

See also 18-Apr-2018 and 16-May-2018 at  
<http://phyloseminar.org/recorded.html>

# Likelihood in Phylogenetics

Workshop on Molecular Evolution  
Woods Hole, Massachusetts

28 May 2023

Paul O. Lewis  
Department of Ecology & Evolutionary Biology



---

# Probability

---

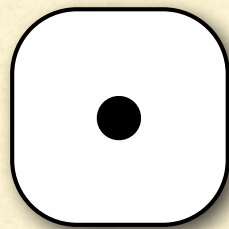


---

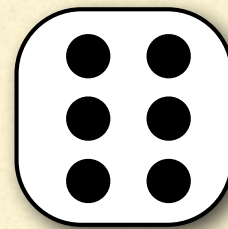
# Probabilities: the AND rule

---

Rolling 2 dice, what is the probability of seeing (simultaneously) a 1 on the first die and a 6 on the second die?



AND



$(1/6)$

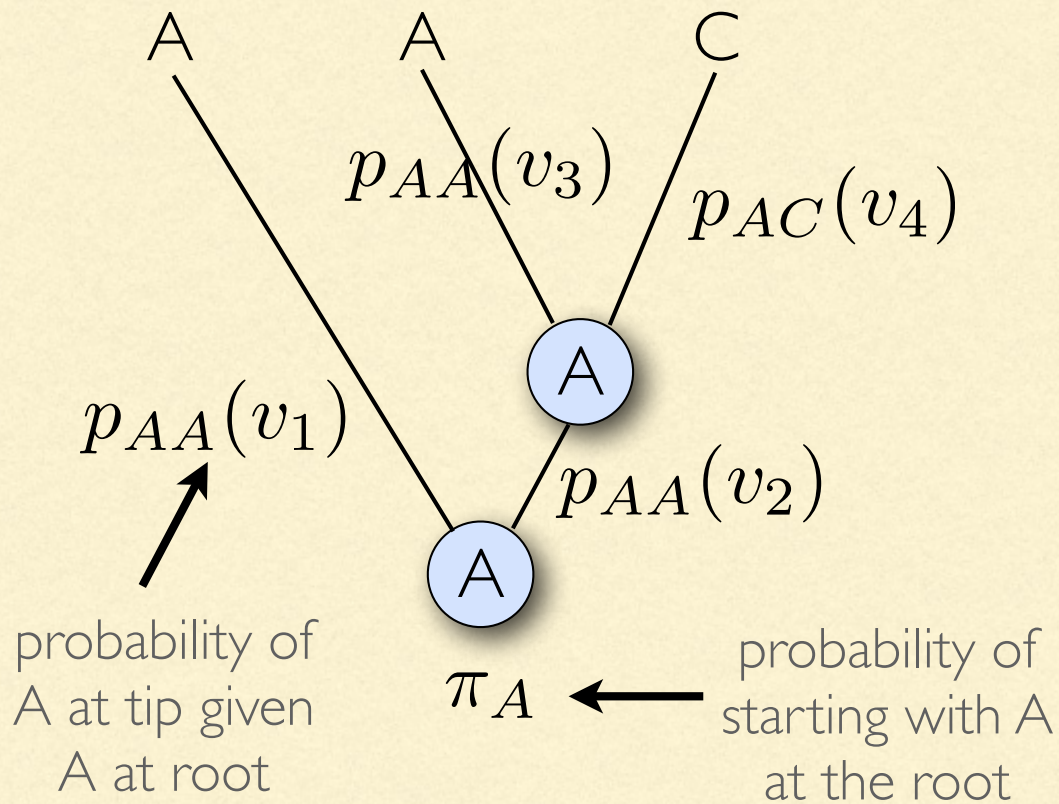
$\times$

$(1/6)$

$=$

$1/36$

# 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.

$$\Pr(A, A, C, A, A) = \pi_A p_{AA}(v_1) p_{AA}(v_2) p_{AA}(v_3) p_{AC}(v_4)$$

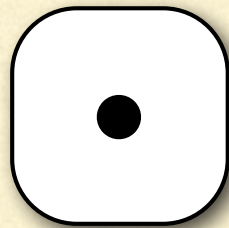


---

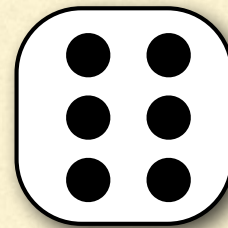
# Probabilities: the OR rule

---

Rolling 1 die, what is the probability of seeing either a 1 or a 6?



OR



$(1/6)$

+

$(1/6)$

=

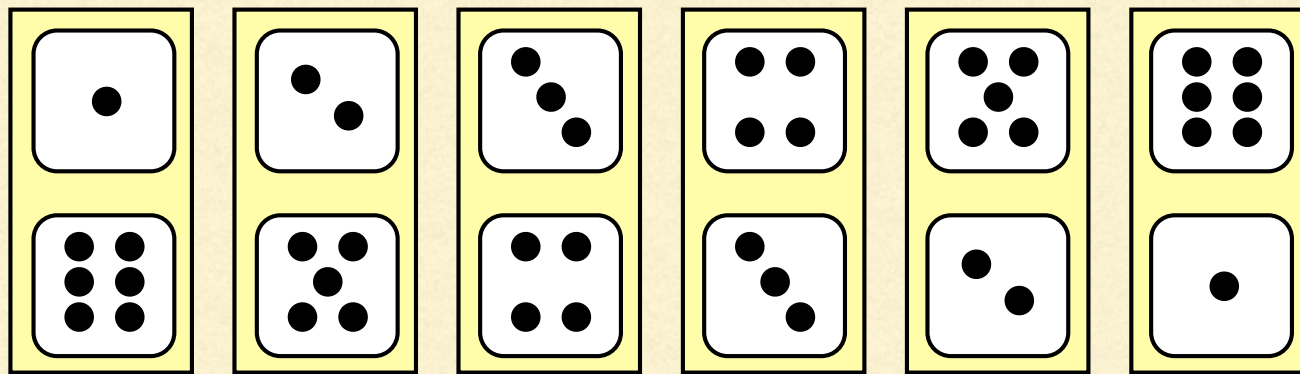
$1/3$

---

# Combining AND and OR

---

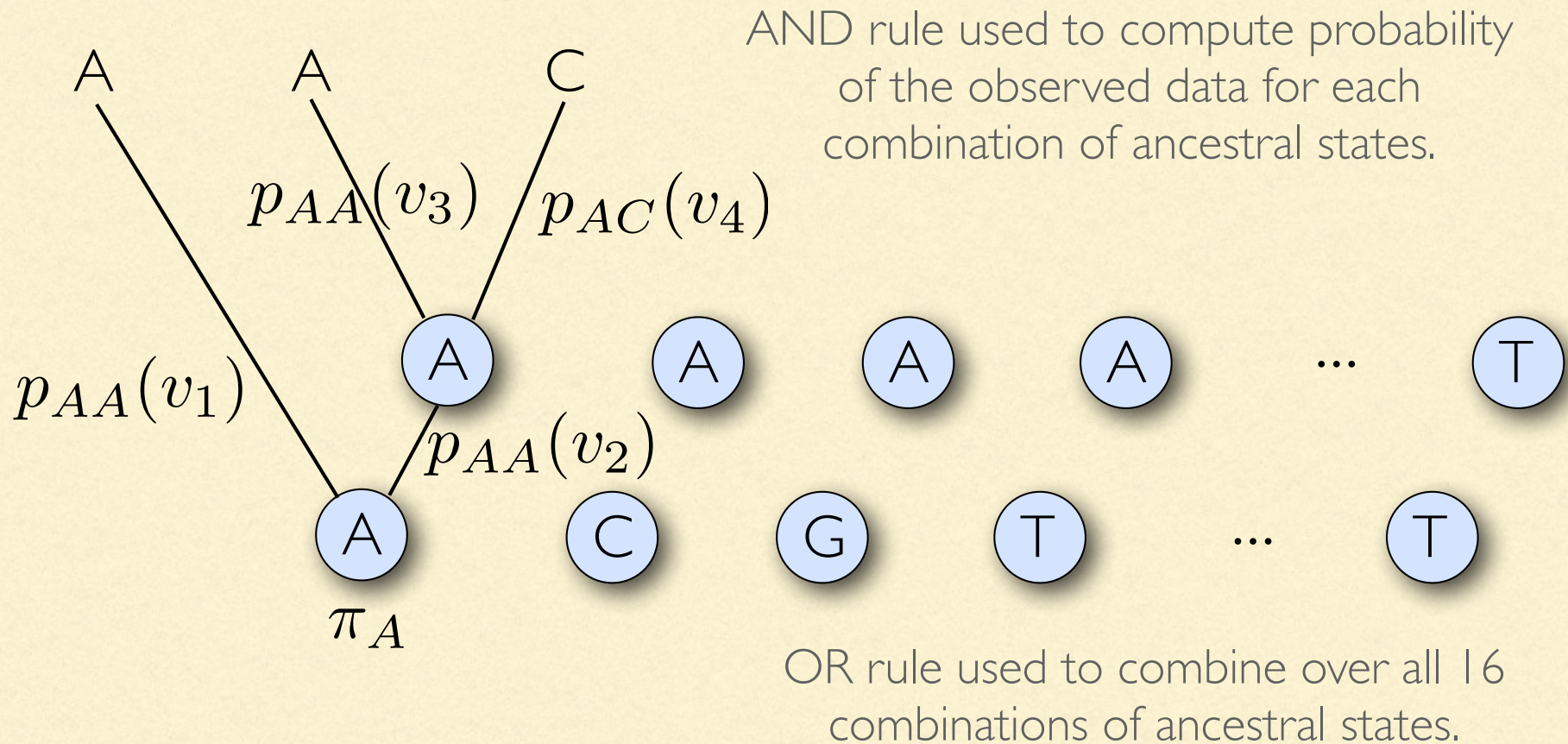
What is the probability that the sum of two dice is 7?



