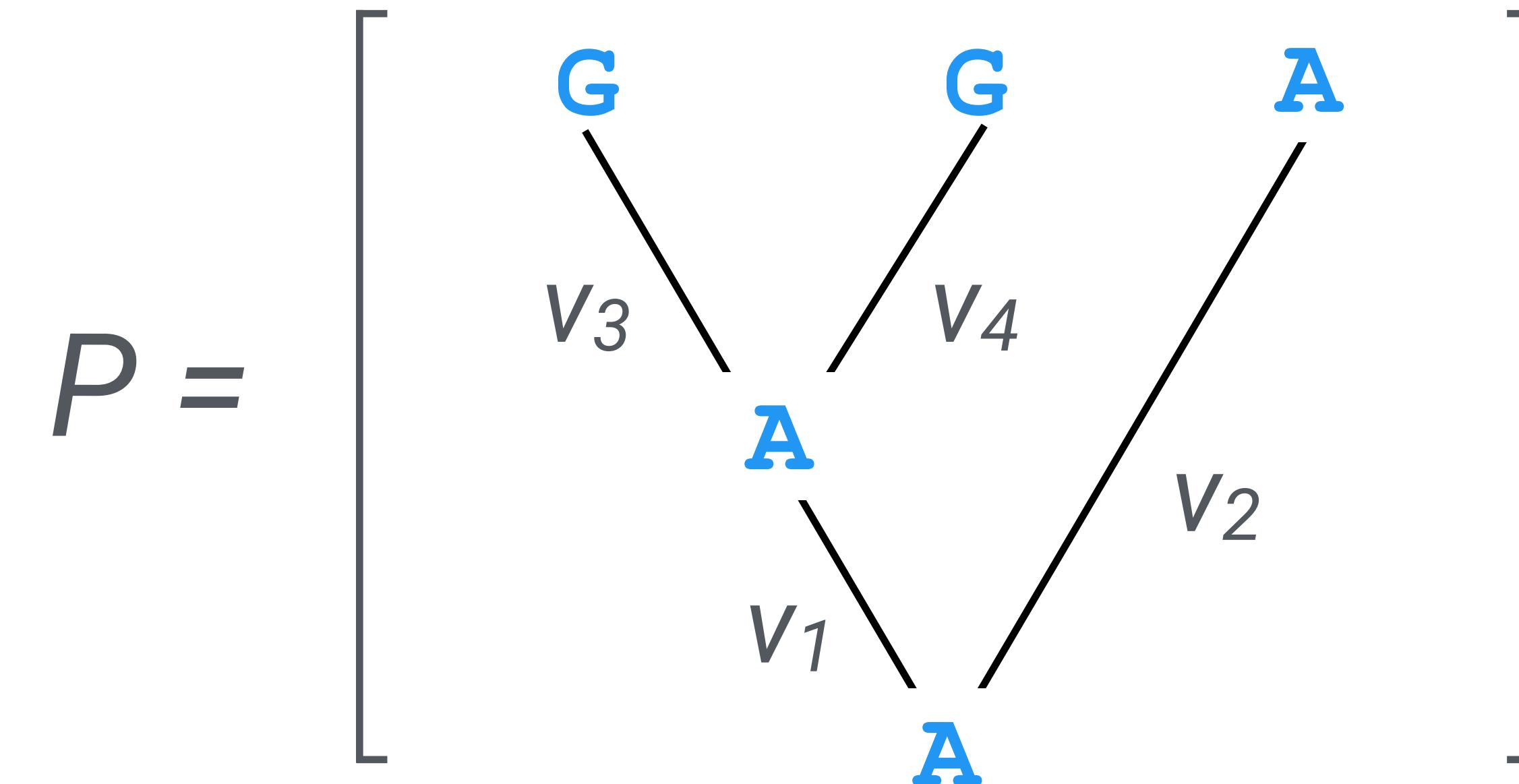
Substitution models

Models of molecular sequence evolution

Also known as substitution / site / character models

They allow us to compute the probability of changing from one state to another over branch length v

Computing the probability of the observed data



Just suppose for now
we know the ancestral
states at internal nodes

$$P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_1)$$

$P_{ij}(v)$ – transition probabilities

Rate matrix

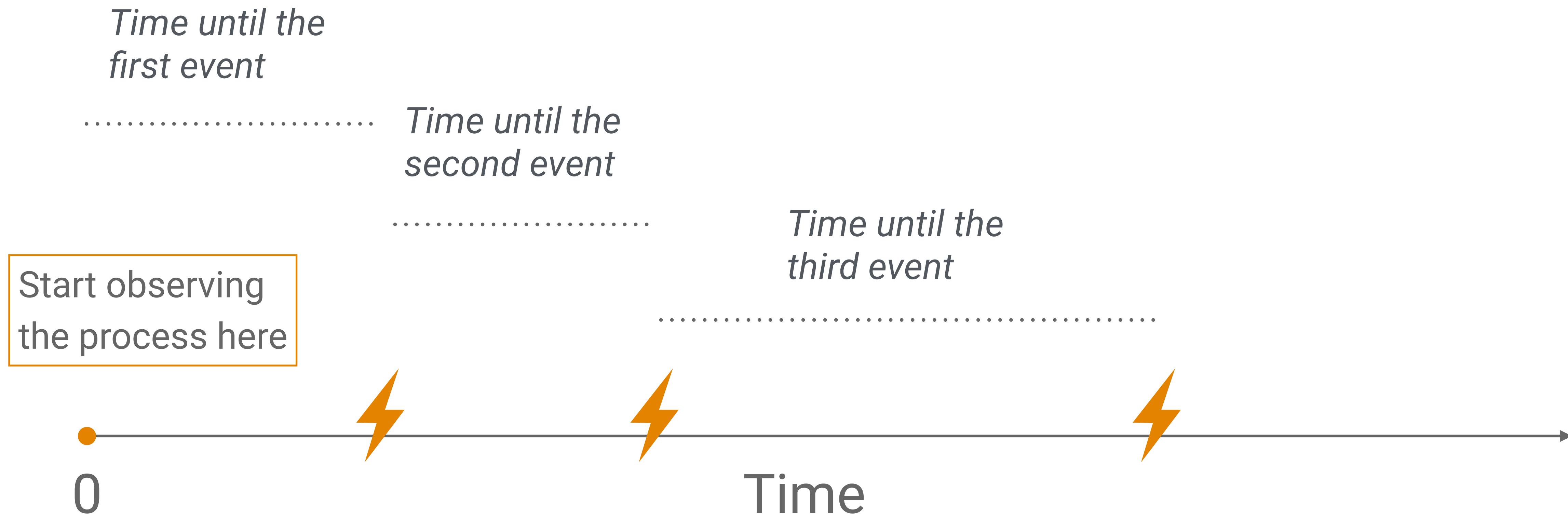
$$Q = \begin{bmatrix} & \text{A} & \text{T} & \text{G} & \text{C} \\ \text{A} & -3\lambda & \lambda & \lambda & \lambda \\ \text{T} & \lambda & -3\lambda & \lambda & \lambda \\ \text{G} & \lambda & \lambda & -3\lambda & \lambda \\ \text{C} & \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}$$

