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

UCONN
UNIVERSITY OF CONNECTICUT

Probability density

The **expected number** of substitutions/site equals the total substitution **rate** multiplied by **time**

v = (subst. rate)(time)

Average of 10 values

$$\text{average} = \frac{1+2+4+3+4+3+2+4+3+4}{10}$$

$$= (0)\left(\frac{0}{10}\right) + (1)\left(\frac{1}{10}\right) + (2)\left(\frac{2}{10}\right) + (3)\left(\frac{3}{10}\right) (4)\left(\frac{4}{10}\right) + (5)\left(\frac{0}{10}\right) + \cdots$$

Expected value of a random variable X

$$E[X] = (0)p(0) + (1)p(1) + (2)p(2) + (3)p(3) + (4)p(4) + \cdots$$

Expected value is same as the simple average if the probabilities used are sample relative frequencies

Expected number of substitutions **if only an instant of time is considered** (only 0 or 1 substitutions possible):

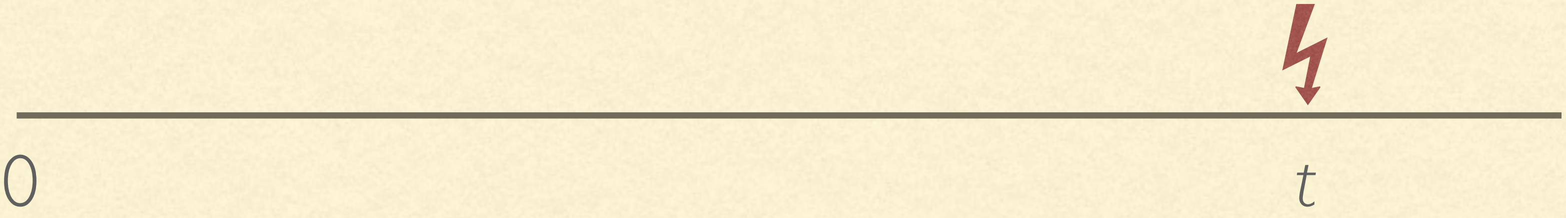
$$E[X] = \cancelto{0}{(0)p(0)} + (1)p(1) = p(1)$$

Note that **probability of a substitution** equals the **expected number of substitutions** in this case

$$p(\text{subst.}) = \underset{\substack{\nearrow \\ \text{substitution rate}}}{\lambda}\underset{\substack{\nwarrow \\ \text{instant of time}}}{dt}$$

Revisiting your simulations from yesterday

Yesterday you drew a uniform random variable *u* (using your 10-sided die) and transformed it to obtain a time distributed as an Exponential distribution with rate lambda



A statistician would write

$$t \sim \text{Exponential}(\lambda)$$

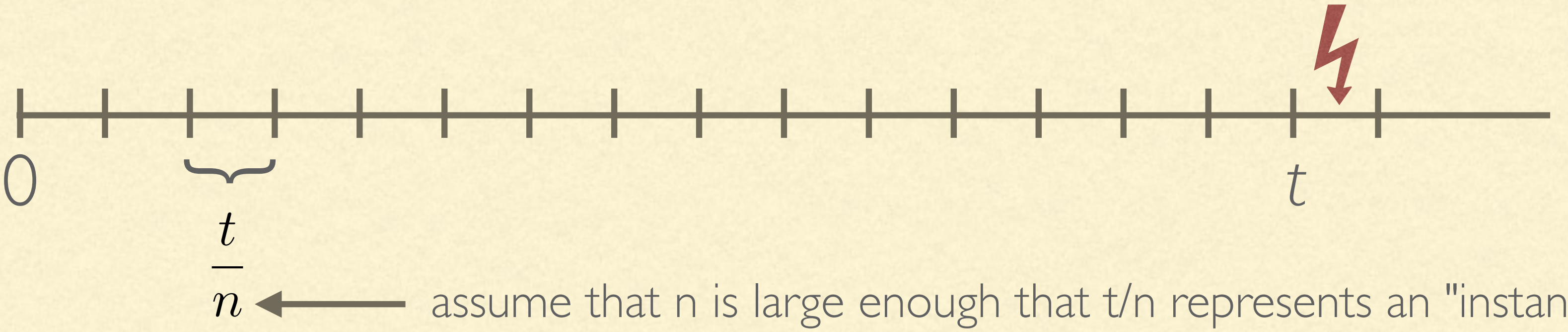
and also would say that the probability **density** of *t* is

$$p(t|\lambda) = \lambda e^{-\lambda t}$$

What is a probability density and where did that e come from?

$$p(t|\lambda) = \lambda e^{-\lambda t}$$

Start by imagining that the interval 0 to *t* is divided into *n* equal segments and our substitution occurs in the very next segment:



assume that n is large enough that t/n represents an "instant" of time

The probability that a substitution falls in any segment is

$$\lambda \frac{t}{n}$$

The probability that there is no substitution in a given segment is thus

$$1 - \lambda \frac{t}{n}$$

The probability that *no substitution* occurred in *any* of the *n* segments is thus

$$\left(1 - \lambda\frac{t}{n}\right)^n \approx e^{-\lambda t}$$

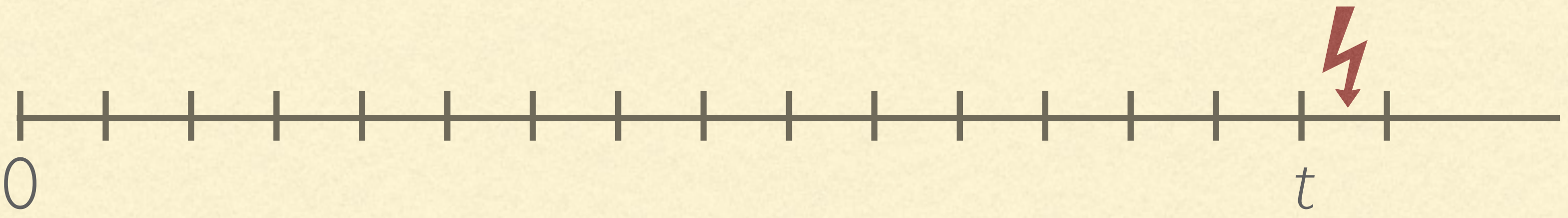
The approximation shown above works well if n is large (and we can set *n* equal to infinity if we like).

More generally,

$$\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = e^a$$

We're using $a = -\lambda t$

You can approximate Euler's constant e by just calculating the above quantity using *a* = 1 and a large value of *n*: for example, for n = 1000000, the formula gives 2.71828047



We now have the probability that *no substitution* occurred in *any* of the *n* segments spanning the interval 0 to t, so all we need now is the probability that a substitution **did occur** in the very next instant of time, which is just $\lambda\ dt$, where *dt* is an infinitesimal time period.

The probability of seeing any substitution at exactly time t is

$$e^{-\lambda t}\,(\lambda\ dt)$$

This can be factored into a probability **density** and *dt*:

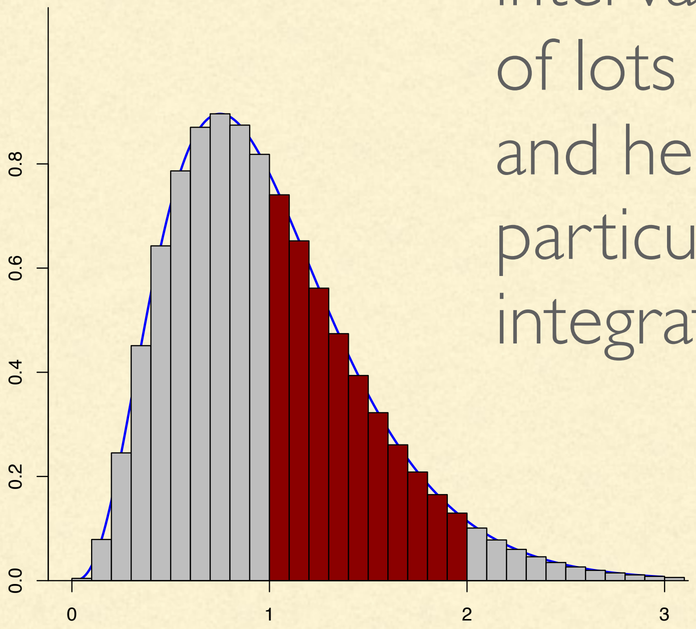
$$\left(\lambda e^{-\lambda t}\right)\ dt$$

A probability density allows you to calculate a probability if multiplied by a time interval, much like the density of gold allows you to compute the mass of gold in a specified volume.

A big difference is that gold has the same density throughout, whereas the probability density is only valid for one value of t; the density continuously changes with t.

