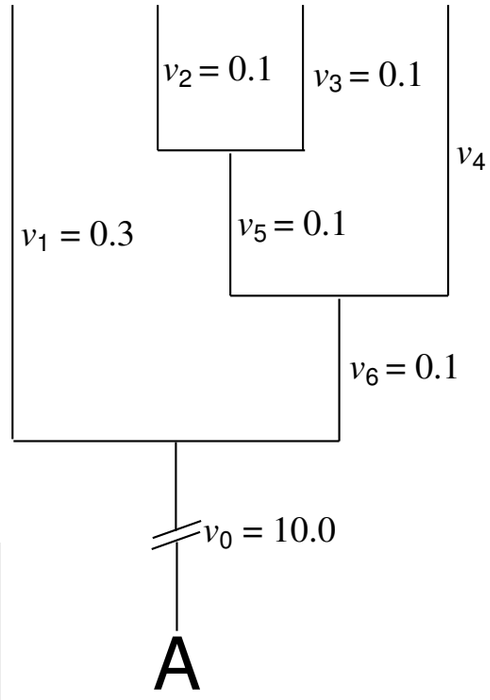
Algorithm to simulate data on a tree

Required: a tree with branch lengths and a mutation transition rate matrix



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

1. $\tau = 0$; nuc $= \{\}$
2. Do forever
3. find λ — if you are at **G** use the diagonal value row of q_G: 0.886
4. draw random number r between 0 and 1 — $r_1 = 0.134$
5. calculate $t = -\log_e(r_1)\frac{1}{\lambda}$ — e.g. $t = -\log(0.134)\frac{1}{0.886} = 0.98521$
6. $\tau = \tau + t$ — $\tau = 0 + 0.98521$
7. if $\tau < v$ — $v = 10.0$

   return current nucleotide nuc
8. calculate change of nucleotide cumulative sum — $s = [0, \frac{q_{GA}}{q_{GG}}, \frac{q_{GA}}{q_{GG}} + \frac{q_{GC}}{q_{GG}}, 1.0]$

   $s = [0, \frac{1.266}{1.519}, \frac{1.266}{1.519} + \frac{0.190}{1.519}, 1.0]$
9. draw random number r between 0 and 1 — $r_2 = 0.912$
10. pick interval in which r_2 lays — it is in the interval $(0.833, 0.958]$ and thus nuc=**C**
11. goto 2

Here is a second round of the example above: we are at **C** now

1. do forever
2. $\lambda = 0.696$ using row q_C
3. $r_1 = 0.449$
4. $t = -\log_e(0.449)\frac{1}{0.696} = 0.49964$
5. $\tau = 0.98521 + 0.49964 = 1.48485$
6. if $(\tau = 0.98521) < 10.0$

   return current nucleotide nuc
7. $s = [0, \frac{0.253}{0.696}, \frac{0.253}{0.696} + \frac{0.127}{0.696}, 1.0]$
8. $r_2 = 0.191$
9. pick interval in which r_2 lays it is in the interval $(0, 0.363]$ and thus nuc=**A**
10. goto 2

$$\begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Details and explanation see mutation model handout

GTR

$$a, b, c, d, e, \pi_A, \pi_C, \pi_G$$
$$a + b + c + d + e + f = 1,$$
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

+all rates are different

TN

$$a = c = d = f, b = e, \pi_A, \pi_C, \pi_G$$
$T_i/T_v, Y/R$ ratio and three frequencies since $\pi_A + \pi_C + \pi_G + \pi_T = 1$

+Purine/Pyrimidins

HKY F84

$$a = f = c = d = 1, b = e, \pi_A, \pi_C, \pi_G$$
T_i/T_v ratio and three frequencies since $\pi_A + \pi_C + \pi_G + \pi_T = 1$

+Transition/Transversion

+uneqal base freq.

F81

$$a = b = c = d = e = f = 1, \pi_A, \pi_C, \pi_G$$
and three frequencies since $\pi_A + \pi_C + \pi_G + \pi_T = 1$

K2P

$$a = f = c = d = 1, b = e,$$
$$\pi_A = \pi_C = \pi_G = \pi_T$$
T_i/T_b ratio

+uneqal base freq.

+Transition/Transversion

JC

$$a = b = c = d = e = f = 1,$$
$$\pi_A = \pi_C = \pi_G = \pi_T$$

The Plan

- Probability review
- Likelihood

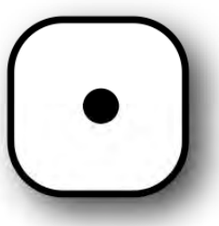
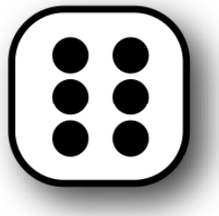
- The AND and OR rules
- Independence of events

- What does it mean?
- Likelihood of a single sequence
- Maximum likelihood distances
- Likelihoods of trees

Combining probabilities

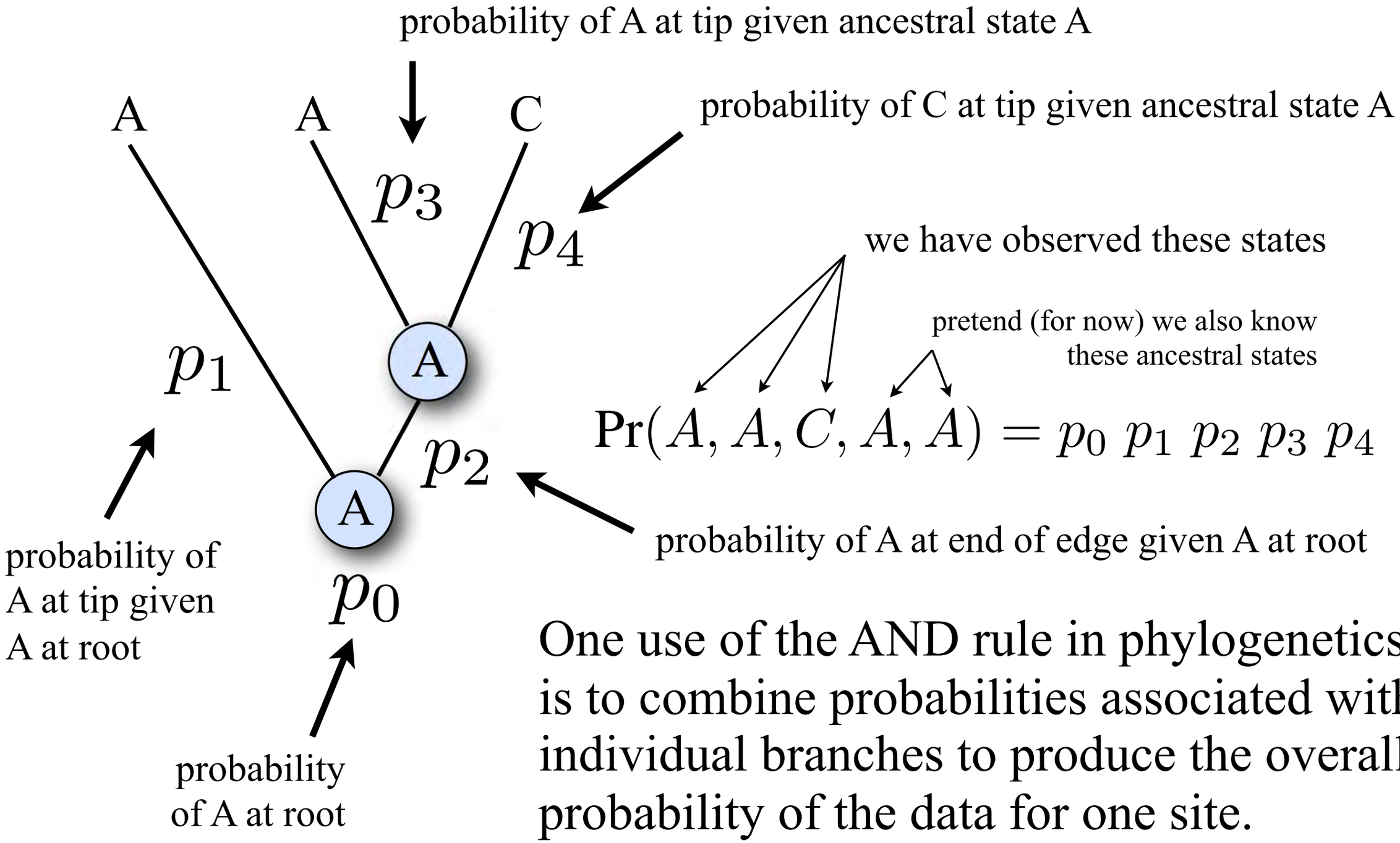
- *Multiply* probabilities if the component events must happen **simultaneously** (i.e. where you would naturally use the word AND when describing the problem)

Using 2 dice, what is the probability of


AND

?

$$(1/6) \times (1/6) = 1/36$$

AND rule in phylogenetics

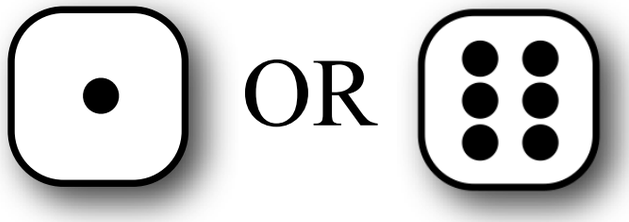


One use of the AND rule in phylogenetics is to combine probabilities associated with individual branches to produce the overall probability of the data for one site.