In this model, we only have one parameter, substitution rate parameter λ

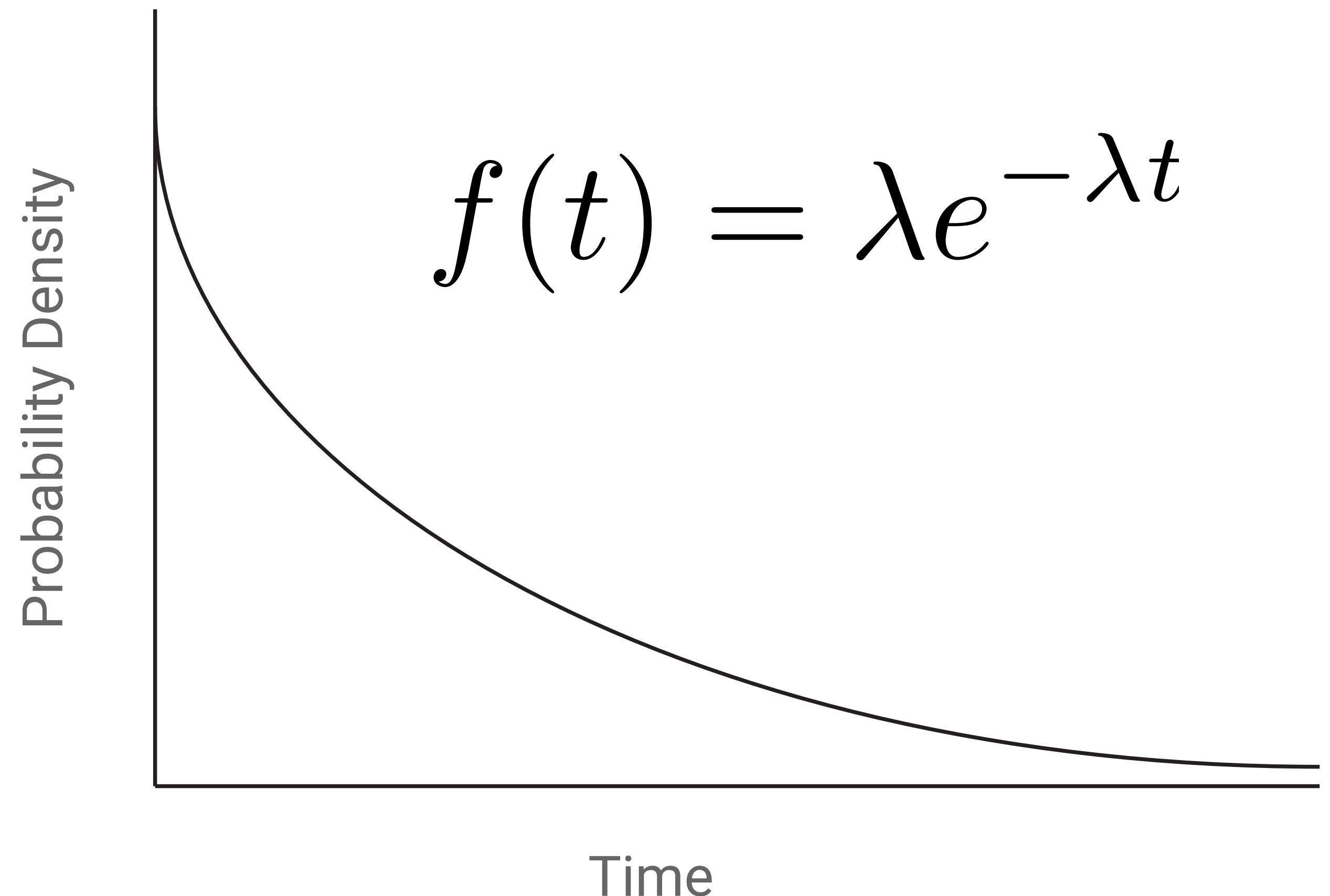
This is the Jukes-Cantor (1969) or JC69 model

Continuous time Markov chain

Nucleotide substitutions (events) occur at a constant rate



The poisson process



The waiting times are
exponentially distributed
random variables

We can use this to calculate
the probability of change
over time (or branch length v)

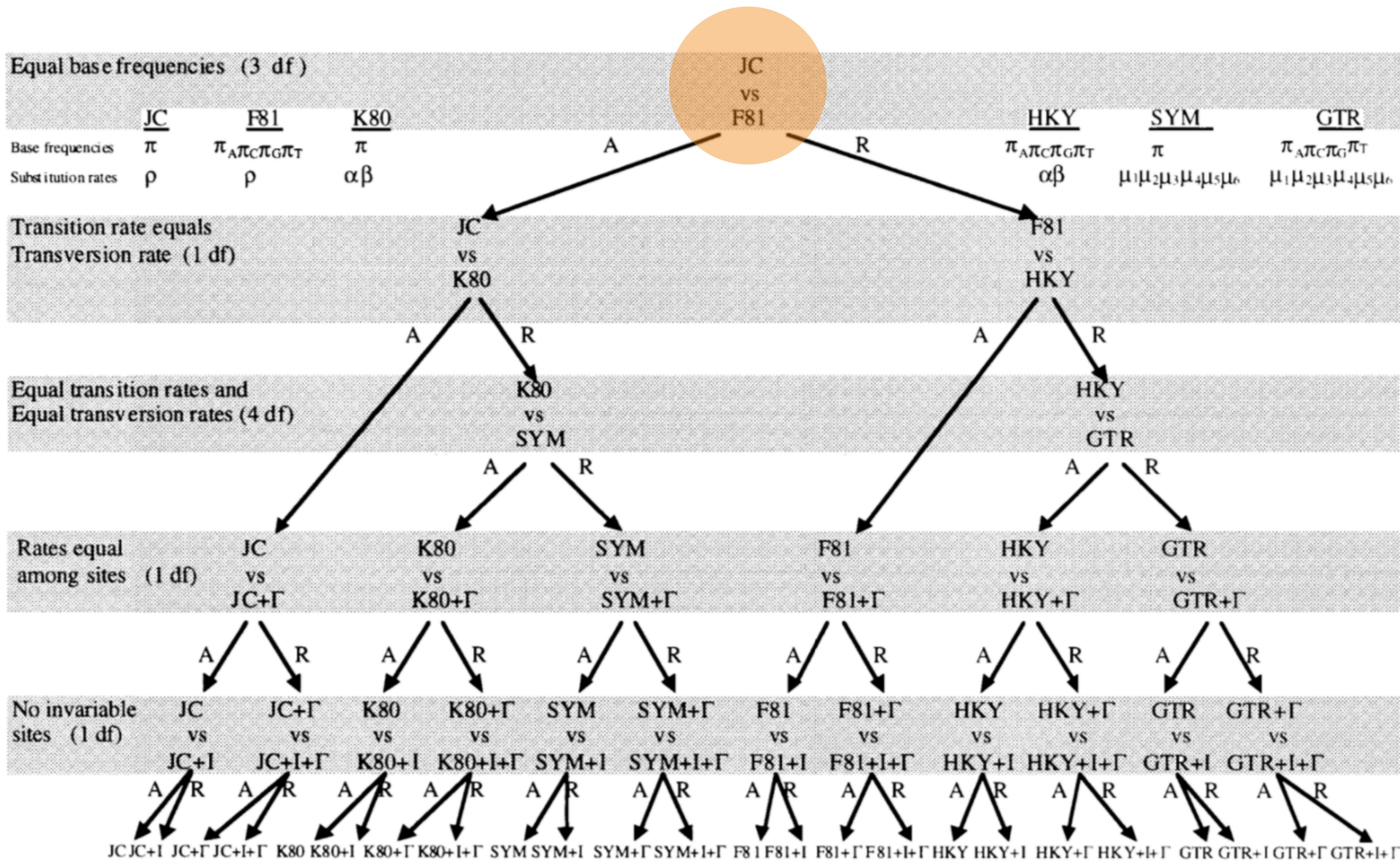
The longer the interval of
time, the more likely we
are to observe change

Exercise

Jukes-Cantor model transition probability applet

Written by Paul Lewis

Other substitution models



Base frequencies

The JC69 model assumes equal transition rates and equal base frequencies

Base frequencies are the proportion of each nucleotide in the dataset

If a given nucleotide appears in our dataset at a low frequency, we are less likely to observe a transition to that state