To compute the probability that t is in the interval 1 to 2 we need to add up the volumes of lots of rectangles each of which has width *dt* and height equal to the density function at a particular value of *t* (i.e. we need to do integration):

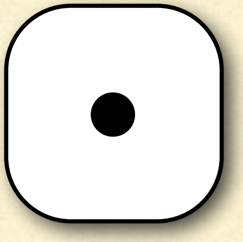
$$\Pr(1 \leq t \leq 2) = \int_0^1 \lambda e^{-\lambda t}\, dt$$



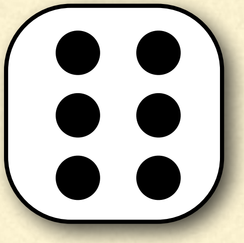
Probability

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?

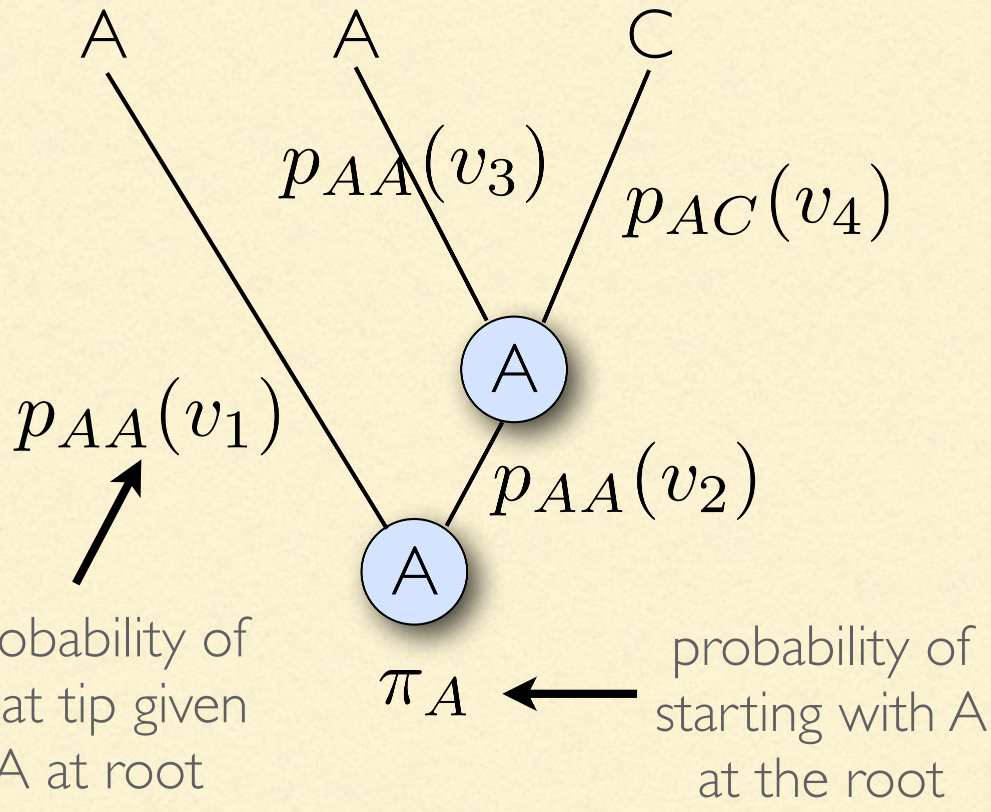


AND



(1/6) × (1/6) = 1/36

AND rule in phylogenetics

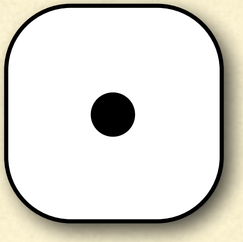

A
A
C
$p_{AA}(v_3)$
$p_{AC}(v_4)$
A
$p_{AA}(v_1)$
$p_{AA}(v_2)$
A
probability of A at tip given A at root
π_A
probability of starting with A at the root

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.

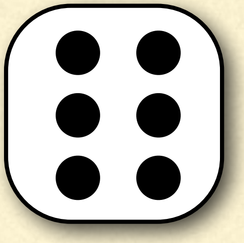
$$\Pr(A, A, C, A, A) = \pi_A \; p_{AA}(v_1) \; p_{AA}(v_2) \; p_{AA}(v_3) \;\; p_{AC}(v_4)$$

Probabilities: the OR rule

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



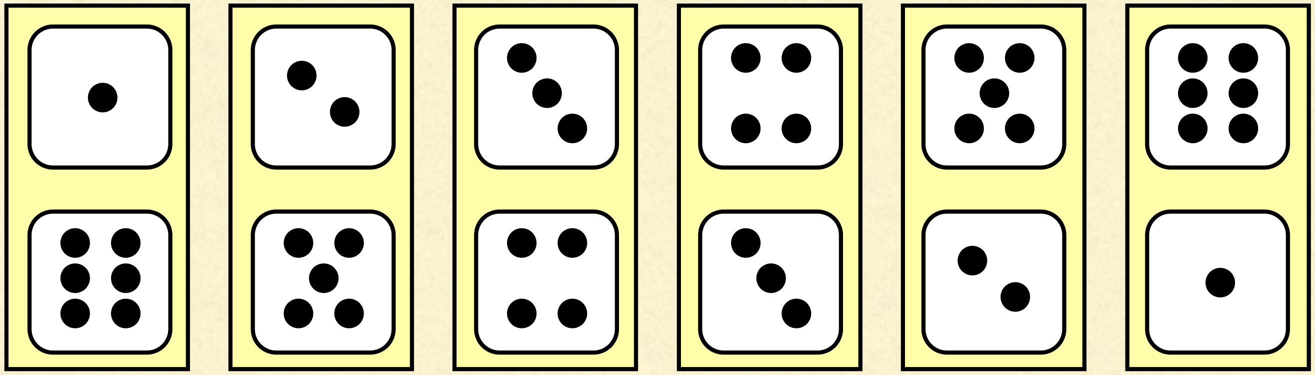
OR



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

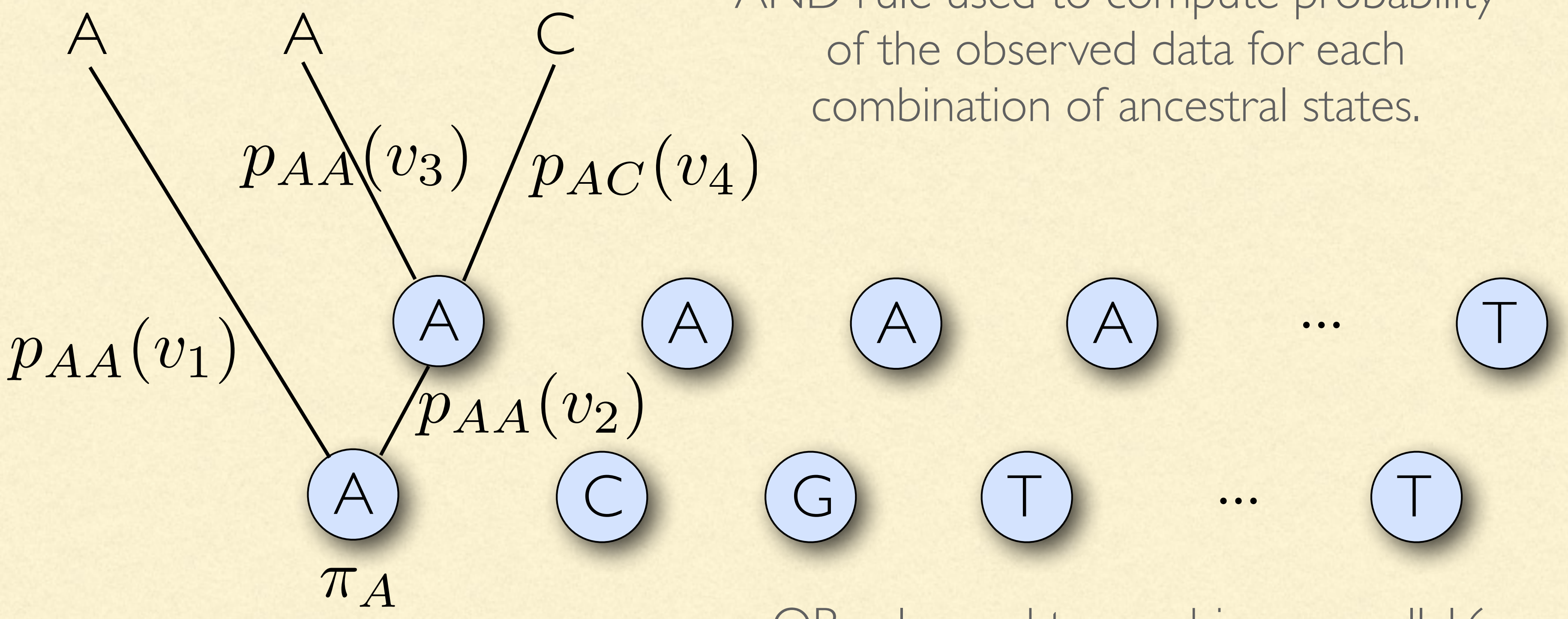
Combining AND and OR

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



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

OR rule used to combine over all 16 combinations of ancestral states.