Combining probabilities

- *Add* probabilities if the component events are **mutually exclusive** (i.e. where you would naturally use the word OR in describing the problem)

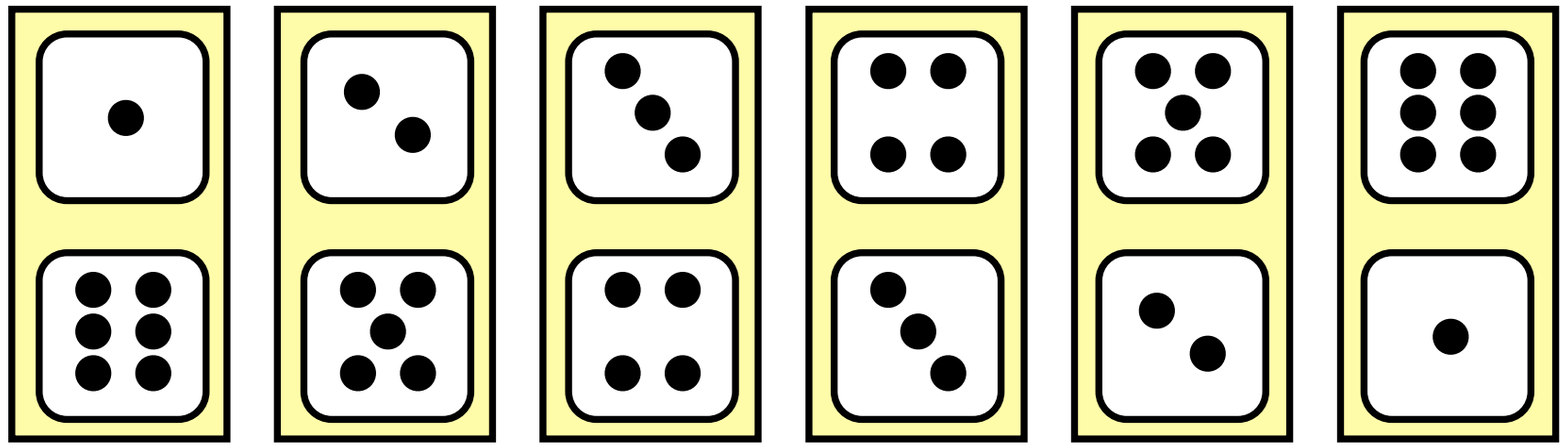
Using one die, what is the probability of

 OR ?

$(1/6) + (1/6) = 1/3$

Combining AND and OR

What is the probability that the sum of two dice is 7?



$$(1/36) + (1/36) + (1/36) + (1/36) + (1/36) + (1/36) = 1/6$$

1 and 6

2 and 5

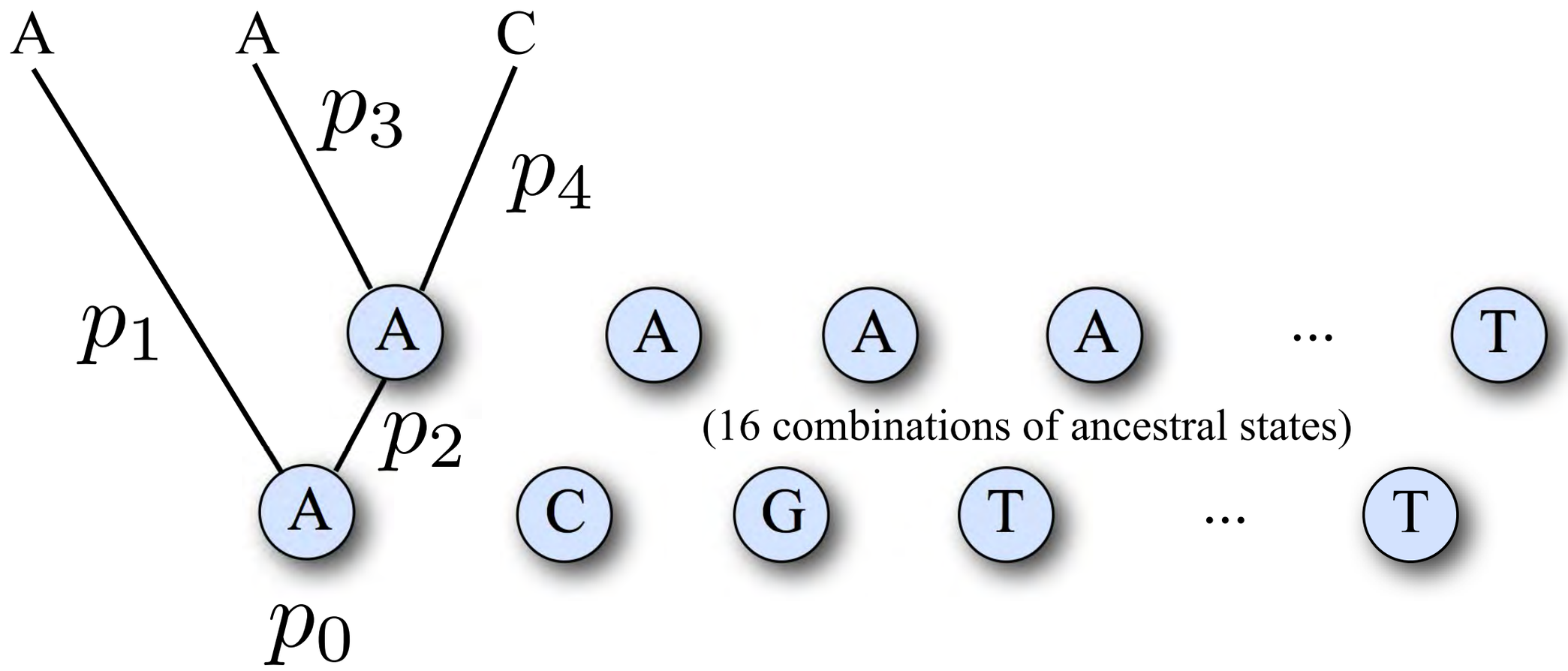
3 and 4

4 and 3

5 and 2

6 and 1

Using both AND and OR in phylogenetics



AND rule used to compute probability of the observed data for *each combination* of ancestral states.

OR rule used to combine different combinations of ancestral states.

Independence

This is always true...

$$\Pr(\text{A and B}) = \Pr(\text{A}) \Pr(\text{B}|\text{A})$$

joint probability

conditional probability

If we can say this...

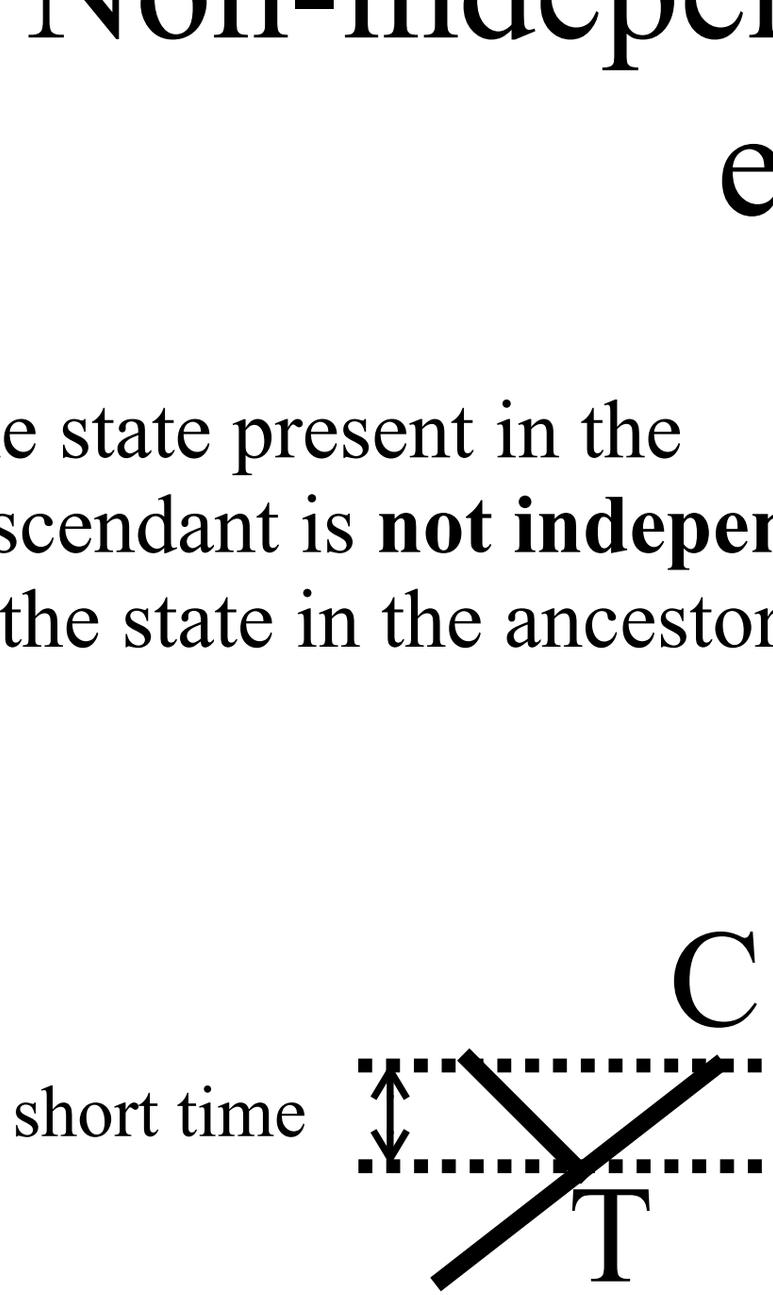
$$\Pr(\text{B}|\text{A}) = \Pr(\text{B})$$