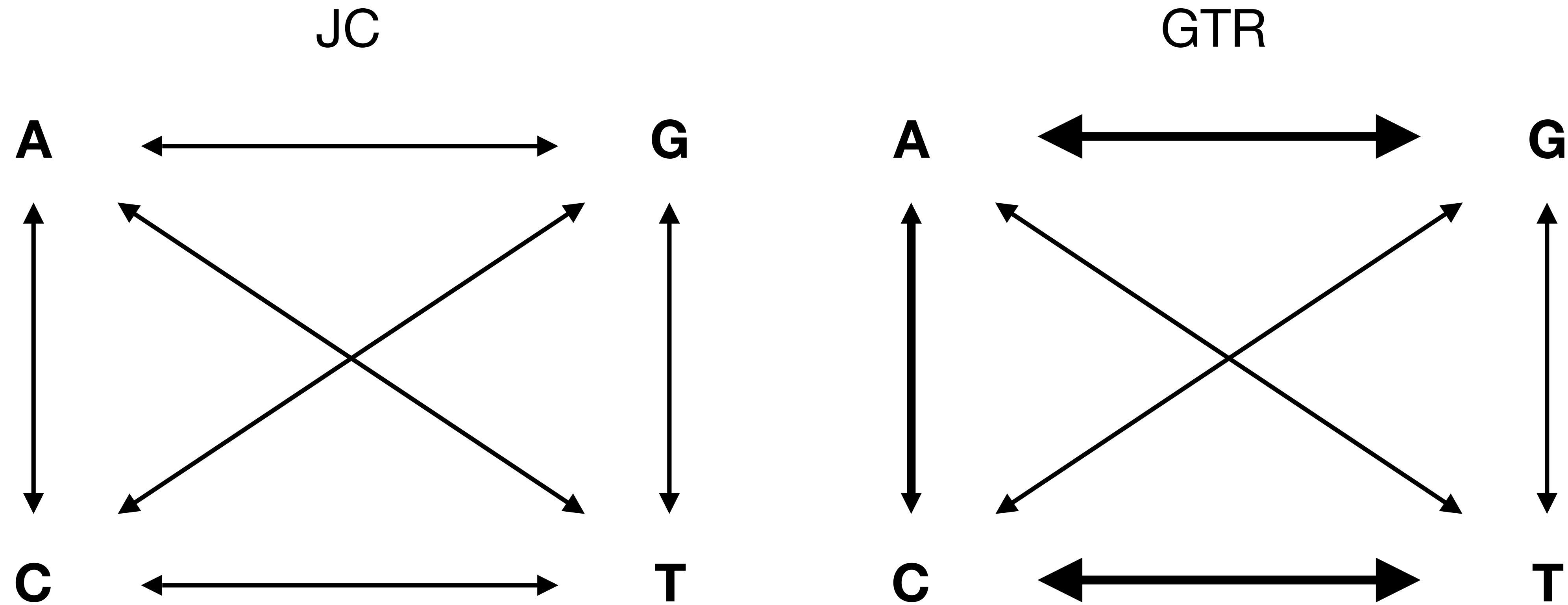
The general time reversible model

$$Q = \begin{pmatrix} * & \mu_{AG}\pi_G & \mu_{AC}\pi_C & \mu_{AT}\pi_T \\ \mu_{GA}\pi_A & * & \mu_{GC}\pi_C & \mu_{GT}\pi_T \\ \mu_{CA}\pi_A & \mu_{CG}\pi_G & * & \mu_{CT}\pi_T \\ \mu_{TA}\pi_A & \mu_{TG}\pi_G & \mu_{TC}\pi_C & * \end{pmatrix}$$

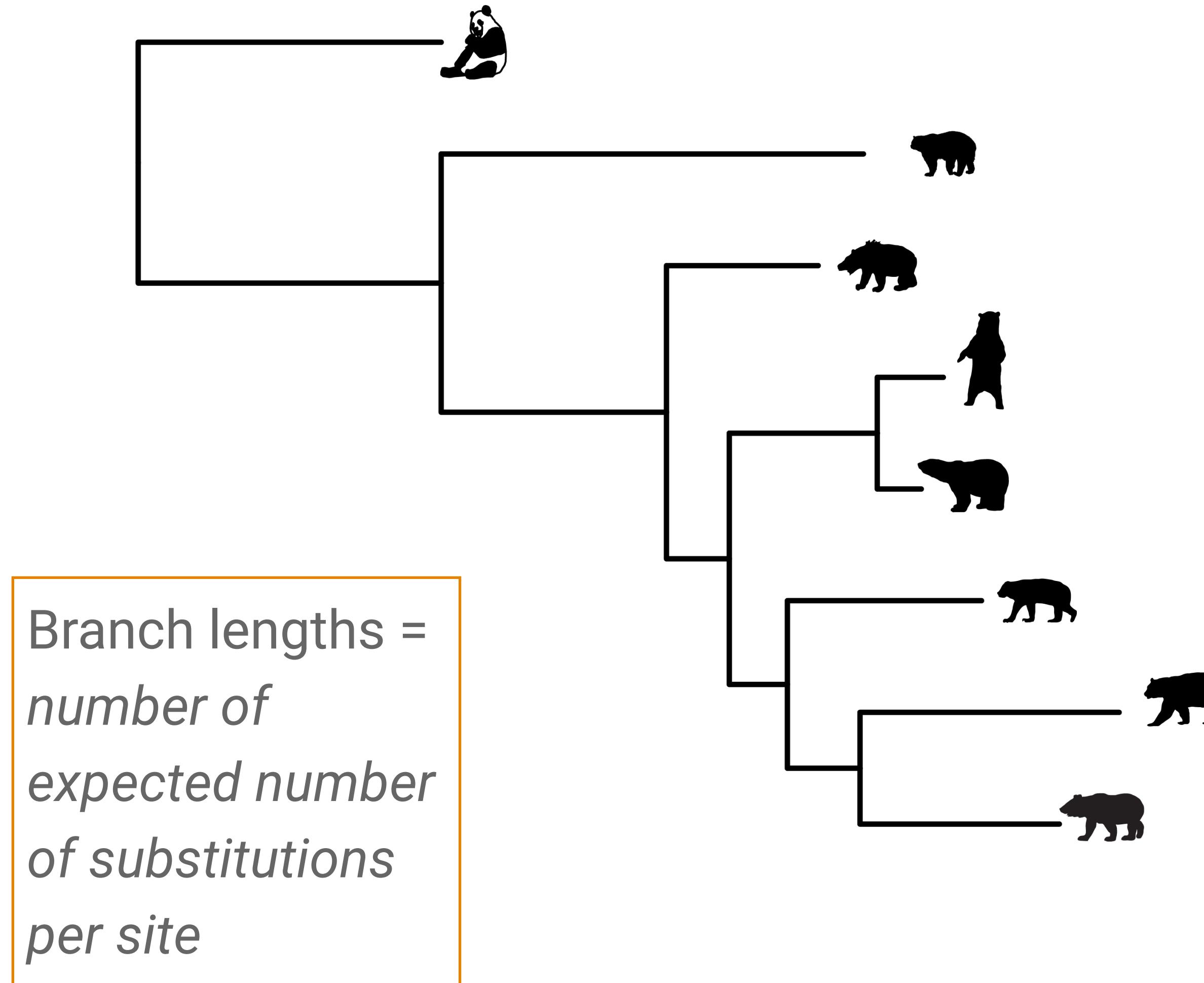
Allows for unequal transition rates (μ) and unequal base frequencies (π)

Note the rates are symmetric – e.g., the rate of change between A and T, is the same in both directions – but the frequency of each character state also affects the probability of change

The JC versus GTR models



Branch lengths



Branch lengths are a product of rate and time

Without temporal information we can only measure relative genetic distance

Felsenstein's pruning algorithm

The following slides are adapted from John Huelsenbeck (c/o Sebastian Höhna)

Computing the probability of the observed data

$$P = \left[\begin{array}{c} G \\ v_3 \\ & \swarrow \\ & A \\ & v_1 \\ & \searrow \\ & A \\ & v_4 \\ & \swarrow \\ & v_2 \\ & \searrow \\ & A \\ & v_1 \\ & \swarrow \\ & v_2 \end{array} \right]$$