Pr(**A,A,C**) = Pr(**A,A,C**,A,A)+ Pr(**A,A,C**,A,C)+...+ Pr(**A,A,C**,T,T)

Independence

$$\mathbf{Pr}(A, B) = \mathbf{Pr}(A)\,\mathbf{Pr}(B)$$

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

Pr(H,H) = Pr(H) Pr(H)

Non-independence

$$\mathbf{Pr}(A, B) = \mathbf{Pr}(A)\,\mathbf{Pr}(B|A)$$

joint probability of A and B

conditional probability of B given A

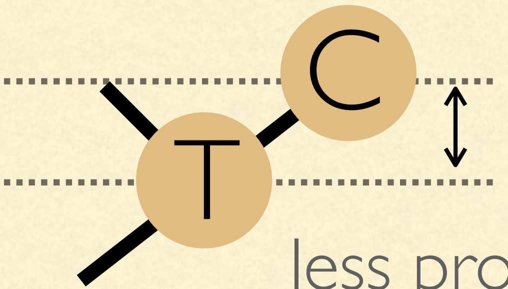
$$\text{Pr(walk to work|sunny)} = 0.99$$

$$\text{Pr(walk to work|raining)} = 0.50$$

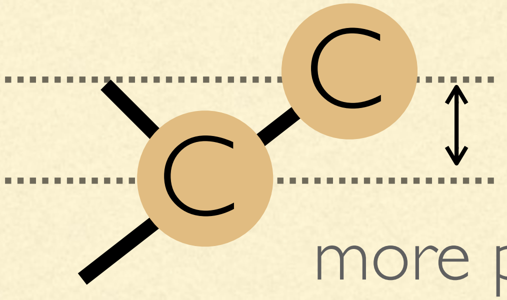
Non-independence in phylogenies

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

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



less probable



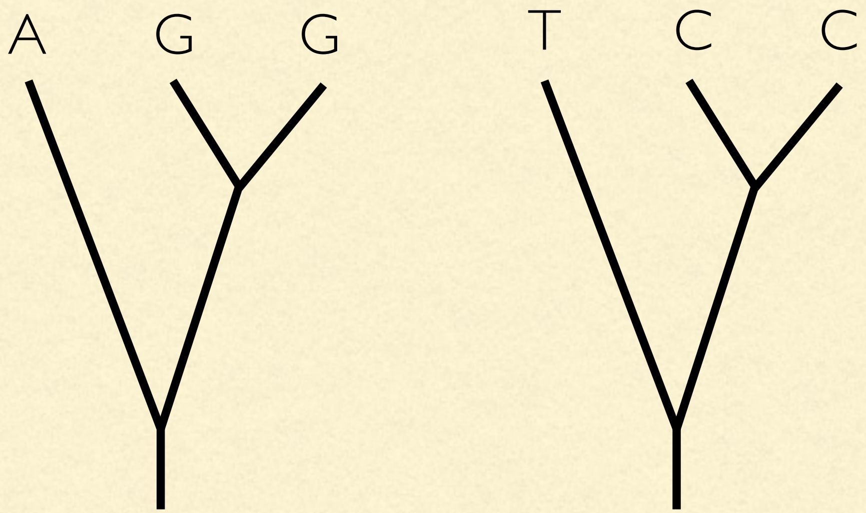
more probable

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

common notation for transition probabilities

Conditional Independence

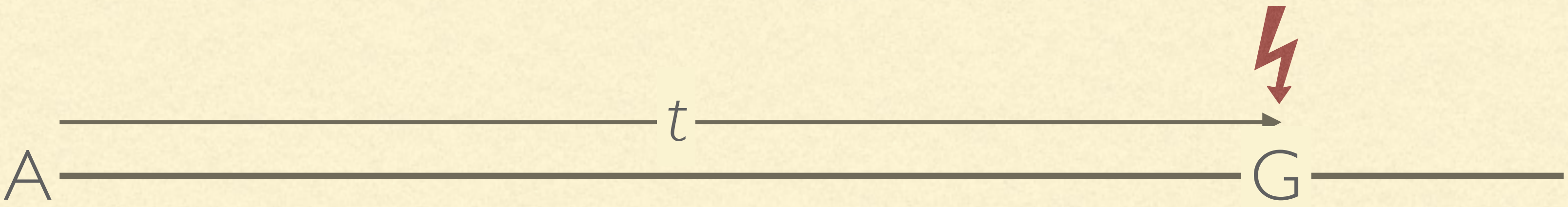
Pr(A,B|C) = Pr(A|C) Pr(B|C)



Pr(AGG,TCC|tree) = Pr(AGG|tree) Pr(TCC|tree)

Back to your simulations...

You first chose a waiting time *t* until the next substitution, and then you used your dice again to choose which nucleotide was actually substituted (G in the case shown below)



To make this choice, we need to calculate the conditional probability of an A→G substitution **given** that a substitution occurred. We can manipulate a formula shown previously to get the conditional probability we want:

$$\Pr(A,B) = \Pr(A)\Pr(B|A) \longrightarrow \Pr(B|A) = \frac{\Pr(A,B)}{\Pr(A)}$$

Replacing A with "substitution" and B with "A→G"

$$\Pr(A \to G|\text{substitution}) = \frac{\Pr(A \to G, \text{substitution})}{\Pr(\text{substitution})}$$

$$= \frac{r_{AG}dt}{r_{AC}dt + r_{AG}dt + r_{AT}dt}$$

$$= \frac{r_{AG}}{\underset{\longleftarrow \text{lambda} \longrightarrow}{r_{AC} + r_{AG} + r_{AT}}}$$

Note that the joint probability Pr(A→G, substitution) equals Pr(A→G) because a substitution must have occurred if an A changed to a G.

The probability of any substitution at exactly time t is λdt, but the rate of any substitution (λ) is just the sum of the rates of all possible substitutions ($r_{AC} + r_{AG} + r_{AT}$)

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

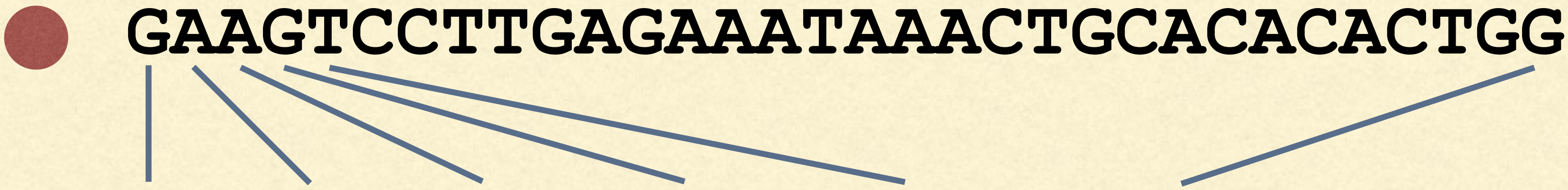
Probabilities of data outcomes given one particular model sum to 1.0

