



Statistical methods for estimating phylogenetic trees **part 1**

Front Matter

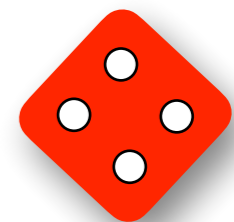
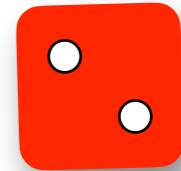
most slides are inspired by or directly lifted
(these will be attributed) from lectures by:

Paul Lewis, Mark Holder, David
Swofford, & John Huelsenbeck

if you wish to see these in person or access their materials
go to [the Workshop on Molecular Evolution at MBL](#) website

Phylogenetics Lecture Plan

- some review and introduction
- some basic probability
- calculating likelihood
- substitution models
- maximum likelihood methods
- Bayesian thinking
- hierarchical models
- divergence-time estimation



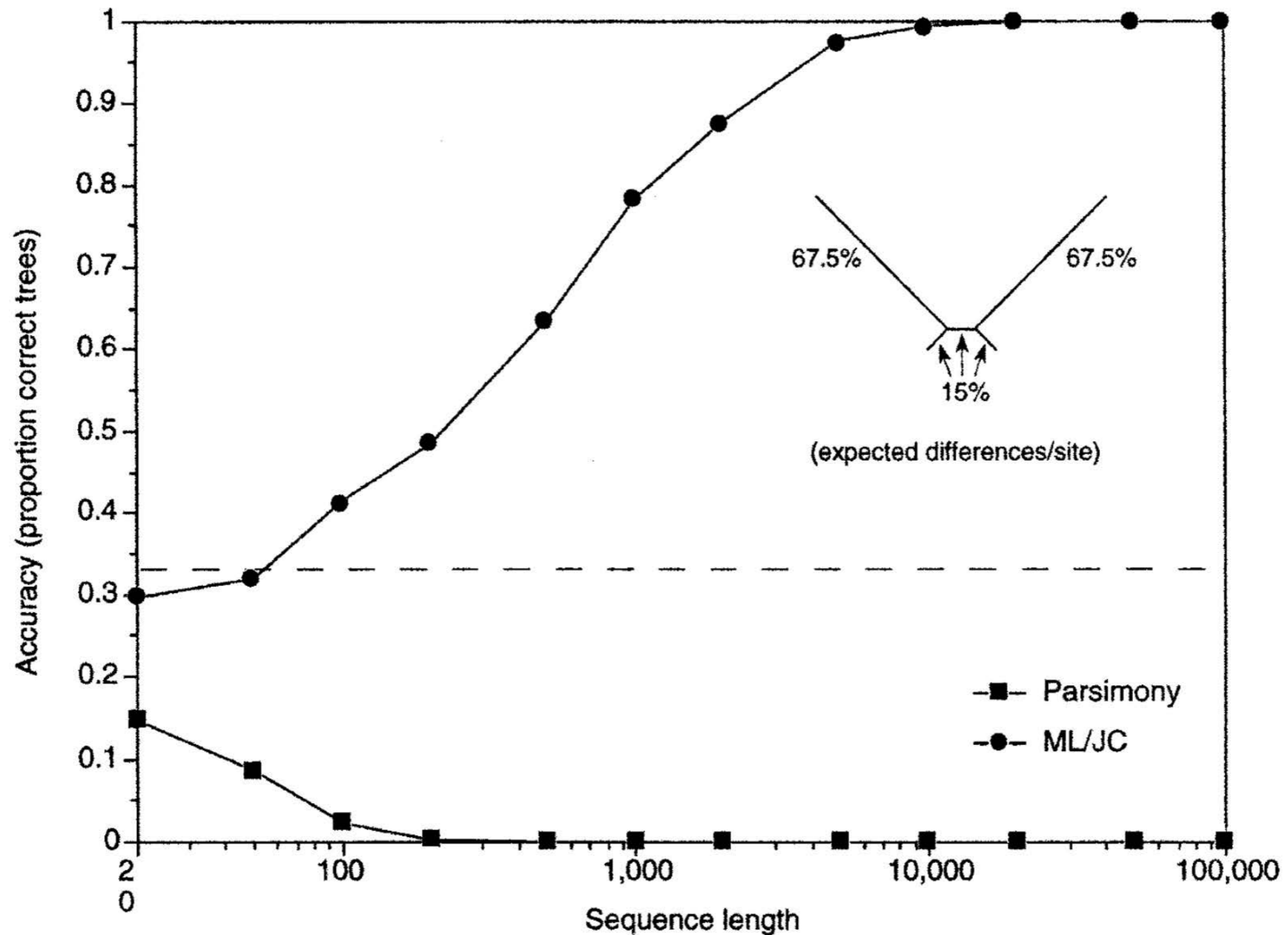
Parsimony can be Inconsistent

remember from last time:

if one feels that consistency is a desirable property for an estimator...

the inconsistency of parsimony is the strongest argument against its use

Statistical Methods are Consistent



Likelihood-based Phylogenetics

statistical methods require a **model** describing the process that generated the data