Just suppose for now
we know the ancestral
states at internal nodes

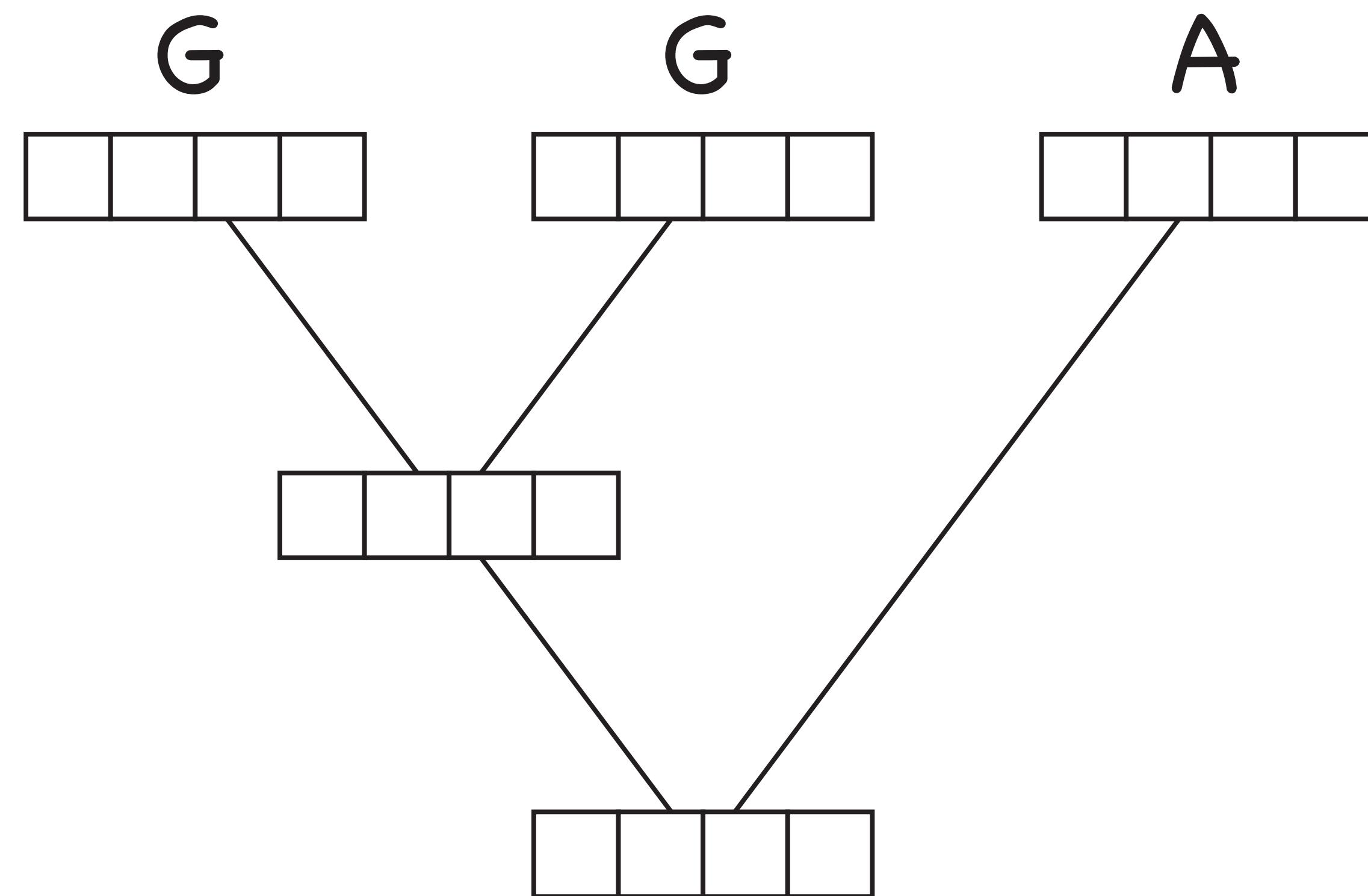
$$\pi_A \times P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_1)$$

$P_{ij}(v)$ – transition probabilities
 π_i – stationary frequencies

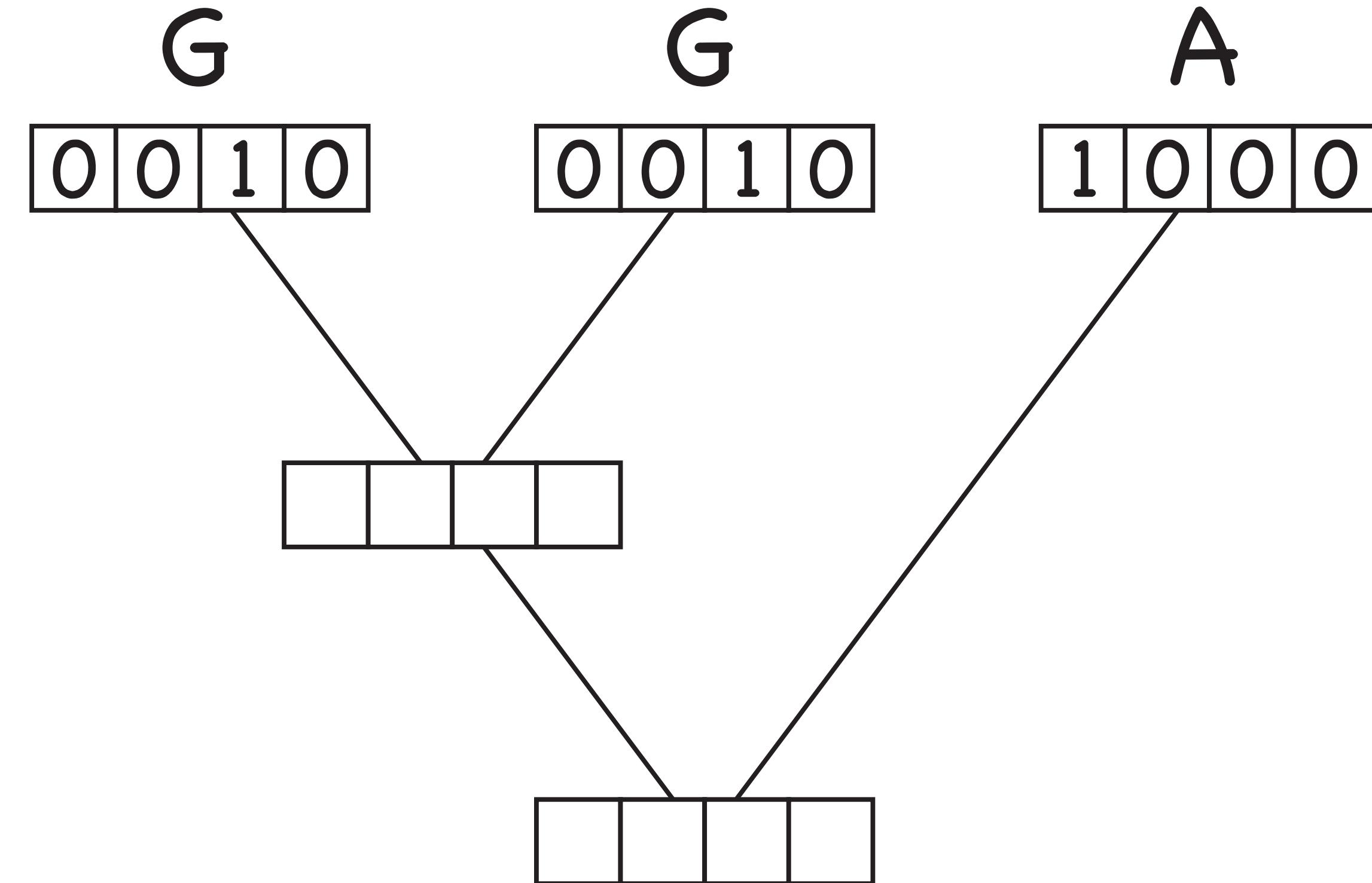
$$\Pr \left[\begin{array}{c} G \\ & A \\ & | \\ & A \\ & | \\ A & & A \end{array} \right] + \Pr \left[\begin{array}{c} G \\ & A \\ & | \\ & C \\ & | \\ A & & A \end{array} \right] + \Pr \left[\begin{array}{c} G \\ & A \\ & | \\ & G \\ & | \\ A & & A \end{array} \right] + \Pr \left[\begin{array}{c} G \\ & A \\ & | \\ & T \\ & | \\ A & & A \end{array} \right] +$$