Probability of the **data** given the model

Likelihood of the **model** given the data

Likelihood of a single vertex

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



GAAGTCCTTGAGAAATAAACTGCACACACTGG

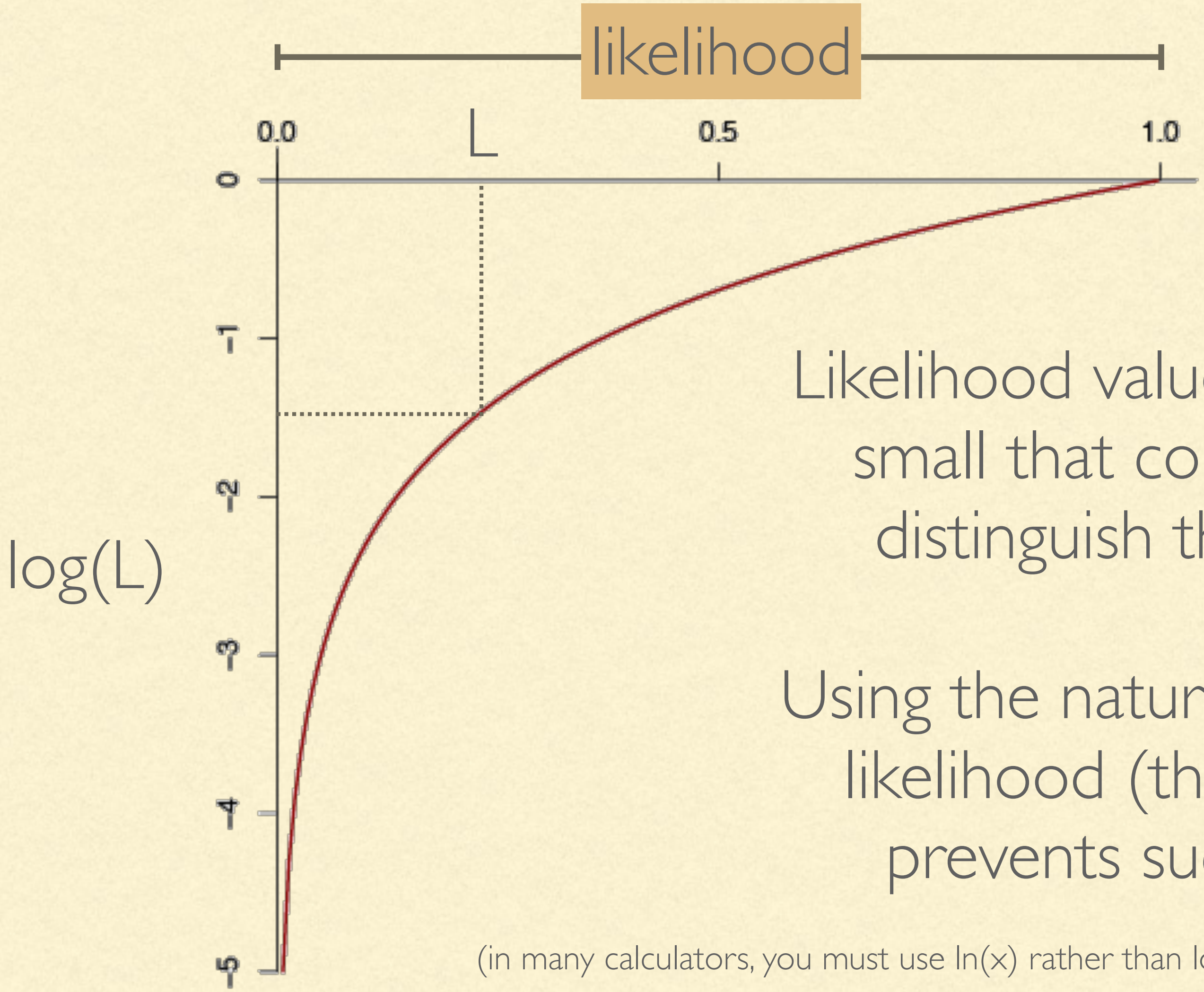
$$L = \Pr(G)\ \Pr(A)\ \Pr(A)\ \Pr(G)\ \Pr(T)\ \cdots\ \Pr(G)$$

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

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

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

Natural logarithm



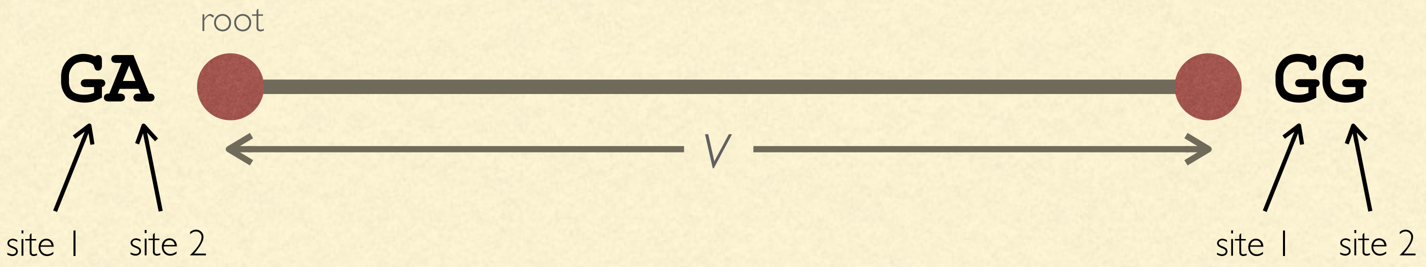
Likelihood values can become so small that computers cannot distinguish them from zero.

Using the natural logarithm of the likelihood (the log-likelihood) prevents such "underflow"

(in many calculators, you must use ln(x) rather than log(x) to take the natural log of the value x)

Likelihood of a single-edge tree

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



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

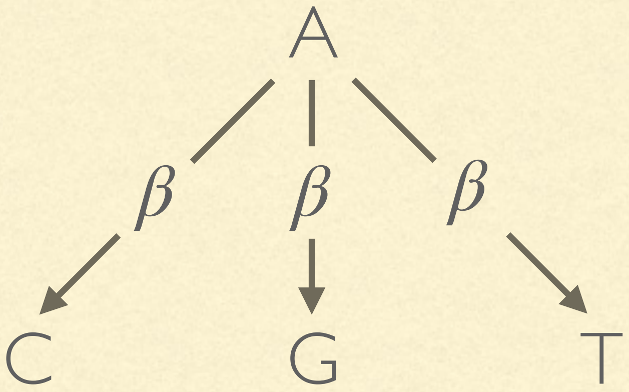
site 1

site 2

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?

expected **number** of substitutions per site	=	substitution **rate** per site	×	**time**
v	=	3β	×	t



3 possible substitutions, each of which happens with rate β

Jukes and Cantor (1969)

JC69 model

to:

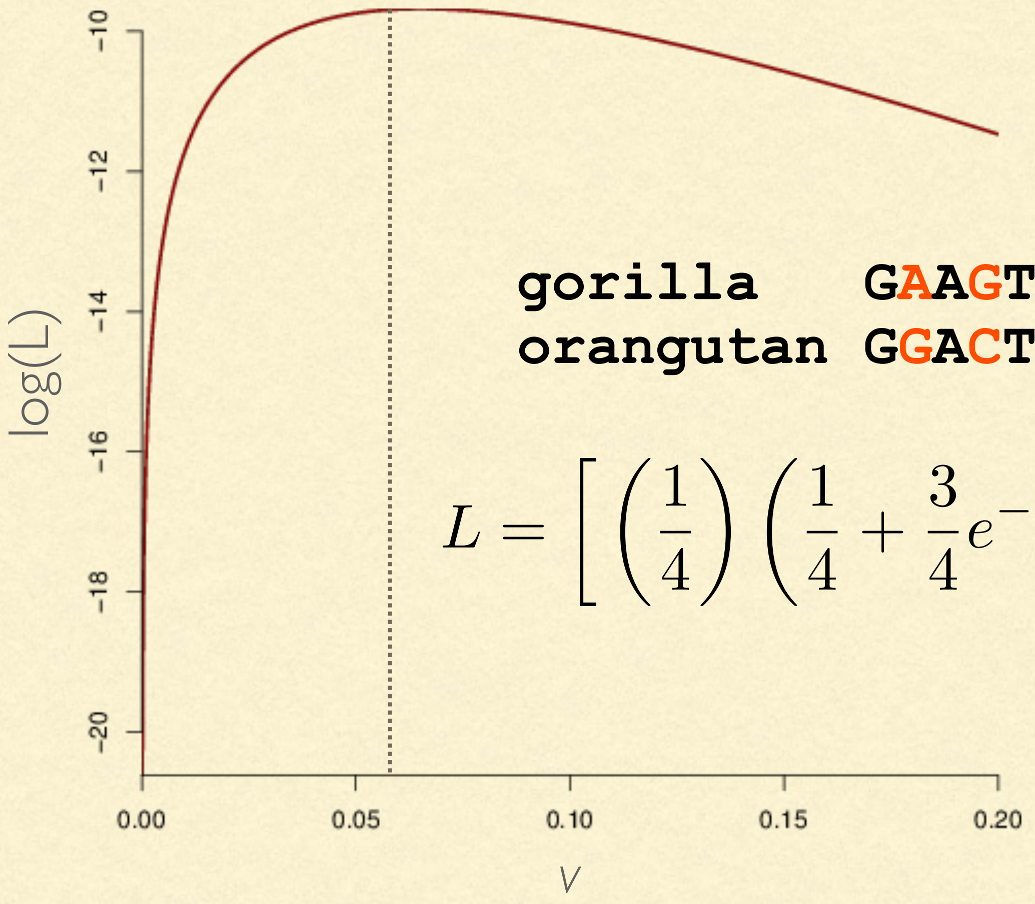
Parameters: β

from:

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

Maximum likelihood estimation

0.065 is the maximum likelihood estimate (MLE) of *v*



```
gorilla    GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan  GGACTCCTTGAGAAATAAACTGCACACACTGG
```

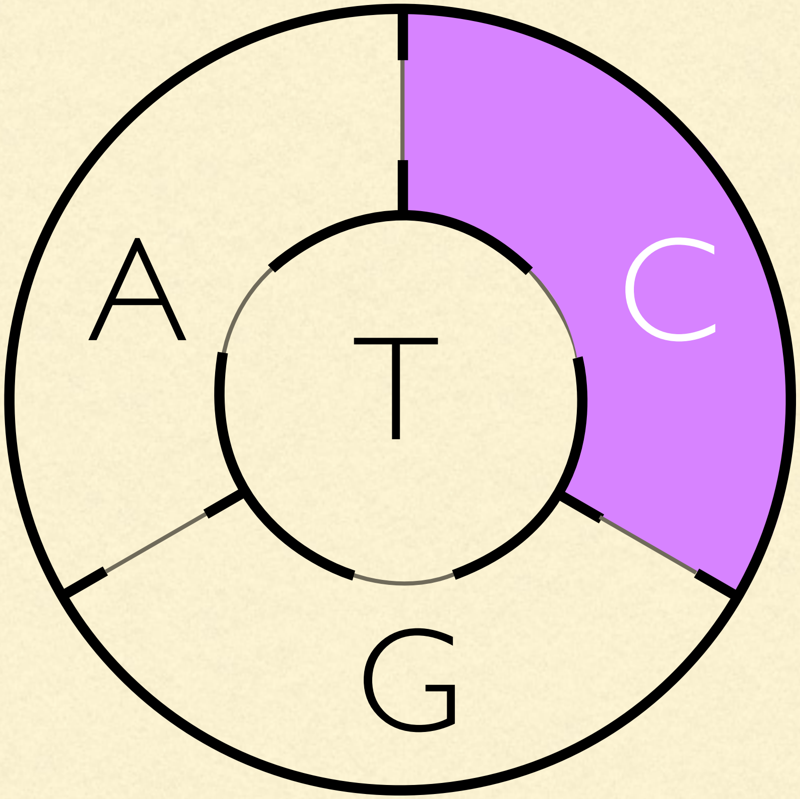
$$L = \left[\left(\frac{1}{4}\right)\left(\frac{1}{4}+\frac{3}{4}e^{-4\nu/3}\right)\right]^{30}\left[\left(\frac{1}{4}\right)\left(\frac{1}{4}-\frac{1}{4}e^{-4\nu/3}\right)\right]^{2}$$

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...

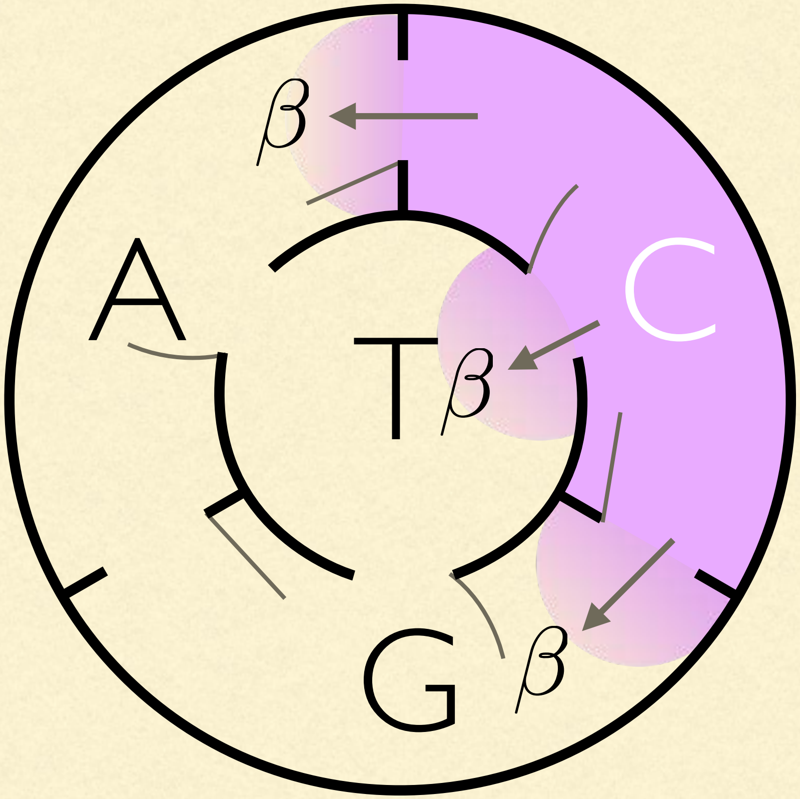
enter room A at rate β

enter room T at rate β

enter room G at rate β

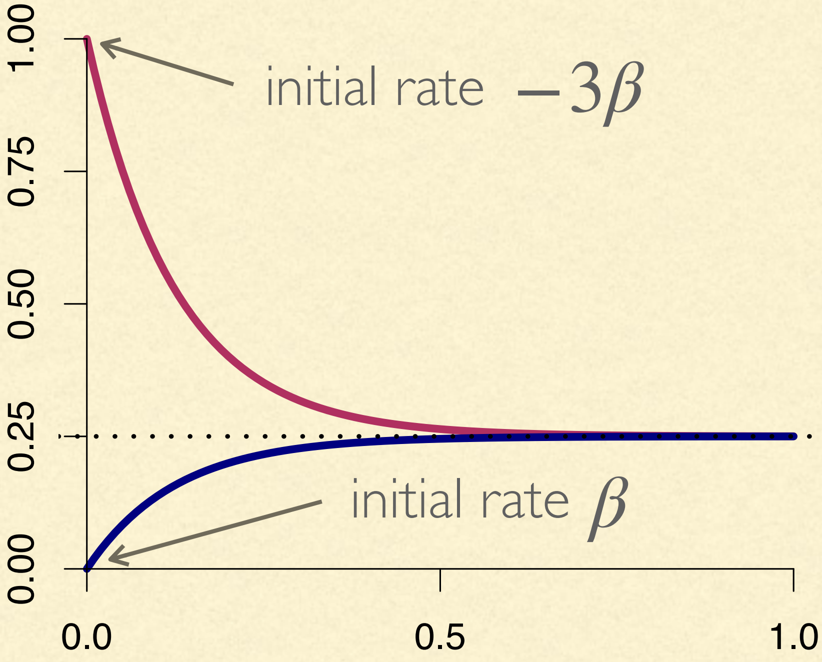
enter room C at rate -3β

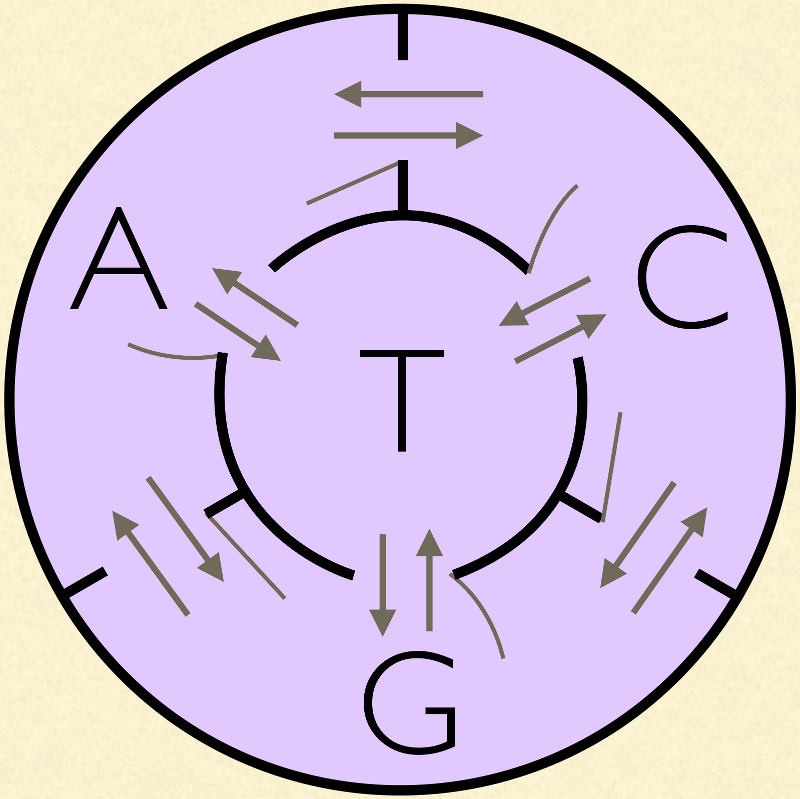
(you could also say they *leave* C at rate 3β)



