Maximum Likelihood in Phylogenetics

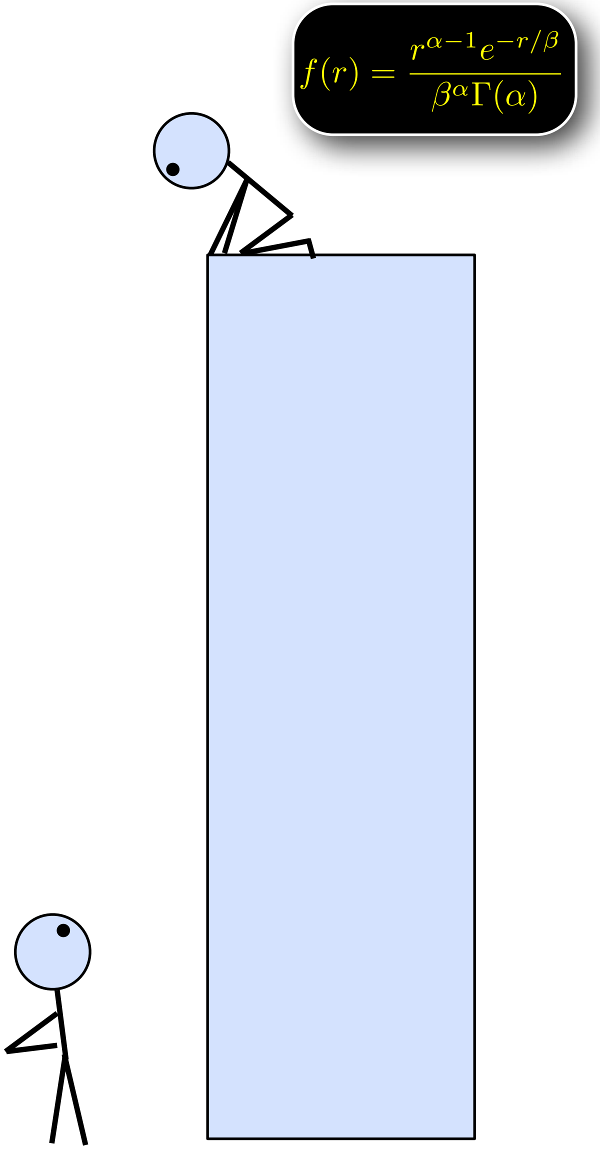
23 January 2013

Workshop on Molecular Evolution
Český Krumlov

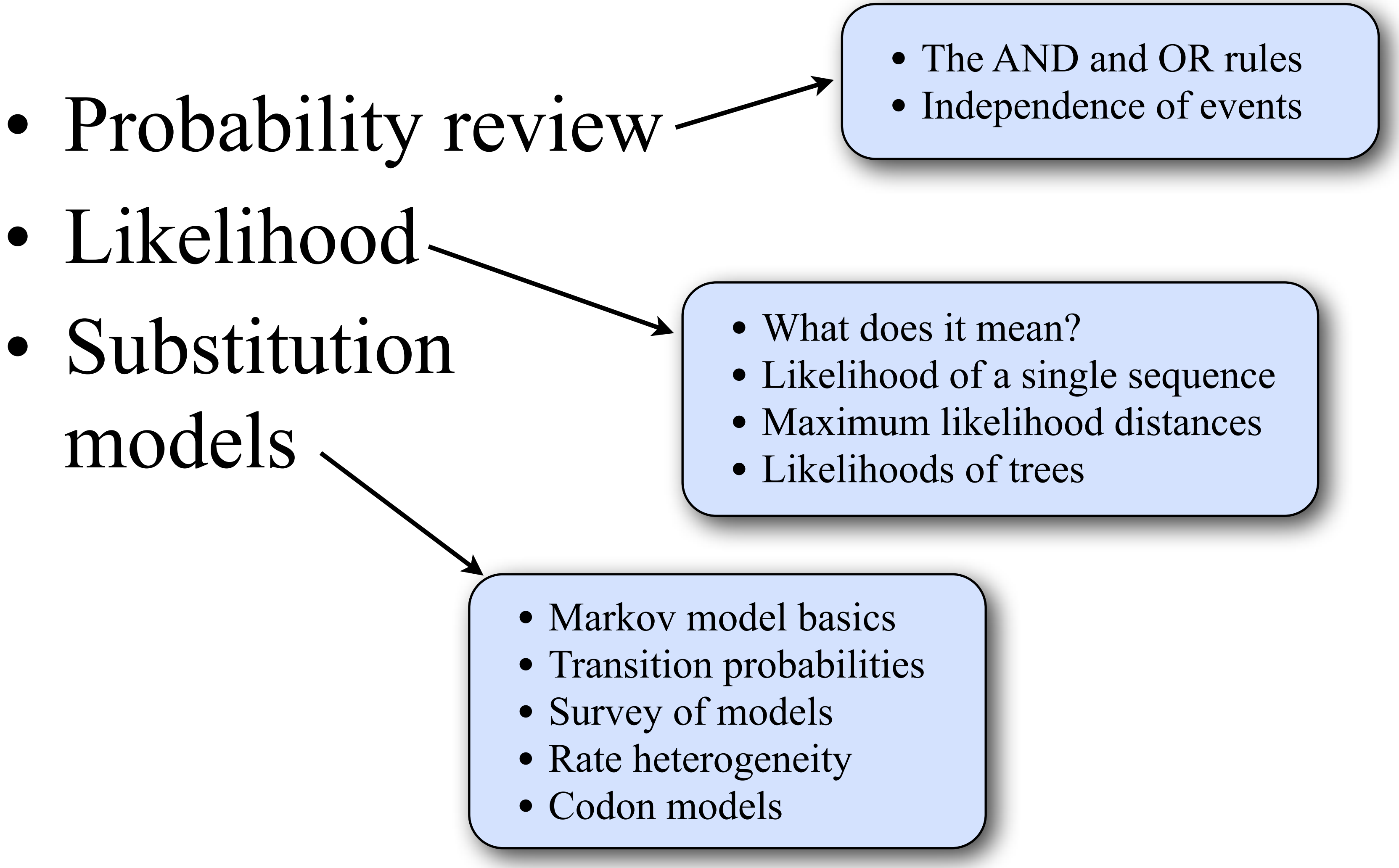
Paul O. Lewis
Department of Ecology & Evolutionary Biology
University of Connecticut, Storrs, CT

Goals

Explain jargon
Increase comfort level
Provide background
In other words...give a hand up


$f(r) = \frac{r^{\alpha-1}e^{-r/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$

The Plan



- Probability review
- Likelihood
- Substitution models

- The AND and OR rules
- Independence of events

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

- Markov model basics
- Transition probabilities
- Survey of models
- Rate heterogeneity
- Codon models

Combining probabilities

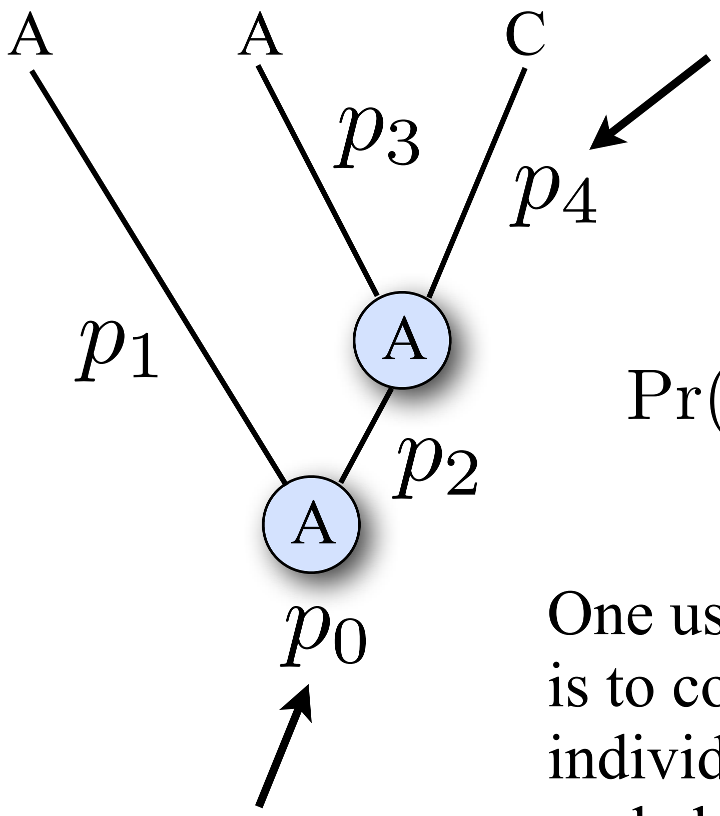
- *Multiply* probabilities if the component events must happen **simultaneously** (i.e. where you would naturally use the word AND when describing the problem)

Using 2 dice, what is the probability of

⚀ AND ⚅ ?

(1/6) × (1/6) = 1/36

AND rule in phylogenetics



probability of C at end of branch given A at beginning

we have observed these states

pretend (for now) we also know these ancestral states

$$\Pr(AAC, AA) = p_0 \; p_1 \; p_2 \; p_3 \; p_4$$

probability of A at 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.

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)

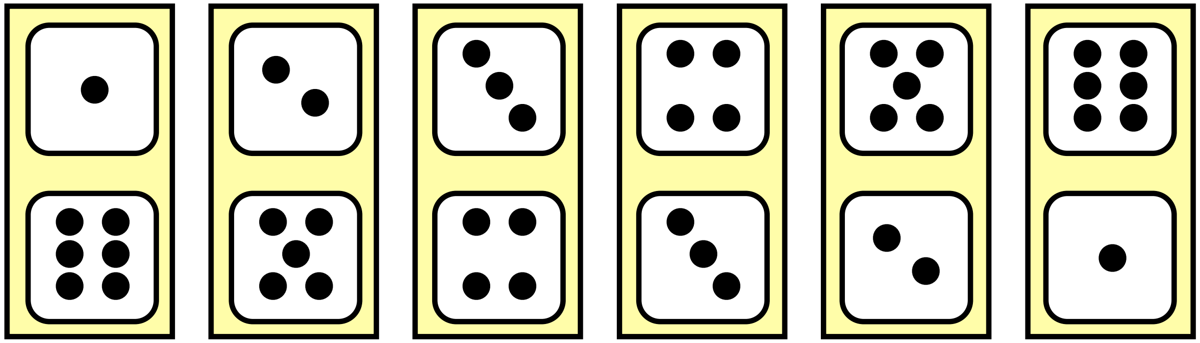
Using one die, what is the probability of

⚀ OR ⚅ ?

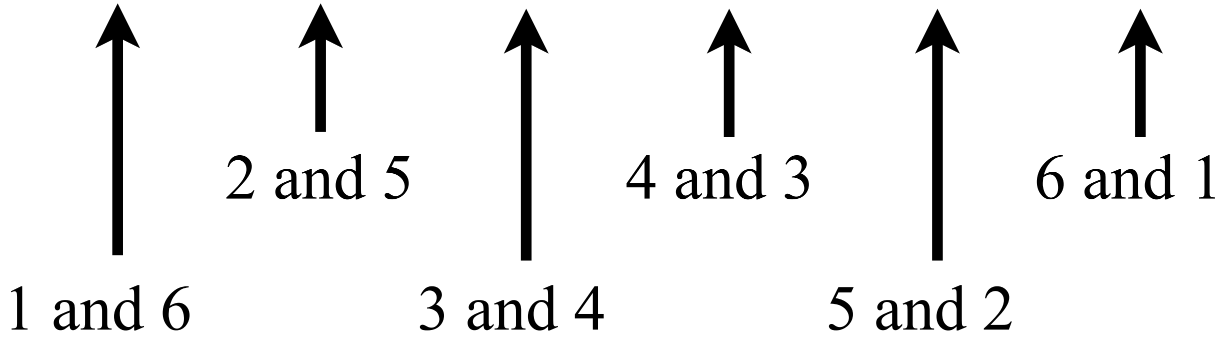
(1/6) + (1/6) = 1/3

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

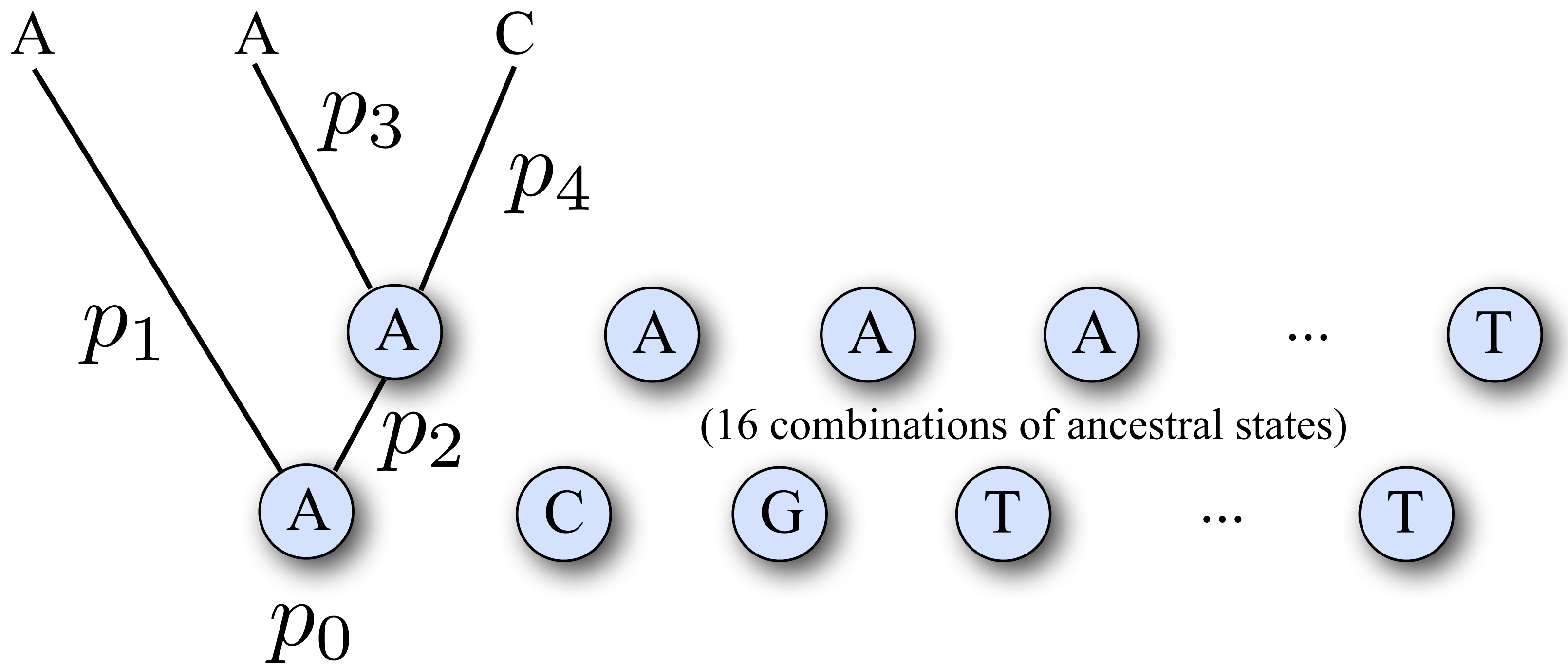
3 and 4

4 and 3

5 and 2

6 and 1

Using both AND and OR in phylogenetics



(16 combinations of ancestral states)

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...