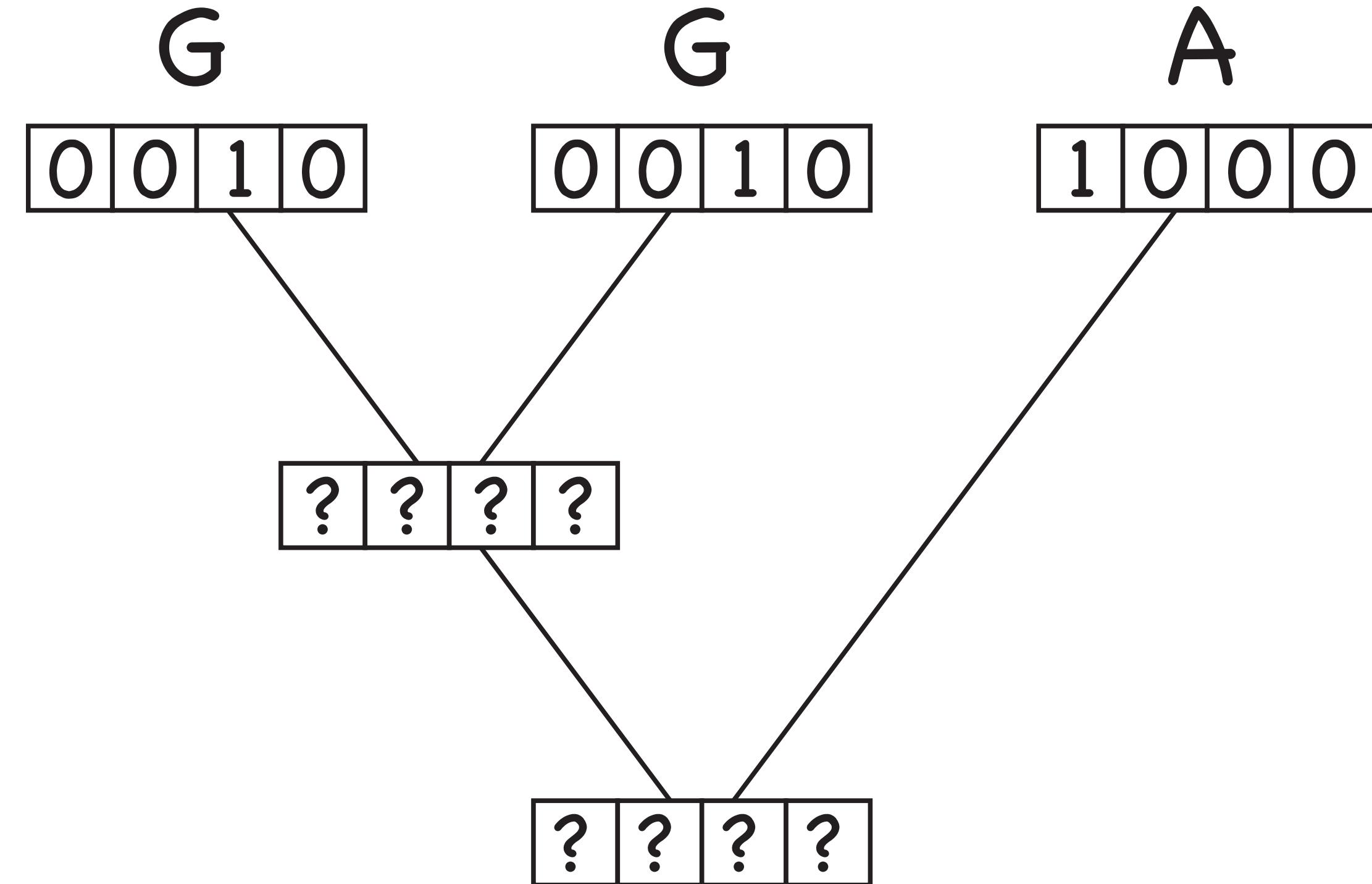
$$\Pr \begin{bmatrix} G & & & A \\ & G & G & \\ & & & A \\ & & A & \end{bmatrix} + \Pr \begin{bmatrix} G & & & A \\ & G & G & \\ & & & A \\ & & C & \end{bmatrix} + \Pr \begin{bmatrix} G & & & A \\ & G & G & \\ & & & A \\ & & G & \end{bmatrix} + \Pr \begin{bmatrix} G & & & A \\ & G & G & \\ & & & A \\ & & T & \end{bmatrix} +$$

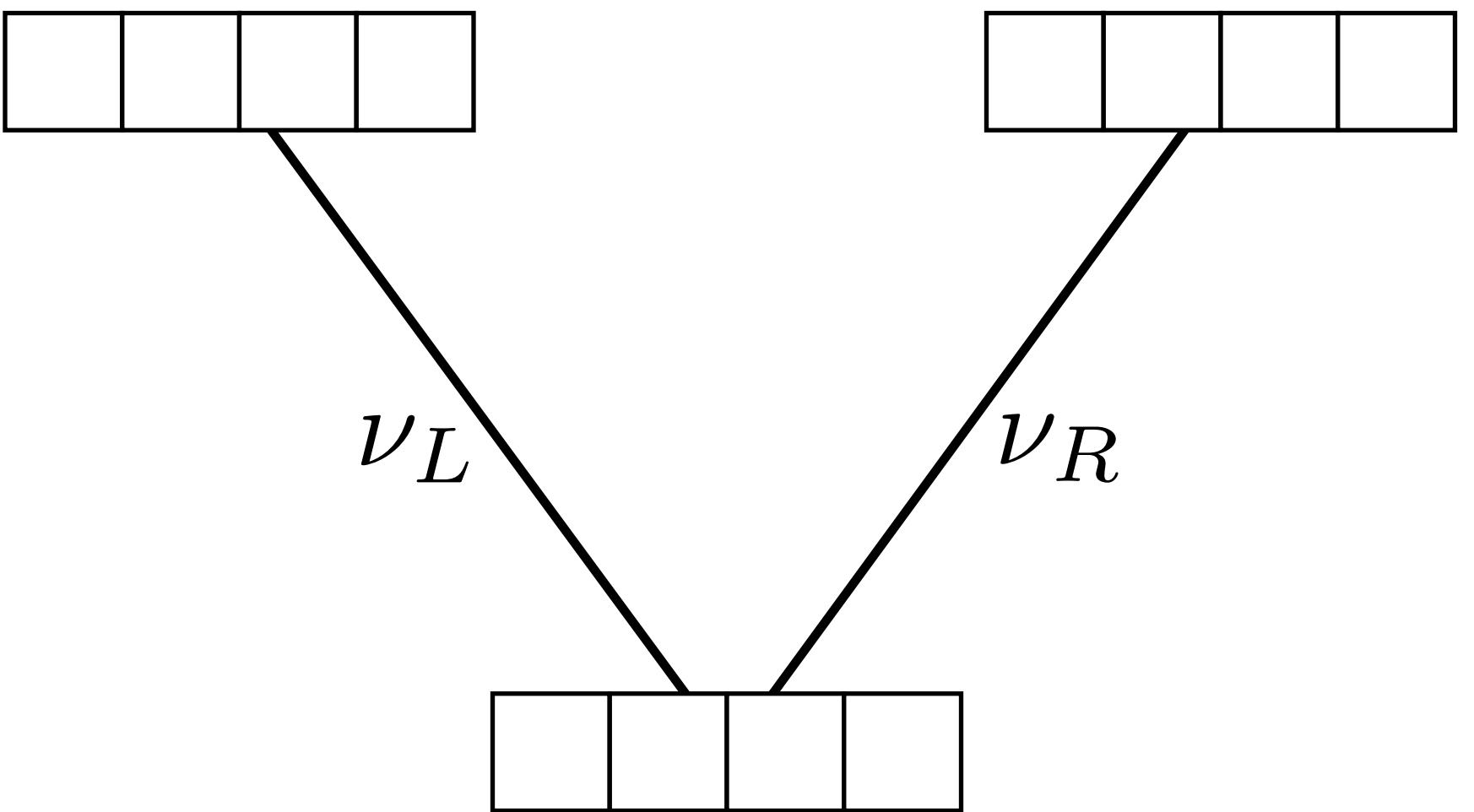
$$\Pr \begin{bmatrix} G \\ T \\ A \end{bmatrix} + \Pr \begin{bmatrix} G \\ T \\ C \end{bmatrix} + \Pr \begin{bmatrix} G \\ T \\ G \end{bmatrix} + \Pr \begin{bmatrix} G \\ T \\ T \end{bmatrix}$$