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

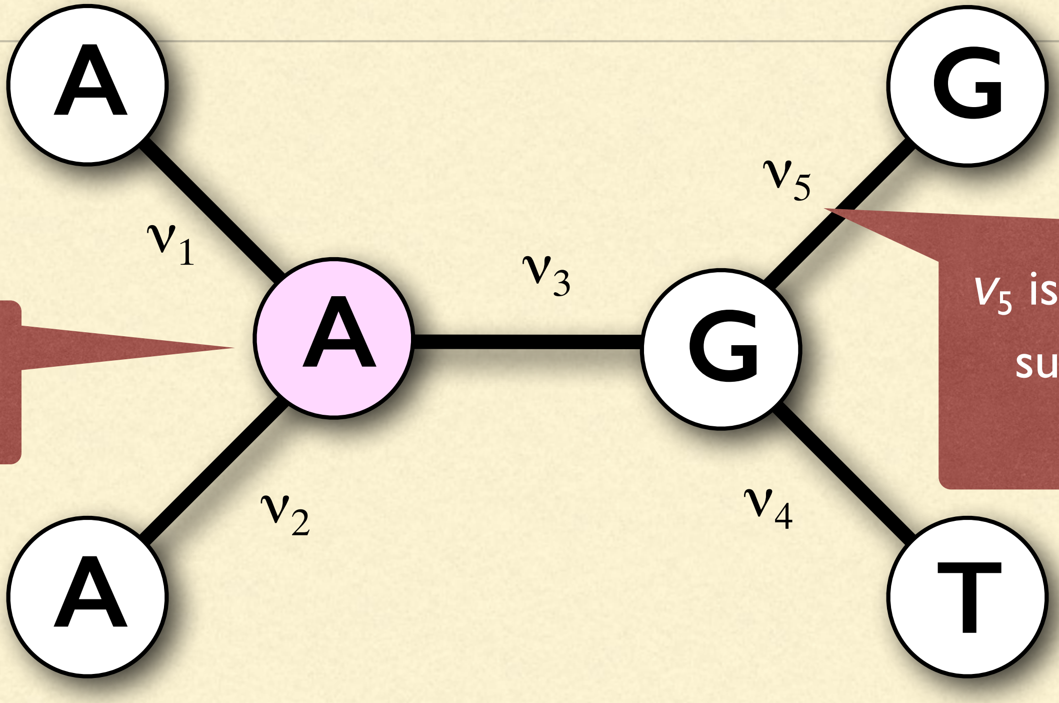
https://plewis.github.io/applets/jc-transition-probabilities/

Sequence data for four taxa

one site

```
Sphagnum  GGCAGCATTTCGAATGACTCCTCAACCTGGAGTACCACCCG...
Asplenium GGCAGCTTTCCGGATGACCCCACAACCCGGAGTACCAGCTG...
Picea     GGCAGCATTCCGAGTAACTCCTCAACCAGGGGTGCCGCCCG...
Avena     GGCAGCATTCCGAGTAACTCCTCAACCTGGGGTTCCGCCGG...
```

Likelihood for tree (one site)



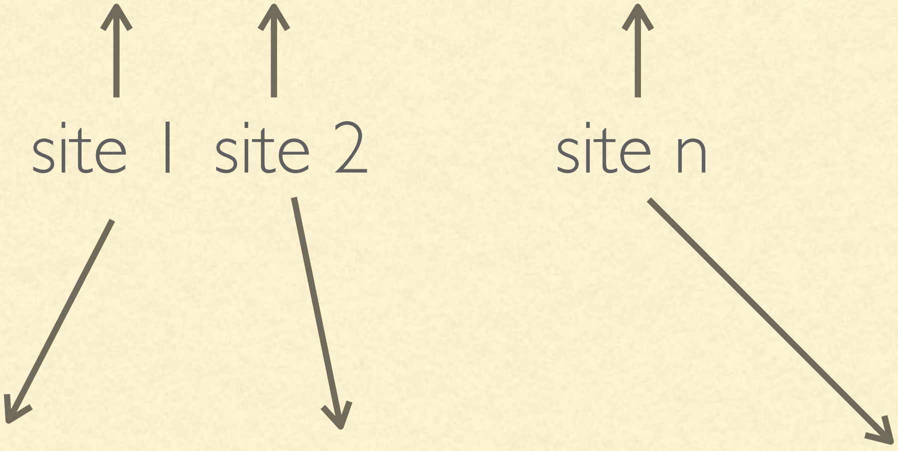
π_A

ν_5 is the expected number of substitutions for just this one branch

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

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

Parameters: β

to:

from:

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

Kimura (1980)

K80 (or K2P) model

Parameters: α, β

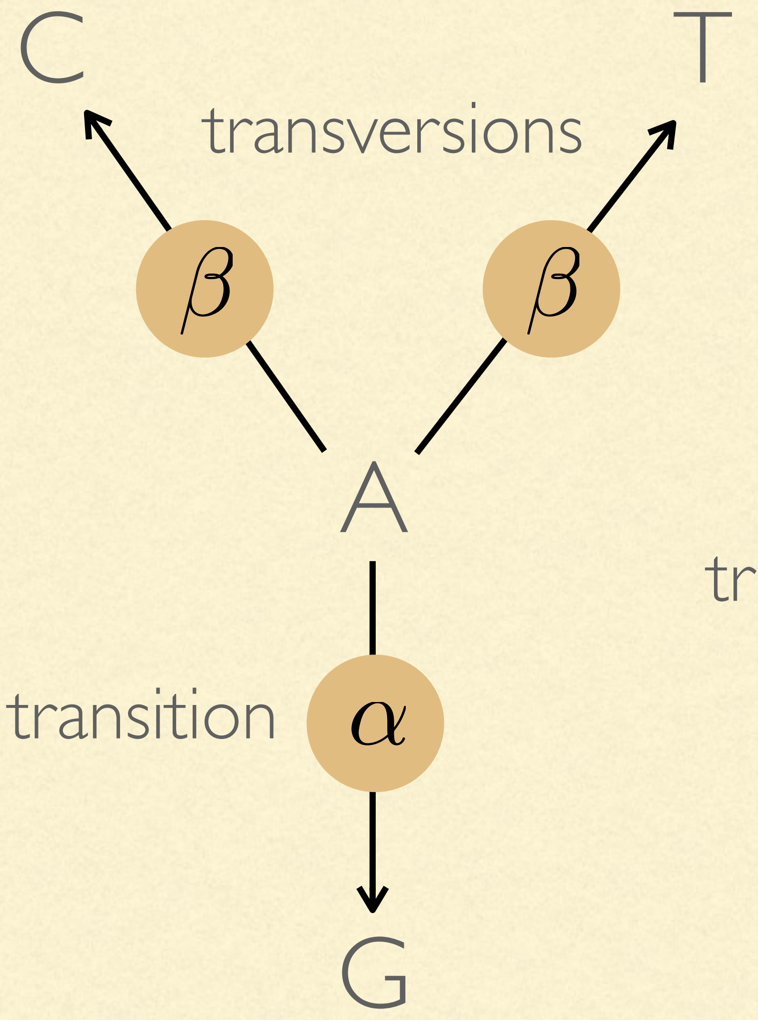
$$
\begin{array}{c|cccc}
 & \text{A} & \text{C} & \text{G} & \text{T} \\
\hline
\text{A} & -\alpha - 2\beta & \beta & \alpha & \beta \\
\text{C} & \beta & -\alpha - 2\beta & \beta & \alpha \\
\text{G} & \alpha & \beta & -\alpha - 2\beta & \beta \\
\text{T} & \beta & \alpha & \beta & -\alpha - 2\beta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

Kimura (1980)

K80 (or K2P) model $\kappa = \alpha/\beta$ Parameters: κ, β

$$\begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & -\beta(\kappa+2) & \beta & \kappa\beta & \beta \\ \text{C} & \beta & -\beta(\kappa+2) & \beta & \kappa\beta \\ \text{G} & \kappa\beta & \beta & -\beta(\kappa+2) & \beta \\ \text{T} & \beta & \kappa\beta & \beta & -\beta(\kappa+2) \end{array}$$

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: μ, π_{A}, π_{C}, π_{G}

$$\begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \left[\begin{array}{cccc} -\mu(1-\pi_A) & \pi_C\mu & \pi_G\mu & \pi_T\mu \\ \pi_A\mu & -\mu(1-\pi_C) & \pi_G\mu & \pi_T\mu \\ \pi_A\mu & \pi_C\mu & -\mu(1-\pi_G) & \pi_T\mu \\ \pi_A\mu & \pi_C\mu & \pi_G\mu & -\mu(1-\pi_T) \end{array}\right] \end{array}$$