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



# Using both AND and OR in phylogenetics



$$\Pr(\mathbf{A}, \mathbf{A}, \mathbf{C}) = \Pr(\mathbf{A}, \mathbf{A}, \mathbf{C}, A, A) + \Pr(\mathbf{A}, \mathbf{A}, \mathbf{C}, A, C) + \dots + \Pr(\mathbf{A}, \mathbf{A}, \mathbf{C}, T, T)$$

---

# Independence

---

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

Probability of flipping a coin twice and  
getting heads both times:

$$\Pr(H, H) = \Pr(H) \Pr(H)$$



# Non-independence

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

↑ joint probability of A and B      ↑ conditional probability of B given A

$$\begin{aligned}\Pr(\text{walk to work}|\text{sunny}) &= 0.99 \\ \Pr(\text{walk to work}|\text{raining}) &= 0.50\end{aligned}$$

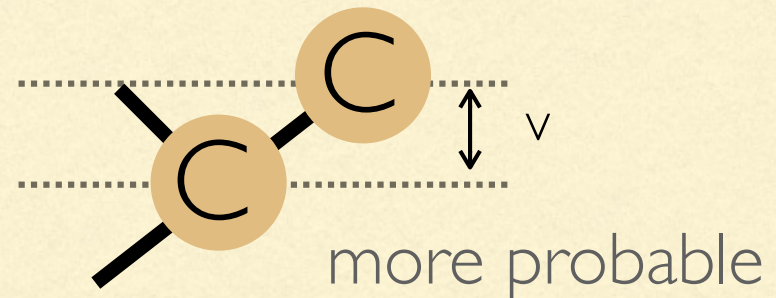
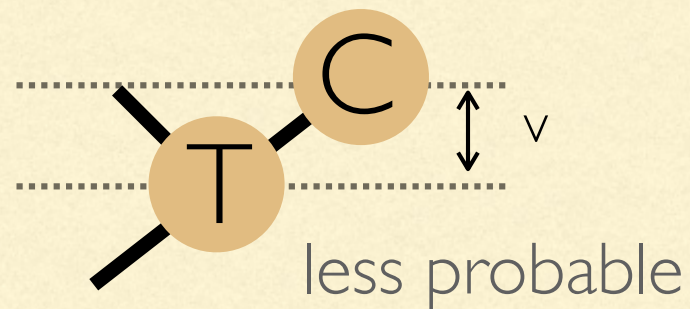
# Non-independence in phylogenies

Normally, for a given rate of substitution and time, the probability of the end state is *dependent* on the starting state

$$p(C|C, v) > p(C|T, v)$$

$$p_{CC}(v) > p_{TC}(v)$$

common notation for  
transition probabilities



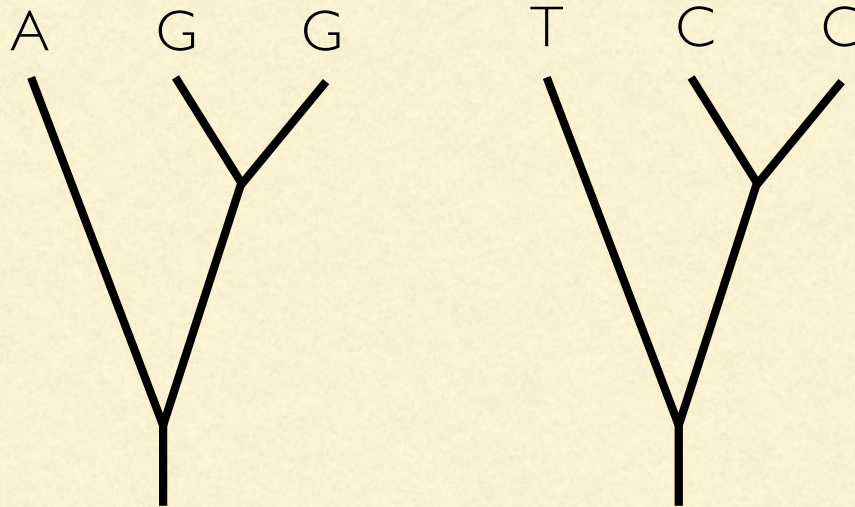


---

# Conditional Independence

---

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



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

---

# Likelihood

---



# Why do we need the term **likelihood**?

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

**Probability** of the **data**  
given the model  
**Likelihood** of the **model**  
given the data

# Likelihood of a single vertex

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

● **GAAGTCCTTGAGAAATAAACTGCACACACTGG**

$$L = \Pr(G) \Pr(A) \Pr(A) \Pr(G) \Pr(T) \cdots \Pr(G)$$

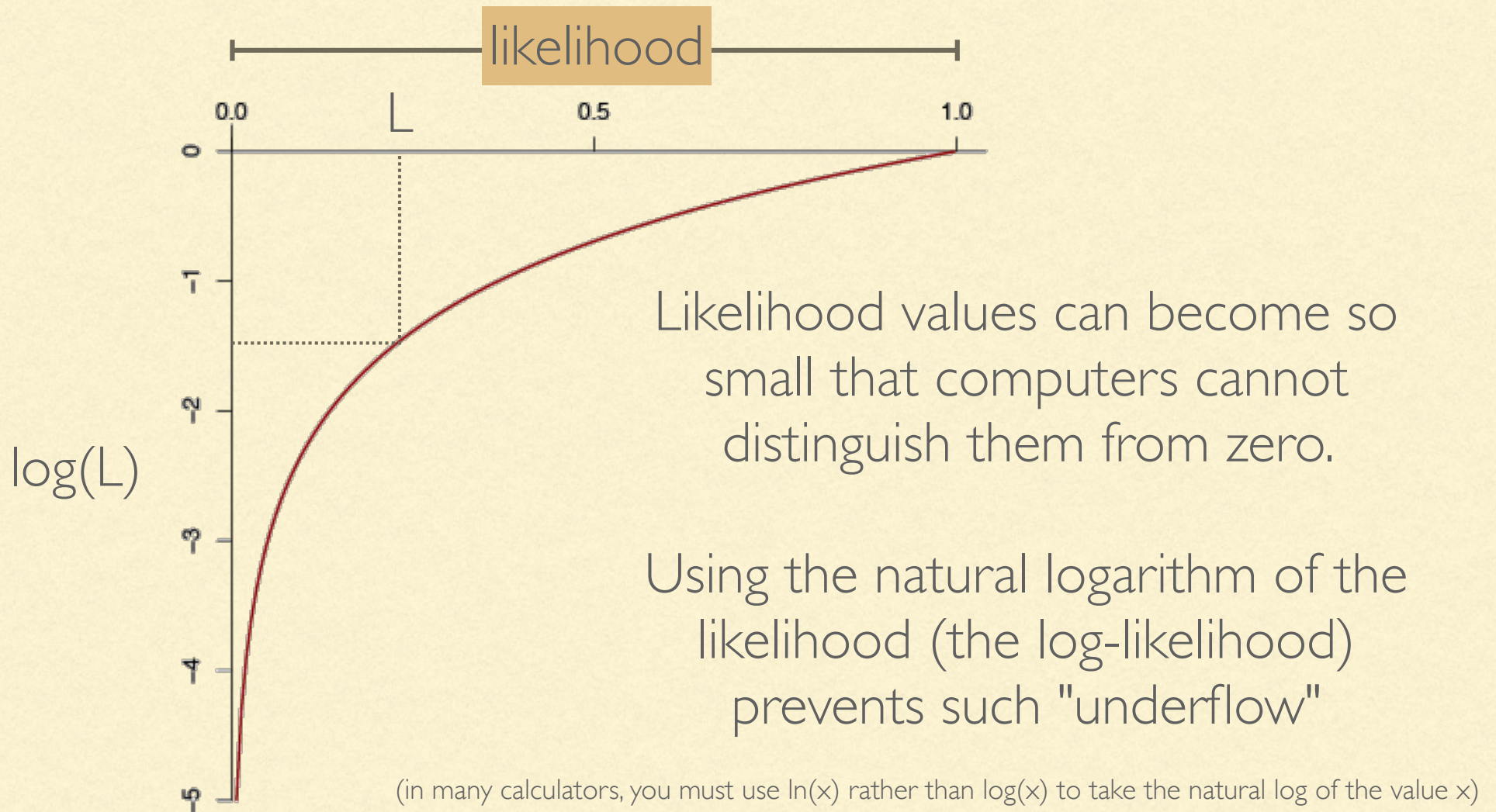
$$L = \pi_G \pi_A \pi_A \pi_G \pi_T \cdots \pi_G$$