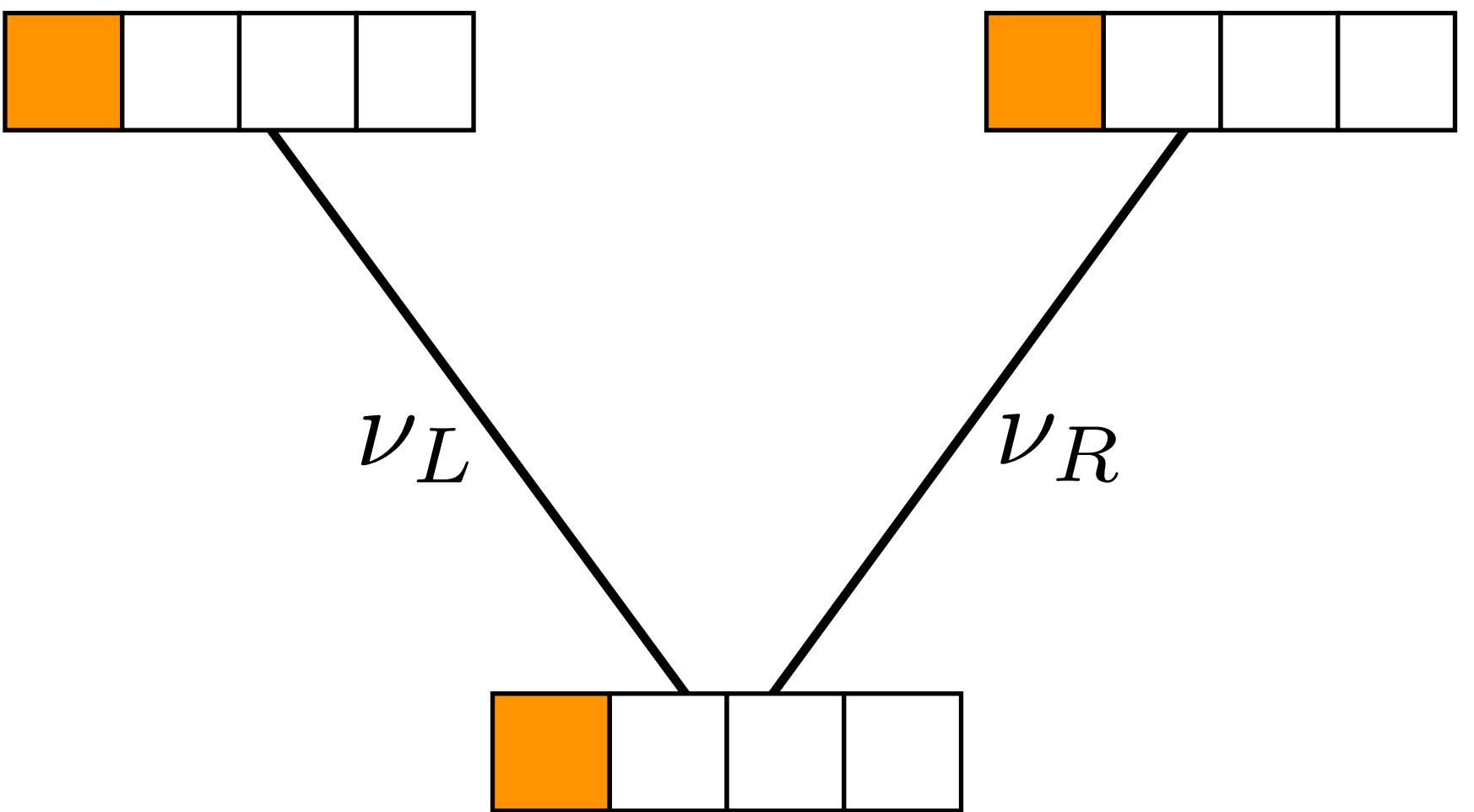
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach.