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

joint probability

conditional probability

If we can say this...

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

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

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

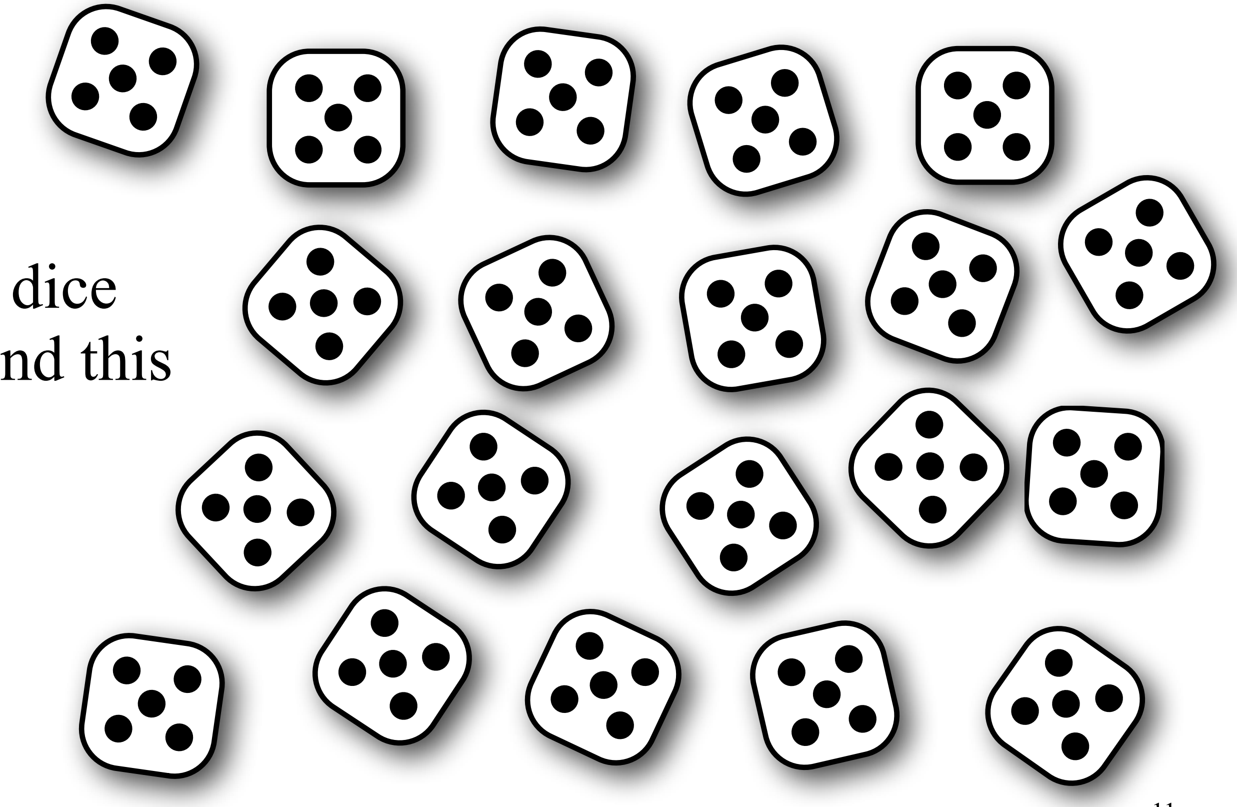
Likelihood

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.

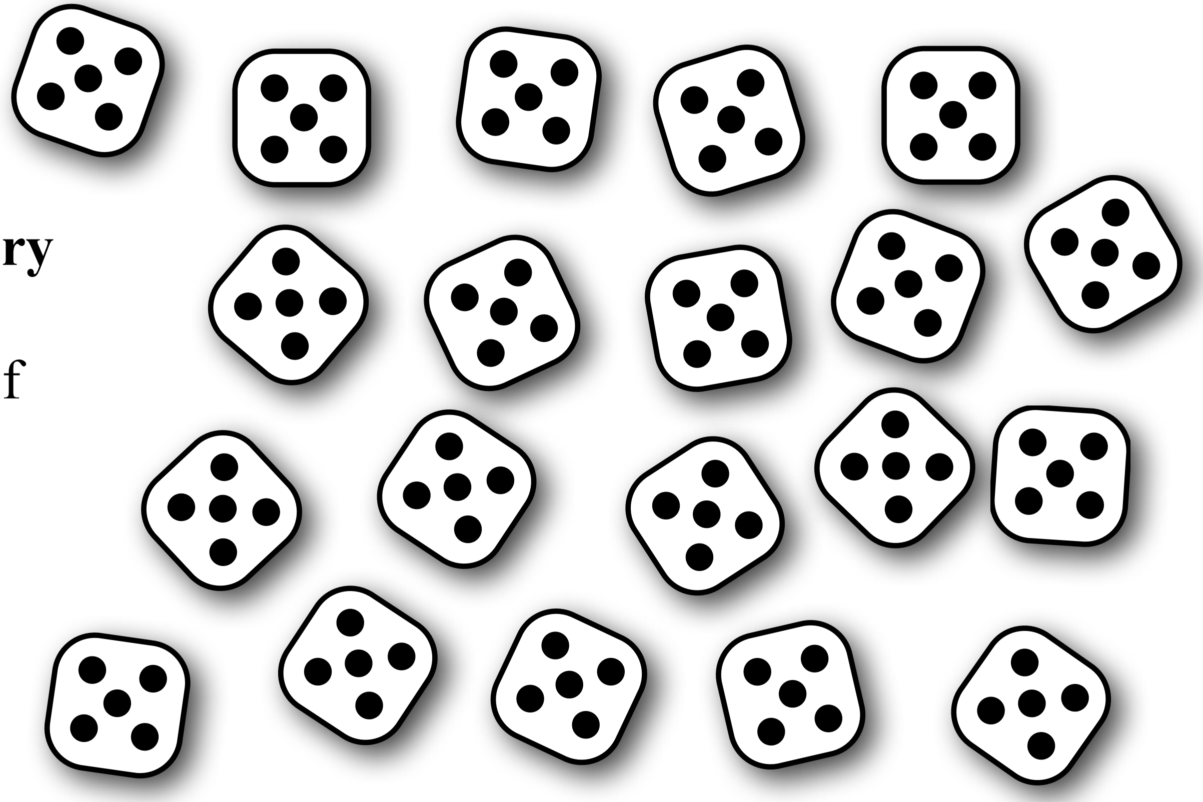
Suppose I threw 20 dice down on the table and this was the result...



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!

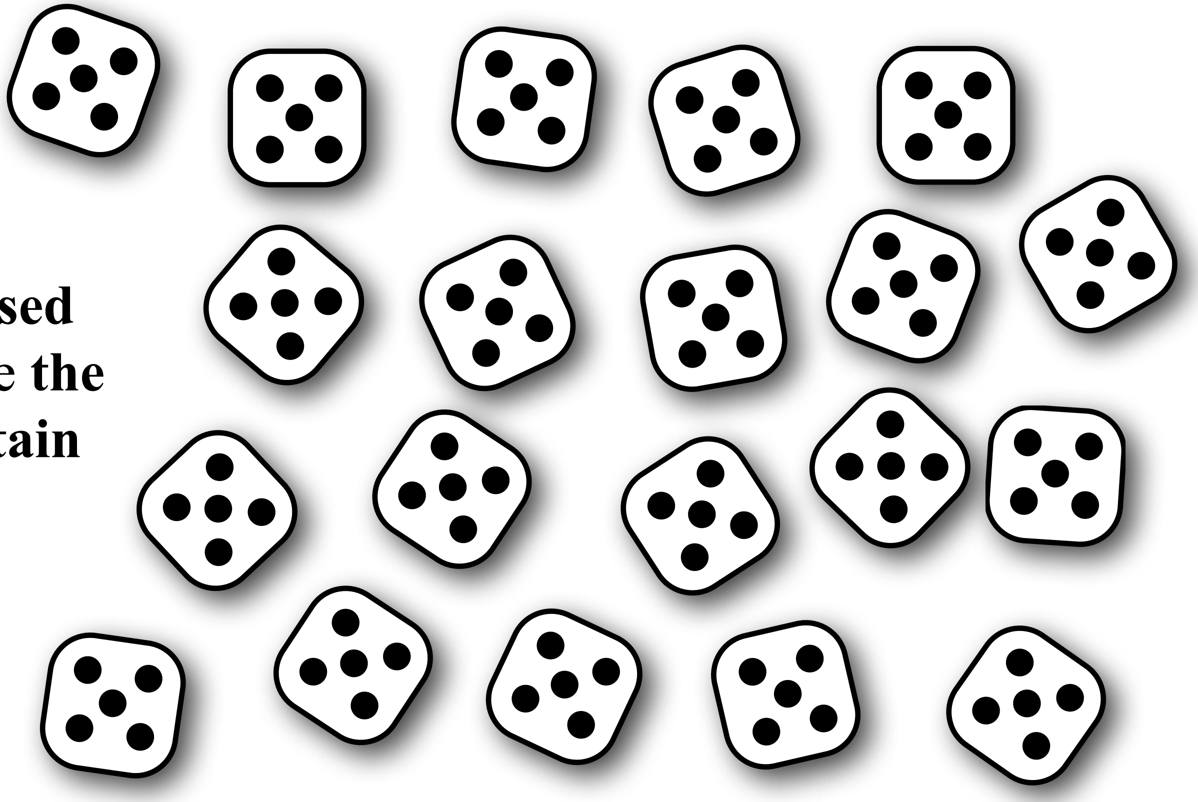


The Trick Dice model

(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	1.0	Not surprised at all

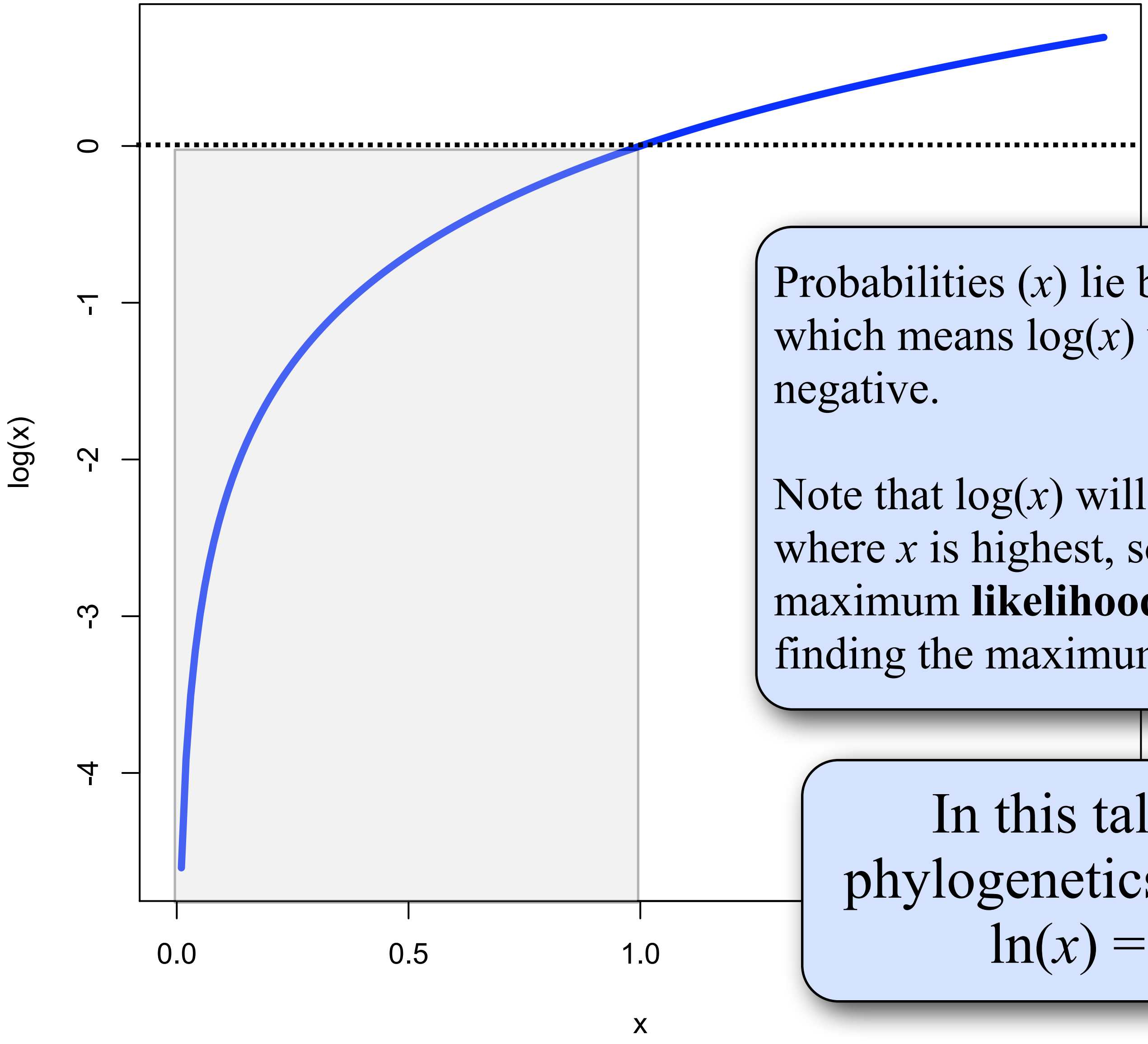
winning model maximizes likelihood
(and thus minimizes surprise)

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 0.05
  - Model 2: branch length is 0.06