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

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

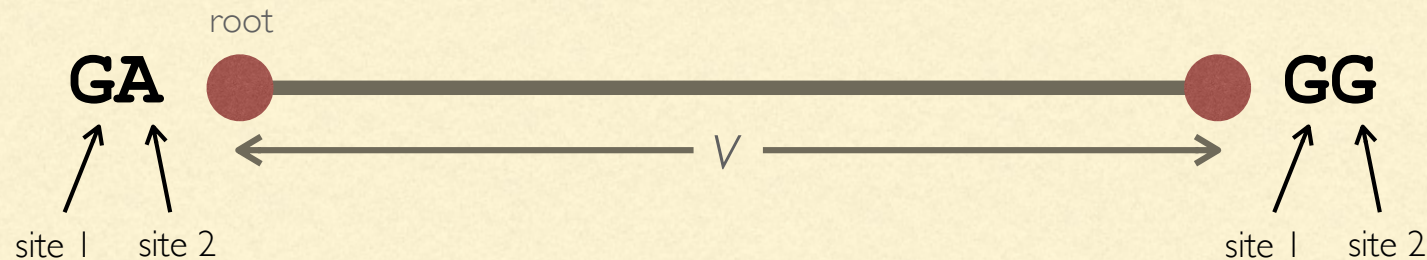


# Natural logarithm



# Likelihood of a single-edge tree

Two nodes have sequence data (but only for two sites)



$$L = \left[ \left( \frac{1}{4} \right) \left( \frac{1}{4} + \frac{3}{4} e^{-4v/3} \right) \right] \left[ \left( \frac{1}{4} \right) \left( \frac{1}{4} - \frac{1}{4} e^{-4v/3} \right) \right]$$

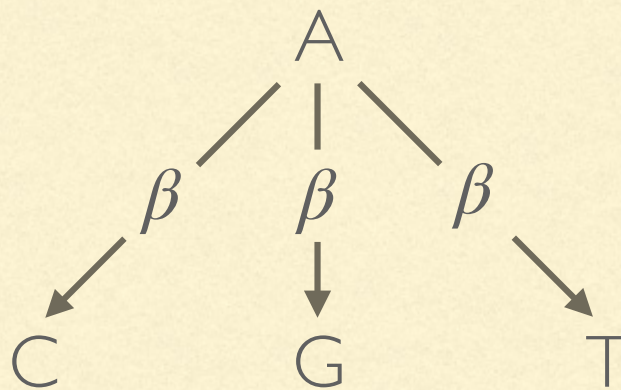
The diagram shows two orange boxes labeled 'site 1' and 'site 2'. Below 'site 1' is a double-headed arrow. Below 'site 2' is a double-headed arrow.

Each **site likelihood** is the probability of the **starting state** at the root ( $1/4$ ) times the **transition probability** (probability of the end state given the starting state)



# What is the edge length $v$ ?

$$\begin{array}{lcl} \text{expected} & & \\ \textbf{number} & \text{substitution} & \\ \text{of} & \textbf{rate} & \times \textbf{time} \\ \text{substitutions} & \text{per site} & \\ \text{per site} & = & \\ v & = & 3\beta \times t \end{array}$$



3 possible substitutions, each of which happens with rate  $\beta$

# Jukes and Cantor (1969)

JC69 model

to:

Parameters:  $\beta$

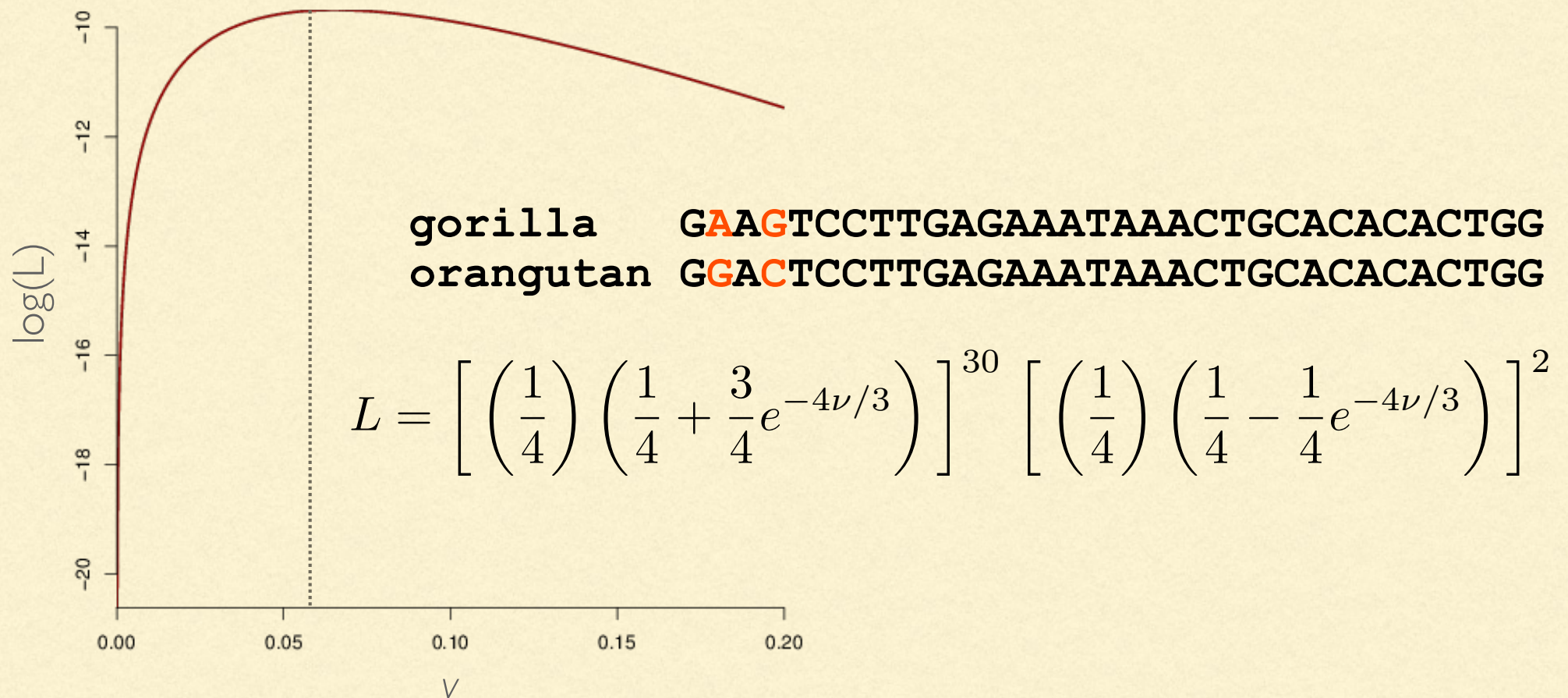
from:

	A	C	G	T
A	$-3\beta$	$\beta$	$\beta$	$\beta$
C	$\beta$	$-3\beta$	$\beta$	$\beta$
G	$\beta$	$\beta$	$-3\beta$	$\beta$
T	$\beta$	$\beta$	$\beta$	$-3\beta$