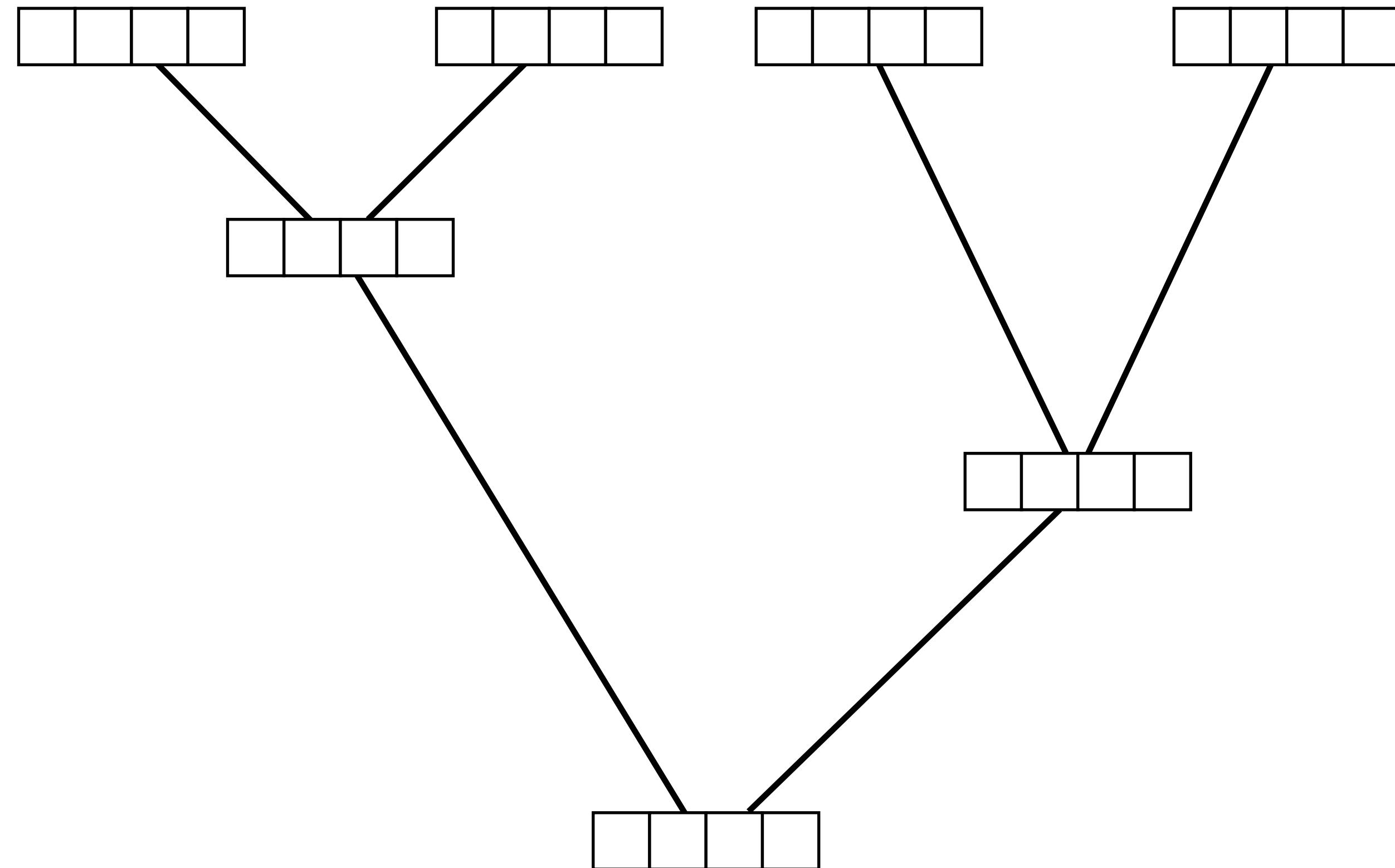


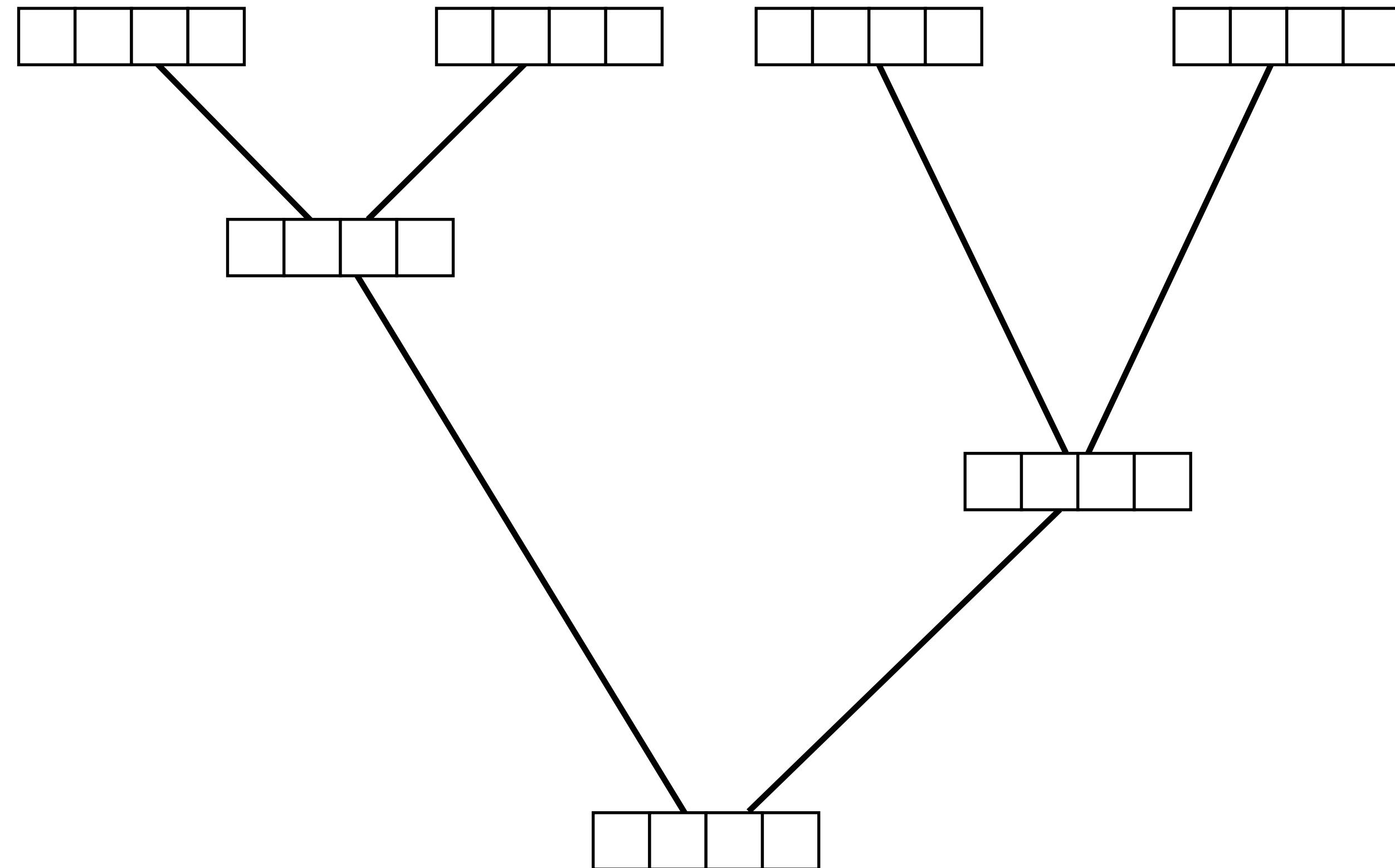


$$\ell_i = \left(\sum_j p_{ij}(\nu_L) \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R) \ell_j^R \right)$$

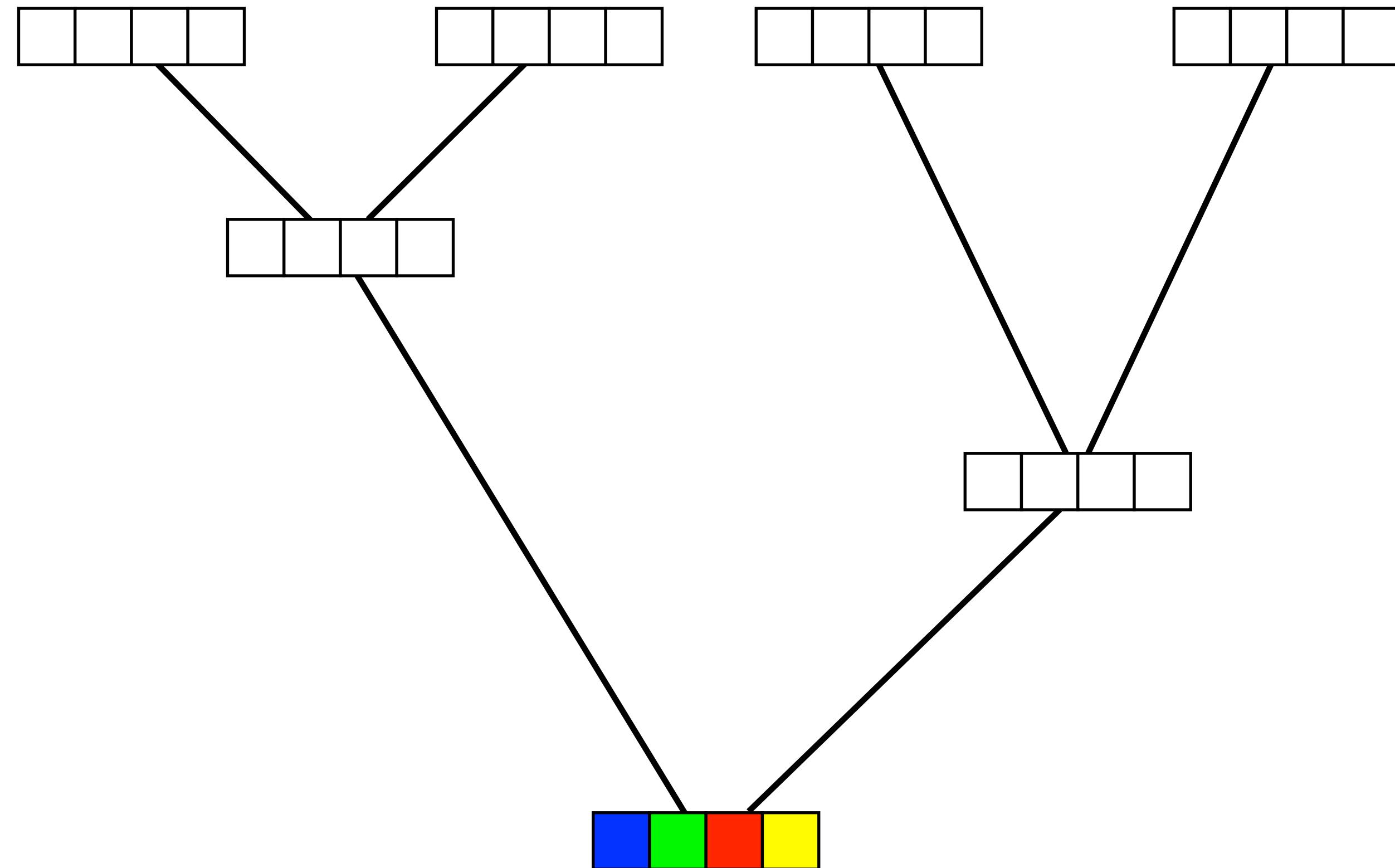


$$\ell_i = \left(\sum_j p_{ij}(\nu_L) \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R) \ell_j^R \right)$$