JC69 is a special case of F81

$$\begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \text{C} & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \text{G} & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \text{T} & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{array}$$

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

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

Hasegawa, Kishino, and Yano (1985)

HKY85 model

Parameters: $\mu, \boxed{\kappa, \pi_\mathrm{A}, \pi_\mathrm{C}, \pi_\mathrm{G}}$

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

these are global parameters (apply to all edge lengths)

$$\begin{array}{c|cccc}
 & \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\
\hline
\mathrm{A} & -\mu\left(\pi_C + \pi_G\kappa + \pi_T\right) & \pi_C\mu & \pi_G\mu\kappa & \pi_T\mu \\
\mathrm{C} & \pi_A\mu & -\mu\left(\pi_A + \pi_G + \pi_T\kappa\right) & \pi_G\mu & \pi_T\mu\kappa \\
\mathrm{G} & \pi_A\mu\kappa & \pi_C\mu & -\mu\left(\pi_A\kappa + \pi_C + \pi_T\right) & \pi_T\mu \\
\mathrm{T} & \pi_A\mu & \pi_C\mu\kappa & \pi_G\mu & -\mu\left(\pi_A + \pi_C\kappa + \pi_G\right)
\end{array}$$

Tavaré (1986)

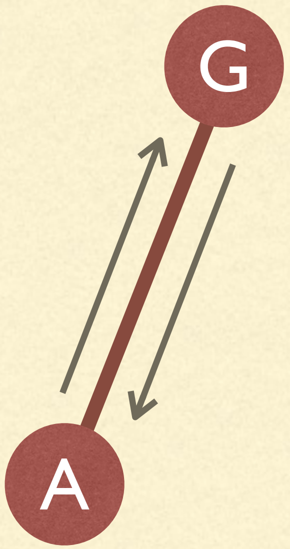
GTR model

Parameters: ?

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

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)   AAAGATTACAGATTAACTTACTATACTCCTGAGTATAAAACTAAAGATACTGACATTTTAGCTGCATTTCGTGTAACTCCA
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.....C
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--
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...........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...........C..T...C.T..C...C.T..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..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..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

```
CACCGGGTCCCCGAGAGCGGGCGCGTGCGCGATCTCACGGACTGACACGTTGACGAGGTTACAGTTGACGTAAAGGAGTGTAGAATGA ATCTATAAAGTAATAATTTTAGTTTGTACATTGCACAAACCTTA
............................................................................C........... .AT..A..GTG..A..AA..T.G.A..TT...A.T..TTTTCCG
................TG..............................................C...........C........... .AT.....TT.TT.T.AAA.T.A.A..TT.A.T.T..TTTTCCG
..........G...............................................C....AC...........C......G.... G.GA.A...AA.T.T........A...TTT.CTTTT.T..T..C
..........................................................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..
..........................................................C.....C....................... ..AG..G.GA....C..C....C...C.TTC.TTT.G...TCCG
..........................................................C.....C...........C........... .AGGGCG.GAA...T..CC...C...C.TT..TTT.GG..TCCG
..........G...............................................C.....C...........C.....C..... .CA.T...G.CG..C.......AAG...TTC.TTT......CCG
..........G.....................................................C...........C........... .CAA....CA.....GC.A...C.G.AG.GCCT.T.GC....CG
..........................................................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

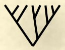
(sites 89-132)

Relative rates have mean 1.0: $r_1 \; p(r_1) + r_2 \; p(r_2) = 1$

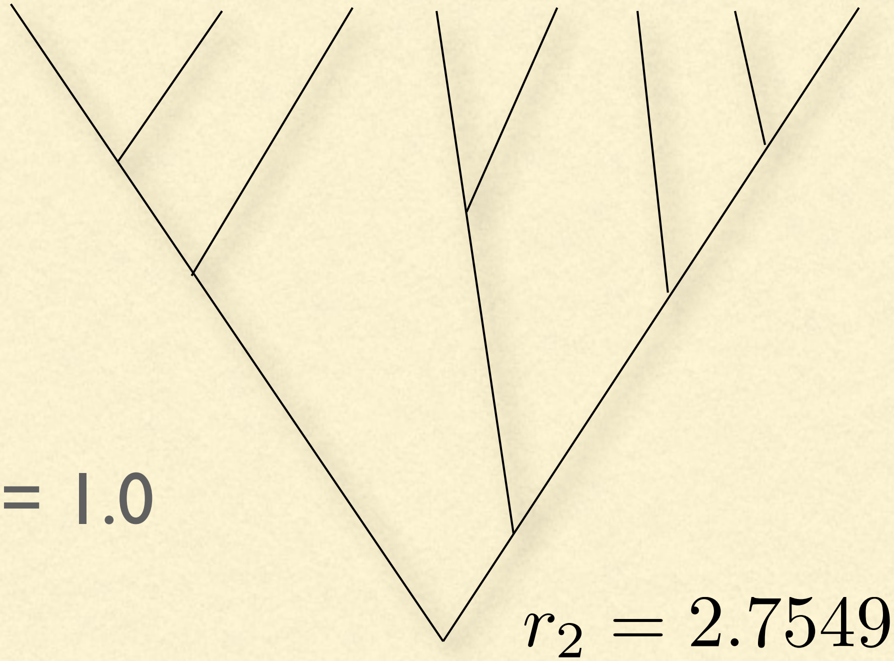
2/3 1/3

Site-specific rates

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



$$r_1 = 0.12255$$



$$r_2 = 2.75490$$

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 *homo*geneity were assumed:

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

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

C C

identity

C T

difference

Site specific rates

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

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

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

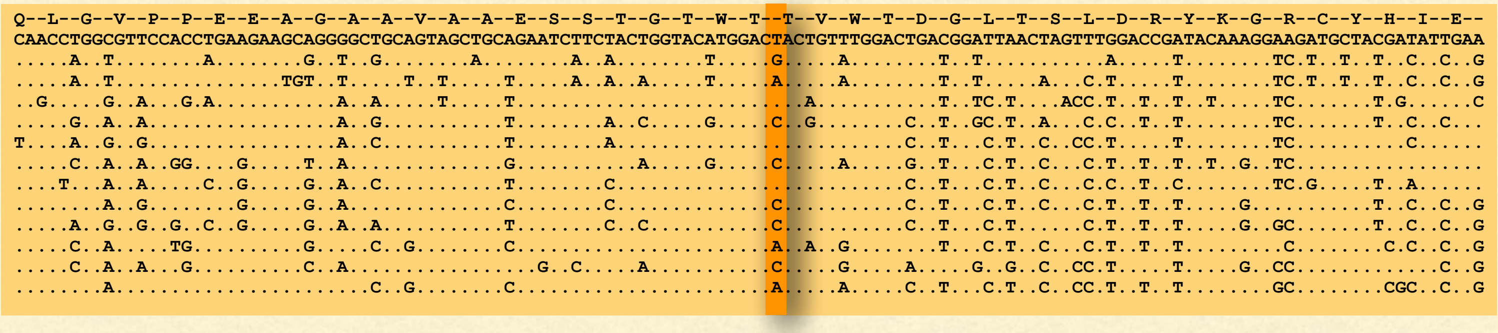
JC69 transition probabilities that would be used for sites in **subset 2**:

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

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

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

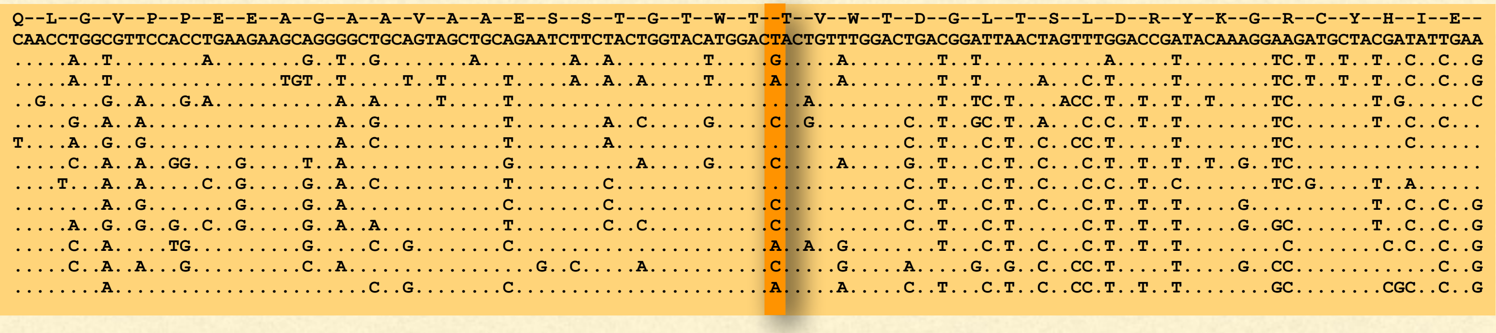
Invariable sites model (Reeves 1992)

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

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

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--
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...........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...........C..T...C.T..C...C.T..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..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..G
........A.......................C..G........C.......................A.....A.....C..T...C.T..C..CC.T..T..T........GC........CGC..C..G
```