...then events A and B are **independent** and we can express the joint probability as the product of Pr(A) and Pr(B)

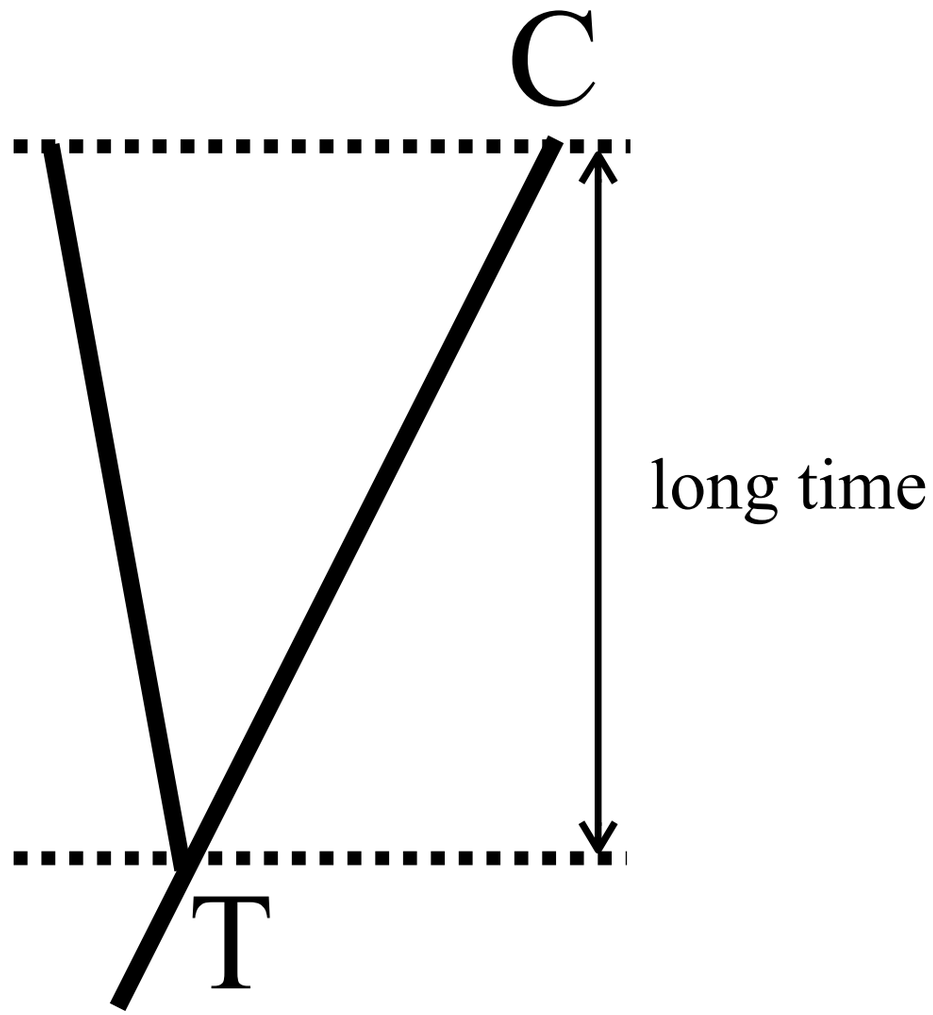
$$\Pr(\text{A and B}) = \Pr(\text{A}) \Pr(\text{B})$$

Non-independence in molecular evolution

The state present in the descendant is **not independent** of the state in the ancestor



less probable



more probable

Conditional Independence

Assume both A and B depend on C:

$$\Pr(A|C) \neq \Pr(A) \quad \Pr(B|C) \neq \Pr(B)$$

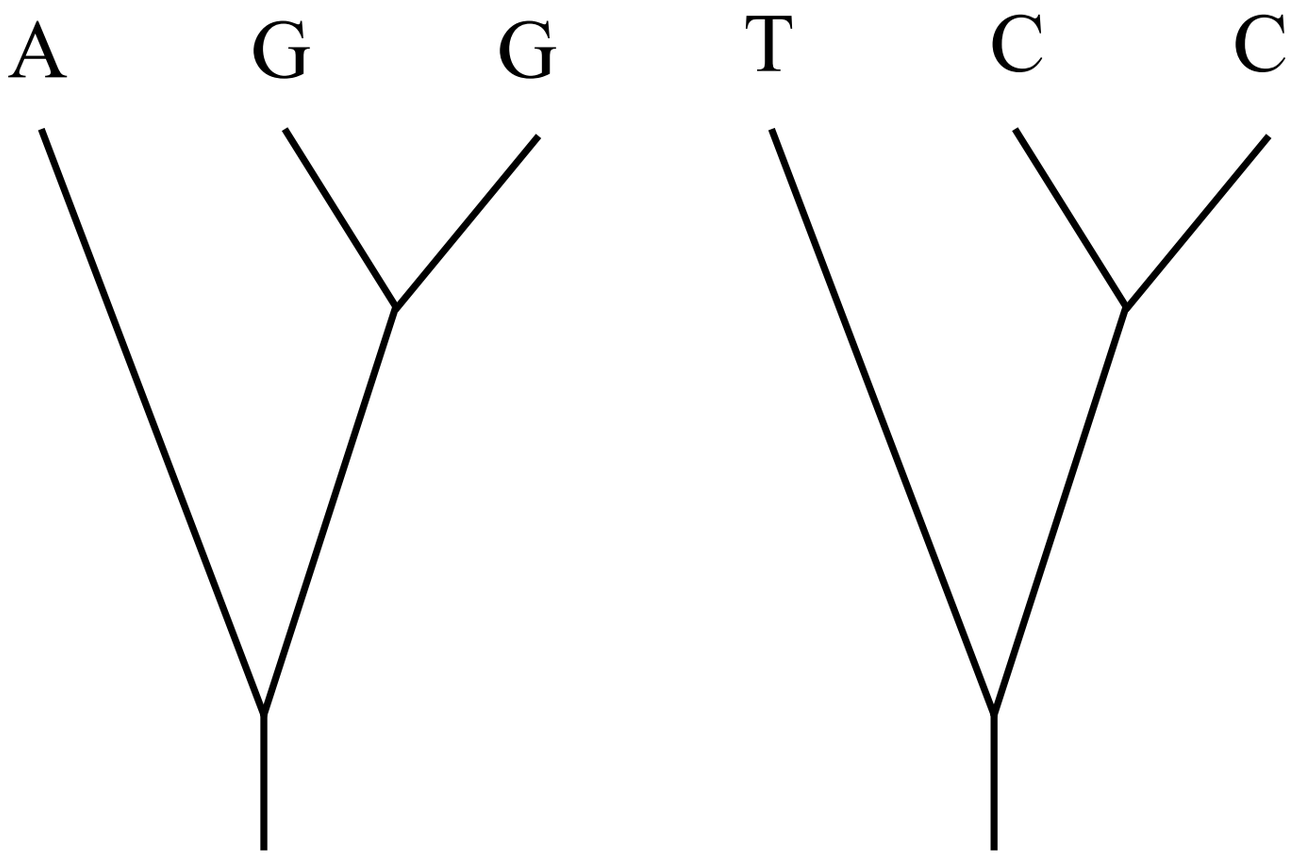
If we can say this...

$$\Pr(B|A,C) = \Pr(B|C)$$

...then events A and B are **conditionally independent** and we can express the joint (conditional) probability as the product of Pr(A|C) and Pr(B|C)

$$\Pr(A \text{ and } B|C) = \Pr(A|C)\Pr(B|C)$$

Conditional independence in molecular evolution



The site data patterns AGG and TCC are assumed by most models to be conditionally independent.