$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$



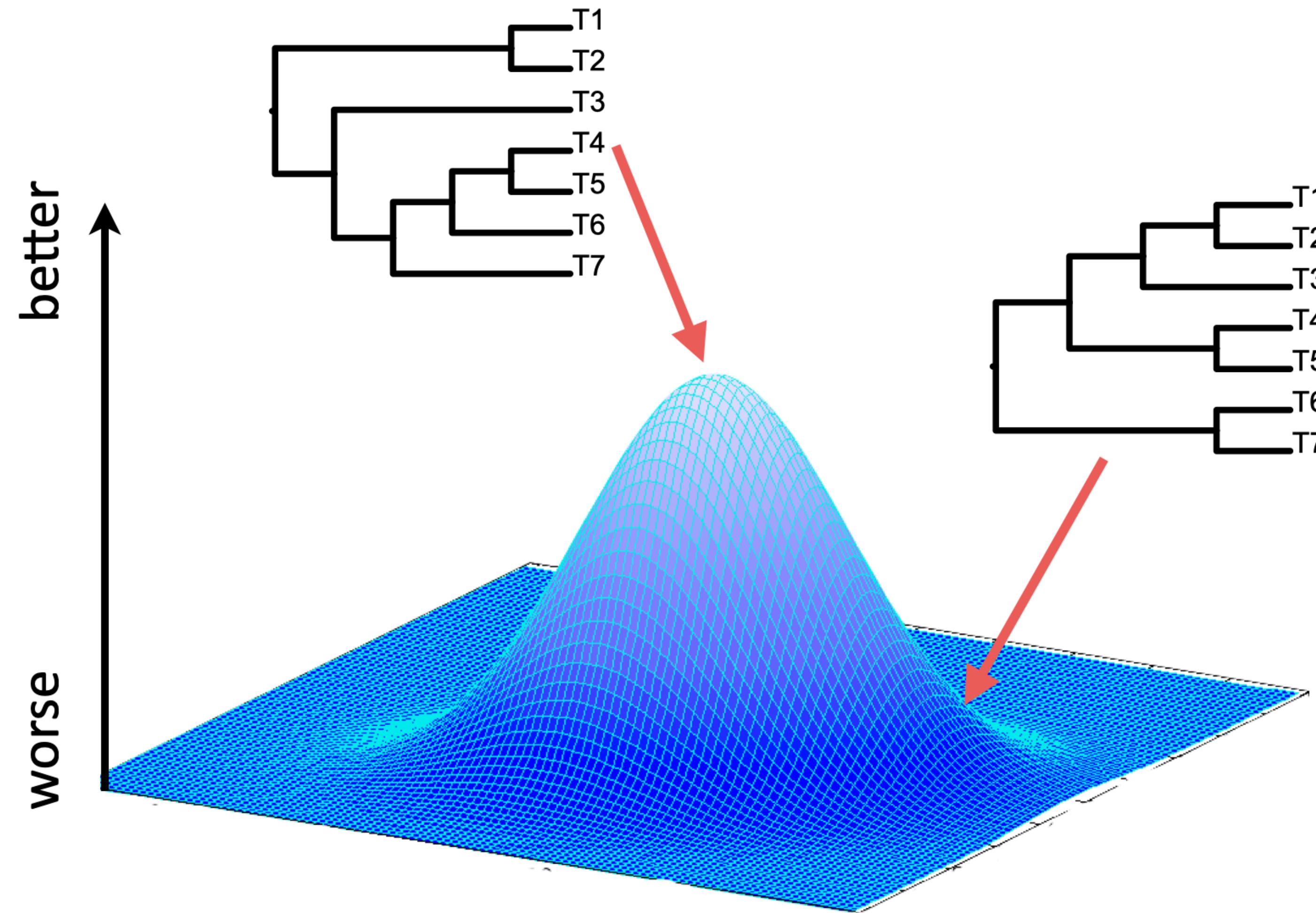
$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

$$\ell_i = \left(\sum_j p_{ij}(\nu_L) \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R) \ell_j^R \right)$$

$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

Maximum likelihood

How do we find the ‘best’ tree?



It depends how you measure ‘best’

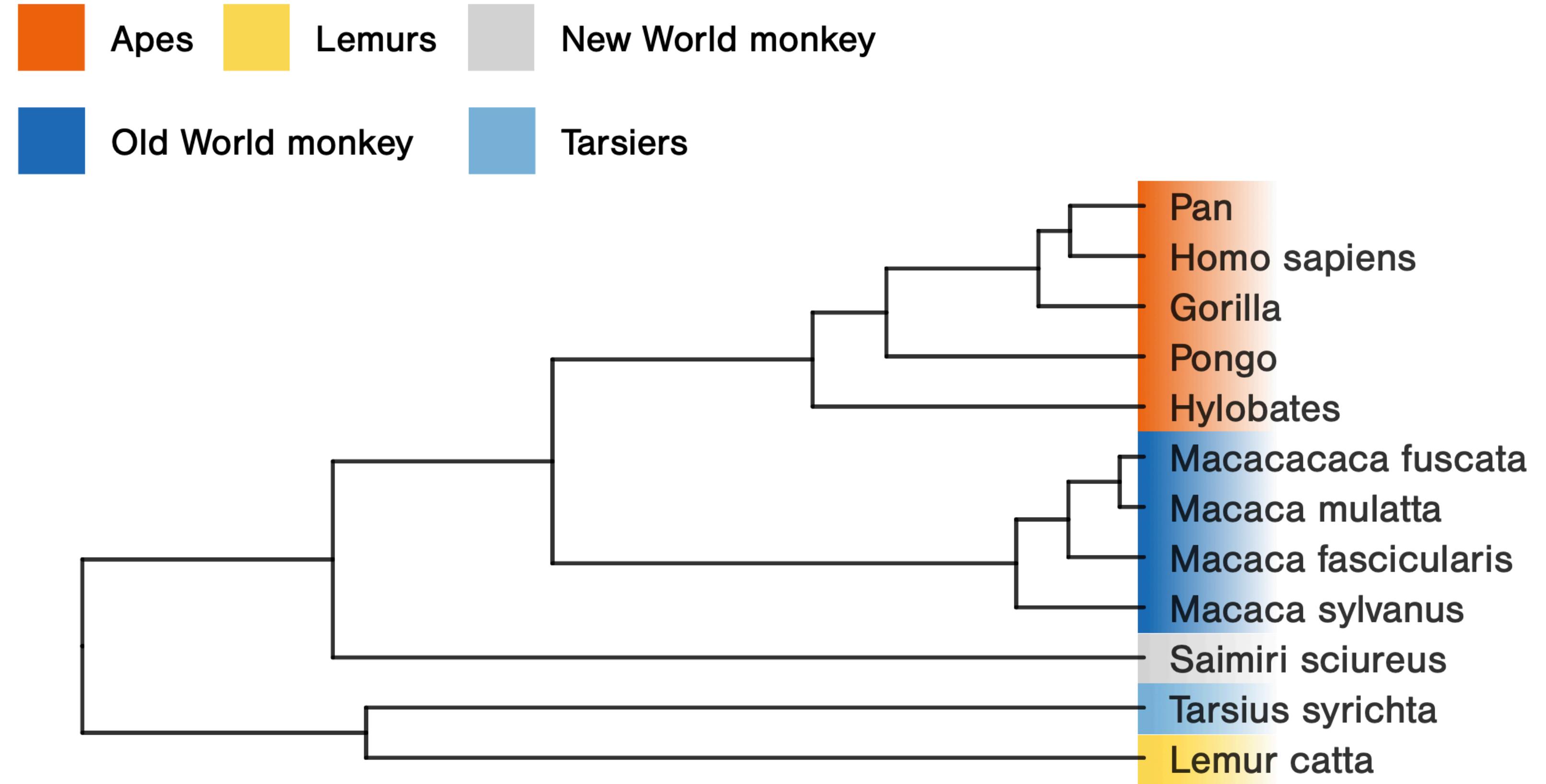
Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Likelihood score (probability), optimised over branch lengths and model parameters
Bayesian inference	Posterior probability, integrating over branch lengths and model parameters

Both maximum likelihood and Bayesian inference are model-based approaches

Note these are not the only approaches to tree-building but they are the most widely used

Exercise

Tree building using likelihood



Homework



Install the software [RevBayes](#)



[Phylogenetics primer part 1: tree terminology and substitution models Paul Lewis](#)

→ See the question guide