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

Likelihoods vs. log-likelihoods

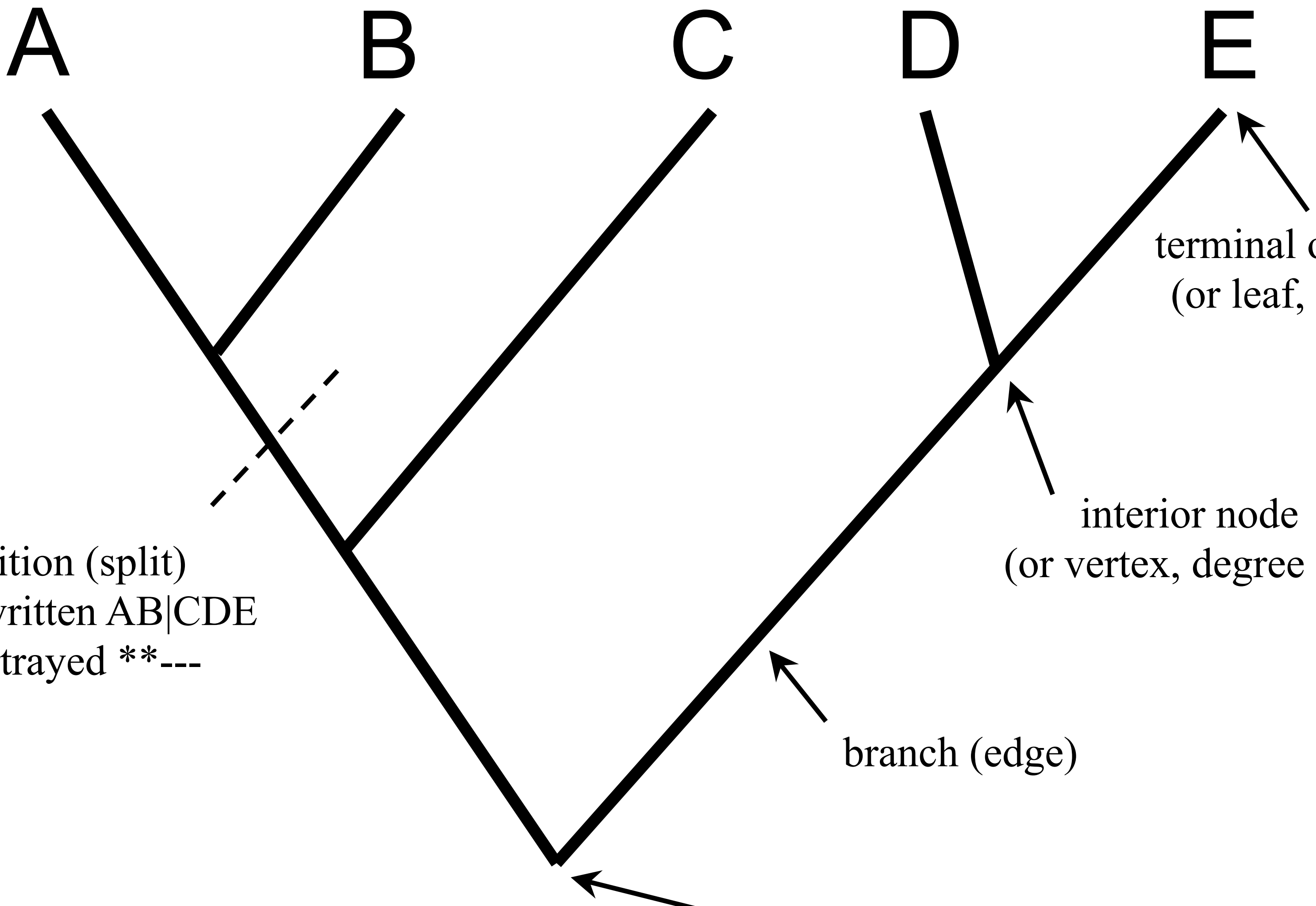


Probabilities (*x*) lie between 0 and 1, which means log(*x*) will always be negative.

Note that log(*x*) will always be highest where *x* is highest, so finding the maximum **likelihood** is equivalent to finding the maximum **log-likelihood**).

In this talk (and in phylogenetics in general), ln(*x*) = log(*x*)

Tree jargon



terminal or tip node
(or leaf, degree 1)

interior node
(or vertex, degree 3+)

bipartition (split)
also written AB|CDE
or portrayed **---

branch (edge)

root node of tree (degree 2)

Likelihood of a single tip node

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.

Model ranking using AIC

The Akaike Information Criterion (AIC) can be used to evaluate whether an **unconstrained** model ("free") fits the data significantly better than a **constrained** version ("equal") of the same model.

Find *maximum* logL under the *unconstrained* model:

$$\begin{aligned}\log L_{\text{free}} &= 12\log(\pi_A) + 7\log(\pi_C) + 7\log(\pi_G) + 6\log(\pi_T)\\ &= 12\log(0.375) + 7\log(0.219) + 7\log(0.219) + 6\log(0.188)\\ &= -43.1\end{aligned}$$

This model has 3 parameters

Find *maximum* logL under the *constrained* model:

$$\begin{aligned}\log L_{\text{equal}} &= 12\log(\pi_A) + 7\log(\pi_C) + 7\log(\pi_G) + 6\log(\pi_T)\\ &= 12\log(0.25) + 7\log(0.25) + 7\log(0.25) + 6\log(0.25)\\ &= -44.4\end{aligned}$$

This model has 0 parameters

Model ranking using AIC

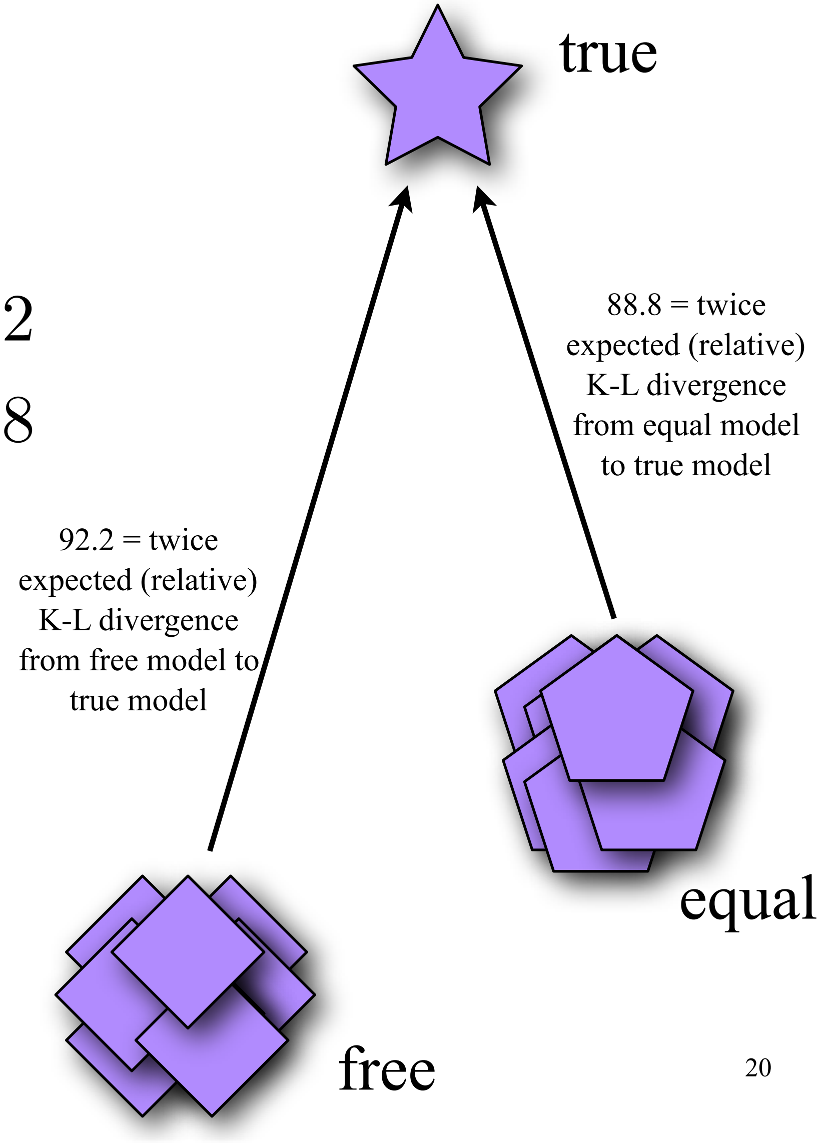
Calculate AIC for each model:

$$AIC = 2k - 2\log(L_{\max})$$

$$AIC_{\mathrm{free}} = 2(3) - 2(-43.1) = 92.2$$

$$AIC_{\mathrm{equal}} = 2(0) - 2(-44.4) = 88.8$$

The constrained model ("equal") is a better choice than the unconstrained model ("free") according to AIC



88.8 = twice expected (relative) K-L divergence from equal model to true model

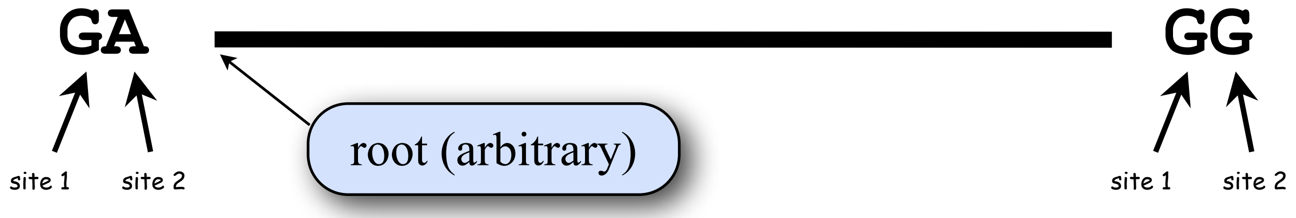
92.2 = twice expected (relative) K-L divergence from free model to true model

(K-L stands for Kullback-Leibler)

Likelihood of the simplest tree

sequence 1
sequence 2

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



$$
\begin{aligned}
L &= L_1 \; L_2 \\
&= \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha t}\right)\right]\left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha t}\right)\right]
\end{aligned}
$$

Pr(G) Pr(G|G, α*t*) Pr(A) Pr(G|A, α*t*)

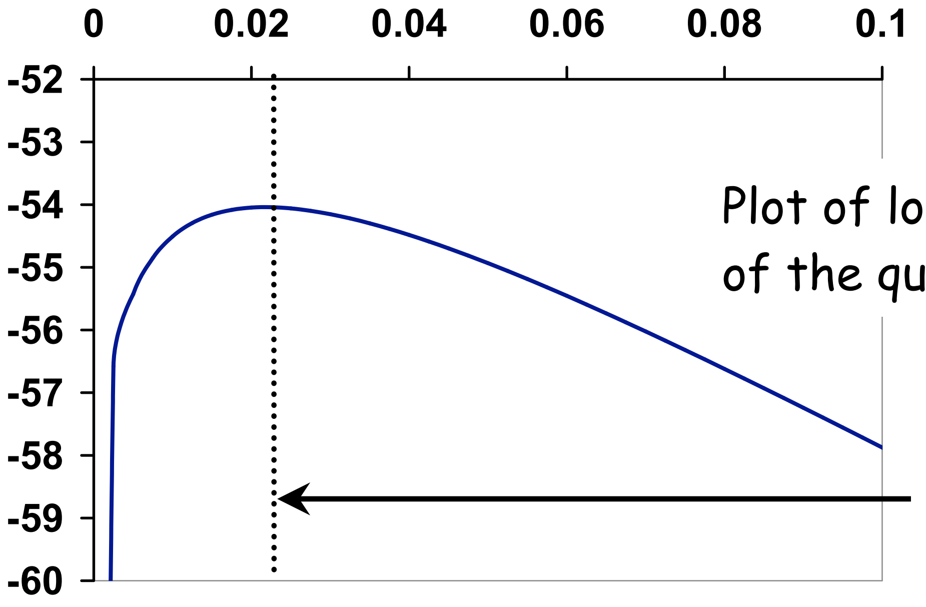
Note that we are NOT assuming independence here