# Maximum likelihood estimation

0.065 is the maximum likelihood estimate (MLE) of  $\nu$



---

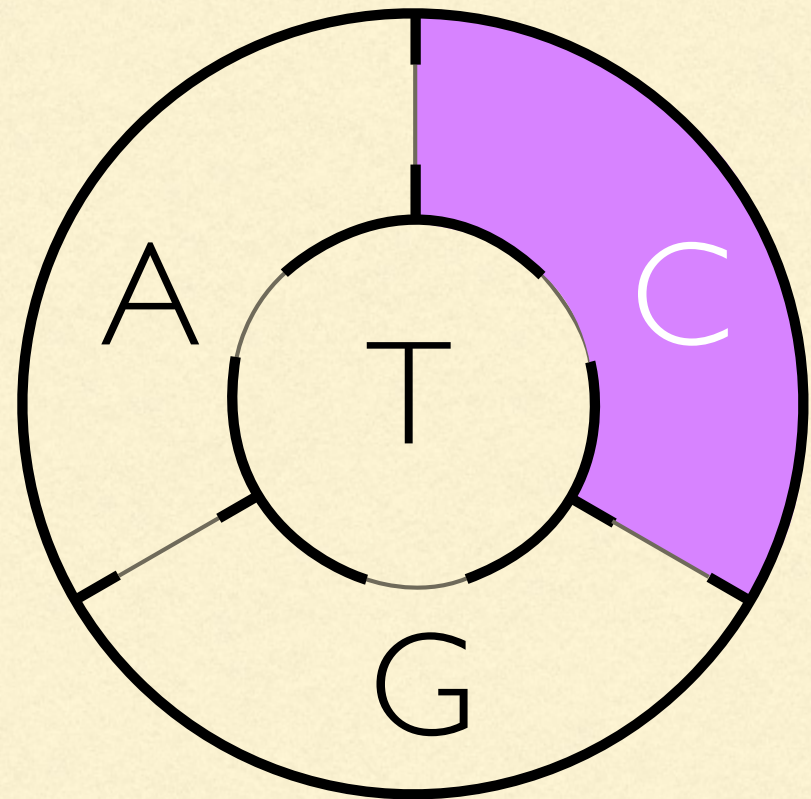
# Equilibrium Frequencies

---

Imagine a bottle of perfume has been spilled in room C.

The doors to the other rooms are closed, so the perfume has, thus far, not been able to spread.

What would happen if we opened all the doors?



Architect: Joe Bielawski



# Equilibrium Frequencies

At the **instant the doors open**, perfume molecules...

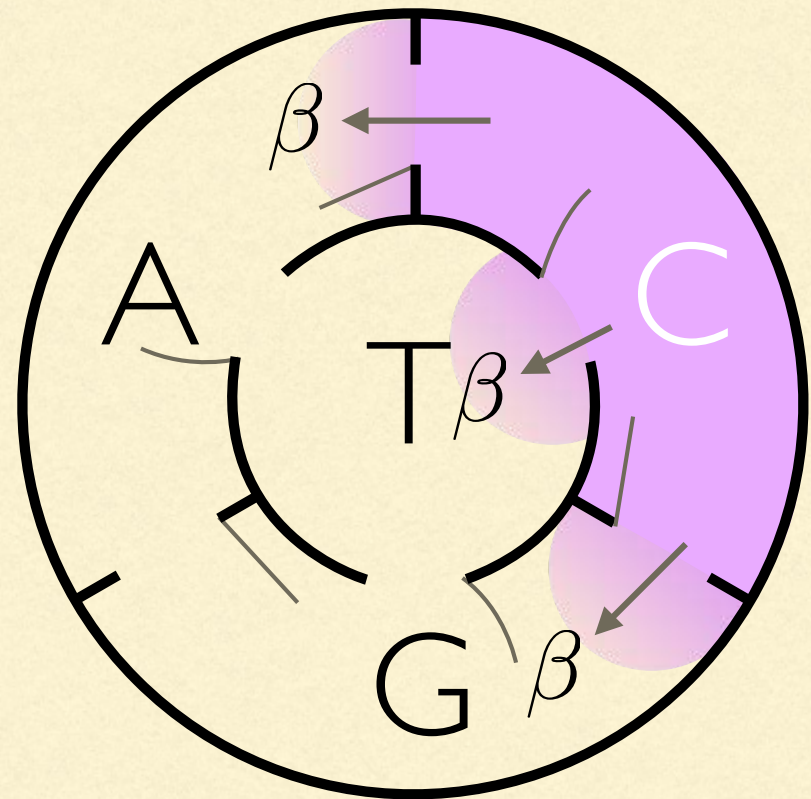
enter room A at rate  $\beta$

enter room T at rate  $\beta$

enter room G at rate  $\beta$

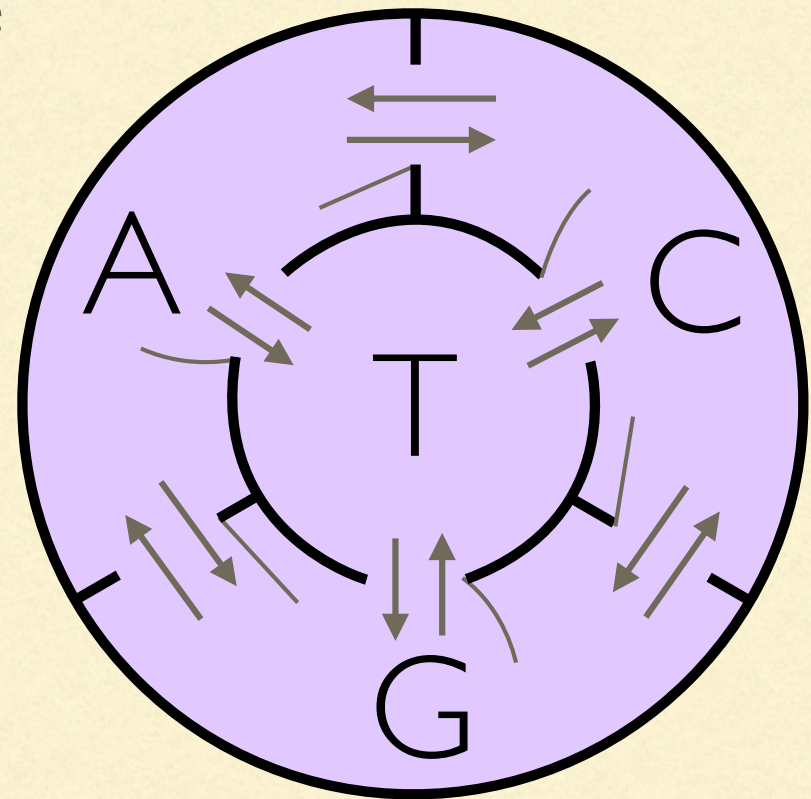
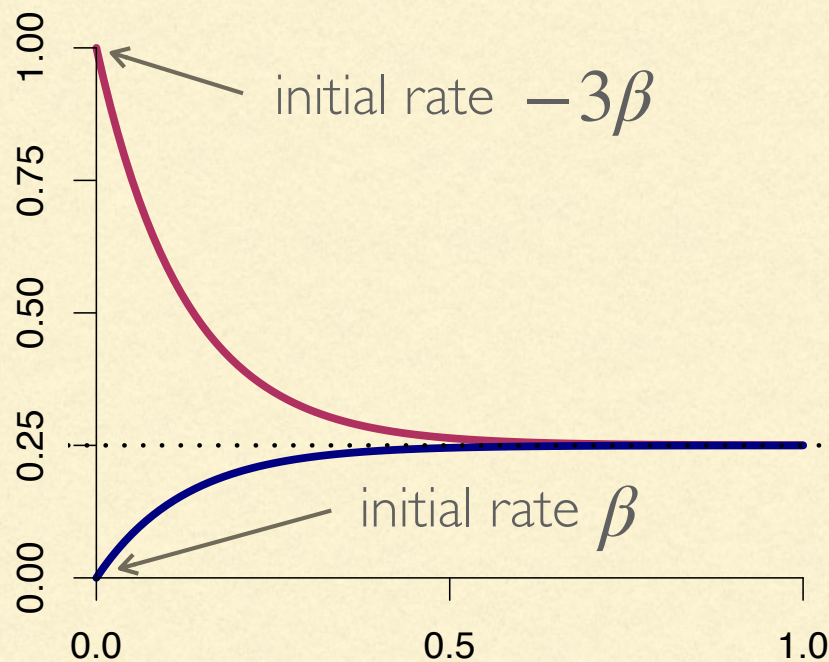
enter room C at rate  $-3\beta$