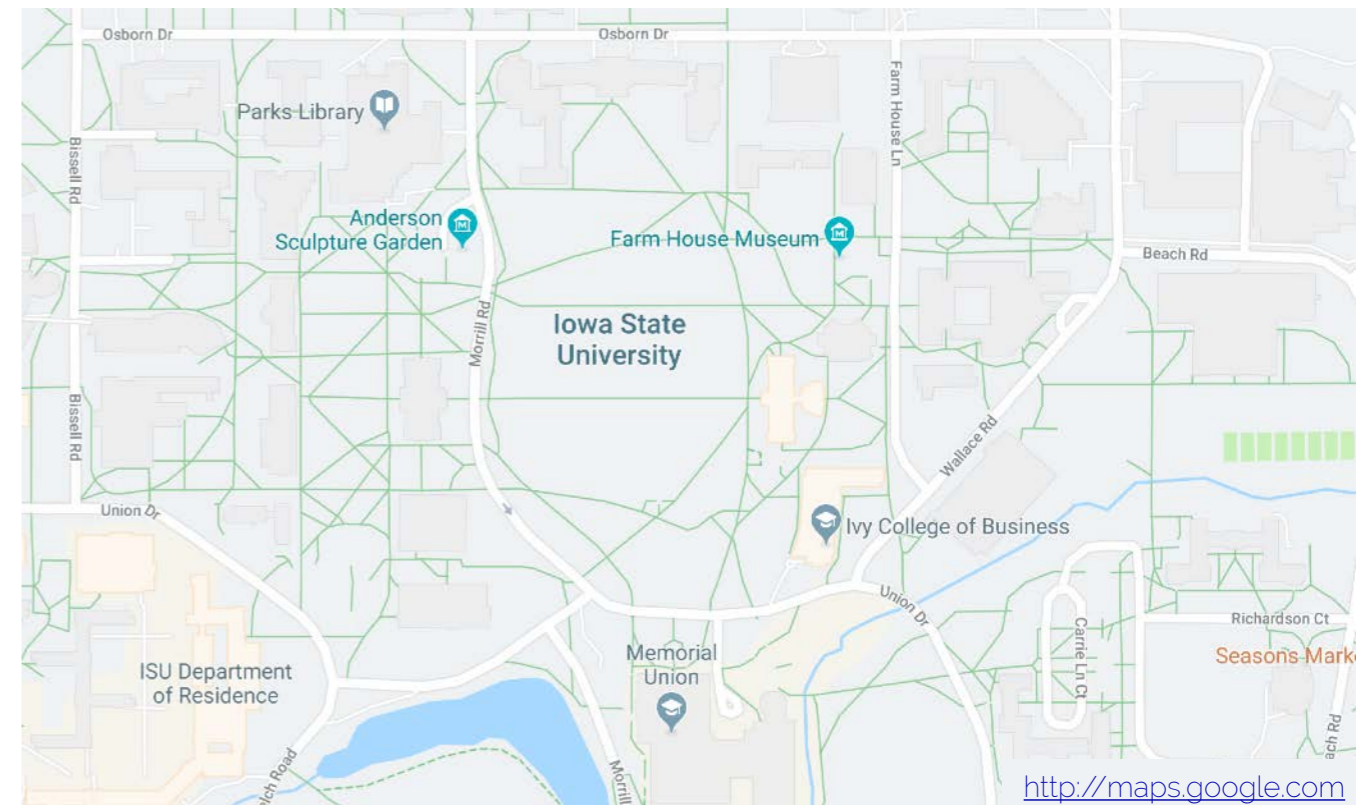
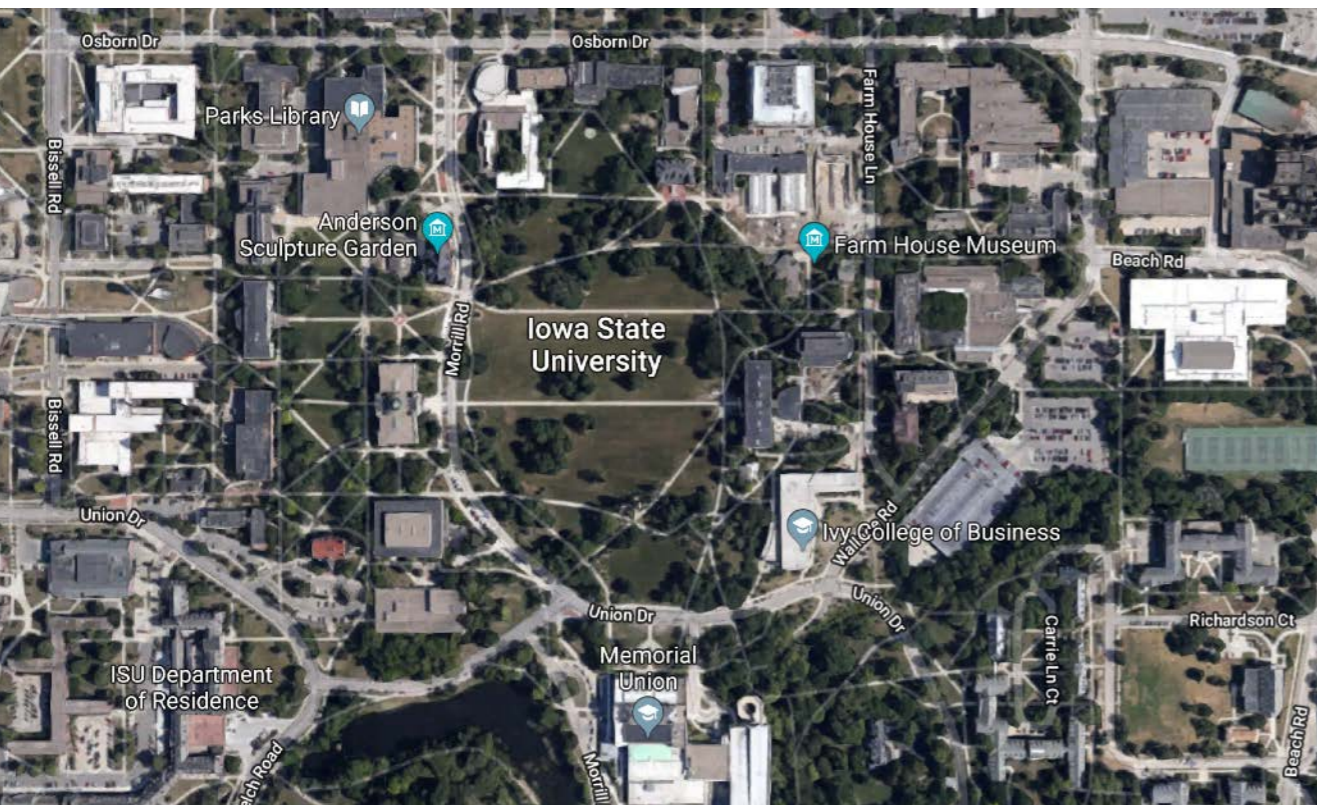
models have specific **assumptions**

"A model is an intentional simplification of a complex situation designed to eliminate extraneous detail in order to focus attention on the essentials of the situation."

(Daniel L. Hartl, 2000)

Models don't Need to Perfectly Capture Reality

"The most that can be expected from any model is that it can supply a useful approximation to reality: **All models are wrong; some models are useful.**" (George E. P. Box, 1987)



Likelihood-based Phylogenetics

if the model is specified correctly,
likelihood-based methods are statistically
consistent, even for Felsenstein Zone trees

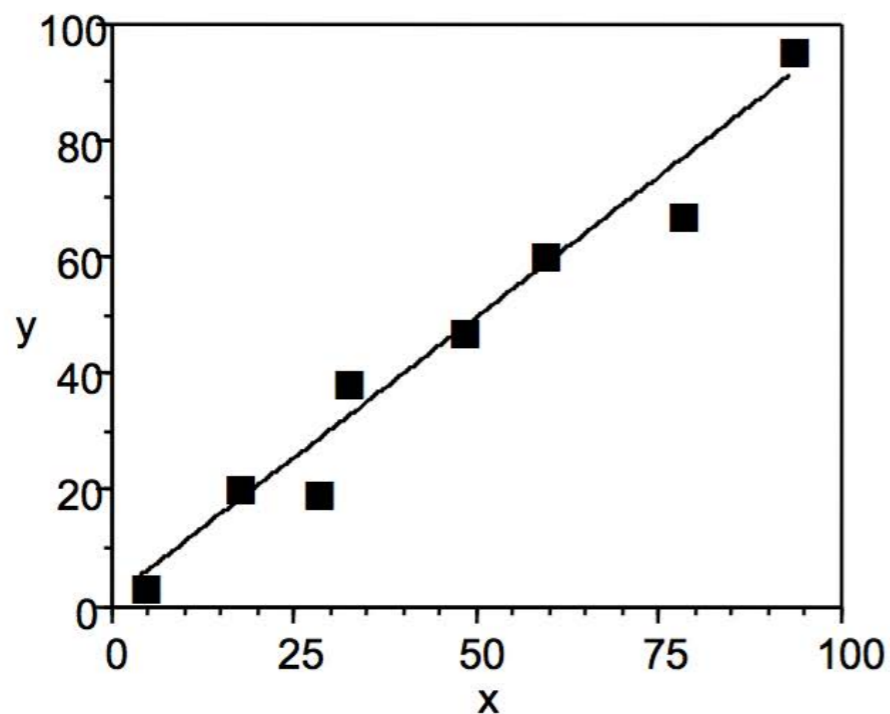
if the model is oversimplified, however,
likelihood methods can be inconsistent, so
we have to choose a *good* model



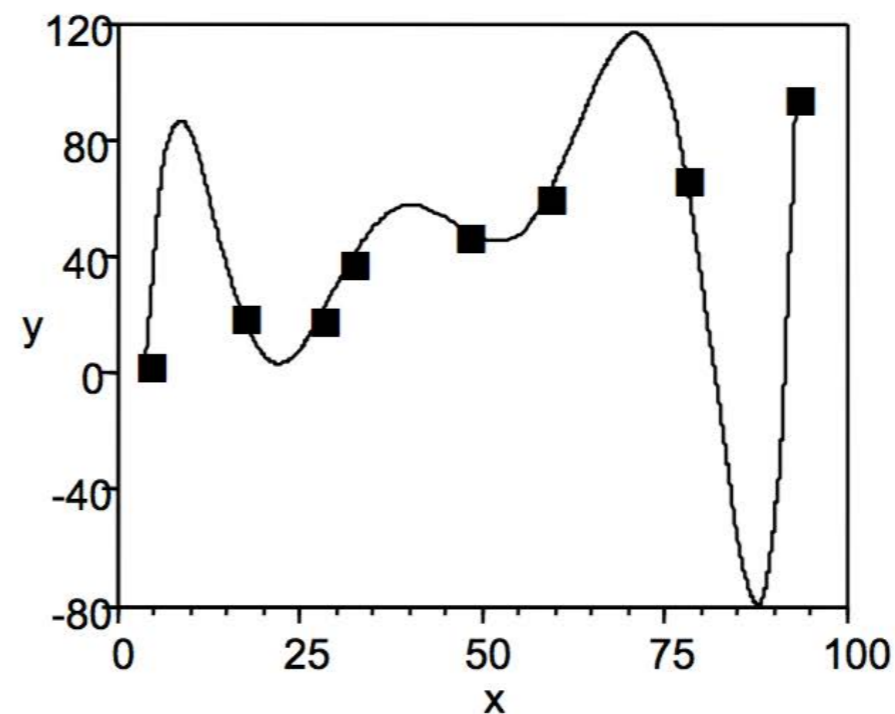
What is a good model?