Maximum likelihood estimation

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

```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```

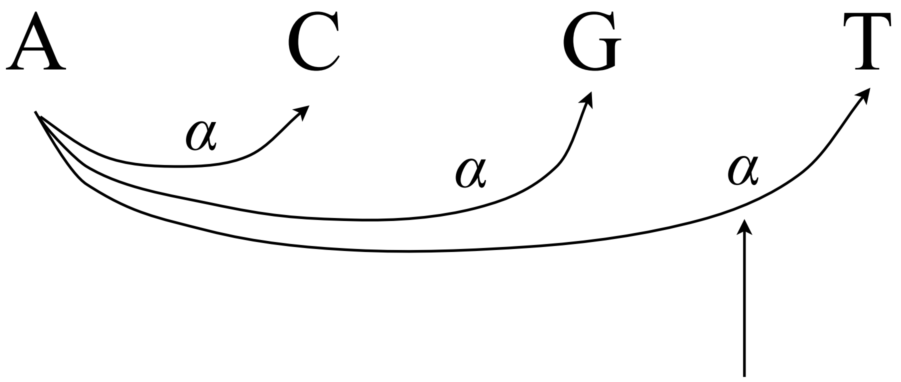
$$L = \left[\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\right]^{30}\left[\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$



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

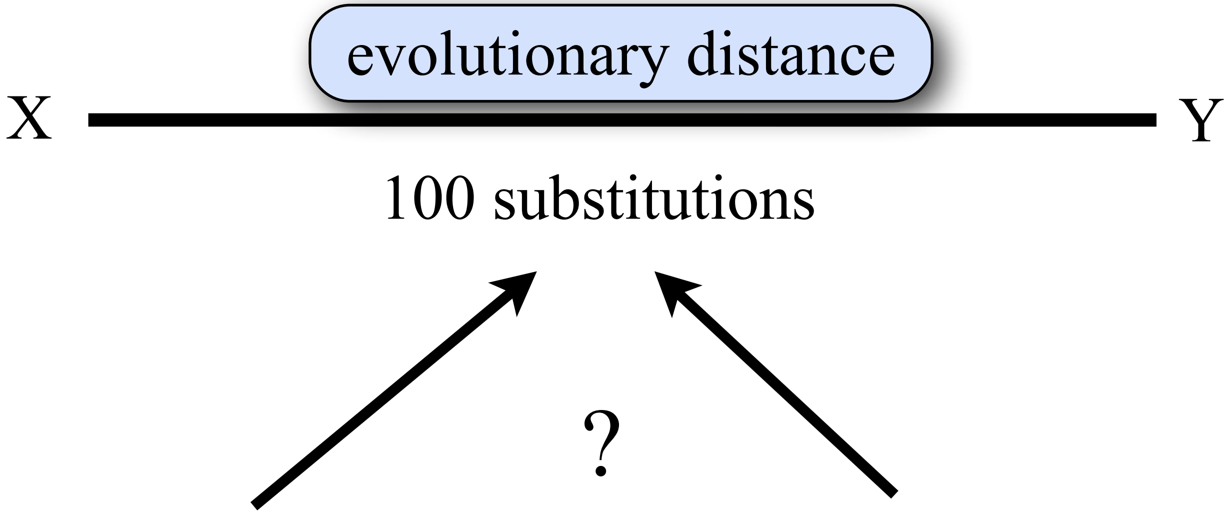


This is the rate at which an existing A changes to a T

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

$$v = 3\alpha t$$

Rate and time are confounded



evolutionary distance

X

Y

100 substitutions

?

$$\left(\frac{\text{1 substitution}}{\text{million years}}\right)\text{100 million years}$$

$$\left(\frac{\text{100 substitutions}}{\text{million years}}\right)\text{1 million years}$$

Later this week you will be introduced to models in which constraints on times can be used to infer rates (and vice versa), but without some extra information or constraints, sequence data allow only estimation of the **number** of substitutions.

A convenient convention

Because rate and time are confounded, it is convenient to arbitrarily standardize things by setting the rate to a value such that **one substitution** is expected to occur in **one unit of time** for each site.

This results in "time" (the length of a branch) being measured in units of **evolutionary distance (expected number of substitutions per site)** rather than years (or some other calendar unit).

evolutionary distance $v = 3\alpha t$

$$v = 3\left(\frac{1}{3}\right)t$$

Setting *α*=1/3 results in *v* equalling *t*

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	$v = \{2\mu\, [\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)]\}\, t$

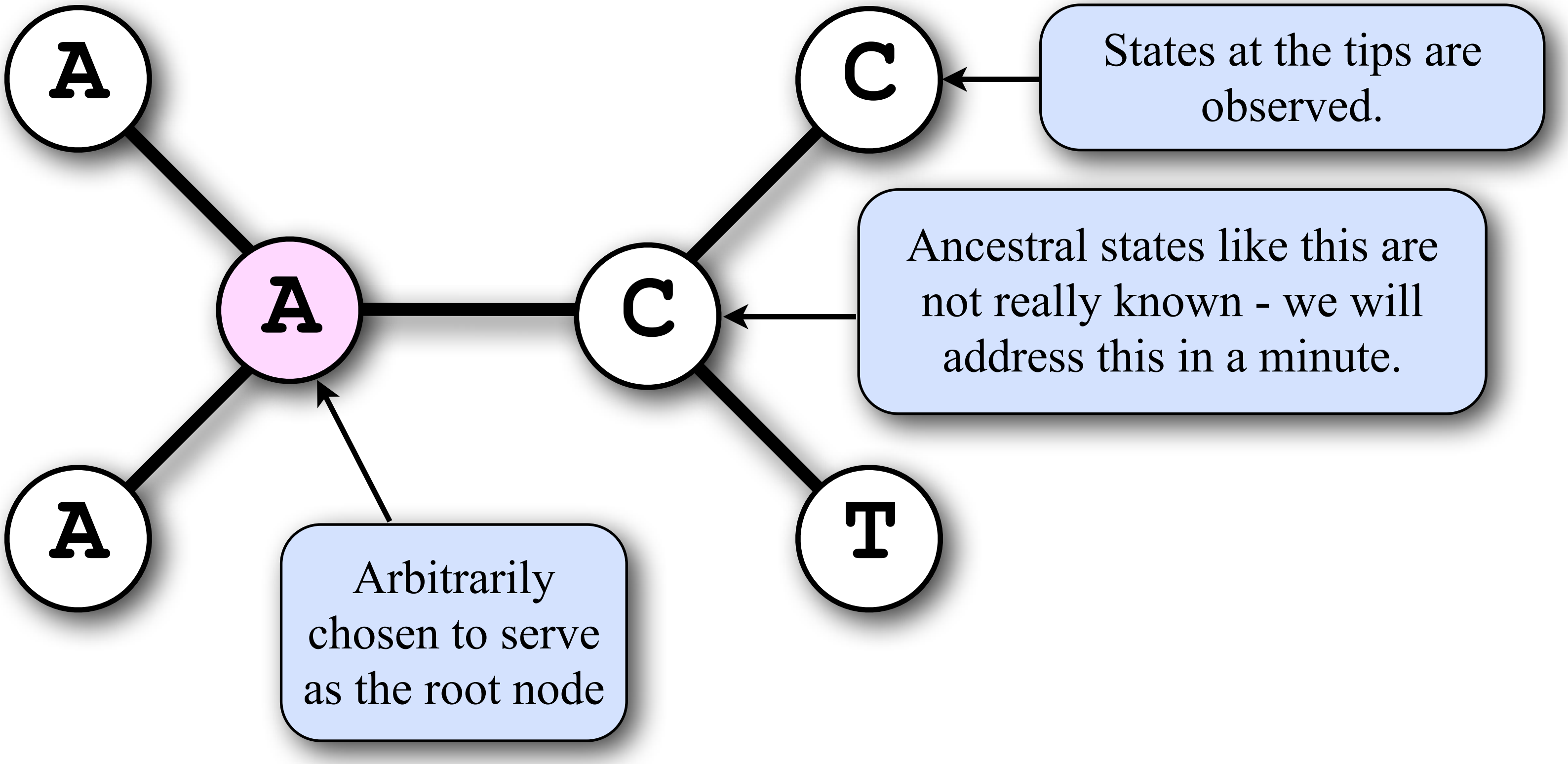
In the formulas above, the overall rate *r* (in curly brackets) is a function of all parameters in the substitution model.

Note that one of the parameters of the substitution model can always be *determined from the branch length* (using our convention that $v = t$).

Typically, all other model parameters are estimated for the *entire tree* (for example, each branch uses the same value of κ)

Likelihood of an unrooted tree

(data shown for only one site)

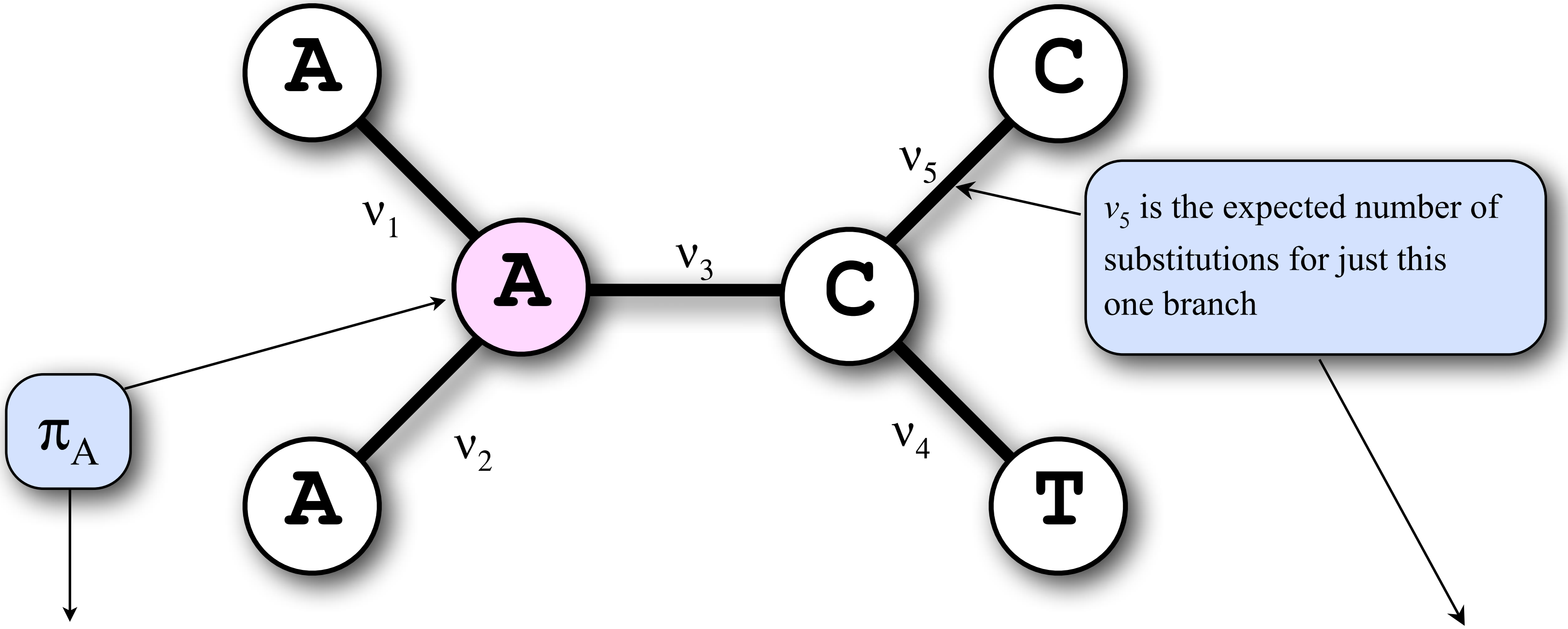


States at the tips are observed.

Ancestral states like this are not really known - we will address this in a minute.

Arbitrarily chosen to serve as the root node

Likelihood for site *k*



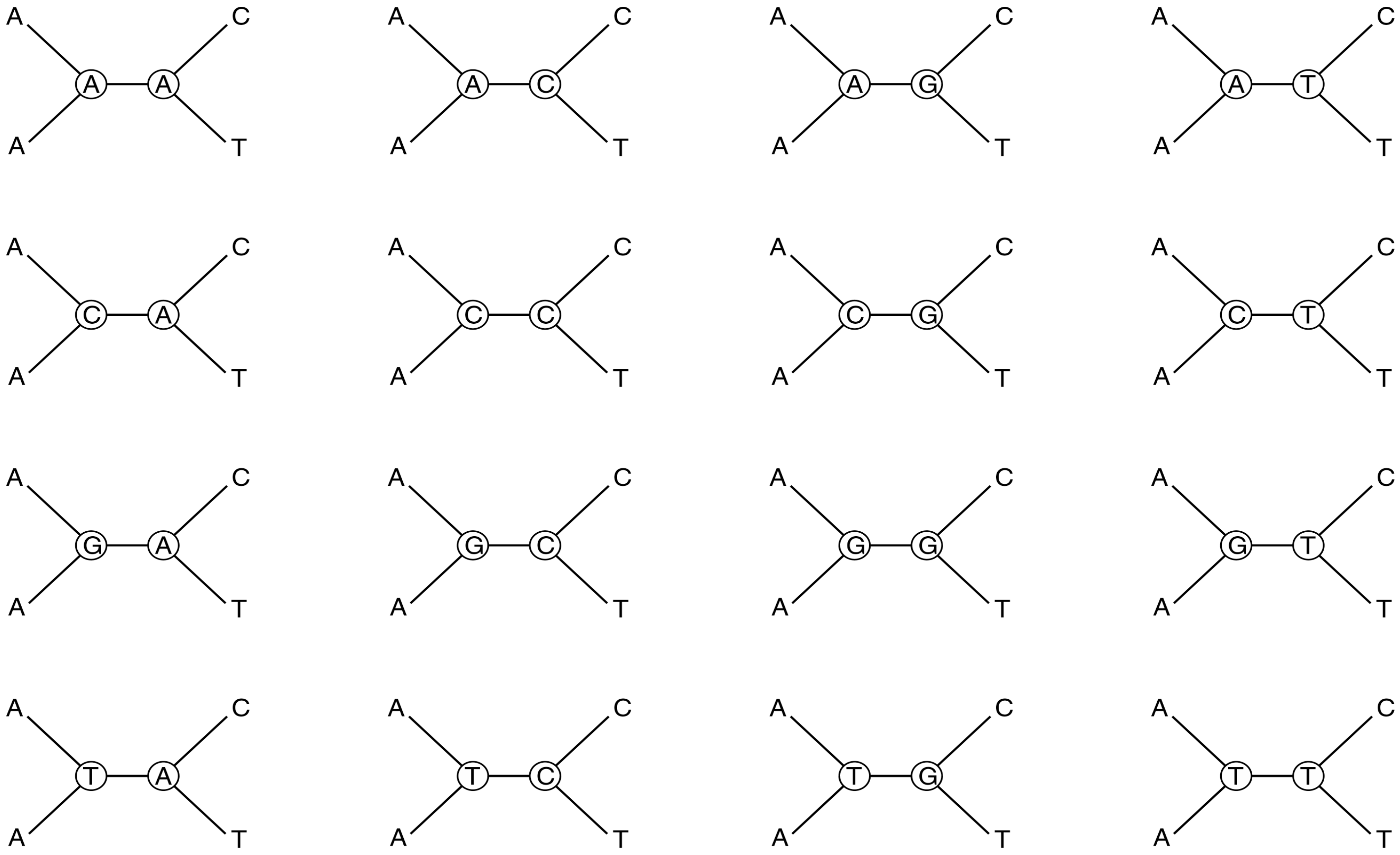
v_5 is the expected number of substitutions for just this one branch

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

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

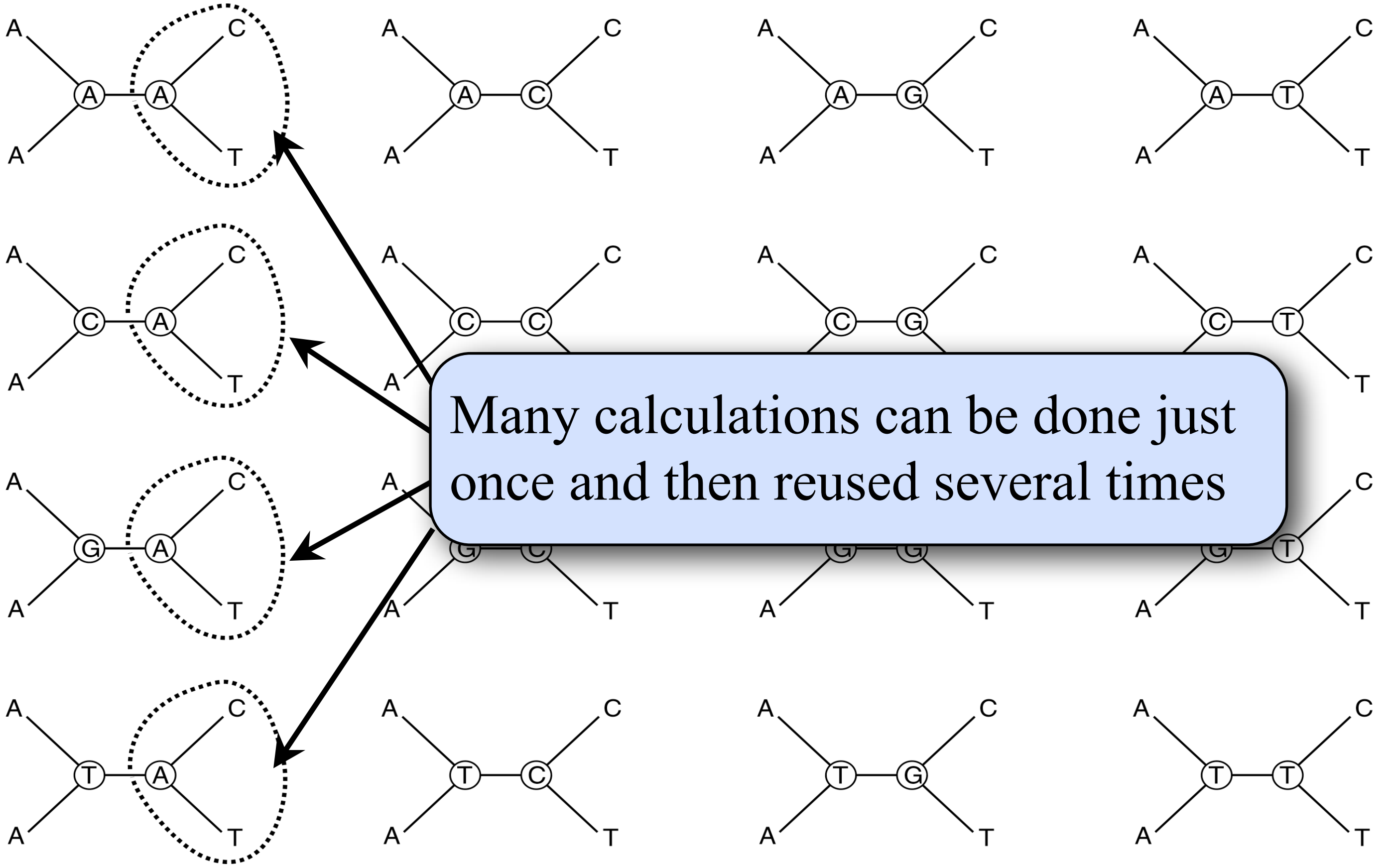
Note use of the AND probability rule

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)



Many calculations can be done just once and then reused several times

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

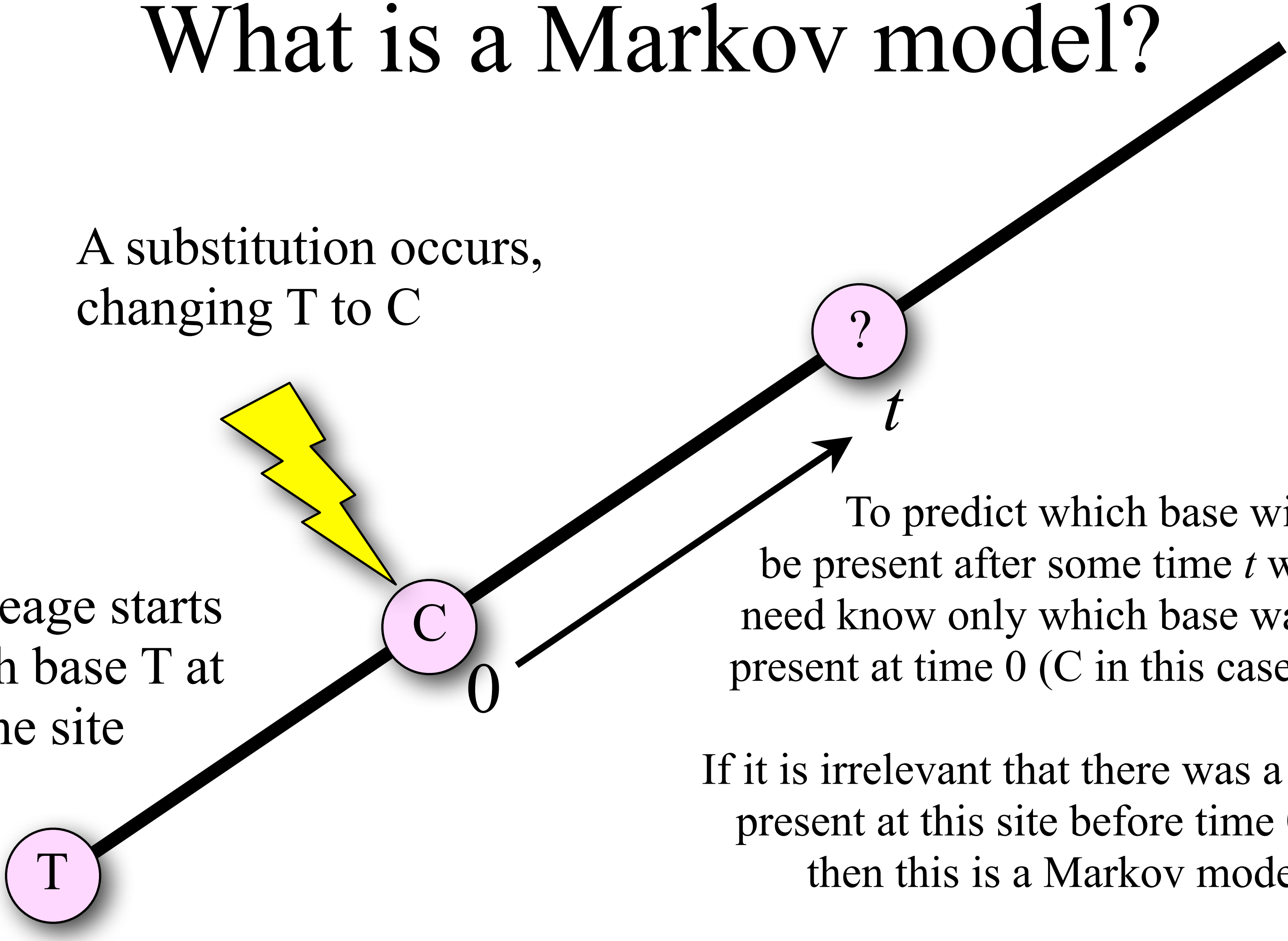
Substitution Models

Jukes-Cantor (JC69) model

- The four bases (A, C, G, T) are expected to be **equally frequent** in sequences ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$)
- Assumes **same rate** for all types of substitution ($r_{A \to C} = r_{A \to G} = r_{A \to T} = r_{C \to G} = r_{C \to T} = r_{G \to T} = \alpha$)
- Usually described as a **1-parameter** model (the parameter being the branch length)
  - Remember, however, that each branch in a tree can have its own length, so there are really as many parameters in the model as there are edges in the tree!
- Assumes substitution is a **Markov** process...

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 *in* H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

What is a Markov model?



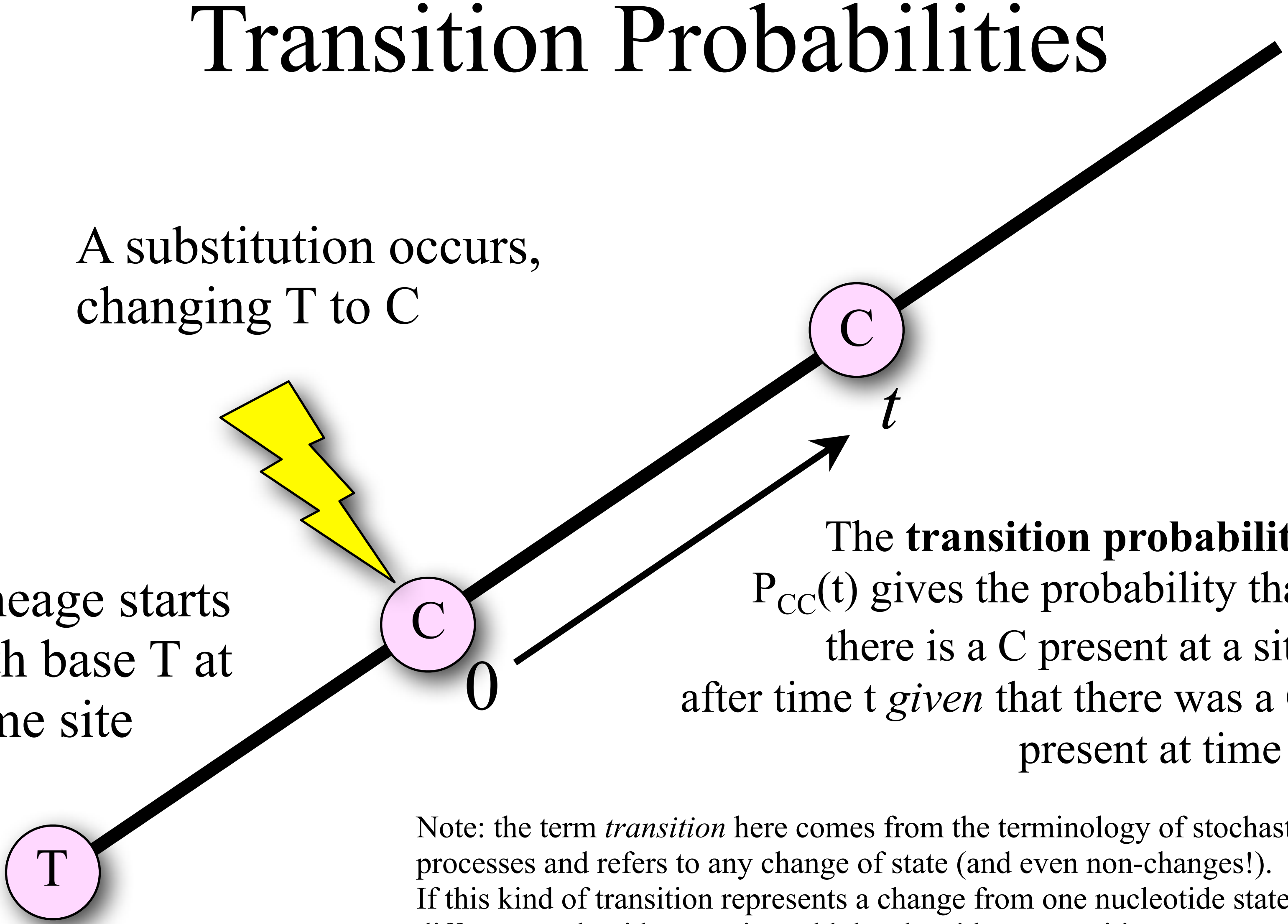
A substitution occurs, changing T to C

Lineage starts with base T at some site

To predict which base will be present after some time *t* we need know only which base was present at time 0 (C in this case).

If it is irrelevant that there was a T present at this site before time 0, then this is a Markov model.

Transition Probabilities



A substitution occurs, changing T to C

Lineage starts with base T at some site

The **transition probability** $P_{CC}(t)$ gives the probability that there is a C present at a site after time t *given* that there was a C present at time 0