(you could also say they *leave* C at rate  $3\beta$ )



# Equilibrium Frequencies

At **equilibrium**, the relative concentration of perfume is **equal** in all rooms



$$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$$



---

# Transition probability demo

---

<https://plewis.github.io/applets/jc-transition-probabilities/>

---

# Sequence data for four taxa

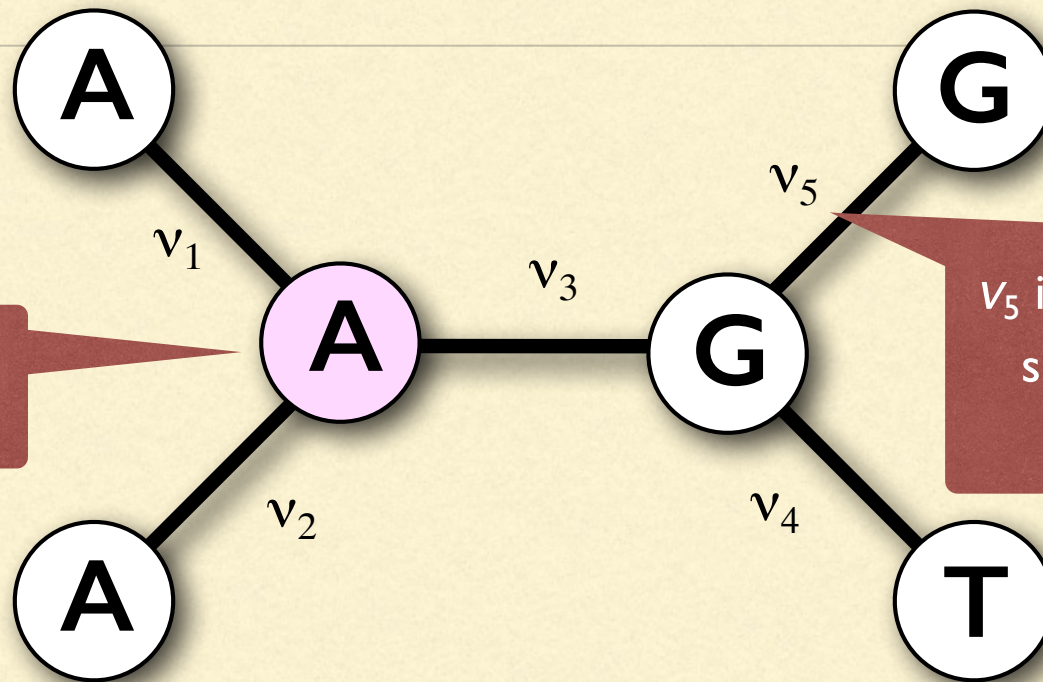
---

one site

Sphagnum	GGCAGCATTTCGAATGACTCCTCAACCTGGAGT	<b>A</b>	TCACCCG...
Asplenium	GGCAGCTTTCCGGATGACCCCAACAACCCGGAGT	<b>A</b>	TCAGCTG...
Picea	GGCAGCATTCCGAGTAACTCCTCAACCAGGGGT	<b>G</b>	TCGCCCG...
Avena	GGCAGCATTCCGAGTAACTCCTCAACCTGGGGT	<b>T</b>	TCGCCGG...



# Likelihood for tree (one site)



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

---

# Total likelihood

---

$$L = L_1 L_2 \cdots L_n$$

↑      ↑      ↑  
site 1   site 2   site n


$$\log L = \log L_1 + \log L_2 + \cdots + \log L_n$$



# Jukes and Cantor (1969)

JC69 model

to:

Parameters:  $\beta$

from:

$$\begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{bmatrix} -3\beta & \beta & \beta & \beta \\ \beta & -3\beta & \beta & \beta \\ \beta & \beta & -3\beta & \beta \\ \beta & \beta & \beta & -3\beta \end{bmatrix}$$

# Kimura (1980)

K80 (or K2P) model

Parameters:  $\alpha, \beta$

$$\begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{bmatrix} -\alpha - 2\beta & \beta & \alpha & \beta \\ \beta & -\alpha - 2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha - 2\beta & \beta \\ \beta & \alpha & \beta & -\alpha - 2\beta \end{bmatrix}$$



# Kimura (1980)

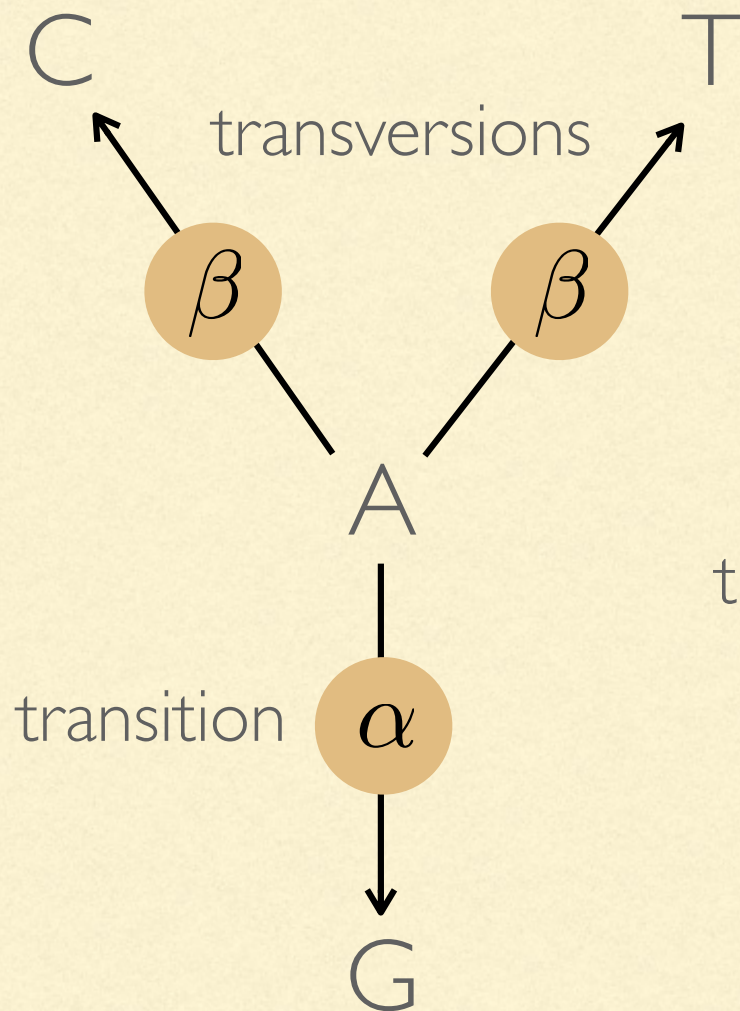
K80 (or K2P) model

$$\kappa = \alpha/\beta$$

Parameters:  $\kappa, \beta$

	A	C	G	T
A	$-\beta(\kappa + 2)$	$\beta$	$\kappa\beta$	$\beta$
C	$\beta$	$-\beta(\kappa + 2)$	$\beta$	$\kappa\beta$
G	$\kappa\beta$	$\beta$	$-\beta(\kappa + 2)$	$\beta$
T	$\beta$	$\kappa\beta$	$\beta$	$-\beta(\kappa + 2)$

# Transition-transversion (rate) ratio



transition rate =  $\alpha$

transversion rate =  $\beta$

assume  $\alpha = \beta$

transition-transversion rate ratio = 1.0

transition-transversion ratio = 0.5



# Felsenstein (1981)

F81 model

Parameters:  $\mu, \pi_A, \pi_C, \pi_G$

	A	C	G	T
A	$-\mu(1 - \pi_A)$	$\pi_C \mu$	$\pi_G \mu$	$\pi_T \mu$
C	$\pi_A \mu$	$-\mu(1 - \pi_C)$	$\pi_G \mu$	$\pi_T \mu$
G	$\pi_A \mu$	$\pi_C \mu$	$-\mu(1 - \pi_G)$	$\pi_T \mu$
T	$\pi_A \mu$	$\pi_C \mu$	$\pi_G \mu$	$-\mu(1 - \pi_T)$