a model that appropriately balances fit of the data with simplicity (parsimony, in a different sense)

i.e., if a simpler model fits the data almost as well as a more complex model, prefer the simpler one



$$y = 1.30 + 0.965x$$
$$(r^2 = 0.963)$$



$$y = -330 + 134x - 15.5x^2 + 0.816x^3$$
$$- 0.0225x^4 + 0.000335x^5$$
$$- 0.00000255x^6 + 0.00000000777x^7$$
$$(r^2 = 1.000)$$

How do I choose a good model?

there are several approaches depending on the type of statistical inference you are using

frequentist: use model selection criteria and tests (AIC, BIC, likelihood ratio test)

Bayesian: compare support for a model using Bayes factors or use mixture models

model selection will be covered in greater detail later in the semester

What is a likelihood?

the likelihood is central to both frequentist (maximum likelihood) and Bayesian inference

the likelihood is the probability of your observed data given a fully specified model

it is a function of the model (θ): given the parameters of θ , $\Pr(\mathbf{X} \mid \theta)$ is the probability of observing the data

$$\mathcal{L}(\theta) = \Pr(\mathbf{X} \mid \theta)$$

$$\mathbf{X} = \{ \text{red die with 1 dot}, \text{red die with 4 dots}, \text{red die with 1 dot} \} \quad \theta = \text{model}$$



Probability Review

probabilities are associated with the outcomes of random processes

the probabilities of all possible outcomes under a given model will sum to **1.0**

$$\text{Pr}(\text{HEADS} | \text{FAIRCOIN}) = 0.5$$

$$\text{Pr}(\text{TAILS} | \text{FAIRCOIN}) = 0.5$$

Outcome	FAIRCOIN model
	0.5
	0.5
	1.0

Combining Probabilities

we **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 *fair* coins, what is the probability of

the **AND** rule



0.5

AND

×



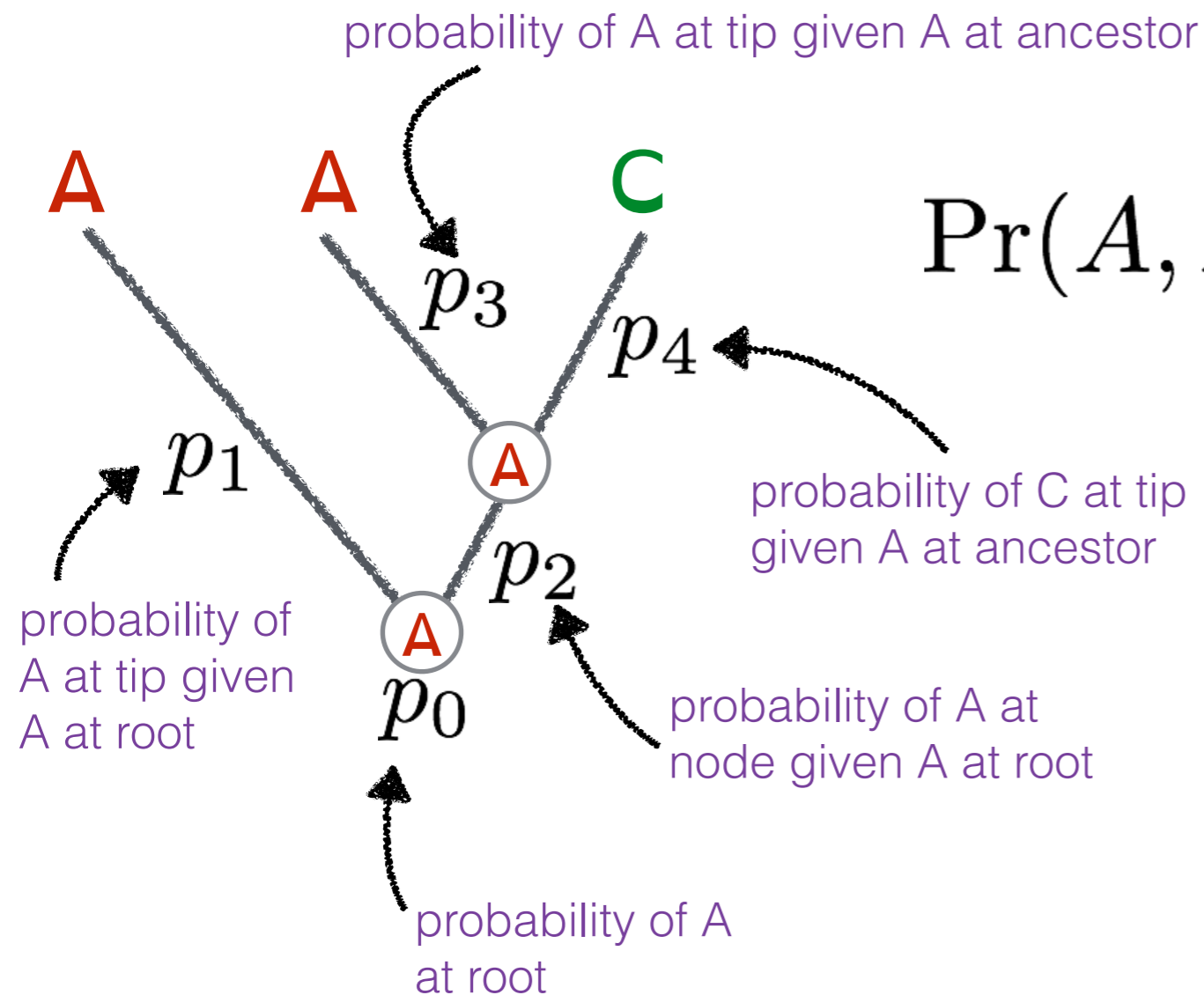
0.5

?

= 0.25

Combining Probabilities 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



$$\Pr(A, A, C, A, A) = p_0 p_1 p_2 p_3 p_4$$

we have observed the states $\{A, A, C, A, A\}$, including the ancestral states (let's pretend for now)