The patterns both depend on the underlying tree (including edge lengths) and the substitution model.

$$\Pr(\text{AGG and TCC}|\text{tree}, \text{model}) = \Pr(\text{AGG}|\text{tree}, \text{model})\ \Pr(\text{TCC}|\text{tree}, \text{model})$$

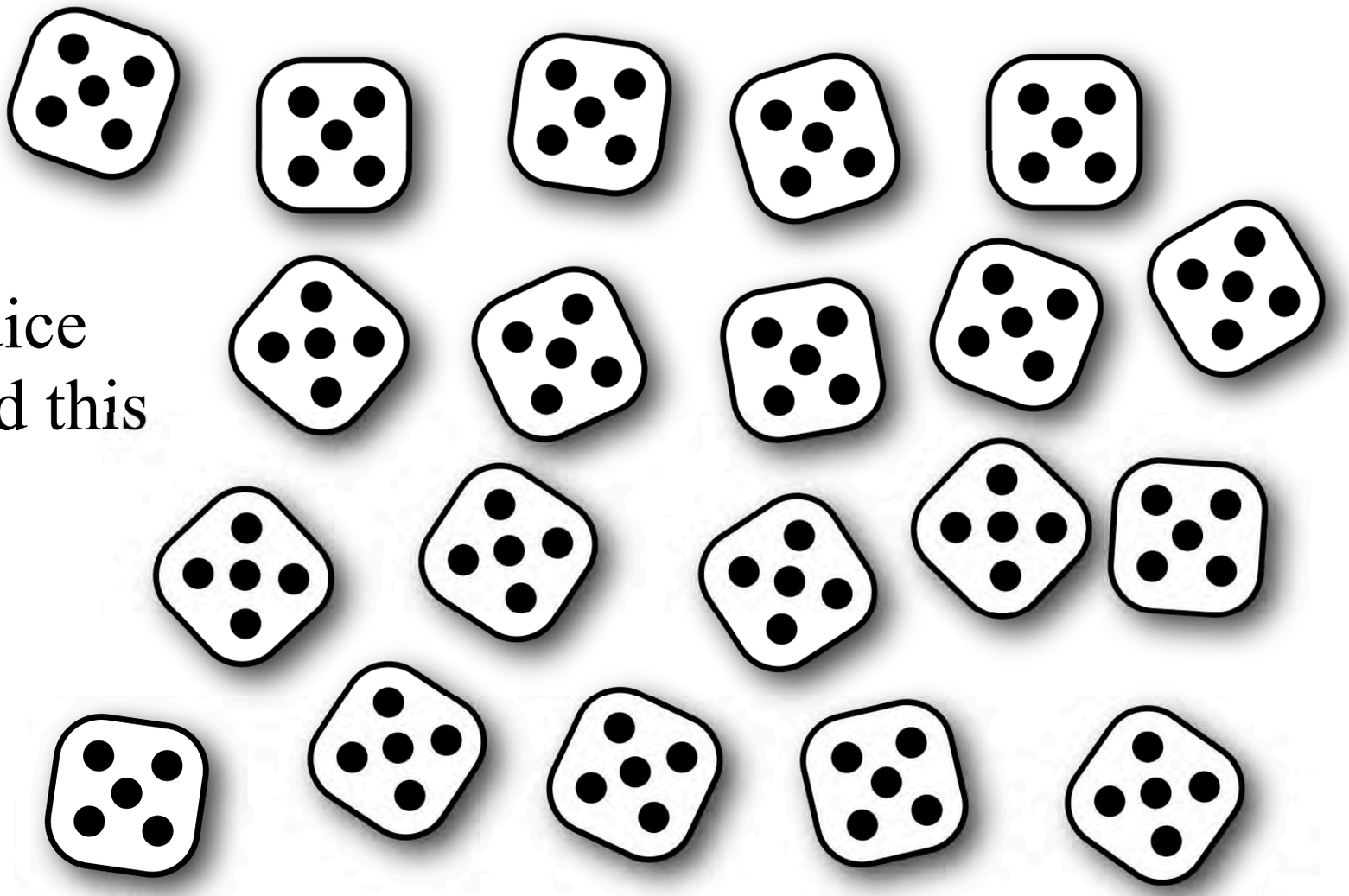
Likelihood

The Likelihood Criterion

The probability of the observations computed using a model tells us how surprised we should be.

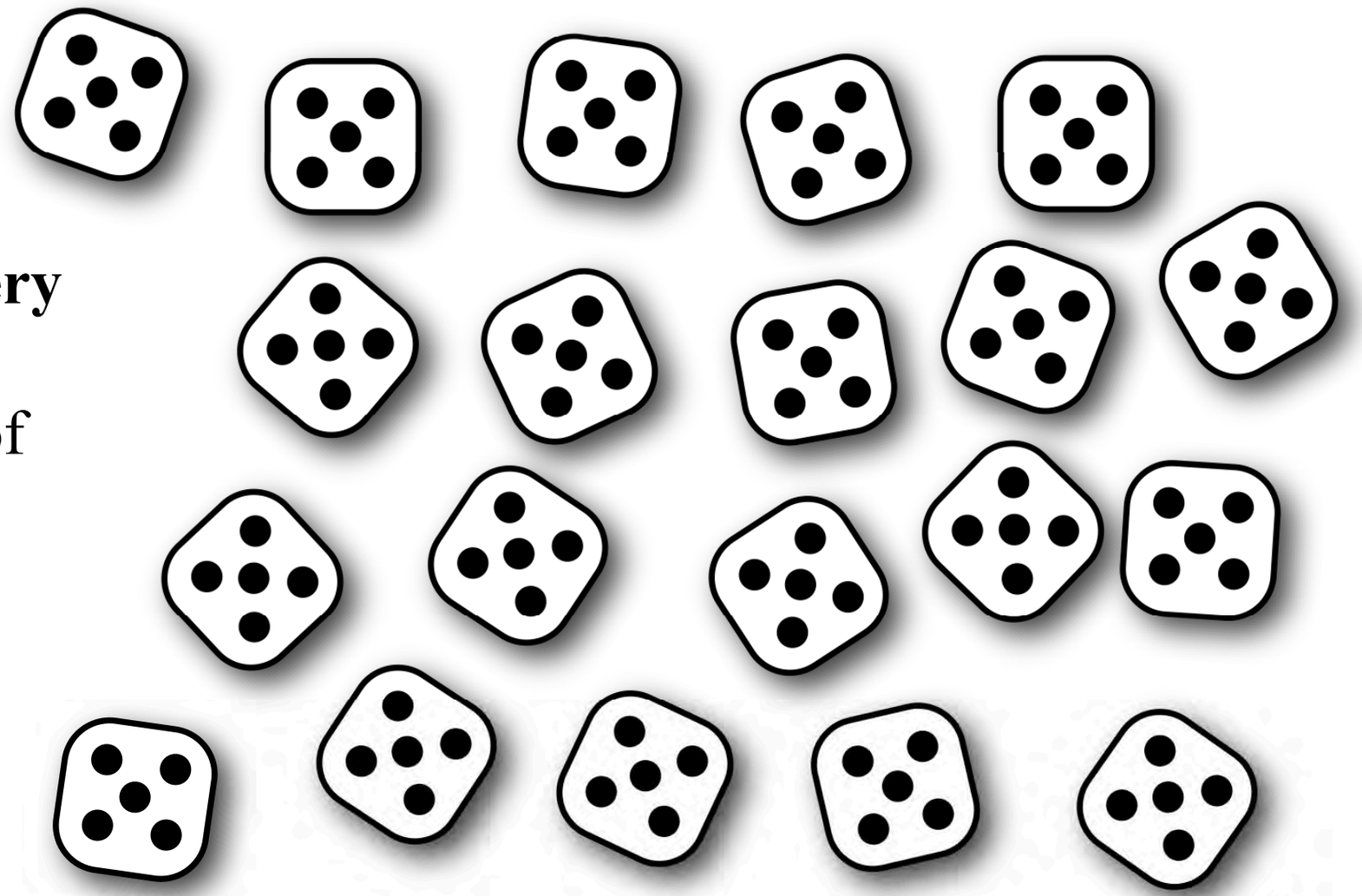
The preferred model is the one that surprises us least.

Suppose I threw 20 dice down on the table and this was the result...



The Fair Dice model

$$\Pr(\text{obs.}|\text{fair dice model}) = \left(\frac{1}{6}\right)^{20} = \frac{1}{3,656,158,440,062,976}$$