---

# JC69 is a special case of F81

---

	A	C	G	T		A	C	G	T
A	$-\frac{3}{4}\mu$	$\frac{1}{4}\mu$	$\frac{1}{4}\mu$	$\frac{1}{4}\mu$	$\left[ \begin{array}{cccc} -3\beta & \beta & \beta & \beta \\ \beta & -3\beta & \beta & \beta \\ \beta & \beta & -3\beta & \beta \\ \beta & \beta & \beta & -3\beta \end{array} \right]$	$-3\beta$	$\beta$	$\beta$	$\beta$
C	$\frac{1}{4}\mu$	$-\frac{3}{4}\mu$	$\frac{1}{4}\mu$	$\frac{1}{4}\mu$		$\beta$	$-3\beta$	$\beta$	$\beta$
G	$\frac{1}{4}\mu$	$\frac{1}{4}\mu$	$-\frac{3}{4}\mu$	$\frac{1}{4}\mu$		$\beta$	$\beta$	$-3\beta$	$\beta$
T	$\frac{1}{4}\mu$	$\frac{1}{4}\mu$	$\frac{1}{4}\mu$	$-\frac{3}{4}\mu$		$\beta$	$\beta$	$\beta$	$-3\beta$

$$\beta = \frac{1}{4}\mu$$



# Hasegawa, Kishino, and Yano (1985)

HKY85 model

Parameters:  $\mu, \kappa, \pi_A, \pi_C, \pi_G$

these are global  
parameters  
(apply to all  
edge lengths)

one parameter in each model is  
associated with the length of an edge

	A	C	G	T
A	$-\mu(\pi_C + \pi_G\kappa + \pi_T)$	$\pi_C\mu$	$\pi_G\mu\kappa$	$\pi_T\mu$
C	$\pi_A\mu$	$-\mu(\pi_A + \pi_G + \pi_T\kappa)$	$\pi_G\mu$	$\pi_T\mu\kappa$
G	$\pi_A\mu\kappa$	$\pi_C\mu$	$-\mu(\pi_A\kappa + \pi_C + \pi_T)$	$\pi_T\mu$
T	$\pi_A\mu$	$\pi_C\mu\kappa$	$\pi_G\mu$	$-\mu(\pi_A + \pi_C\kappa + \pi_G)$

# Tavaré (1986)

GTR model

Parameters: ?

$$\begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{bmatrix} \text{A} & \text{C} & \text{G} & \text{T} \\ \text{—} & \pi_C \mu \textcircled{a} & \pi_G \mu \textcircled{b} & \pi_T \mu \textcircled{c} \\ \pi_A \mu a & \text{—} & \pi_G \mu \textcircled{d} & \pi_T \mu \textcircled{e} \\ \pi_A \mu b & \pi_C \mu d & \text{—} & \pi_T \mu \textcircled{f} \\ \pi_A \mu c & \pi_C \mu e & \pi_G \mu f & \text{—} \end{bmatrix}$$

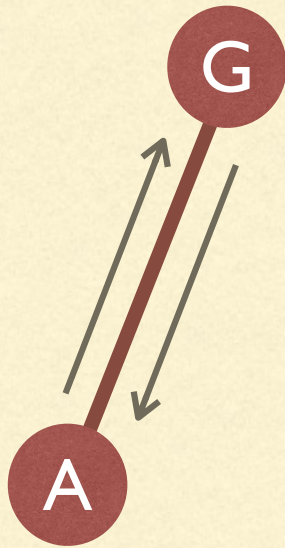
exchangeability  
parameters are  
circled



---

# Time reversibility

---



Time reversibility means...

$$\Pr(A) \Pr(G|A, v) = \Pr(G) \Pr(A|G, v)$$

Time reversibility allows any point on the tree to serve as the root, and thus has some practical advantages, but time reversibility is not a requirement for substitution models used in phylogenetics

---

# Rate heterogeneity

---



# Green plant rbcL gene

First 88 amino acids (translation is for *Zea mays*)

M--S--P--Q--T--E--T--K--A--S--V--G--F--K--A--G--V--K--D--Y--K--L--T--Y--Y--T--P--E--Y--E--T--K--D--T--D--I--L--A--A--F--R--V--T--P--	
Chara	(green alga; land plant lineage) AAAGATTACAGATTAACTTACTATACTCCTGAGTATAAACTAAAGATACTGACATTTTAGCTGCATTTCGTGTAAGTCCA
Chlorella	(green alga) .....C...C.T.....T..CC..C.A....C.....T...C.T..A..G..C...A.G.....T
Volvox	(green alga) .....TC.T....A....C..A....C...GT.GTA....C.....C.....A.....A.G.....
Conocephalum	(liverwort) .....TC.....T.....G..T...G.....G..T.....A.....A.AA.G.....T
Bazzania	(moss) .....T.....C..T....G....A...G.G..C....G..A..T....G..A.....A.G.....C
Anthoceros	(hornwort) .....T.....CC.T....C....T..CG.G..C..G.....T....G..A..G.C.T.AA.G.....T
Osmunda	(fern) .....TC...G...C.....C..T...G.G..C..G.....T....G..A...C...AA.G.....T
Lycopodium	(club "moss") .GG.....C.T..C.....T....G..C....A..C..T....C.G..A.....AA.G.....T
Ginkgo	(gymnosperm; Ginkgo biloba) .....G.....T.....A..C....C.....T..C..G..A....C..A.....T
Picea	(gymnosperm; spruce) .....T.....A..C.G..C.....G..T....G..A....C..A.....T
Iris	(flowering plant) .....G.....T.....T..CG...C.....T..C..G..A....C..A.....T
Asplenium	(fern; spleenwort) .....TC..C.G....T..C..C..C..A..C..G..C.....C..T..C..G..A..T..C..GA.G..C...
Nicotiana	(flowering plant; tobacco) .....G...A..G....T.....CC...C..G.....T..A..G..A....C..A.....T

Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--	
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAAGTATTTGGACCGATACAAAGGAAGATGCTACGATATTGAA	
.....A..T.....A.....G..T..G.....A.....A.....T.....G.....T..T.....A.....T.....TC..T..T..C..C..G	
.....A..T.....TGT..T....T..T....T....A..A..A....T....A....A....T..T....A..C.T....T.....TC..T..T..C..C..G	
..G.....G..A...G.A.....A..A....T....T.....A.....T..TC.T...ACC.T..T..T..T....TC.....T.G.....C	
.....G..A..A.....A..G.....T.....A..C....G....C..G.....C..T..GC.T..A...C.C..T..T.....TC.....T..C..C...	
T...A..G..G.....A..C.....T.....A.....C..T..C.T..C..CC.T....T.....TC.....C.....	
.....C..A..A..GG...G....T..A.....G.....A....G....C....A....G..T..C.T..C..C.T..T..T..T..G..TC.....	
....T...A..A....C..G....G..A..C.....T.....C.....C..T..C.T..C..C.C..T..C.....TC.G....T..A.....	
.....A..G.....G....G..A.....C.....C.....C..T..C.T..C..C.C..T..T..T..G.....T..C..C..G	
.....A..G..G..C..G....G..A..A.....T.....C..C.....C.....C..T..C.T..C..C.T..T..T..G..GC.....T..C..C..G	
.....C..A....TG.....G....C..G....C.....A..A..G.....T.....C.T..C..C.T..T..T.....C.....C..C..G	
.....C..A..A..G.....C..A.....G..C....A.....C.....G....A.....G..G..C..CC.T....T....G..CC.....C..G	
.....A.....C..G.....C.....A.....A.....C..T..C.T..C..CC.T..T..T.....GC.....CGC...C..G	

All 4 bases are  
observed at  
some sites...