Combining Probabilities

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

using 1 *fair* die, what is the probability of

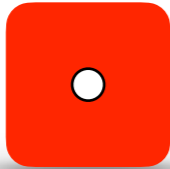










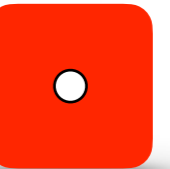
the OR rule

	OR		?
$1/6$	+	$1/6$	$= 1/3$

Combining AND & OR

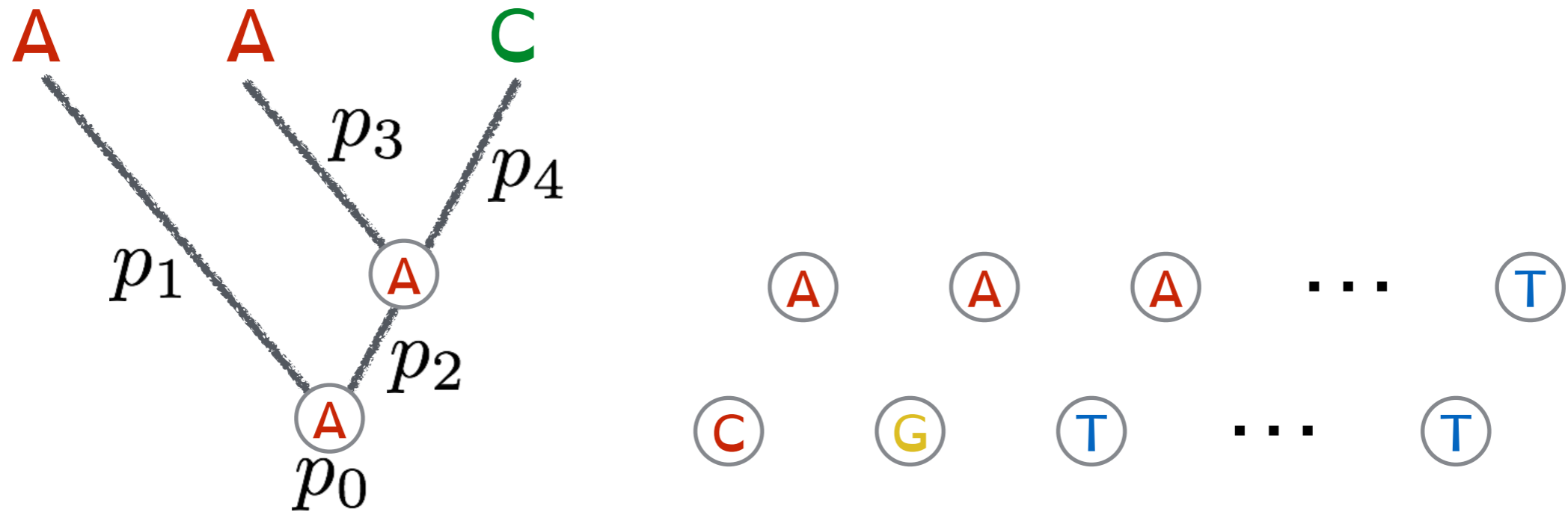
What is the probability that the sum of two fair dice is 7?

die 1
AND
die 2

	OR		OR		OR		OR		OR		
											
$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$	
\times		\times		\times		\times		\times		\times	
$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$	
$=$		$=$		$=$		$=$		$=$		$=$	
$\frac{1}{36}$	+	$\frac{1}{36}$	+	$\frac{1}{36}$	+	$\frac{1}{36}$	+	$\frac{1}{36}$	+	$\frac{1}{36}$	$= \frac{1}{6}$

Using **AND** & **OR** in Phylogenetics

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



the **OR** rule used to combine different combinations of ancestral states

$$\Pr(A, A, C, A, C) + \Pr(A, A, C, A, G) + \Pr(A, A, C, A, T) + \dots + \Pr(A, A, C, A, A)$$

Conditional Probability



the probability of an outcome given that another event has occurred or given a set of assumptions (model)

What is the probability of tails given given that the coin is fair?

What is the probability of tails given given that the coin has two tails?

$$\Pr(\text{heads} \mid \text{FAIRCOIN}) = 0.5$$

$$\Pr(\text{heads} \mid \text{TWOTAILS}) = 1.0$$

Outcome	FAIRCOIN model	TWOTAILS model
	0.5	0.0
	0.5	1.0

Independence

it is always true that:

$$\underbrace{\Pr(A \text{ and } B)}_{\text{joint probability}} = \Pr(A) \underbrace{\Pr(B \mid A)}_{\text{conditional probability}}$$

if we can say this ...

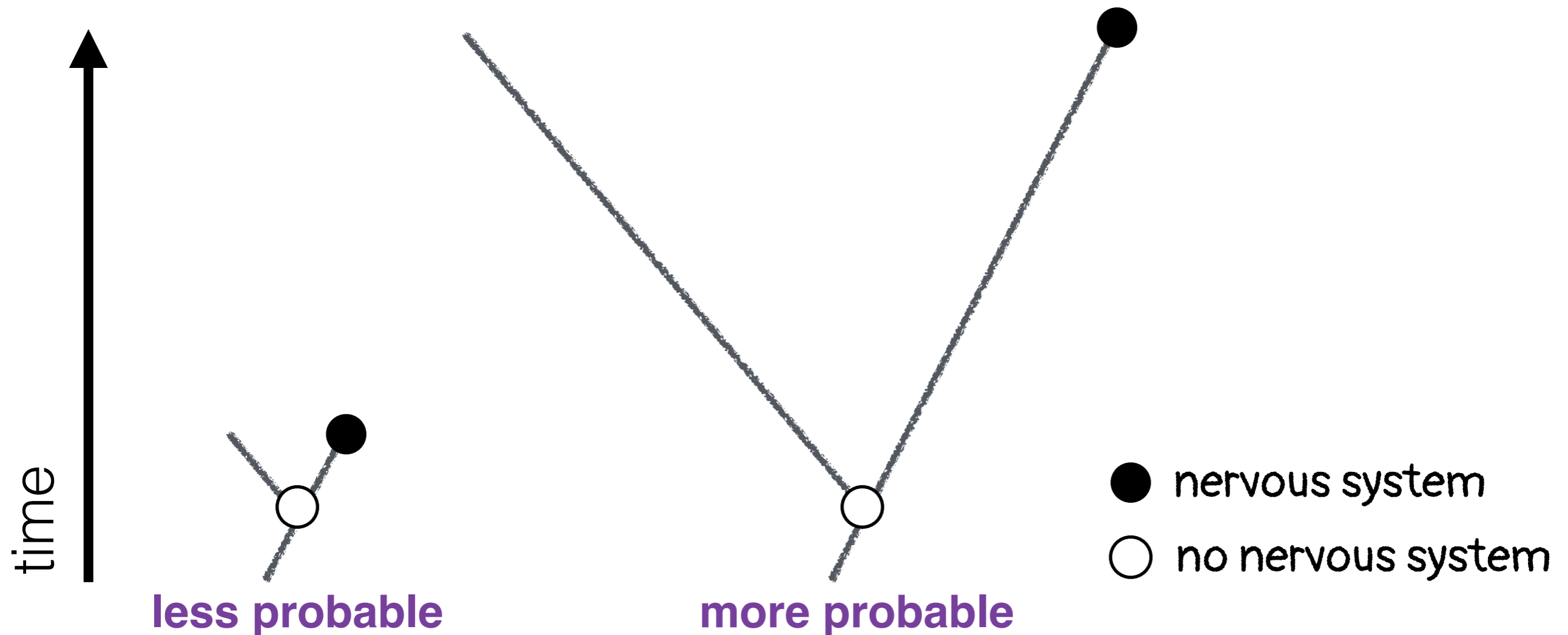
$$\Pr(B \mid A) = \Pr(B)$$

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

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

Non-independence in Evolution

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



Conditional Independence

assume that both A and B depend on C :

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

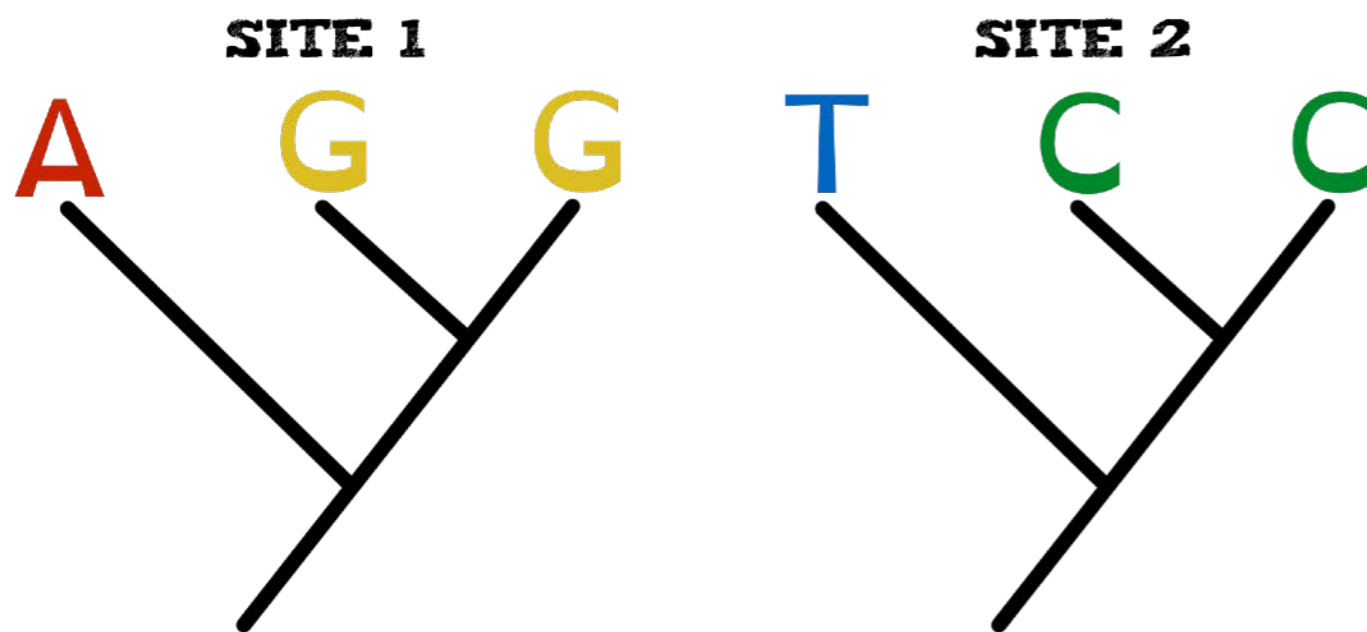
if we can say this ...