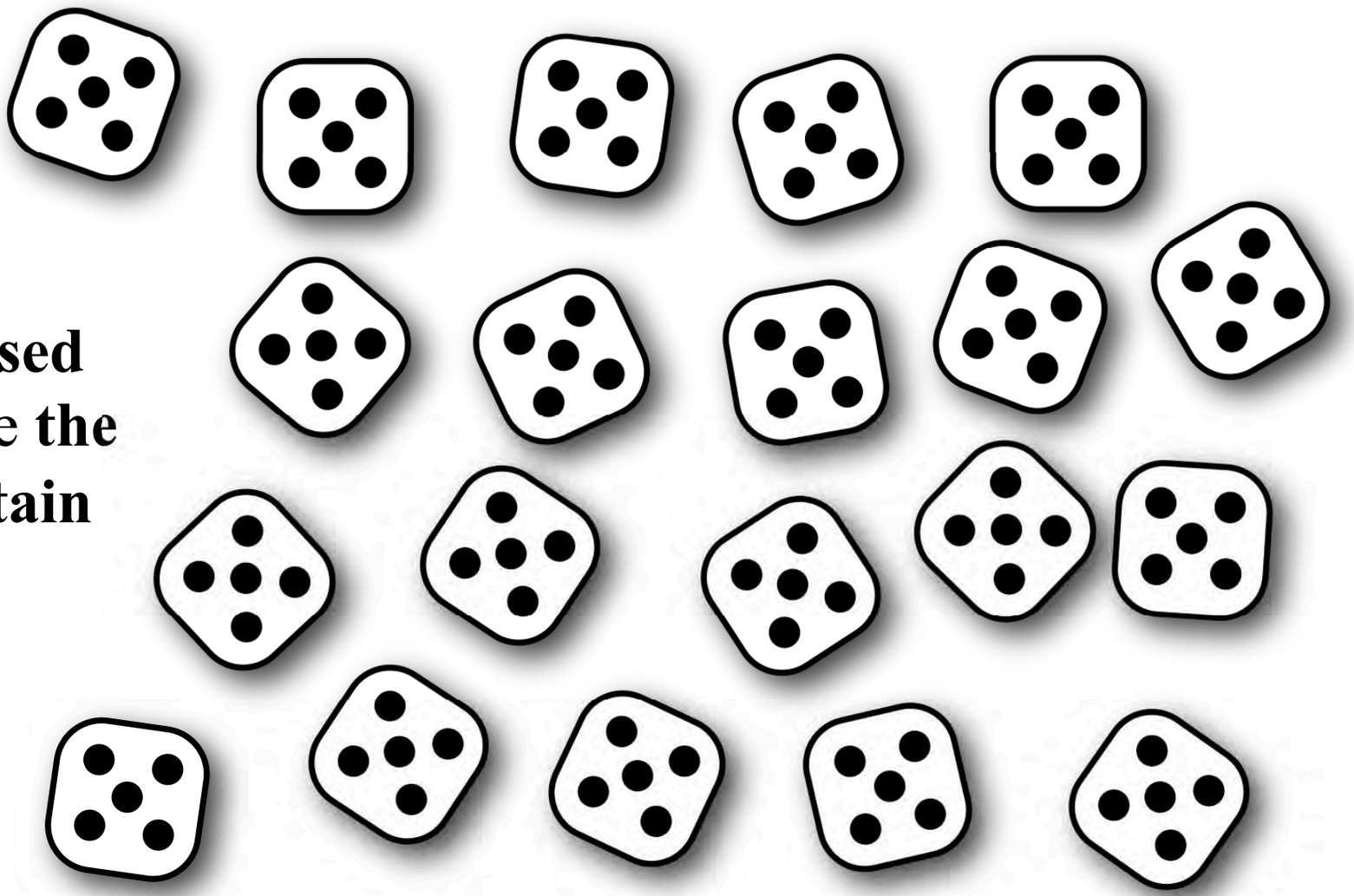
You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!

The Trick Dice model

(assumes dice each have 5 on every side)

$$\Pr(\text{obs.}|\text{trick dice model}) = 1^{20} = 1$$

You should **not be surprised at all** at this result because **the observed outcome is certain** under this model



Results

Model	Likelihood	Surprise level
Fair Dice	$\frac{1}{3{,}656{,}158{,}440{,}062{,}976}$	Very, *very*, ***very*** surprised
Trick Dice	1	Not surprised at all

winning model maximizes likelihood (and thus minimizes surprise)

Likelihood: why a new term?

Outcome	Fair coin model	Two-heads model
H	0.5	1
T	0.5	0
	1	1

Likelihoods of models given one particular data outcome are *not* expected to sum to 1.0

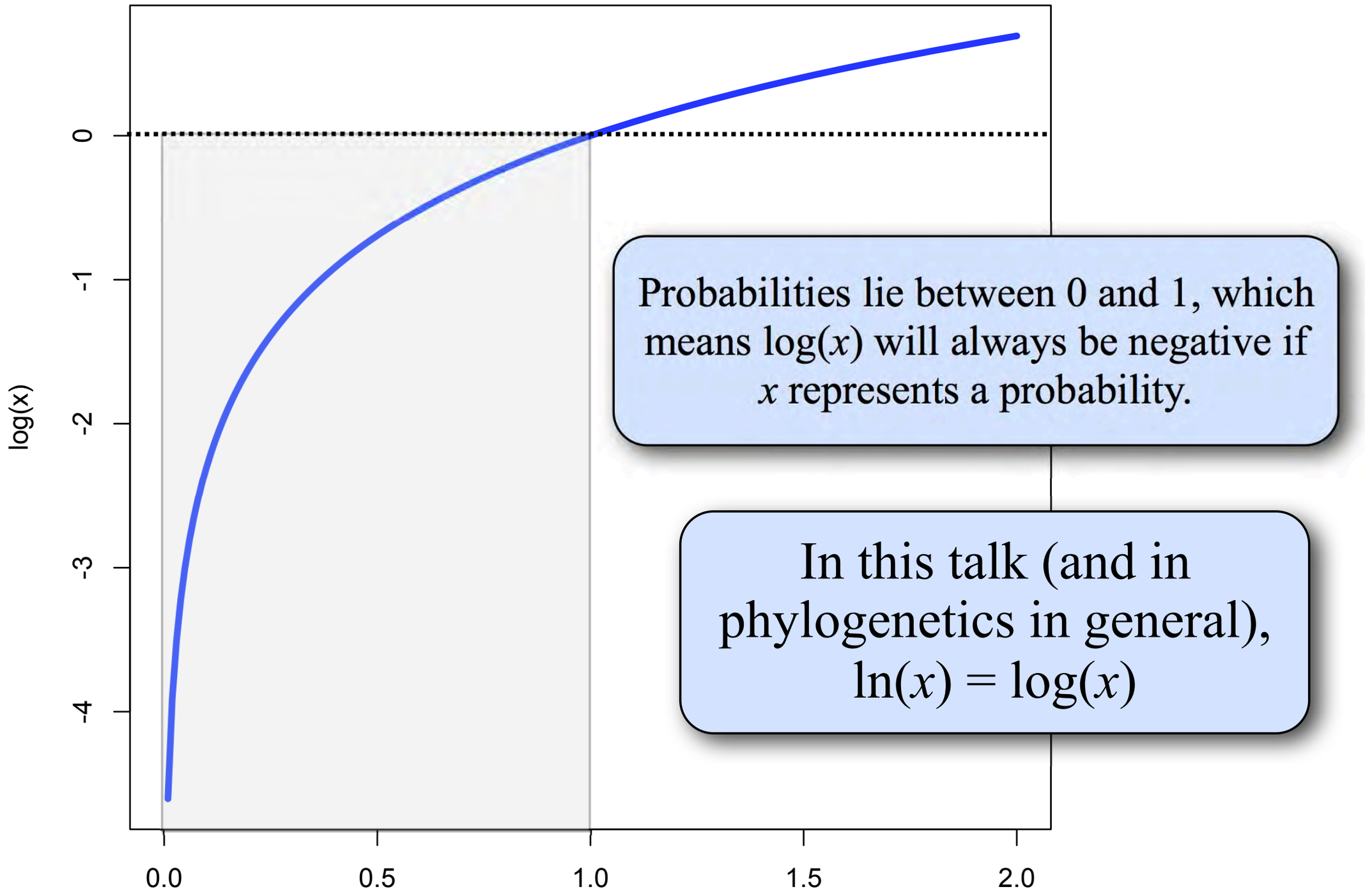
Probabilities of data outcomes given one particular model sum to 1.0

Likelihood and model comparison

- Analyses using likelihoods ultimately involve **model comparison**
- The models compared can be **discrete** (as in the fair vs. trick dice example)
- More often the models compared differ **continuously**:
  - Model 1: branch length is 0.01
  - Model 2: branch length is 0.02
  - Model 3: branch length is 0.03

Rather than having an infinity of models, we instead think of the branch length as a **parameter** within one model

Likelihoods vs. log-likelihoods



Likelihood calculated from a single sequence

$$\Pr(A) = \pi_A$$
$$\Pr(C) = \pi_C$$
$$\Pr(G) = \pi_G$$
$$\Pr(T) = \pi_T$$

First 32 nucleotides of the ψη-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G$$
$$= \pi_A^{12} \pi_C^{7} \pi_G^{7} \pi_T^{6}$$

Note that we are assuming independence among sites here

$$\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

We can already see by eye-balling this that a model allowing **unequal** base frequencies will **fit better** than a model that assumes **equal** base frequencies because there are about twice as many As as there are Cs, Gs and Ts.