Note: the term *transition* here comes from the terminology of stochastic processes and refers to any change of state (and even non-changes!).
If this kind of transition represents a change from one nucleotide state to a different nucleotide state, it could thus be either a transition-type or a transversion-type substitution.

Jukes-Cantor transition probabilities

Here is the probability that a site starting in state T will end up in state G after time *t* when the individual substitution rates are all α:

$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\alpha t}\right)$$

The JC69 model has only one unknown quantity: αt

(The symbol *e* represents the base of the natural logarithms: its value is 2.718281828459045...)

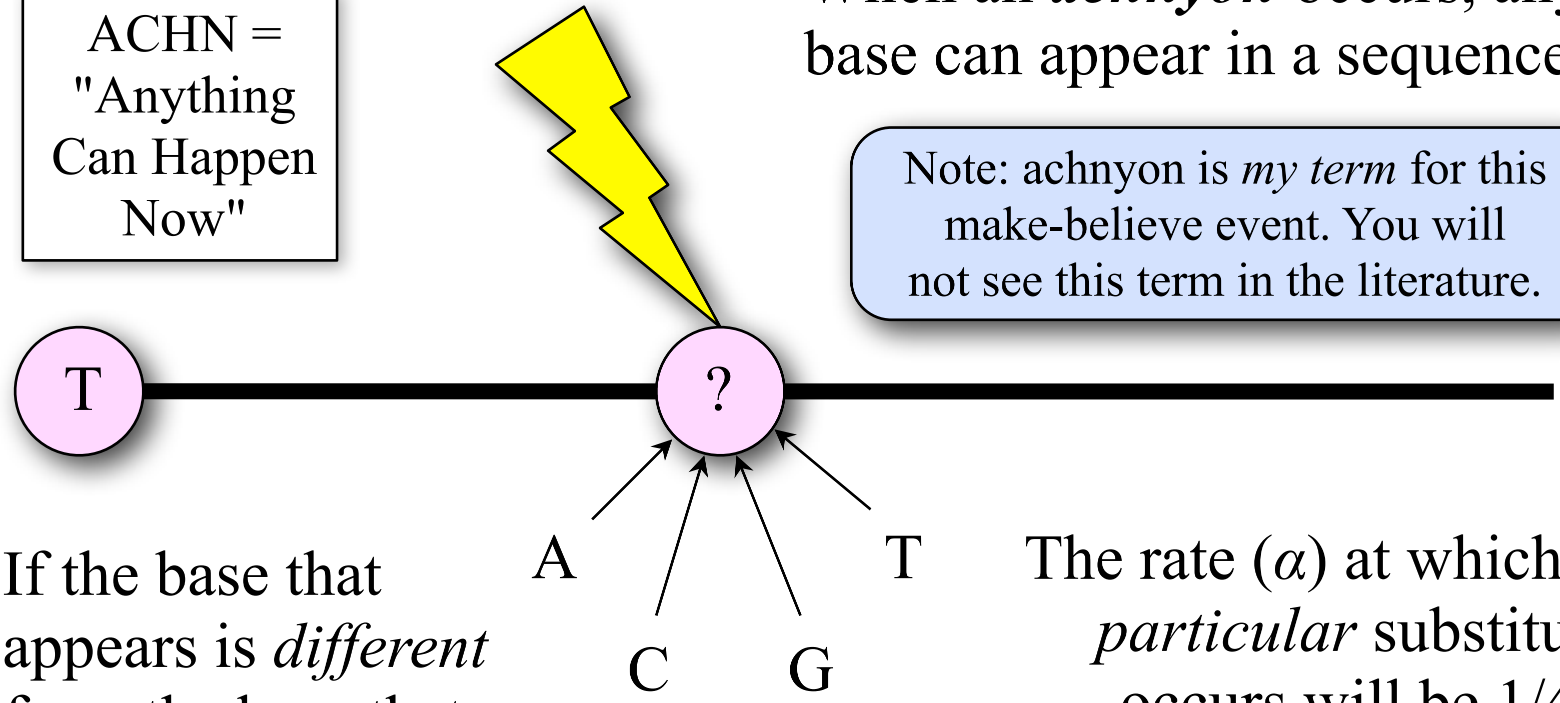
Where does a transition probability formula such as this come from?

"ACHNyons" vs. substitutions

ACHN = "Anything Can Happen Now"

When an ***achnyon*** occurs, any base can appear in a sequence.

Note: achnyon is *my term* for this make-believe event. You will not see this term in the literature.



If the base that appears is *different* from the base that was already there, then a **substitution** event has occurred.

The rate (α) at which any *particular* substitution occurs will be 1/4 the achnyon rate (μ). That is, $\alpha = \mu/4$ (or $\mu = 4\alpha$)

Deriving a transition probability

Calculate the probability that a site currently T will change to G over time *t* when the rate of this particular substitution is α:

Pr(zero achnyons) = $e^{-\mu t}$ (Poisson probability of zero events)

Pr(at least 1 achnyon) = $1 - e^{-\mu t}$

Pr(last achnyon results in base G) = $\frac{1}{4}$

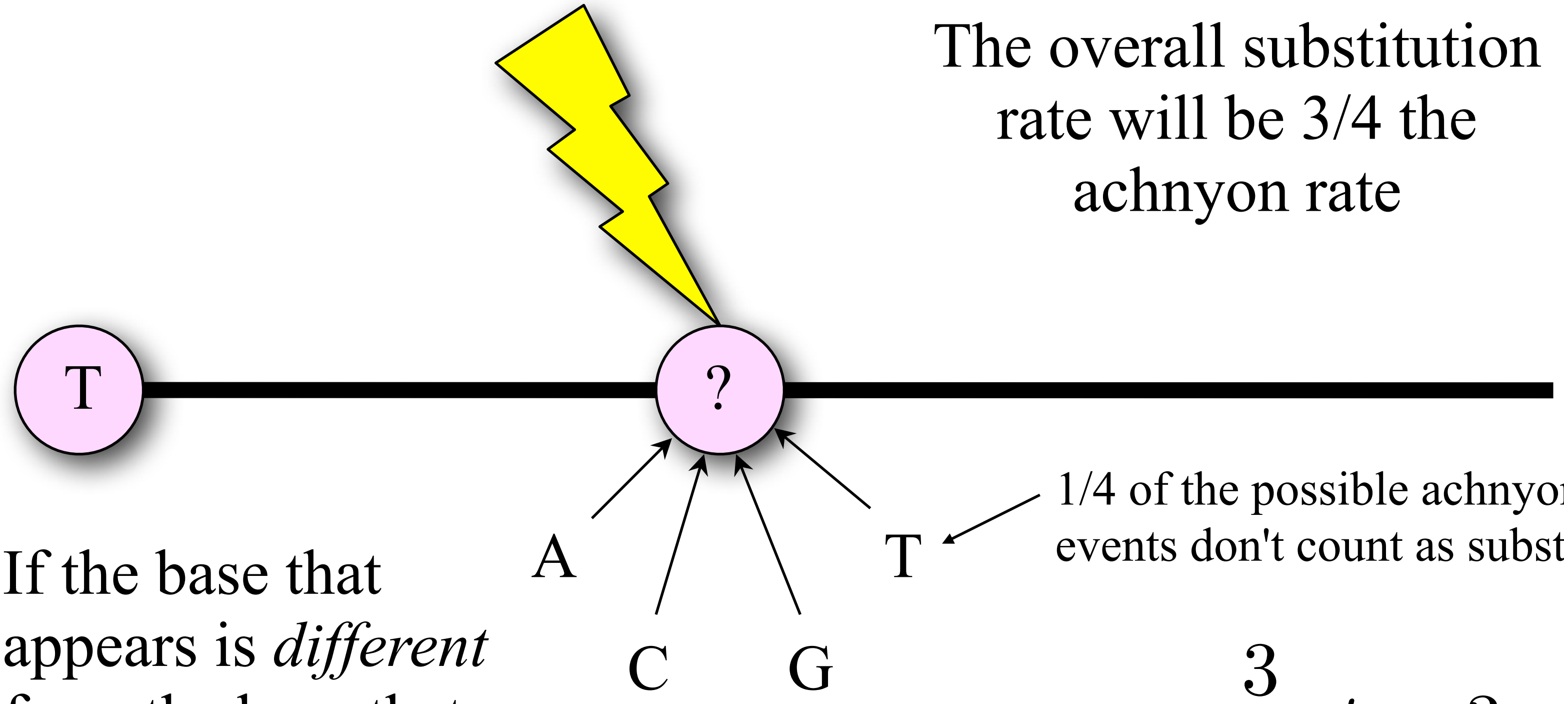
Pr(end in G | start in T) = $\frac{1}{4}\left(1 - e^{-\mu t}\right)$

Remember that the rate (α) of any particular substitution is one fourth the achnyon rate (μ):

$$P_{GT}(t) = \tfrac{1}{4}\left(1 - e^{-4\alpha t}\right)$$

Expected number of substitutions

The overall substitution rate will be 3/4 the achnyon rate



1/4 of the possible achnyon events don't count as substitutions

If the base that appears is *different* from the base that was already there, then a **substitution** event has occurred.

$$\nu = \frac{3}{4}\mu t = 3\alpha t$$

Transition Probabilities: Remarks

$$\begin{array}{rcl}
P_{TA}(t) & = & \frac{1}{4}(1-e^{-4\alpha t}) \\
P_{TC}(t) & = & \frac{1}{4}(1-e^{-4\alpha t}) \\
P_{TG}(t) & = & \frac{1}{4}(1-e^{-4\alpha t}) \\
P_{TT}(t) & = & \frac{1}{4}(1-e^{-4\alpha t}) \\
\hline
 & & 1-e^{-4\alpha t}
\end{array}$$

These should add to 1.0 because T *must* change to something!

Doh! Something must be wrong here...

Transition Probabilities: Remarks

$$
\begin{aligned}
P_{TA}(t) &= \frac{1}{4}(1 - e^{-4\alpha t}) \\
P_{TC}(t) &= \frac{1}{4}(1 - e^{-4\alpha t}) \\
P_{TG}(t) &= \frac{1}{4}(1 - e^{-4\alpha t}) \\
P_{TT}(t) &= \frac{1}{4}(1 - e^{-4\alpha t}) \boxed{+ e^{-4\alpha t}}
\end{aligned}
$$

Forgot to account for the possibility of *no* acnyons over time *t*

Equilibrium frequencies

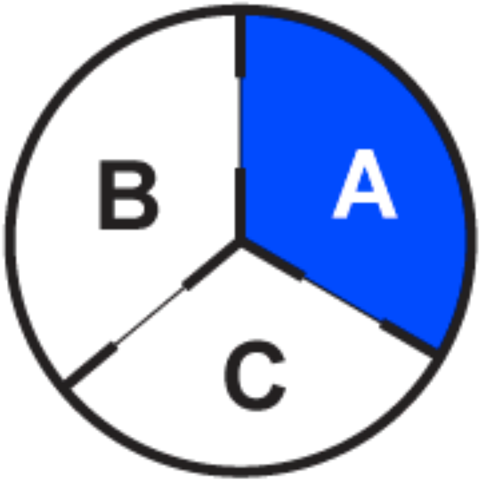
- The JC69 model assumes that the frequencies of the four bases (A, C, G, T) are equal
- The equilibrium relative frequency of each base is thus 0.25
- Why are they called *equilibrium* frequencies?

Equilibrium Frequencies

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

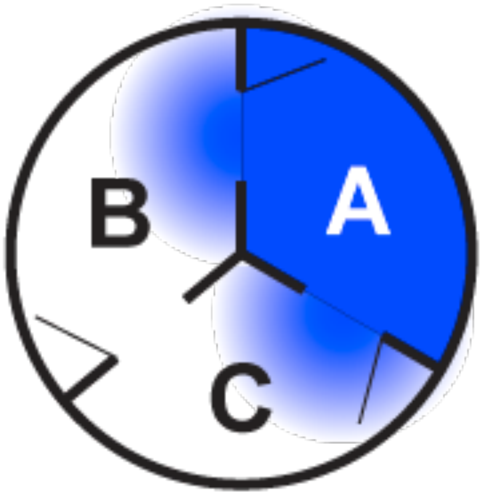
The doors to the other rooms are closed, so the perfume has, thus far, not been able to spread.

What would happen if we opened all the doors?



Equilibrium Frequencies

If the doors are suddenly opened, the perfume would begin diffusing from the area of highest concentration to lowest.



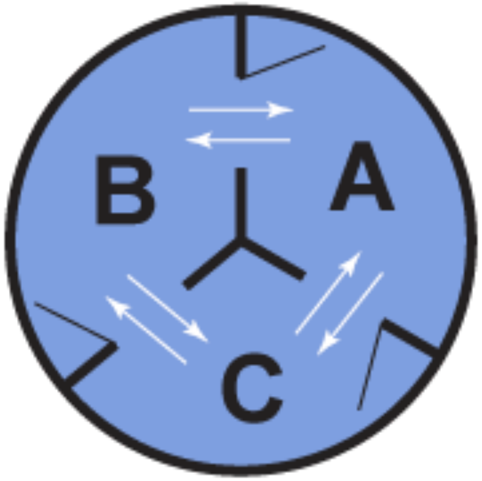
Molecules of perfume go both ways through open doors, but more pass one way than another, leading to a net flow from room A to rooms B and C.

In the instant that the doors are opened, A is losing perfume molecules at *twice the rate* each of the other rooms is gaining molecules. As diffusion progresses, however, the rate of loss from A drops, approaching an equilibrium.

Equilibrium Frequencies

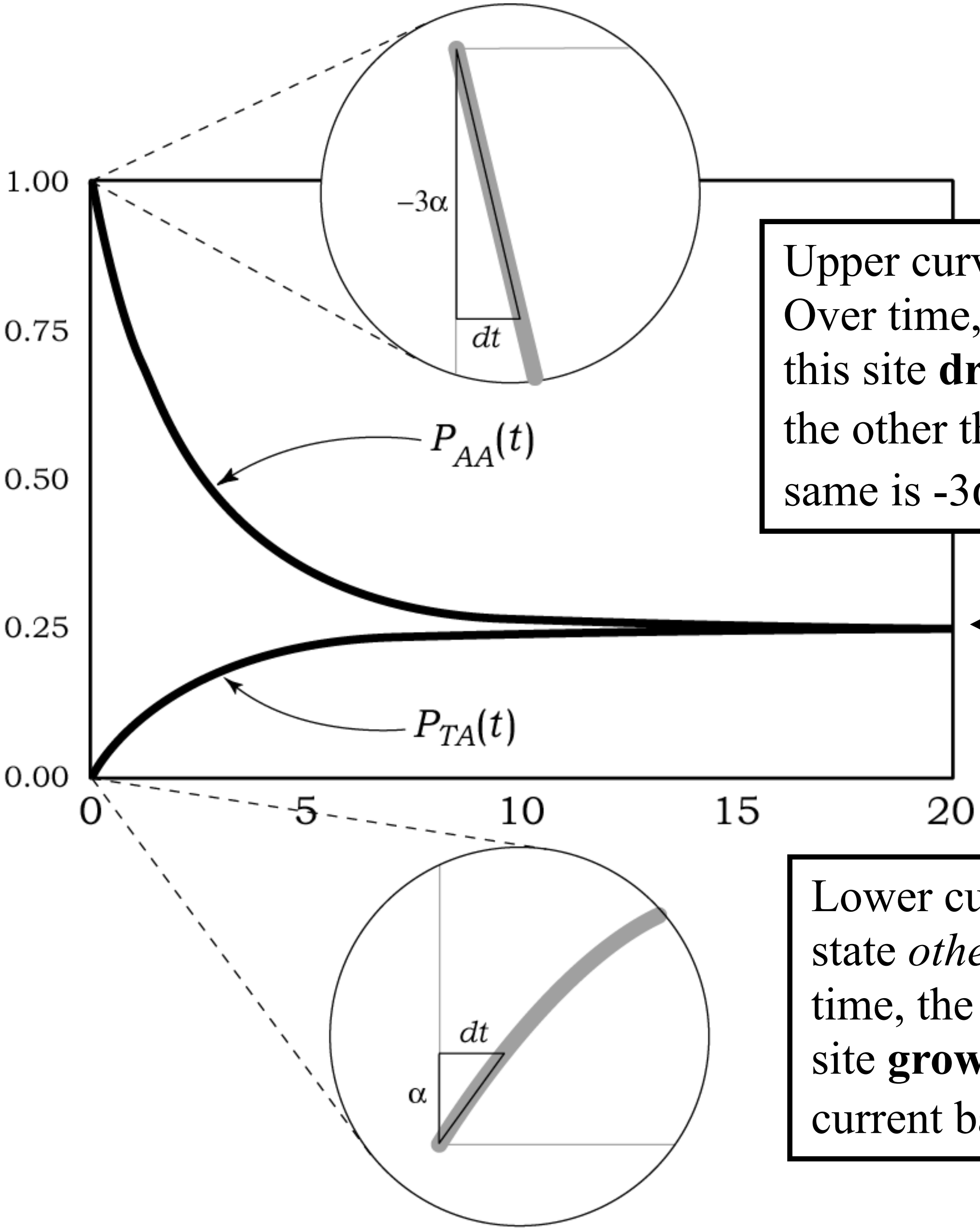
Eventually, all four rooms have essentially the same concentration of perfume.