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

...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 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

$$\Pr(AGG \text{ and } TCC | \mathcal{T}, \theta) = \Pr(AGG | \mathcal{T}, \theta) \Pr(TCC | \mathcal{T}, \theta)$$

$$\mathcal{T} = \text{tree}, \quad \theta = \text{model}$$

The Likelihood Criterion

the probability of the observations computed under a given model tells us how surprised we should be

The preferred model is the one that surprises us least.

suppose I threw
20 dice and
ended up with
this result

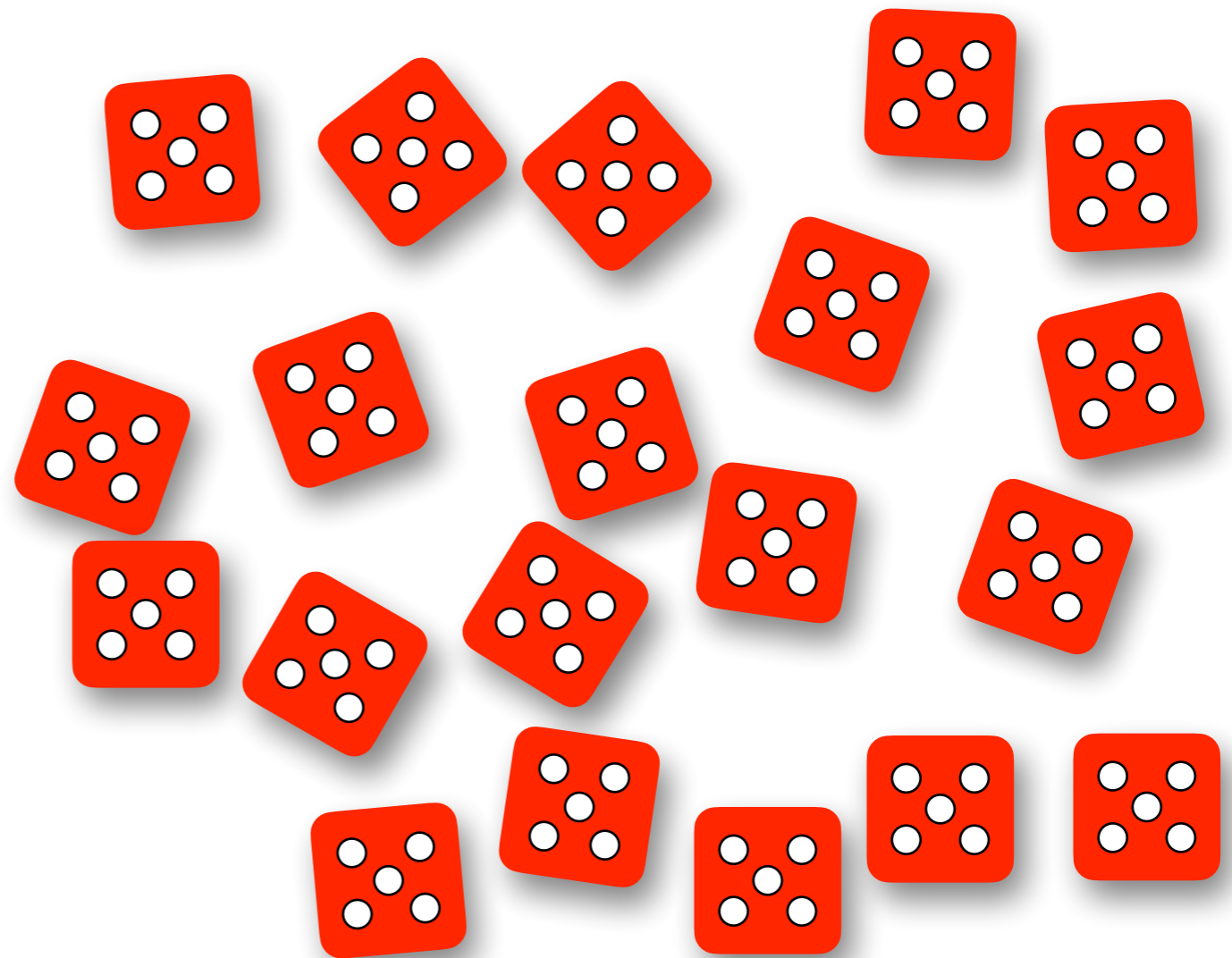


The Fair Dice Model

the AND rule

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

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 all dice only have  on every side)



$$\Pr(\text{obs.} \mid \text{TRICKDICE}) = 1^{20} = 1$$

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





Results

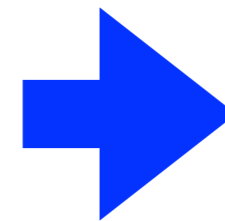
computing the $\Pr(X \mid \theta)$ allows us to know how surprised we should be by our data

Model	Likelihood	Surprise level	Emoji
Fair dice	$\frac{1}{3,656,158,440,062,976}$	very, very, <i>very</i> surprised <i>wtf!</i>	
Trick dice	1.0	not surprised at all <i>meh</i>	

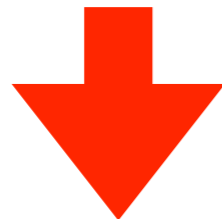
the winning model maximizes the likelihood
(and thus minimizes the surprise)

Likelihood vs. Probability

Outcome	FAIRCOIN model	TWOTAILS model
	0.5	0.0
	0.5	1.0
	1.0	1.0



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 & 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 = 0.01
- model 2: branch length = 0.02
- model 3: branch length = 0.03