Jukes-Cantor (JC) allows for a single parameter and has a transition matrix

$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

The base frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ are all the same and 0.25. There are only two types of changes possible, either one does not change or one changes. This results in two probabilities:

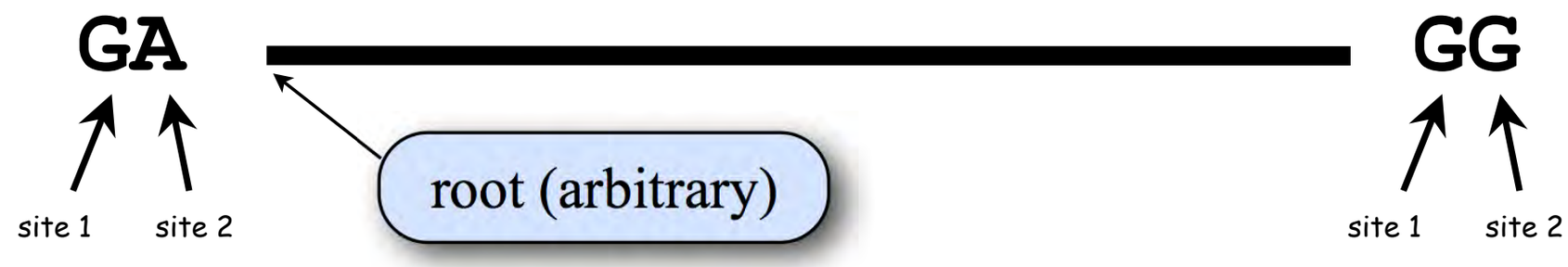
$$\text{Prob}(t)_{ii} = \frac{1}{4} + \frac{3}{4}e^{-\mu t} \tag{19}$$

$$\text{Prob}(t)_{ij} = \frac{1}{4} - \frac{1}{4}e^{-\mu t} \tag{20}$$

Likelihood of the simplest tree

sequence 1 ——— sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:



$$L = L_1 \, L_2$$

$$= \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)\right] \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\right]$$

Pr(G) Pr(G|G, αt) Pr(A) Pr(G|A, αt)

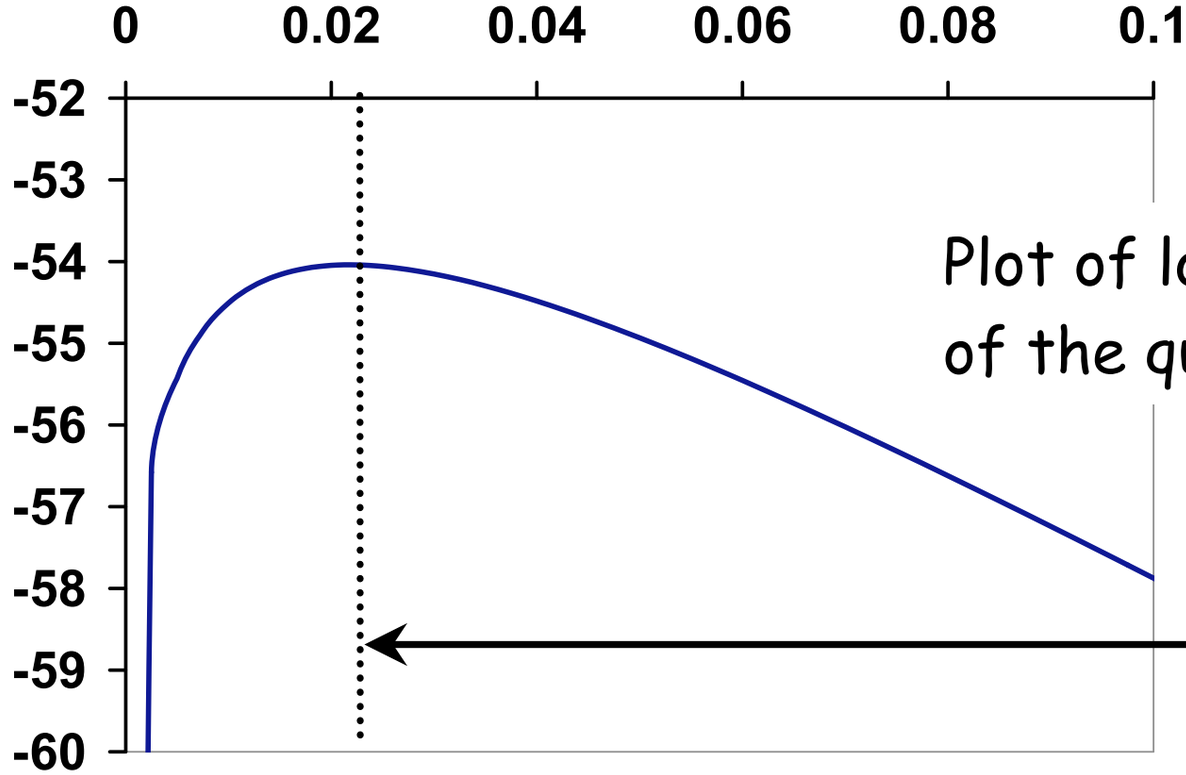
Note that we are NOT assuming independence here

Maximum likelihood estimation

First 32 nucleotides of the ψη-globin gene of gorilla and orangutan:

```
gorilla    GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan  GGACTCCTTGAGAAATAAACTGCACACACTGG
```

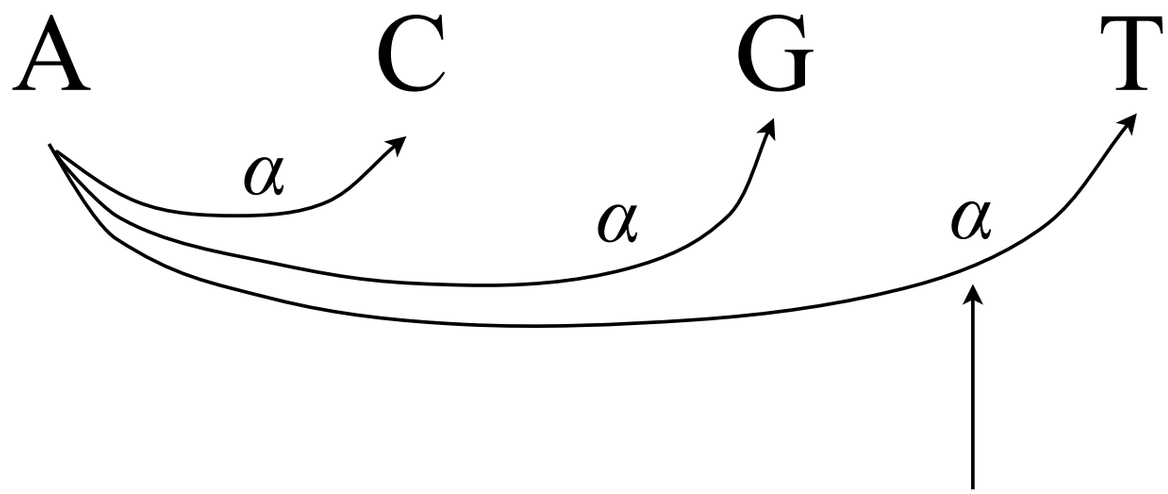
$$L = \left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\right]^{30}\left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$



Plot of log-likelihood as a function of the quantity αt

Maximum likelihood estimate (MLE) of αt is 0.021753

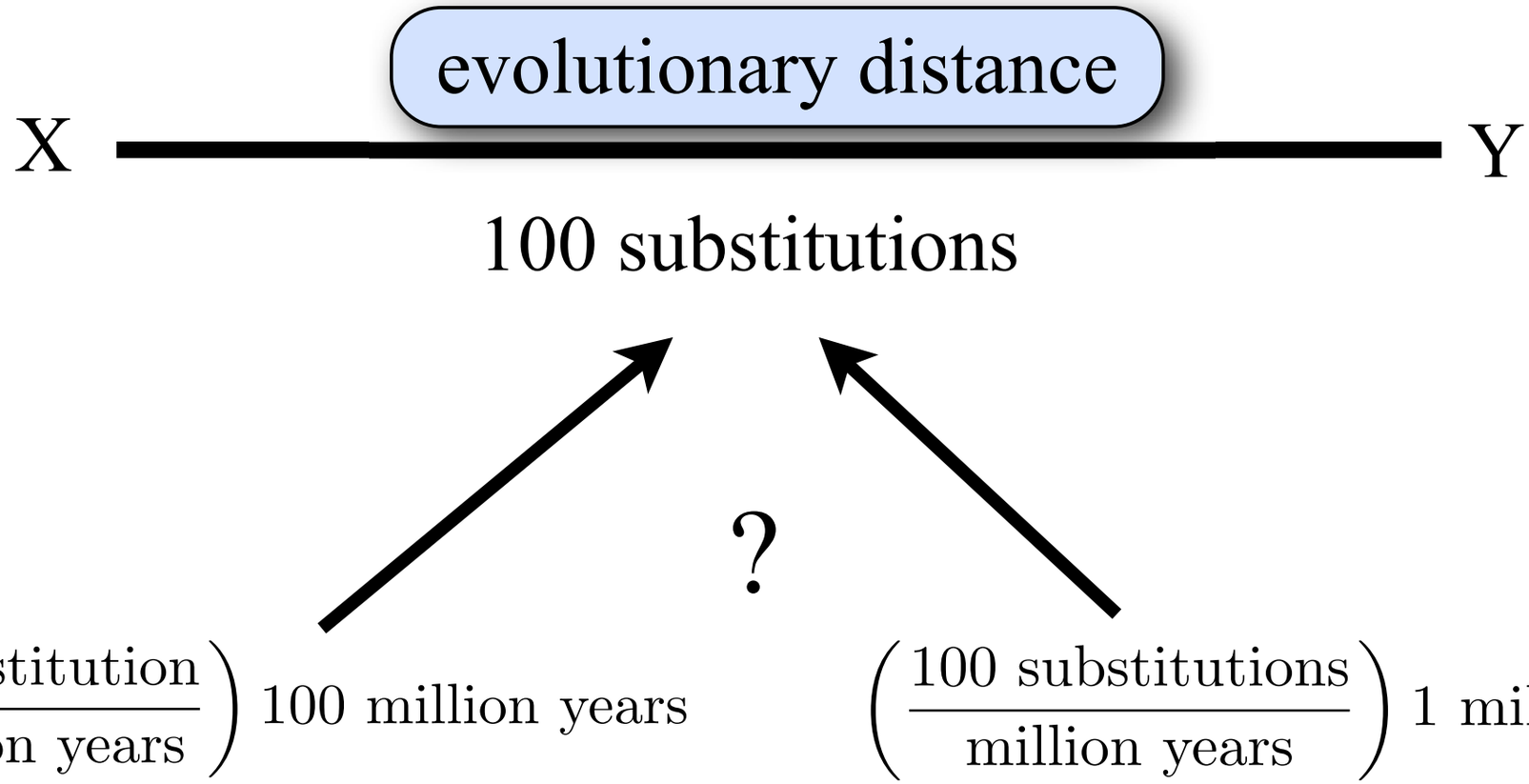
number of substitutions = rate × time



Overall substitution rate is 3α, so the expected number of substitutions (v) is

$$v = 3\alpha t$$

Rate and time are confounded



On Tuesday, Tracy Heath will introduce models that allow separate estimation of rates and times, but without extra information/constraints, sequence data allow only estimation of the **number** of substitutions.