...while at other  
sites, only 1 base  
is observed

# Site-specific rates

Each defined subset (e.g. 1st+2nd pos. versus 3rd pos.) has its own relative rate

CACCGGGTCCCCGAGAGCGGGCGCGTGC	CGATCTCACGGACTGACACGTTGACGAGGTTACAGTTGACGTAAAGGAGTGTAGAATGA	ATCTATAAAGTAATAATTTTAGTTTGTACATTGCACAAACCTTA
.....TG.....	.....C.....	.AT..A..GTG..A..AA..T.G.A..TT...A.T..TTTCCG
.....G.....	.....C.....AC.....C.....G.....	.AT.....TT.TT.T.AAA.T.A.A..TT.A.T.T..TTTCCG
.....T.....	.....C.....C.....C.....	G.GA.A...AA.T.T.....A...TTT.CTTT.T..T..C
.....G.....	.....C.....C.....C.....	.GAA....AG...T..AC.G.CG..CGTTA.CTT..T..TCC.
...T.....	.....C.....C.....C.....	.AGG....AC...T..A.....C.TTCCT.T..T...C..
.....G.....	.....C.....C.....C.....	.CAAG.G.TA...G...A.G.C.A.G.TTC.TTTTGT.....
...T.....	.....C.....C.....C.....	..AA.CG.GAC...T..C.....C.TTC.CTC..TG.TA..
.....G.....	.....C.....C.....C.....	..AG..G.GA...C..C...C...C.TTC.TTT.G...TCCG
.....G.....	.....C.....C.....C.....	.AGGGCG.GAA...T..CC...C...C.TT..TTT.GG..TCCG
.....G.....	.....C.....C.....C.....	.CA.T...G.CG..C.....AAG...TTC.TTT.....CCG
.....G.....	.....C.....C.....C.....	.CAA....CA....GC.A...C.G.AG.GCCT.T.GC...CG
.....G.....	.....C.....C.....C.....CG.....	..A.....CG..C.....A.A.C.TTCCTTT..G...CCG

$r_1$  applies to subset 1  
1st+2nd codon positions  
(sites 1 - 88)

$r_2$  applies to subset 2  
3rd codon positions  
(sites 89-132)

Relative rates have mean 1.0: 
$$\underbrace{r_1}_{2/3} p(r_1) + \underbrace{r_2}_{1/3} p(r_2) = 1$$



# Site-specific rates

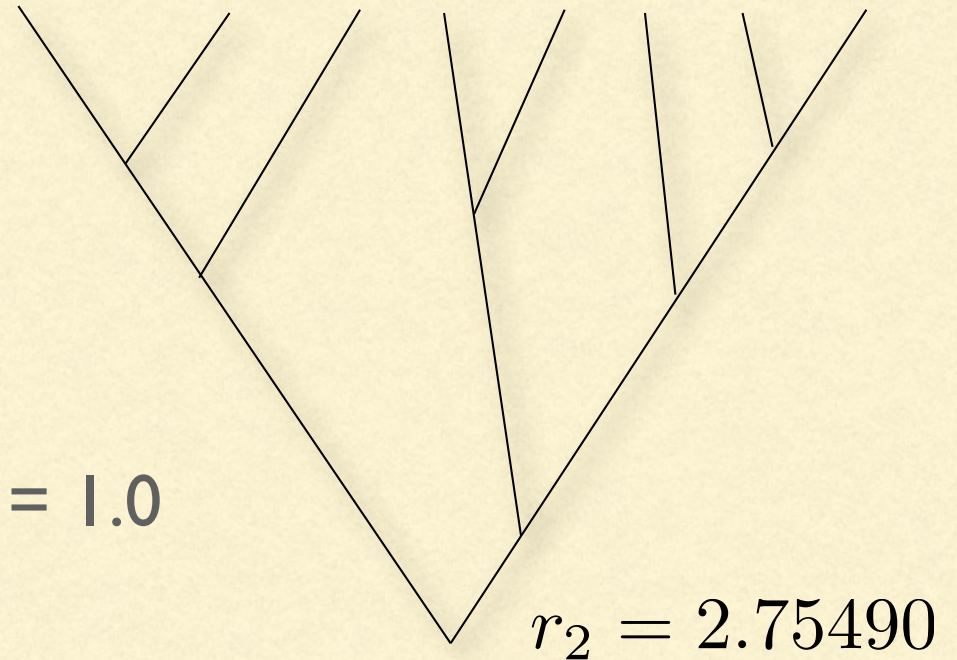
$$L = \underbrace{p(\mathbf{y}_1|r_1) \cdots p(\mathbf{y}_{88}|r_1)}_{\text{1st+2nd codon positions}} \underbrace{p(\mathbf{y}_{89}|r_2) \cdots p(\mathbf{y}_{132}|r_2)}_{\text{3rd codon positions}}$$



$$r_1 = 0.12255$$

mean relative rate:

$$(0.12255)(2/3) + (2.75490)(1/3) = 1.0$$



---

# Site-specific rates

---

JC69 transition probabilities that would be used for every site if rate *homogeneity* were assumed:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\beta t} \quad \text{C} \xrightarrow{\text{identity}} \text{C}$$
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \quad \text{C} \xrightarrow{\text{difference}} \text{T}$$



---

# Site specific rates

---

JC69 transition probabilities that would be used for sites in **subset 1**:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4r_1\beta t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4r_1\beta t}$$

JC69 transition probabilities that would be used for sites in **subset 2**:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4r_2\beta t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4r_2\beta t}$$

# Mixture models

All  $k$  relative rates applied to every site

```
Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAACTAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA
...A..T.....A.....G..T..G.....A.....A..A.....T..G.....A.....T..T.....A.....T.....TC..T..T..T..C..C..G
...A..T.....TGT..T.....T..T.....T.....A..A..A.....T.....A.....A.....T..T.....A.....C..T.....T.....TC..T..T..T..C..C..G
..G.....G..A..G..A.....A..A.....T.....T.....A.....A.....T..TC..T.....ACC..T..T..T..T.....TC.....T..G.....C
...G..A..A.....A..G.....T.....A..C.....G.....C..G.....C..T..GC..T..A.....C..C..T..T.....TC.....T..C..C...
T...A..G..G.....A..C.....T.....A.....A.....C.....T.....C..T..C..T..C..CC..T.....T.....TC.....C.....
...C..A..A..GG...G.....T..A.....G.....A.....G.....C.....A.....G..T..C..T..C.....C..T..T..T..G..TC.....
...T..A..A.....C..G.....G..A..C.....T.....C.....C.....C.....T..C..T..C.....C..C..T..C.....TC..G.....T..A.....
...A..G.....G.....G..A.....C.....C.....C.....C.....C.....C.....T..C..T..C.....C..T..T..T..G.....TC.....C..C..G
...A..G..G..G..C..G.....G..A..A.....T.....C..C.....C.....C.....T..C..T.....C..T..T..T..G..GC.....T..C..C..G
...C..A.....TG.....G.....C..G.....C.....A..A..G.....T.....C..T..C.....C..T..T..T.....C.....C..C..C..G
...C..A..A.....G.....C..A.....G.....C.....A.....C.....G.....A.....G..G..C..CC..T.....T.....G..CC.....C..C..G
...A.....C.....C..G.....C.....A.....A.....C.....T.....C..T..C..CC..T..T..T.....GC.....CGC..C..G
```

site  $i$

$$L_i = p(\mathbf{y}_i|r_1)p(r_1) + p(\mathbf{y}_i|r_2)p(r_2) + \cdots + p(\mathbf{y}_i|r_k)p(r_k)$$

Common examples  $\left\{ \begin{array}{l} \text{Invariable sites (I) model} \\ \text{Discrete Gamma (G) model} \end{array} \right.$



---

# Invariable sites model (Reeves 1992)

---

$$L_i = p(\mathbf{y}_i|r_1)p(r_1) + p(\mathbf{y}_i|r_2)p(r_2)$$

$$L_i = p(\mathbf{y}_i|0.0)p_{\text{invar}} + p(\mathbf{y}_i|r_2)(1 - p_{\text{invar}})$$

# Discrete Gamma model (Yang 1994)

No relative rate is exactly 0.0, and all are equally probable

```
Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA
...A.T.....A.....G.T.G.....A.....A.A.....T.....G.....A.....T.T.....A.....T.....TC.T.T.T.C.C..G
...A.T.....TGT..T...T.T...T...A.A.A...T...A...A.....T.T...A...C.T...T.....TC.T.T.T.C.C..G
...G...G.A..G.A.....A.A...T...T...A.....A.....T.TC.T...ACC.T..T.T.T...TC.....T.G.....C
...G.A.A.....A.G.....T.....A.C...G...C...G...C.T.GC.T.A...C.C..T.T...TC.....T.C.C..
T...A.G.G.....A.C.....T.....A.....A.....C.T..C.T.C.CC.T...T.....TC.....C.....
...C.A.A.GG...G...T.A.....G.....A...G...C...A...G.T...C.T.C...C.T..T.T..G.TC.....
...T.A.A...C.G...G.A.C.....T.....C.....C.T..C.T.C...C.C..T.C...TC.G...T.A.....
...A.G...G...G.A.....C.....C.....C.....C.T..C.T.C.C.C.T..T..G.....T.C.C..G
...A.G.G.G.C.G...G.A.A.....T.....C.C.....C.....C.T..C.T..C.T..T..G.GC.....T.C.C..G
...C.A...TG.....G...C.G...C.....A.A.G...T...C.T.C...C.T..T..C.....C.C.C..G
...C.A.A.G.....C.A.....G.C...A.....C...G...A...G..G.C.CC.T...T..G.CC.....C.G
...A.....C.G.....C.....A.....C.T..C.T.C.CC.T..T..GC.....CGC..C..G
```