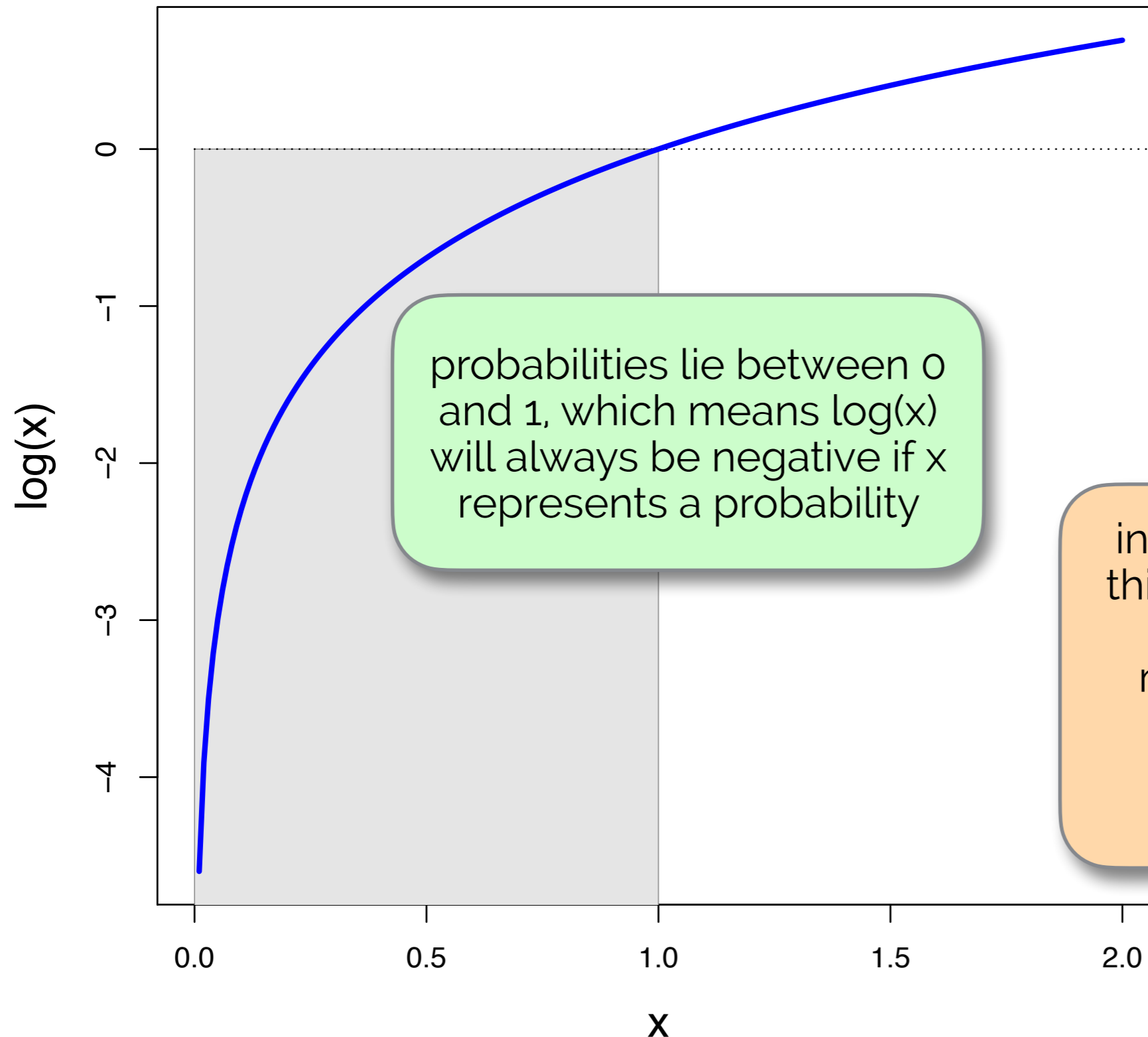
rather than having an infinity of models, we instead think of the branch length as a parameter within one model

Using the Log Likelihood

phylogenetic models often involve a lot of observations (DNA sites) and we have to invoke the AND rule many, many times, thus taking the product of lots of small numbers

this leads to a problem called [underflow](#), where a value is too small (i.e., too close to zero) to be represented by a computer

Using the Log Likelihood



probabilities lie between 0 and 1, which means $\log(x)$ will always be negative if x represents a probability

$$\log(xy) = \log(x) + \log(y)$$
$$\log\left(\frac{x}{y}\right) = \log(x) - \log(y)$$
$$\log(x^n) = n \log(x)$$

in phylogenetics and in this class when we refer to “log” we typically mean the natural log (unless stated otherwise)
 $\log(x) = \ln(x)$

Likelihood of a Single DNA Sequence

first 32 nucleotides of the $\psi\eta$ -globin gene of gorilla:

G A A G T C C T T G A G A A A T A A A C T G C A C A C A C T G G

$$\begin{aligned}\mathcal{L} &= \pi_G \pi_A \pi_A \pi_G \pi_T \pi_C \pi_C \pi_T \pi_T \pi_G \pi_A \pi_G \pi_A \pi_A \pi_A \pi_T \pi_A \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_A \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G \\ &= \pi_A^{12} \pi_C^7 \pi_G^7 \pi_T^6\end{aligned}$$

**we assume sites are independent*

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

$$\Pr(A) = \pi_A$$

$$\Pr(C) = \pi_C$$

$$\Pr(G) = \pi_G$$

$$\Pr(T) = \pi_T$$

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.

Likelihood of the Simplest Tree

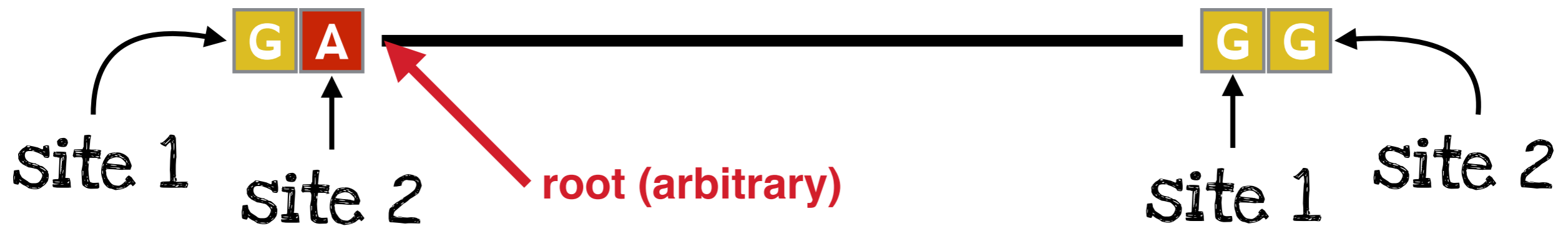


sequence 1



sequence 2

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



$$\mathcal{L} = \mathcal{L}_1 \mathcal{L}_2$$

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

$$\text{Pr}(\text{G})$$

$$\text{Pr}(\text{G} | \text{G}, \alpha t)$$

$$\text{Pr}(\text{A})$$

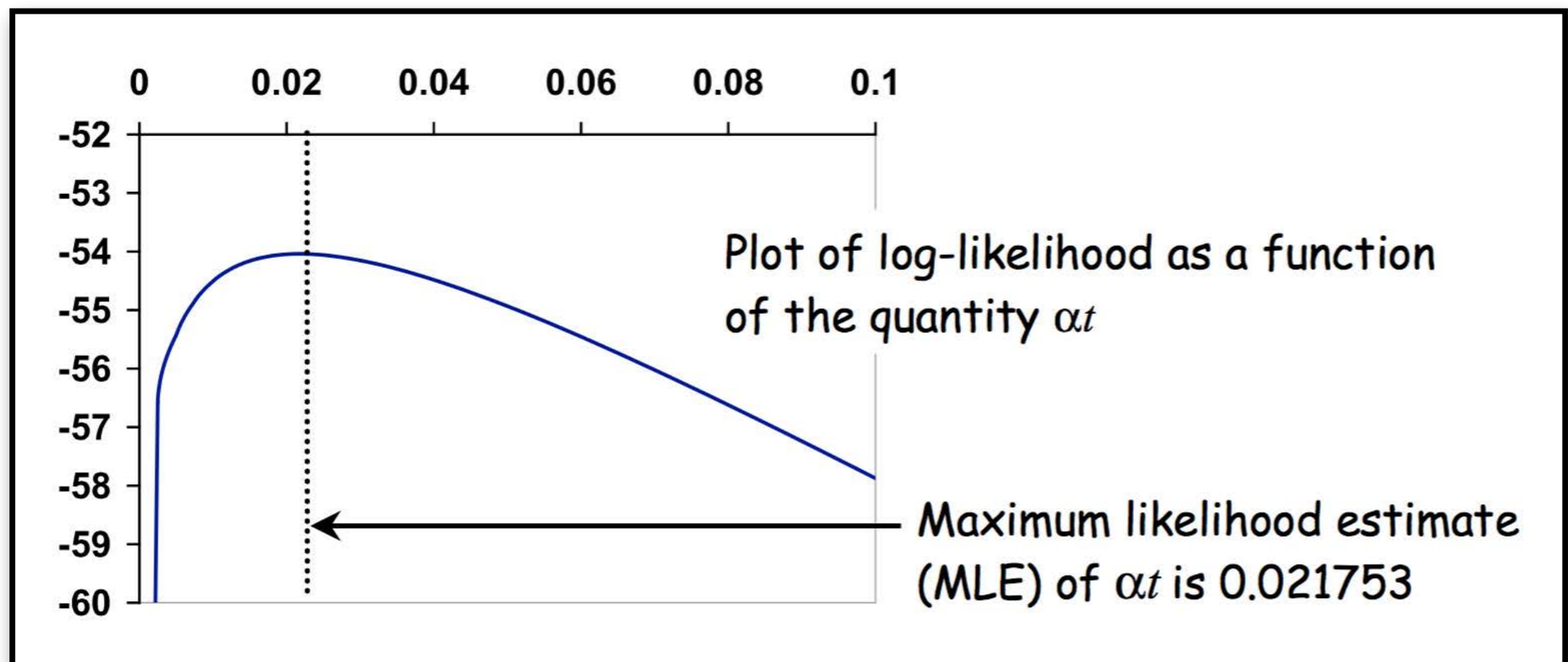
$$\text{Pr}(\text{G} | \text{A}, \alpha t)$$