Evolutionary distances for several common models

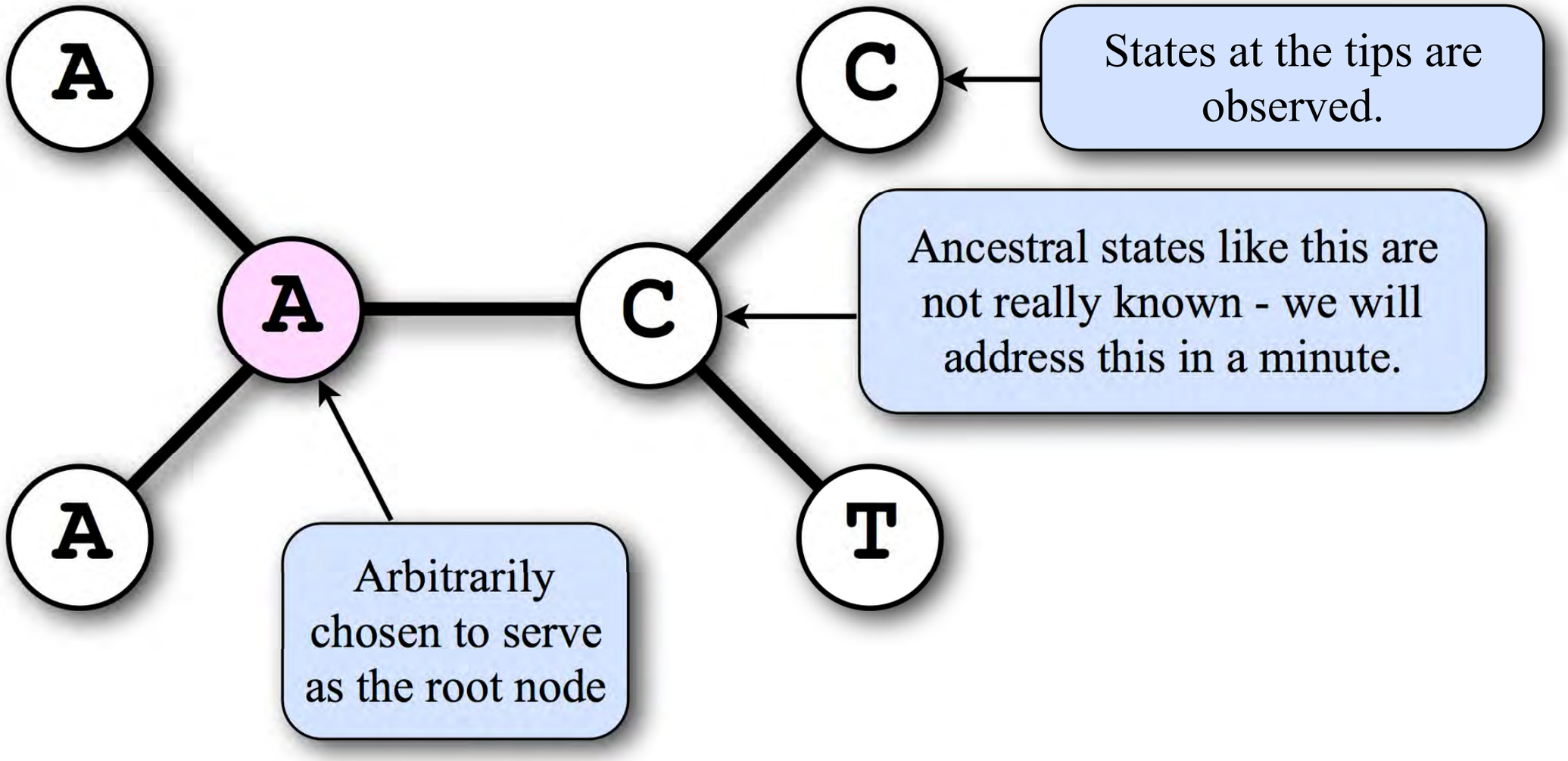
Model	Expected no. substitutions: $v = \{r\}t$
JC69	$v = \{3\alpha\}\, t$
F81	$v = \{2\mu(\pi_R\pi_Y + \pi_A\pi_G + \pi_C\pi_T)\}\, t$
K80	$v = \{\beta(\kappa + 2)\}\, t$
HKY85	$v = \{2\mu\left[\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)\right]\}\, t$

In the formulas above, the overall rate *r* (in curly brackets) is a function of all parameters in the substitution model.

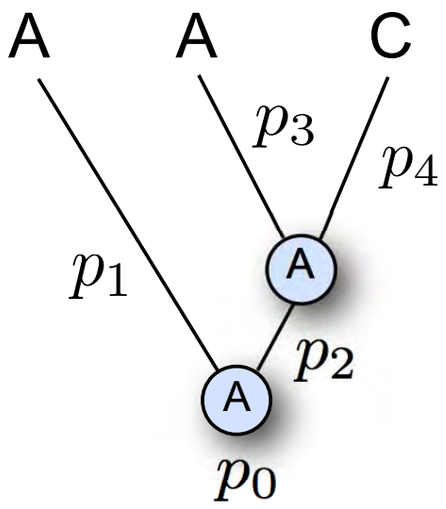
One substitution model parameter is always determined from the edge length; the others are usually global (i.e. same value applies to all edges).

Likelihood of an unrooted tree

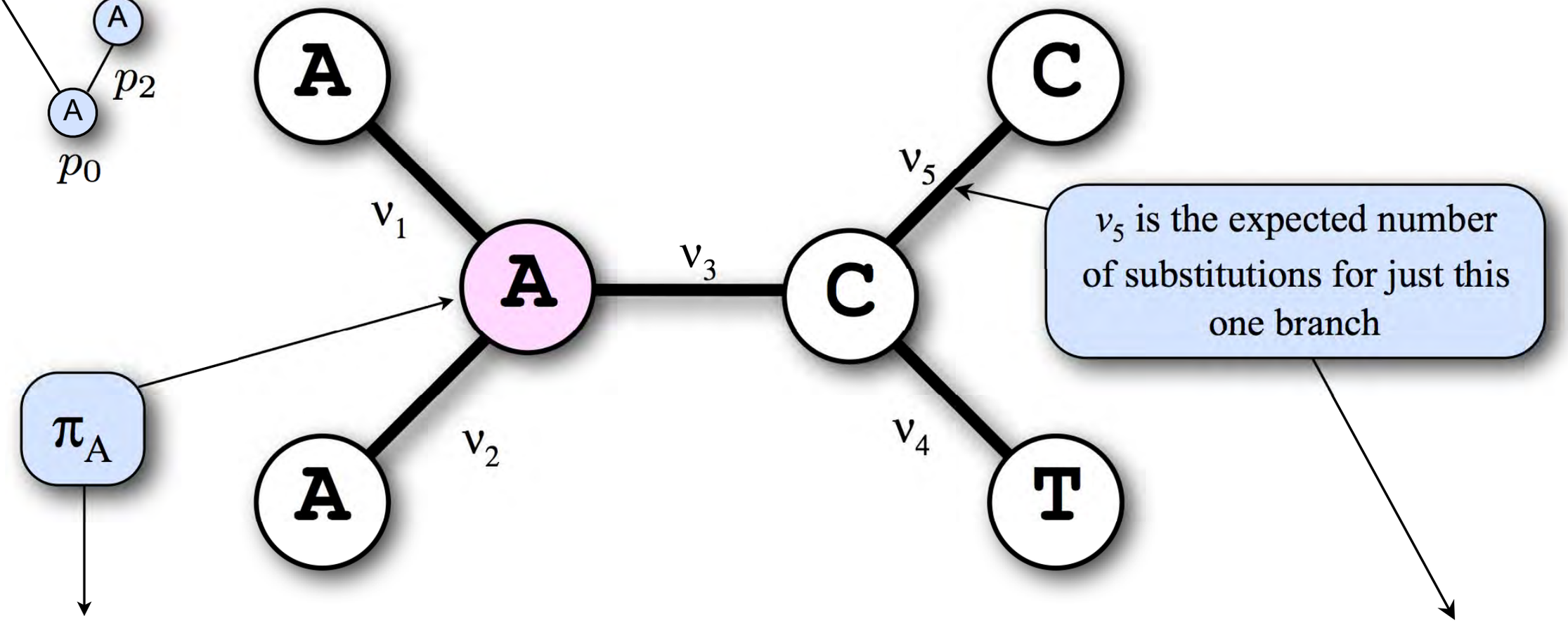
(data shown for only one site)