Molecules still move through doors, but now the rates are the same in all directions.



Back to sequence evolution: assume a sequence began with only A nucleotides (a poly-A sequence). Over time, substitution would begin converting some of these As to Cs, Gs, and Ts, just as the perfume diffused into adjacent rooms.

Pr(A|A) and Pr(A|T) as a function of time



Upper curve assumes we started with A at time 0. Over time, the probability of still seeing an A at this site **drops** because rate of changing to one of the other three bases is 3α (so rate of staying the same is -3α).

The equilibrium relative frequency of A is 0.25

Lower curve assumes we started with some state *other* than A (T is used here). Over time, the probability of seeing an A at this site **grows** because the rate at which the current base will change into an A is α.

JC69 rate matrix

1 parameter:
α

$$\begin{array}{cccccc}
 & & & \text{To} & & \\
 & & A & C & G & T \\
 & A & -3\alpha & \alpha & \alpha & \alpha \\
\text{From} & C & \alpha & -3\alpha & \alpha & \alpha \\
 & G & \alpha & \alpha & -3\alpha & \alpha \\
 & T & \alpha & \alpha & \alpha & -3\alpha
\end{array}$$

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 *in* H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

K80 (or K2P) rate matrix

2 parameters:
α
β

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

transition rate   transversion rate

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120.

K80 rate matrix
(looks different, but actually the same)

2 parameters:
κ
β

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

All I've done is re-parameterize the rate matrix, letting κ equal the *transition/transversion rate ratio* ⟶ $\kappa = \dfrac{\alpha}{\beta}$

Note: the K80 model is identical to the JC69 model if $\kappa = 1$ ($\alpha = \beta$)

F81 rate matrix

4 parameters:

μ

π_A

π_C

π_G

$$\begin{array}{c|cccc} & A & C & G & T \\ \hline A & -\mu(1-\pi_A) & \pi_C\mu & \pi_G\mu & \pi_T\mu \\ C & \pi_A\mu & -\mu(1-\pi_C) & \pi_G\mu & \pi_T\mu \\ G & \pi_A\mu & \pi_C\mu & -\mu(1-\pi_G) & \pi_T\mu \\ T & \pi_A\mu & \pi_C\mu & \pi_G\mu & -\mu(1-\pi_T) \end{array}$$

Note: the F81 model is identical to the JC69 model if all base frequencies are equal

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

HKY85 rate matrix

5 parameters:
κ
β
π_A
π_C
π_G

$$\begin{array}{c|cccc}
 & A & C & G & T \\
\hline
A & - & \pi_C\beta & \pi_G\beta\kappa & \pi_T\beta \\
C & \pi_A\beta & - & \pi_G\beta & \pi_T\beta\kappa \\
G & \pi_A\beta\kappa & \pi_C\beta & - & \pi_T\beta \\
T & \pi_A\beta & \pi_C\beta\kappa & \pi_G\beta & -
\end{array}$$

A dash means equal to negative sum of other elements on the same row

Note: the HKY85 model is identical to the F81 model if $\kappa = 1$. If, in addition, all base frequencies are equal, it is identical to JC69.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 21:160-174.

F84 vs. HKY85

F84 model:

μ rate of process generating *all types of substitutions*

$k\mu$ rate of process generating *only transitions*

Becomes F81 model if $k = 0$

HKY85 model:

β rate of process generating *only transversions*

$\kappa\beta$ rate of process generating *only transitions*

Becomes F81 model if $\kappa = 1$

F84 first used in Felsenstein’s PHYLIP package in 1984, first published by: Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution 29: 170-179.

GTR rate matrix

9 parameters:

π_A
π_C
π_G
a
b
c
d
e
μ

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

Identical to the F81 model if $a = b = c = d = e = f = 1$. If, in addition, all the base frequencies are equal, GTR is identical to JC69. If $a = c = d = f = \beta$ and $b = e = \kappa\beta$, GTR becomes the HKY85 model.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 20:86-93.

Rate Heterogeneity

Green Plant rbc*L*

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

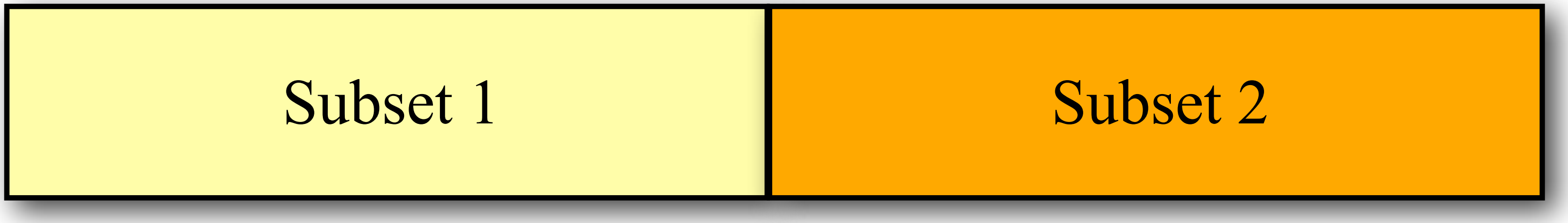
```
M--S--P--Q--T--E--T--K--A--S--V--G--F--K--A--G--V--K--D--Y--K--L--T--Y--Y--T--P--E--Y--E--T--K--D--T--D--I--L--A--A--F--R--V--T--P--
Chara           (green alga; land plant lineage)   AAAGATTACAGATTAACTTACTATACTCCTGAGTATAAAACTAAAGATACTGACATTTTAGCTGCATTTCGTGTAACTCCA
Chlorella       (green alga)                       .....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 four bases are observed at some sites...