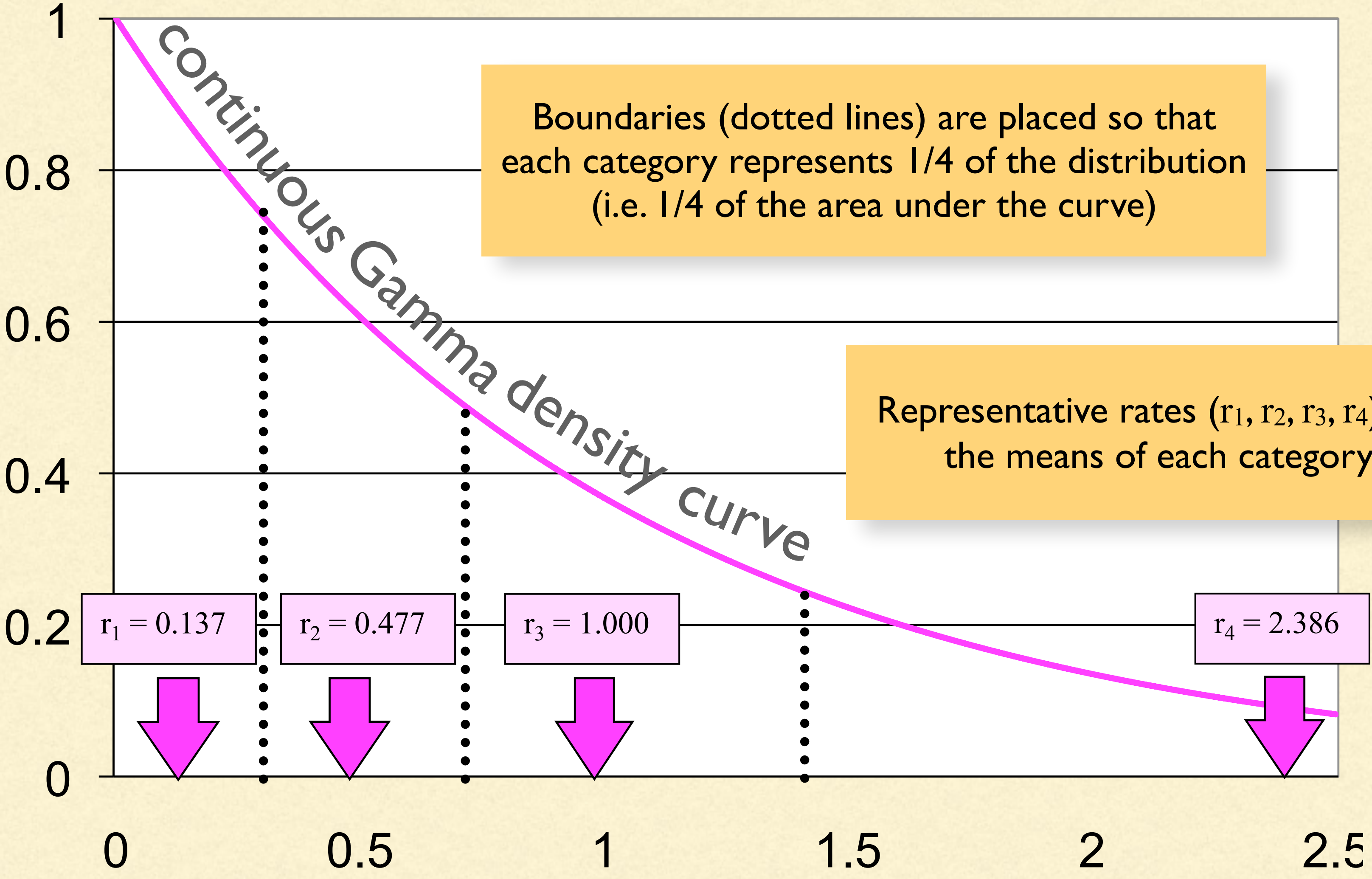
site i

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

Relative rates are determined by a discrete gamma distribution

Number of rate categories can vary (4 used here)

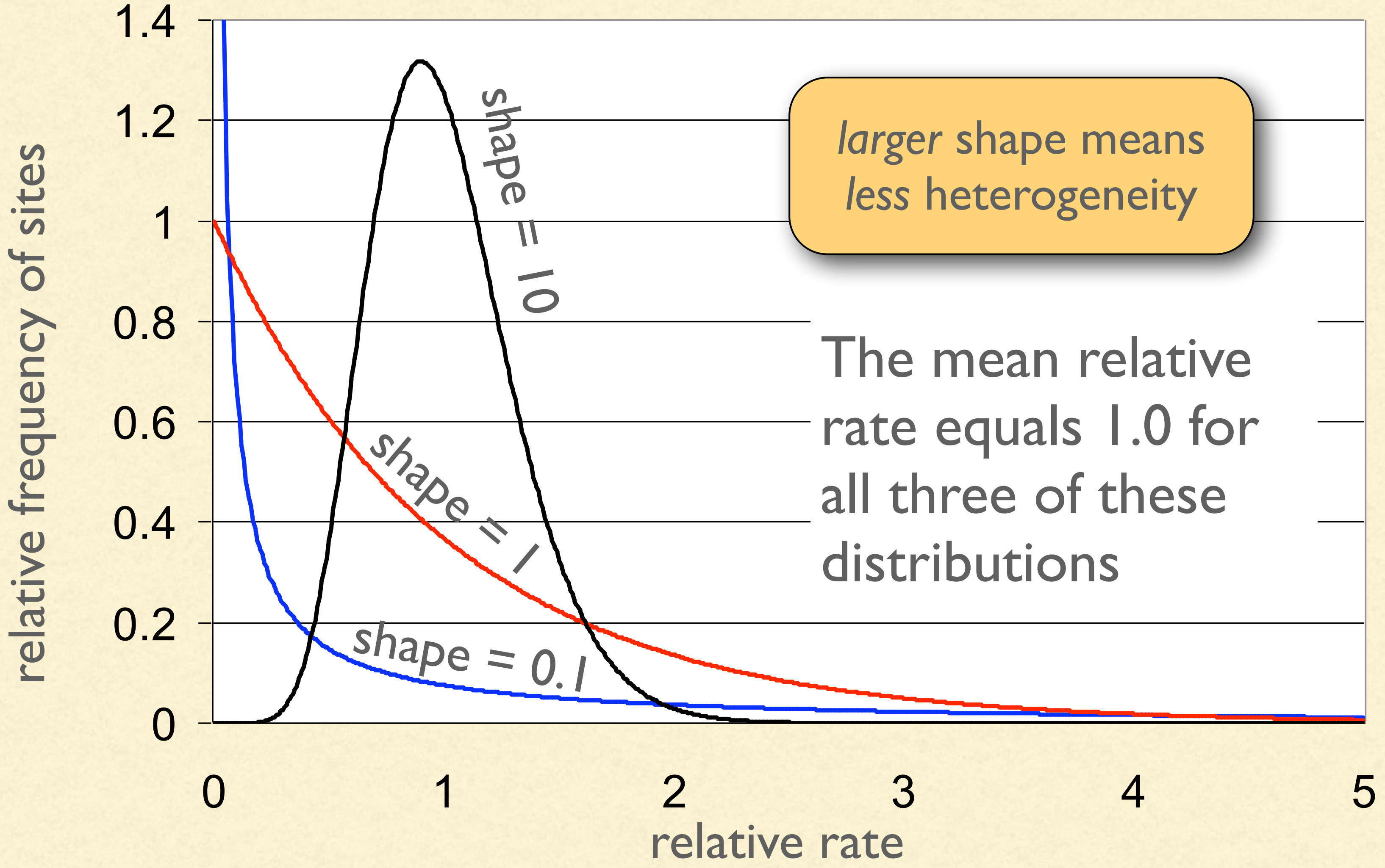
Relative rates in 4-category case



Boundaries (dotted lines) are placed so that each category represents 1/4 of the distribution (i.e. 1/4 of the area under the curve)

Representative rates (r_1, r_2, r_3, r_4) are the means of each category

Gamma distributions



larger shape means *less* heterogeneity

The mean relative rate equals 1.0 for all three of these distributions

~ Coffee Break ~