Maximum Likelihood Estimation

first 32 nucleotides of the $\psi\eta$ -globin gene of gorilla & orangutan:

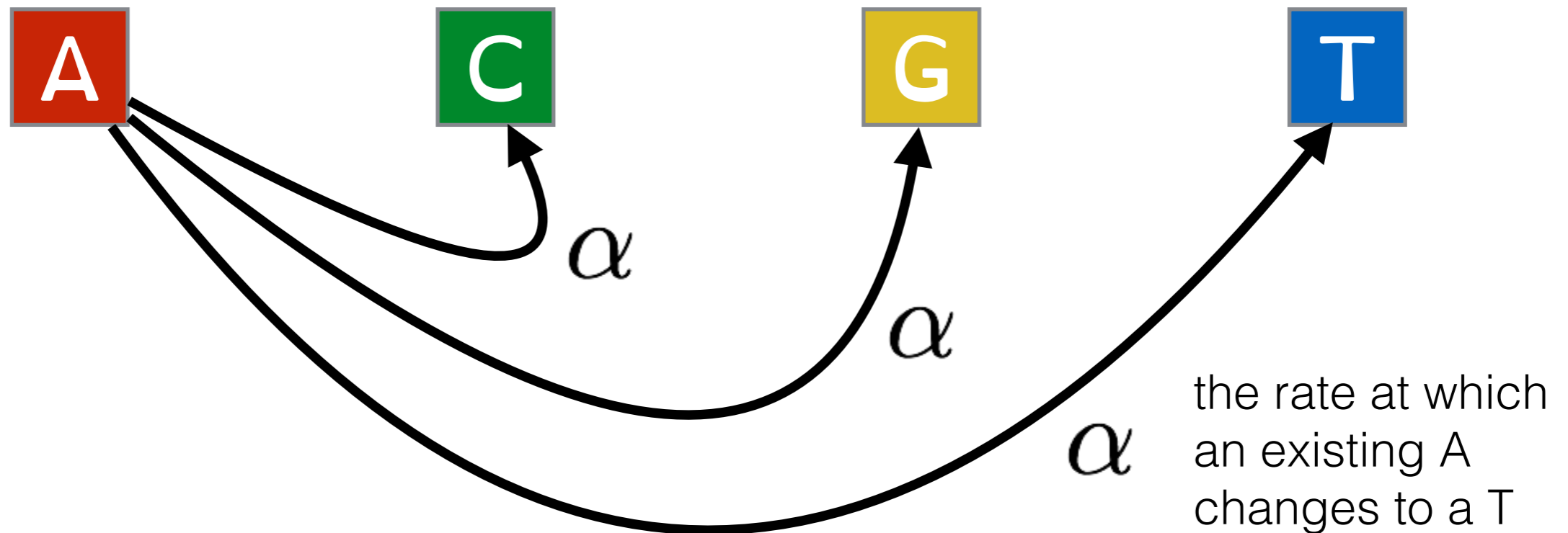
gorilla **G****A****A****G**TCCTTGAGAAATAAACTGCACACACTGG
orangutan **G****G****A****C**TCCTTGAGAAATAAACTGCACACACTGG

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



Substitution Rate

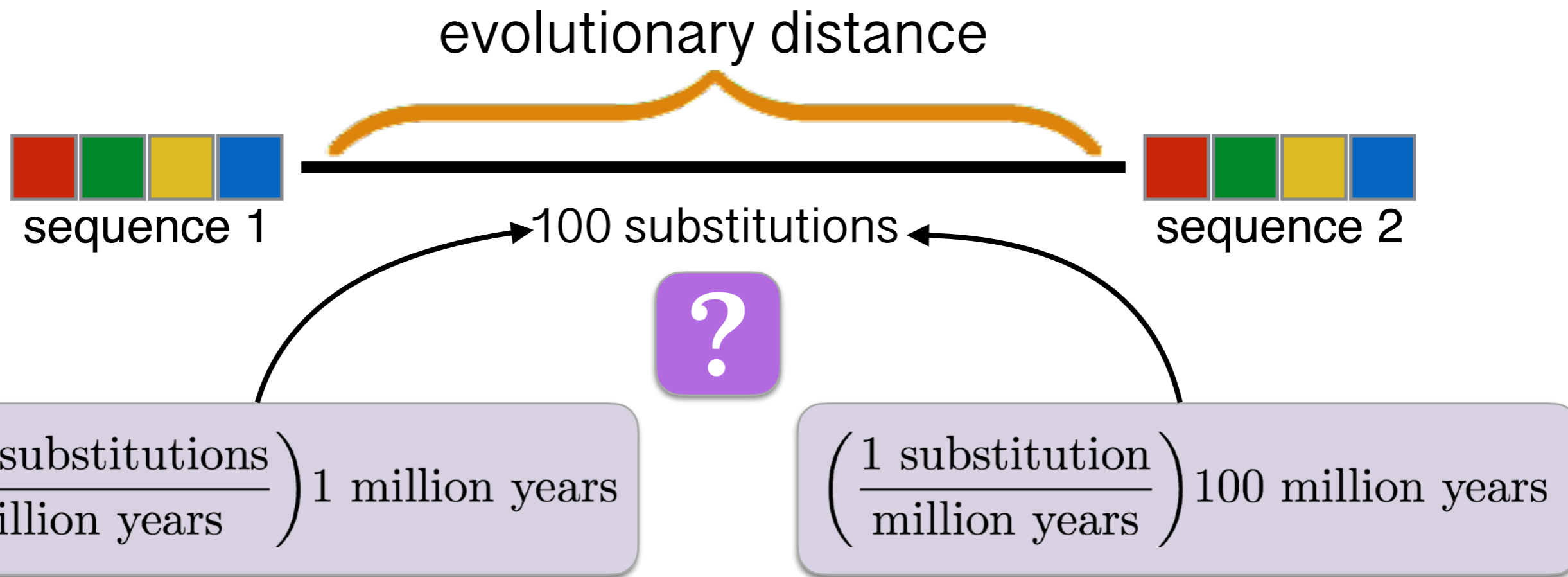
number of substitutions = substitution rate \times time



the overall substitution rate is 3α , so the expected number of substitutions (ν) is:

$$\nu = 3\alpha t$$

Rate and Time are Confounded



sequence data can only provide information about the evolutionary distance as **rate × time**, we cannot identify the absolute rate or time