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

Site-specific rates

Each defined subset (e.g. gene, codon position) has its own relative rate



Subset 1

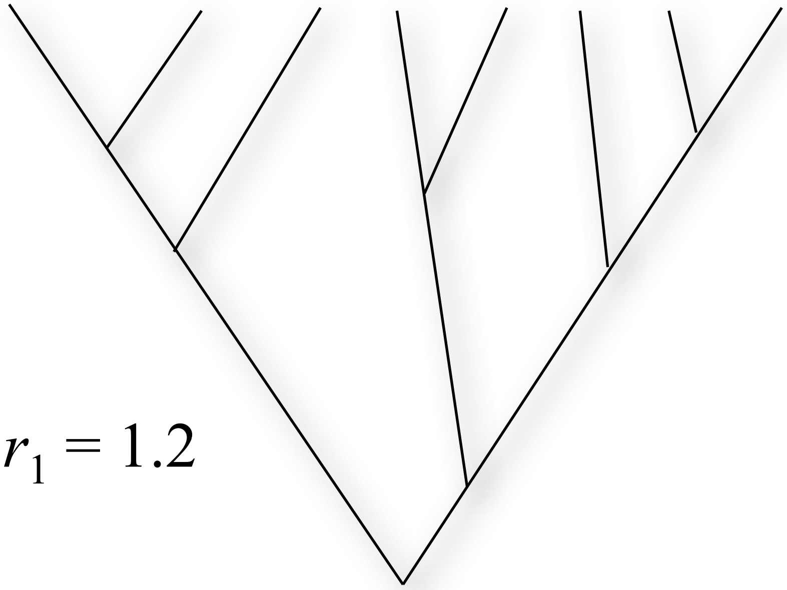
Subset 2

r_1 applies to subset 1
(e.g. sites 1 - 1000)

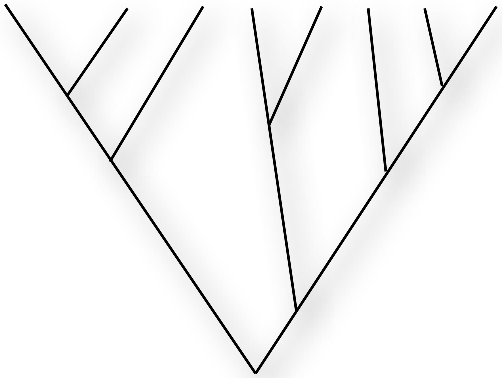
r_2 applies to subset 2
(e.g. sites 1001-2000)

Site-specific rates

$$L = \underbrace{\Pr(D_1|r_1)\cdots\Pr(D_{1000}|r_1)}_{\text{Gene 1}}\ \underbrace{\Pr(D_{1001}|r_2)\cdots\Pr(D_{2000}|r_2)}_{\text{Gene 2}}$$



$r_1 = 1.2$



$r_2 = 0.8$

Site-specific rates

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

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

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

Site specific rates

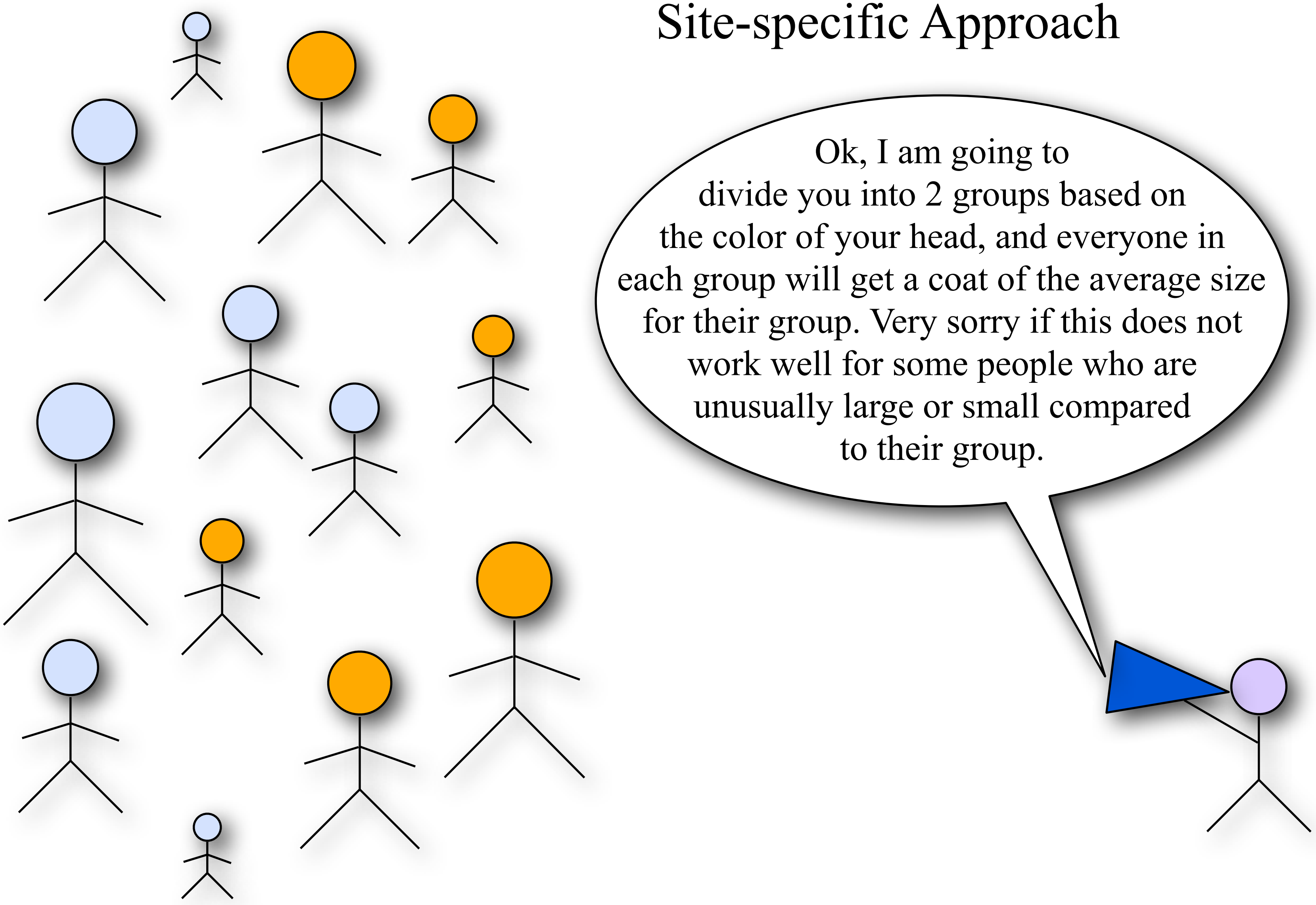
JC69 transition probabilities that would be used for sites in **gene 1**:

$$\begin{aligned} P_{ii}(t) &= \frac{1}{4} + \frac{3}{4} e^{-4 r_1 \alpha t} \\ P_{ij}(t) &= \frac{1}{4} - \frac{1}{4} e^{-4 r_1 \alpha t} \end{aligned}$$

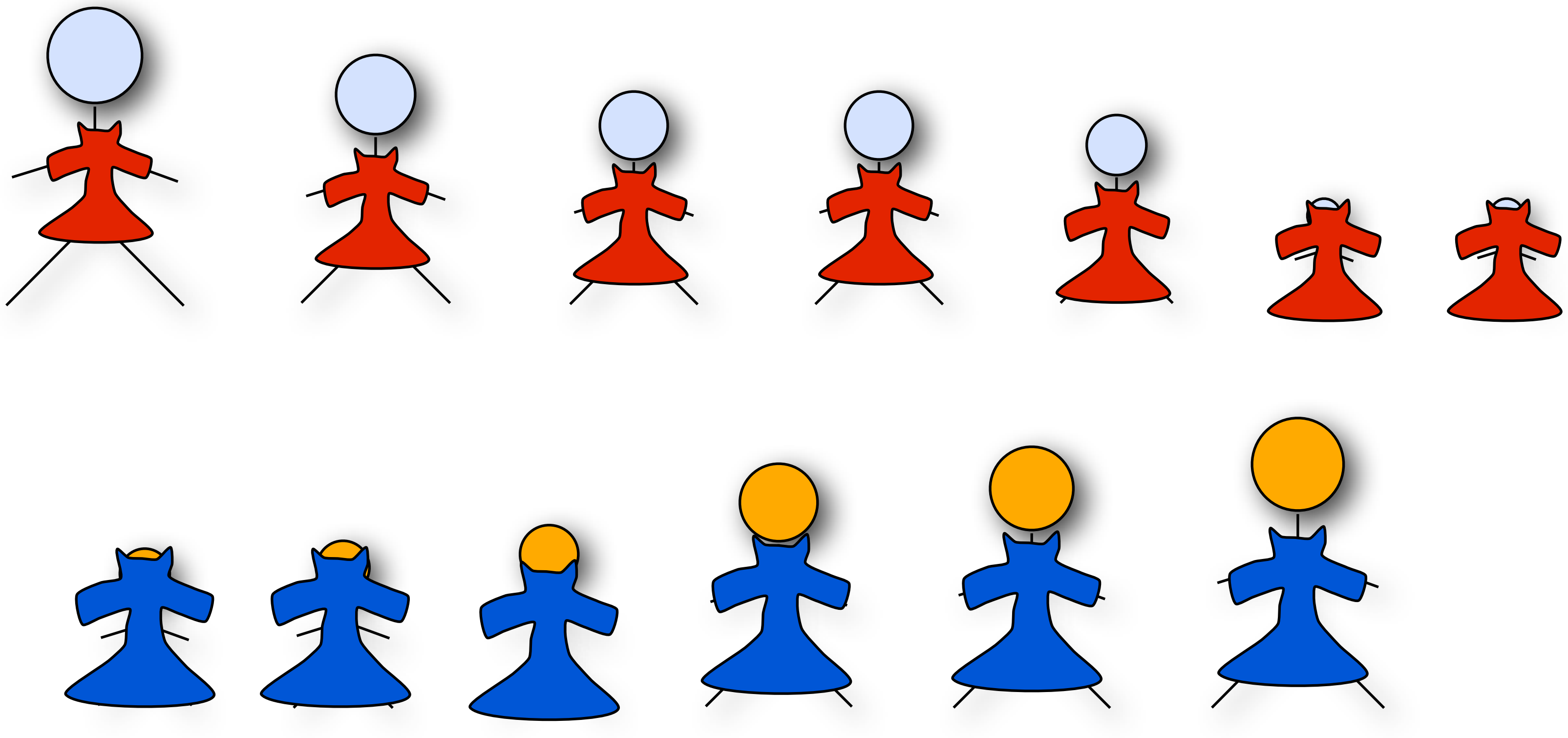
JC69 transition probabilities that would be used for sites in **gene 2**:

$$\begin{aligned} P_{ii}(t) &= \frac{1}{4} + \frac{3}{4} e^{-4 r_2 \alpha t} \\ P_{ij}(t) &= \frac{1}{4} - \frac{1}{4} e^{-4 r_2 \alpha t} \end{aligned}$$

Site-specific Approach


Ok, I am going to divide you into 2 groups based on the color of your head, and everyone in each group will get a coat of the average size for their group. Very sorry if this does not work well for some people who are unusually large or small compared to their group.