site  $i$

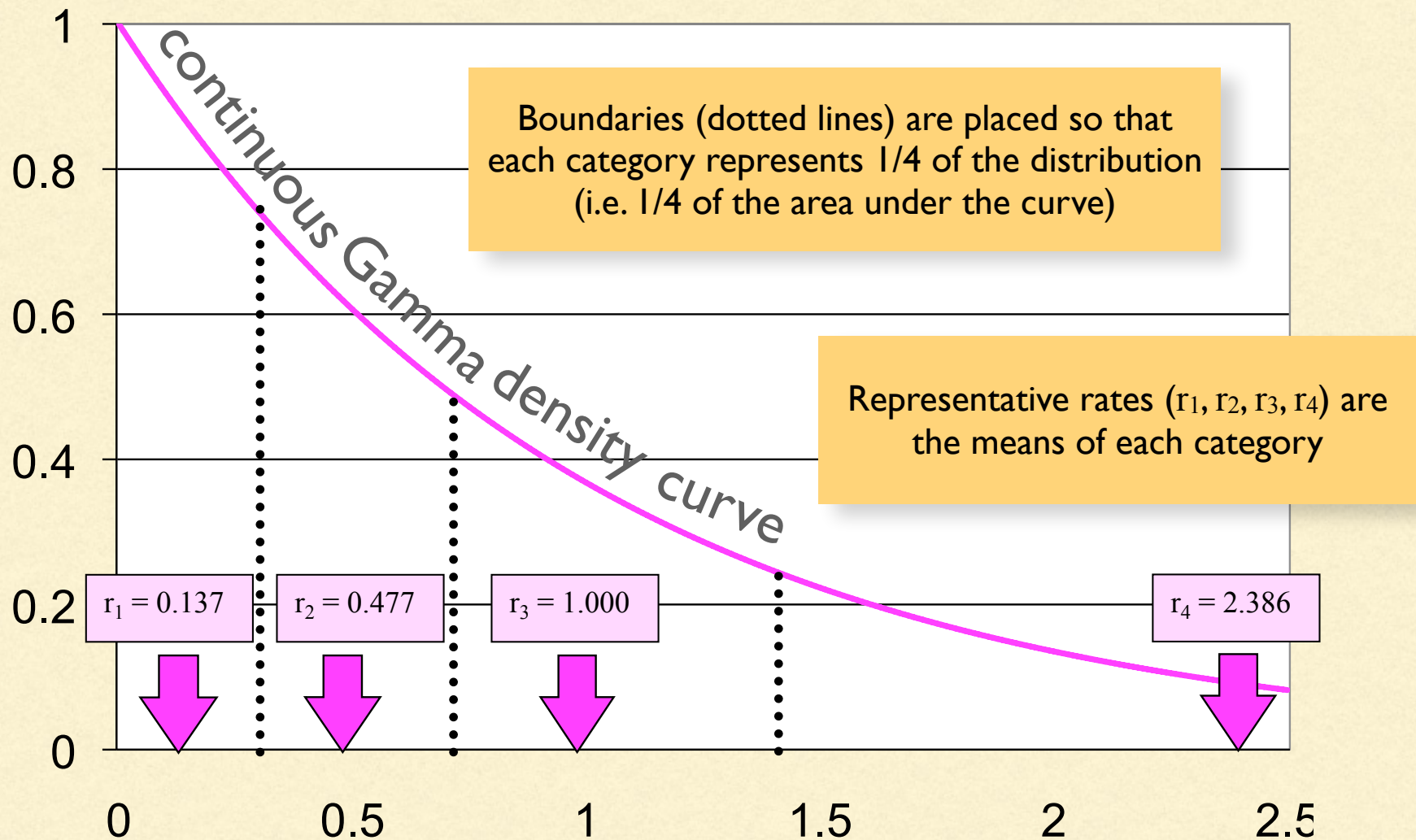
$$L_i = p(\mathbf{y}_i|r_1) \left(\frac{1}{4}\right) + p(\mathbf{y}_i|r_2) \left(\frac{1}{4}\right) + p(\mathbf{y}_i|r_3) \left(\frac{1}{4}\right) + p(\mathbf{y}_i|r_4) \left(\frac{1}{4}\right)$$

Relative rates are determined by a discrete gamma distribution

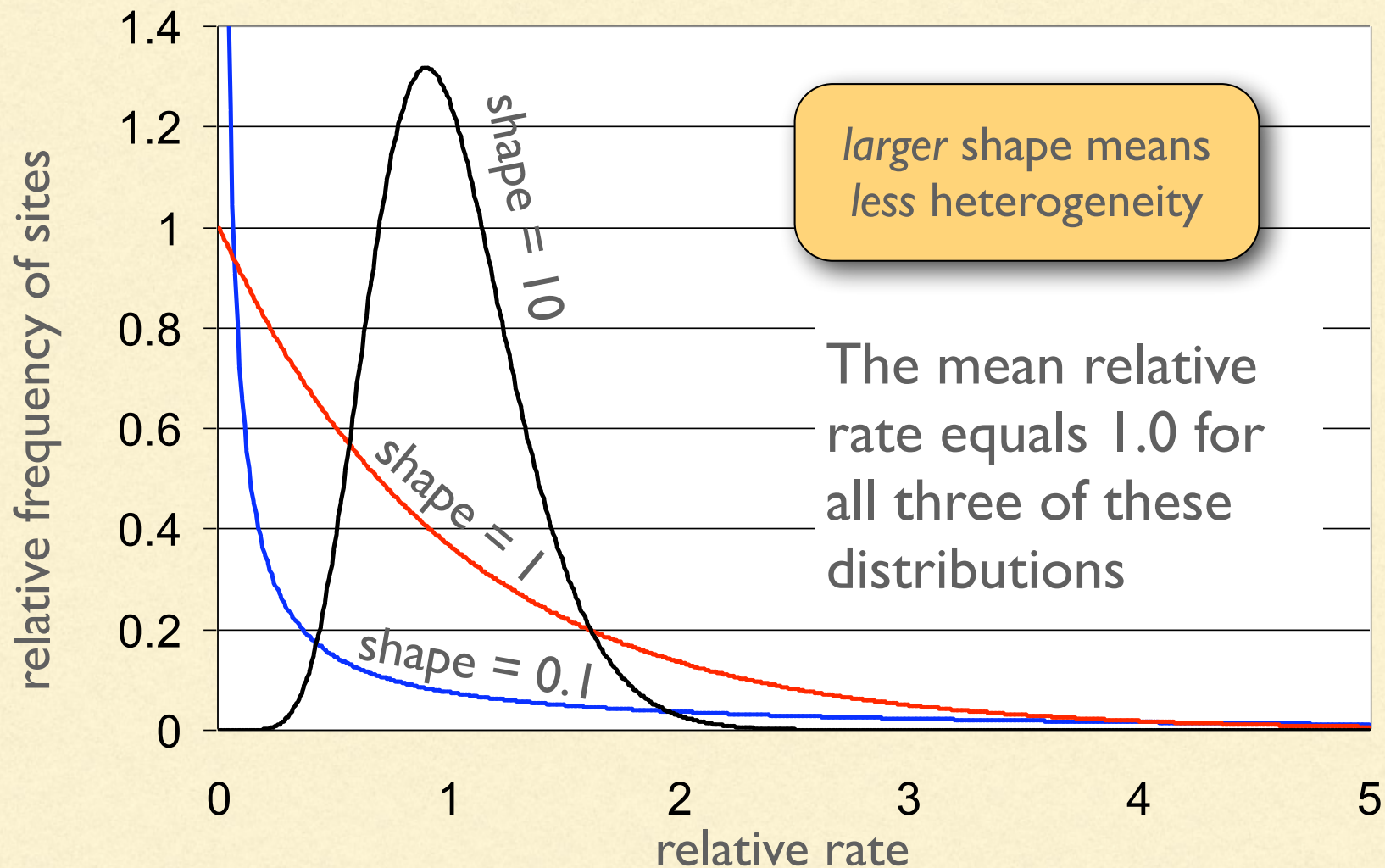
Number of rate categories can vary (4 used here)



# Relative rates in 4-category case



# Gamma distributions







---

~ Coffee Break ~

---