From slide 6



Likelihood for site k



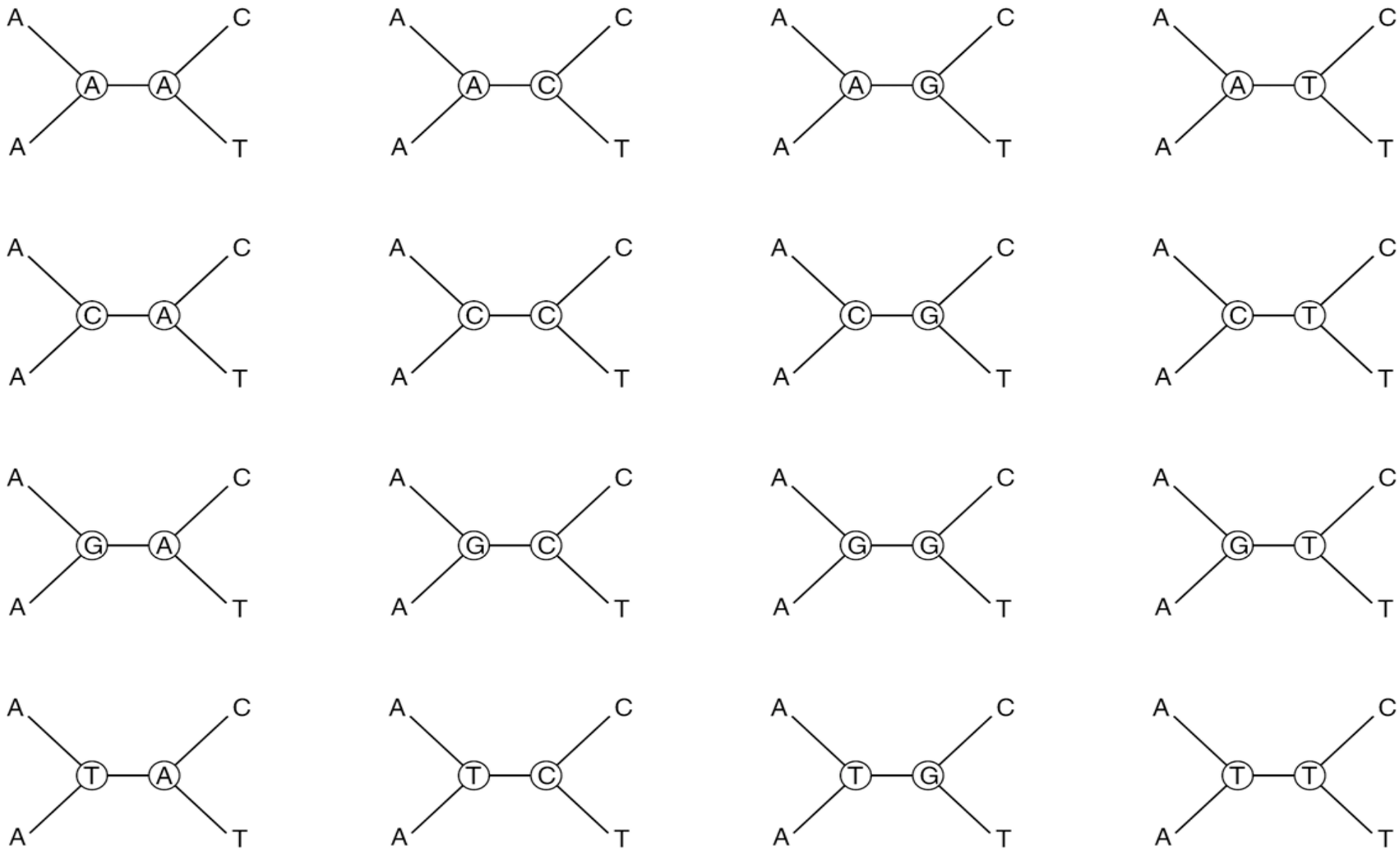
v_5 is the expected number of substitutions for just this one branch

$$L_k = \frac{1}{4}\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_1/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_2/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_3/3}\right]\left[\frac{1}{4}-\frac{1}{4}e^{-4\nu_4/3}\right]\left[\frac{1}{4}+\frac{3}{4}e^{-4\nu_5/3}\right]$$

$P_{AA}(v_1)$ $P_{AA}(v_2)$ $P_{AC}(v_3)$ $P_{CT}(v_4)$ $P_{CC}(v_5)$

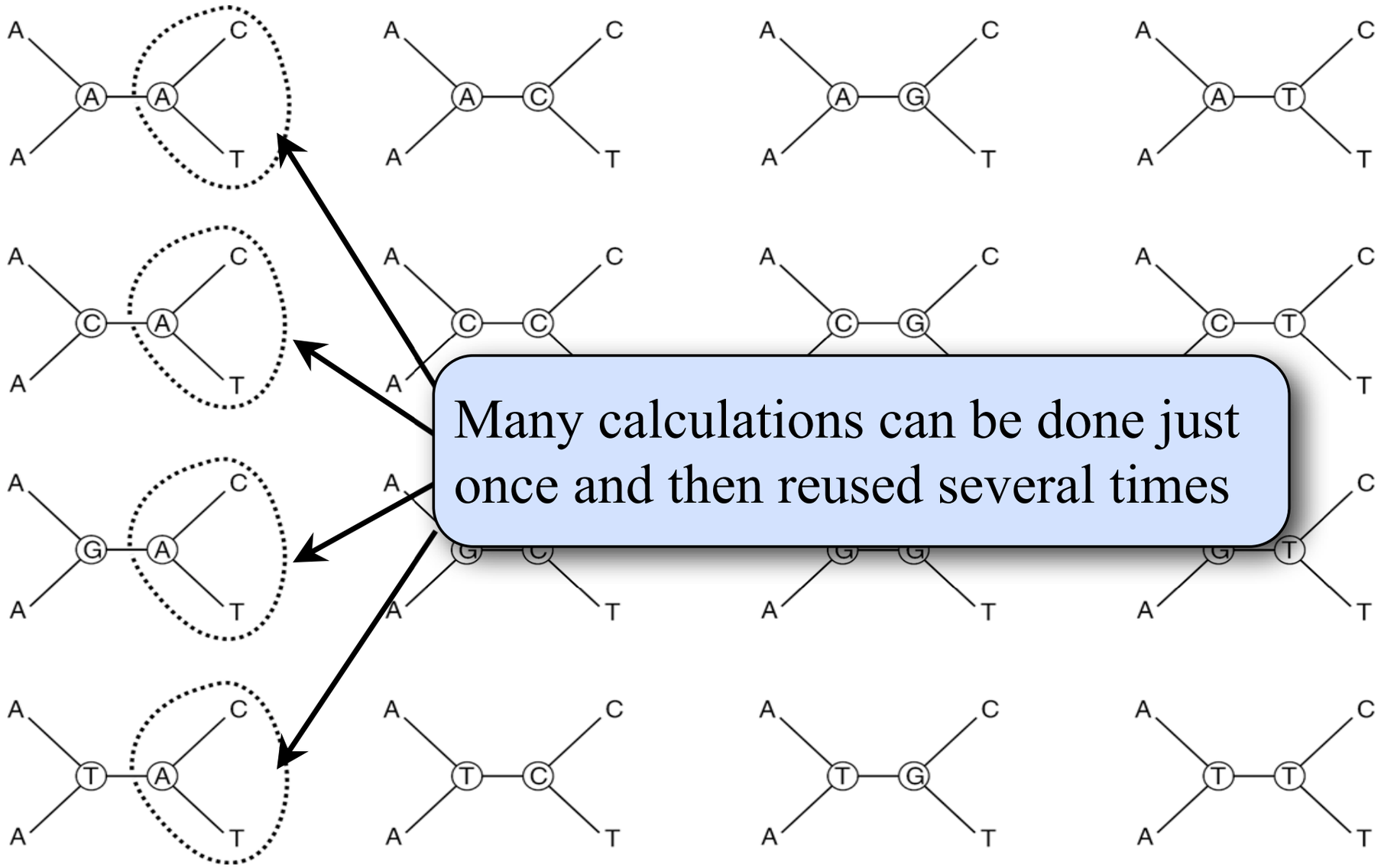
Note use of the AND probability rule

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them



Note use of the OR probability rule

Pruning algorithm
(same result, less time)



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

Algorithm 1 Likelihood downpass algorithm

for all i **do**

$h_i^{(q)} \leftarrow 0$

for all j **do**

$h_i^{(q)} \leftarrow h_i^{(q)} + p_{ij}^{(q)} g_j^{(q)}$

end for

end for

for all i **do**

$h_i^{(r)} \leftarrow 0$

for all j **do**

$h_i^{(r)} \leftarrow h_i^{(r)} + p_{ij}^{(r)} g_j^{(r)}$

end for

end for

for all i **do**

$g_i^{(p)} \leftarrow h_i^{(q)} h_i^{(r)}$

end for

More explanation will follow