we will cover how to estimate substitution rates and divergence times later

Evolutionary Distances



sequence 1



sequence 2

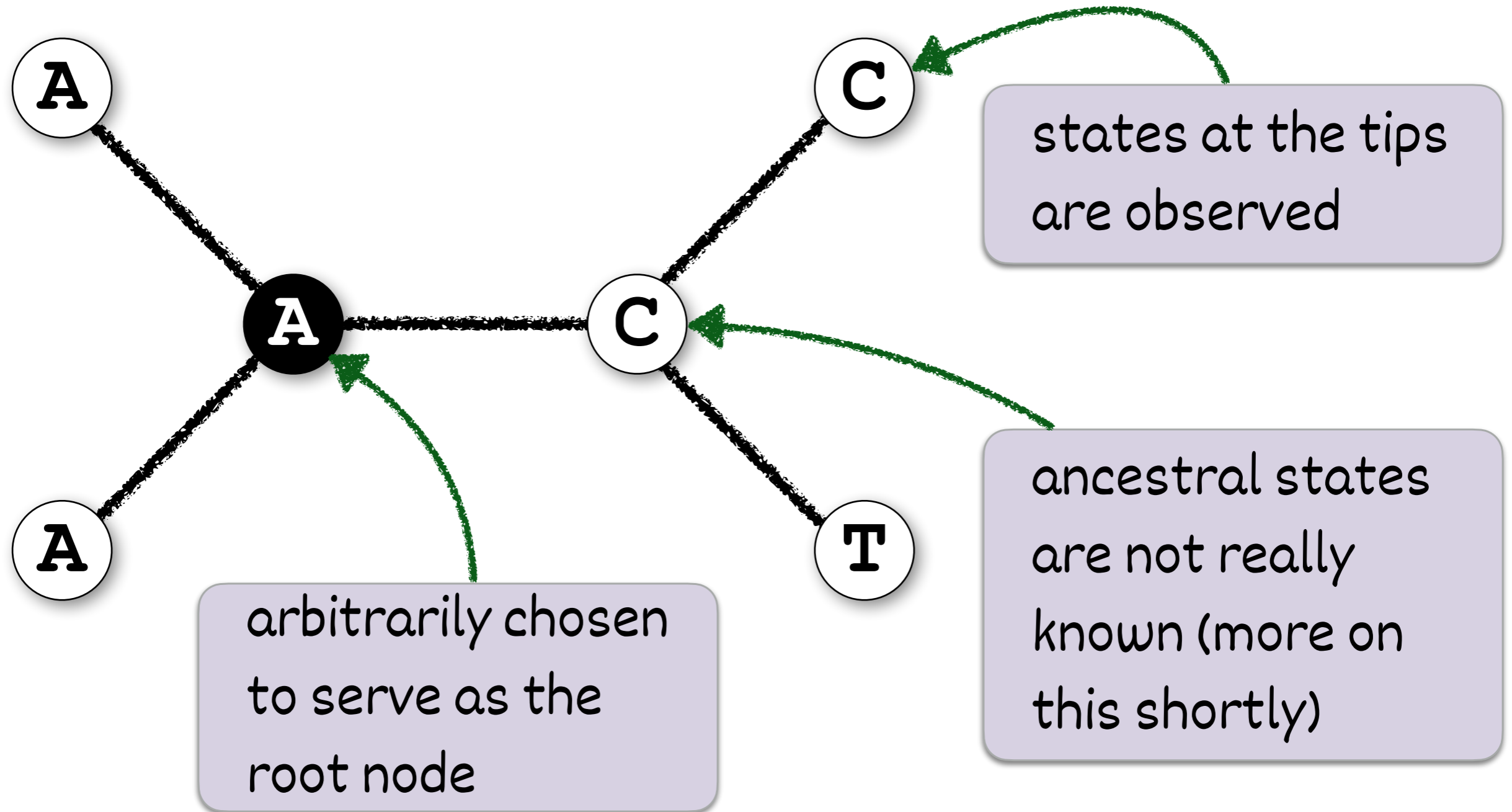
model	expected number of substitutions: $\nu = \{r\}t$
JC69	$\nu = \{3\alpha\}t$
F81	$\nu = \{2\mu(\pi_R\pi_Y + \pi_A\pi_G + \pi_C\pi_T)\}t$
K80	$\nu = \{\beta(\kappa + 2)\}t$
HKY	$\nu = \{2\mu[\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)]\}t$

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

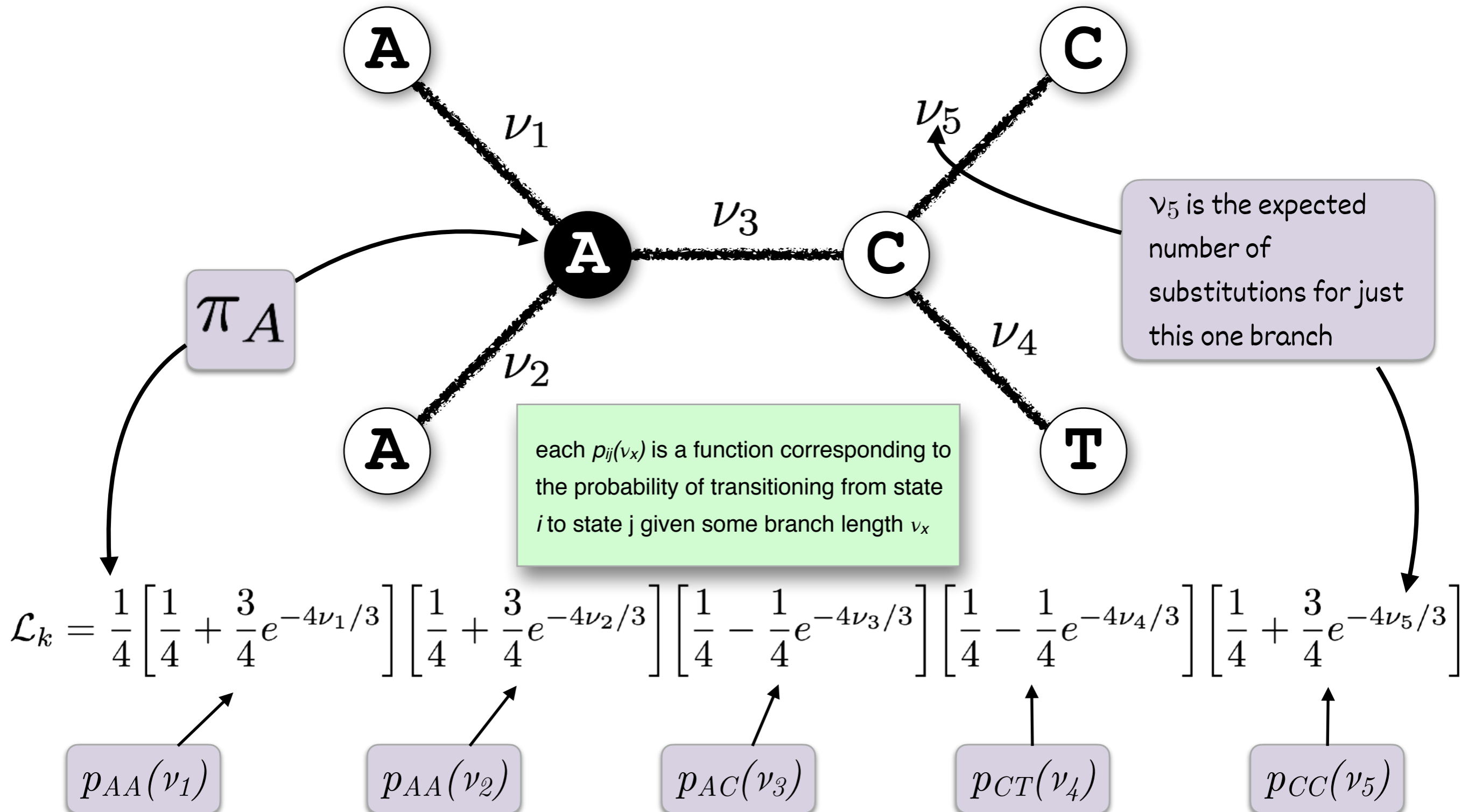
one substitution model parameter is always determined from the edge length; the others are usually global (i.e. same value applies to all edges)

Likelihood of an Unrooted Tree

(data shown only for 1 site)



Likelihood for a Single Site k



the AND rule !