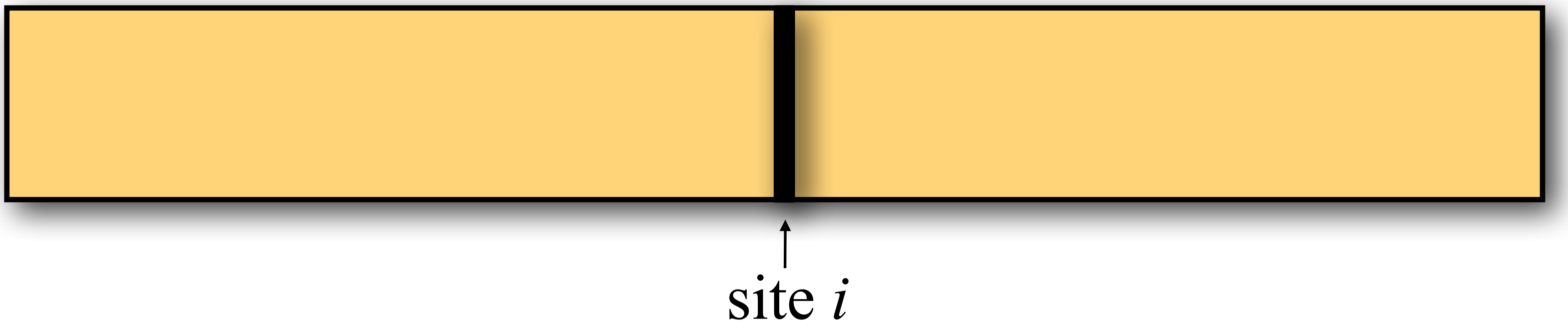
Site-specific Approach



Pro: costs less: need to buy just one coat for every person
Con: every person in a group has to wear the same size coat, so the fit will be poor for some people if they are much bigger or smaller than the average size for the group in which they have been placed

Mixture Models

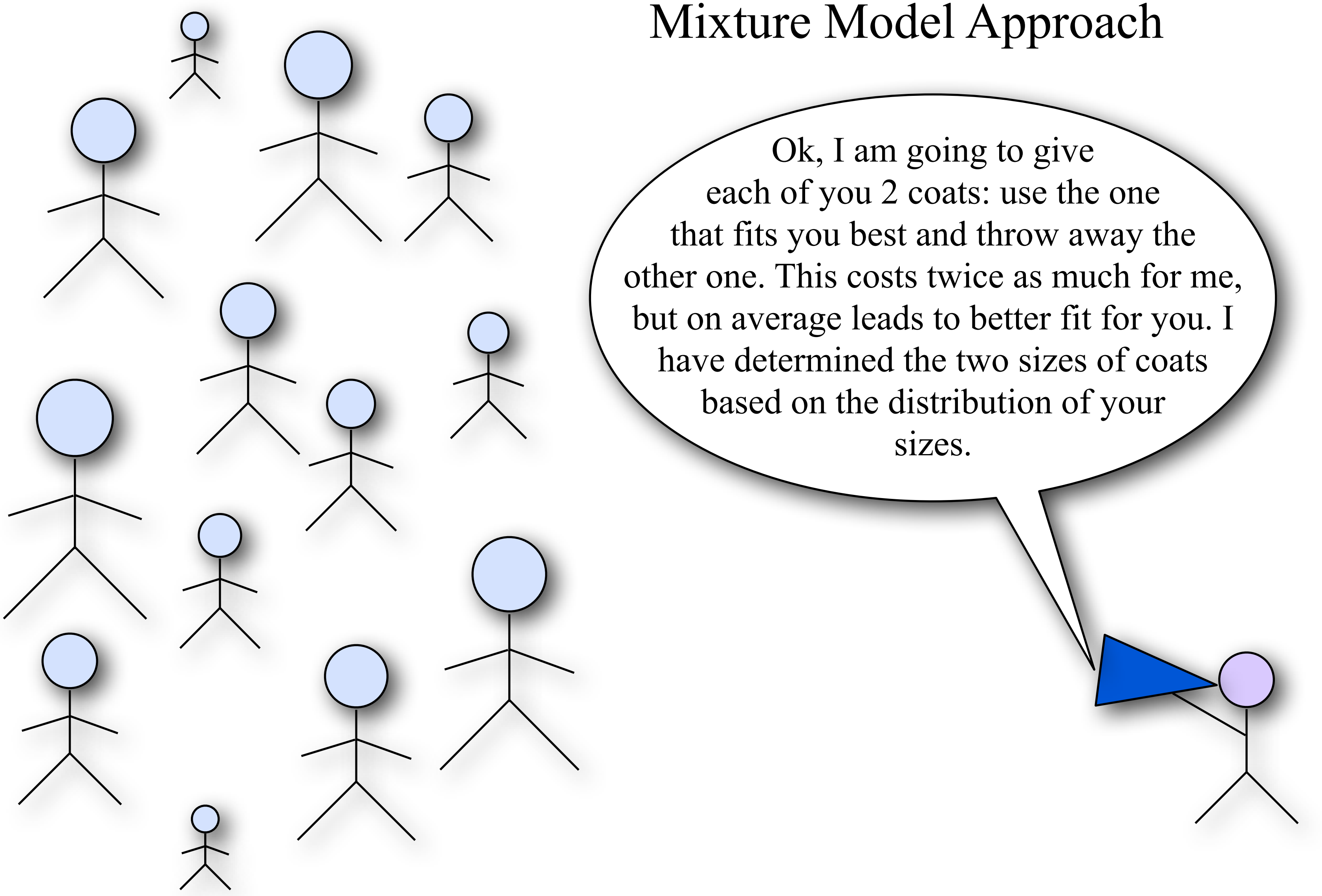
All relative rates applied to every site



$$L_i = \Pr(D_i|r_1)\Pr(r_1) + \Pr(D_i|r_2)\Pr(r_2)$$

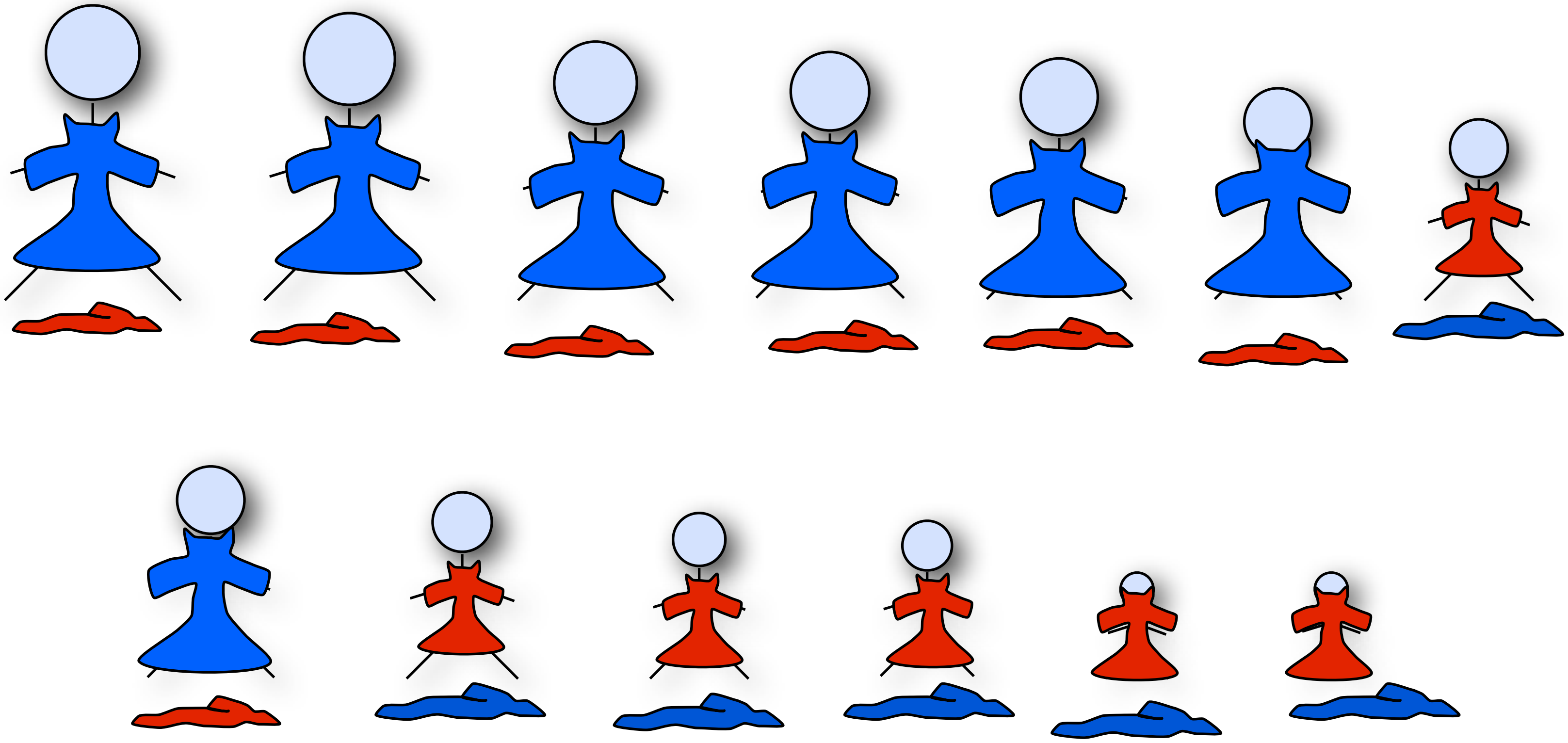
$$\text{Common examples} \begin{cases} \text{Invariable sites (I) model} \\ \text{Discrete Gamma (G) model} \end{cases}$$

Mixture Model Approach



Ok, I am going to give each of you 2 coats: use the one that fits you best and throw away the other one. This costs twice as much for me, but on average leads to better fit for you. I have determined the two sizes of coats based on the distribution of your sizes.

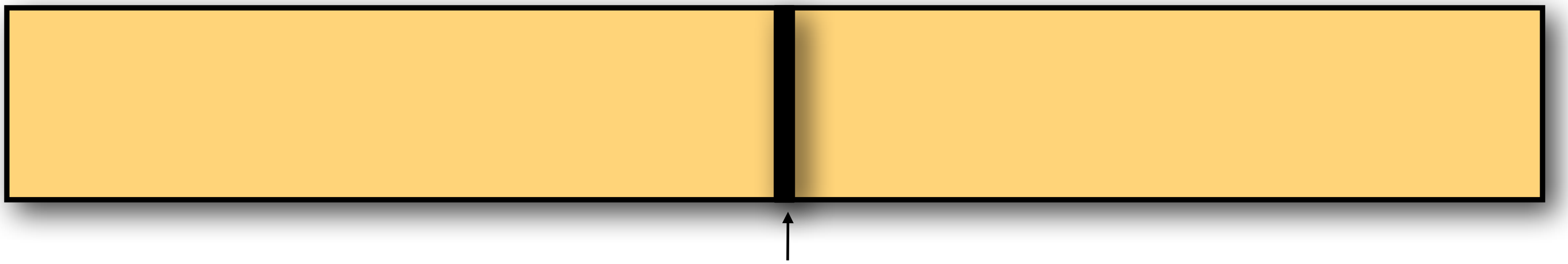
Mixture Model Approach



Pro: every person experiences better fit because they can choose the size coat that fits best
Con: costs more because two coats much be provided for each person

Invariable Sites Model

A fraction p_{invar} of sites are assumed to be invariable (i.e. rate = 0.0)



site i

$$L_i = \Pr(D_i|r_1)p_{\text{invar}} + \Pr(D_i|r_2)(1 - p_{\text{invar}})$$

$$r_1 = 0.0$$

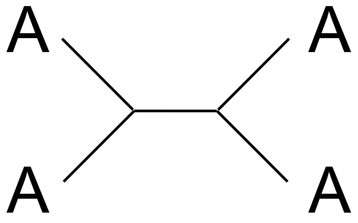
$$r_2 = \frac{1}{1 - p_{\text{invar}}}$$

Allows for the possibility that any given site could be variable or invariable

Reeves, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. Journal of Molecular Evolution 35:17-31.

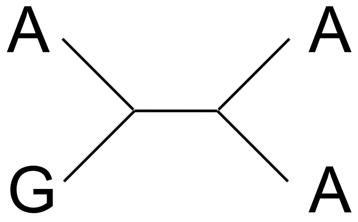
Invariable sites model

If site *i* is a *constant* site, both terms will contribute to the site likelihood:



$$L_i = \Pr(D_i|0.0)p_{\text{invar}} + \Pr(D_i|r_2)(1 - p_{\text{invar}})$$

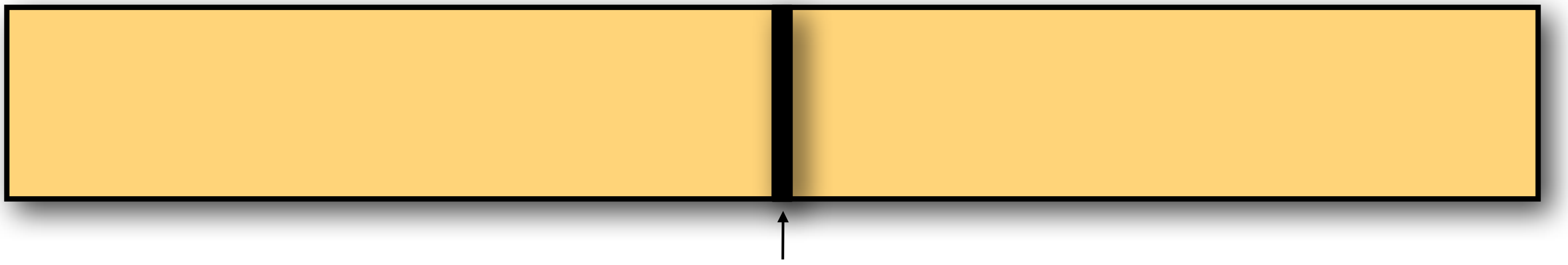
If site *i* is a *variable* site, there is no way to explain the data with a zero rate, so the first term is zero:



$$L_i = \cancel{\Pr(D_i|0.0)p_{\text{invar}}} + \Pr(D_i|r_2)(1 - p_{\text{invar}})$$

Discrete Gamma Model

No relative rates are exactly 0.0, and all are equally probable



site *i*

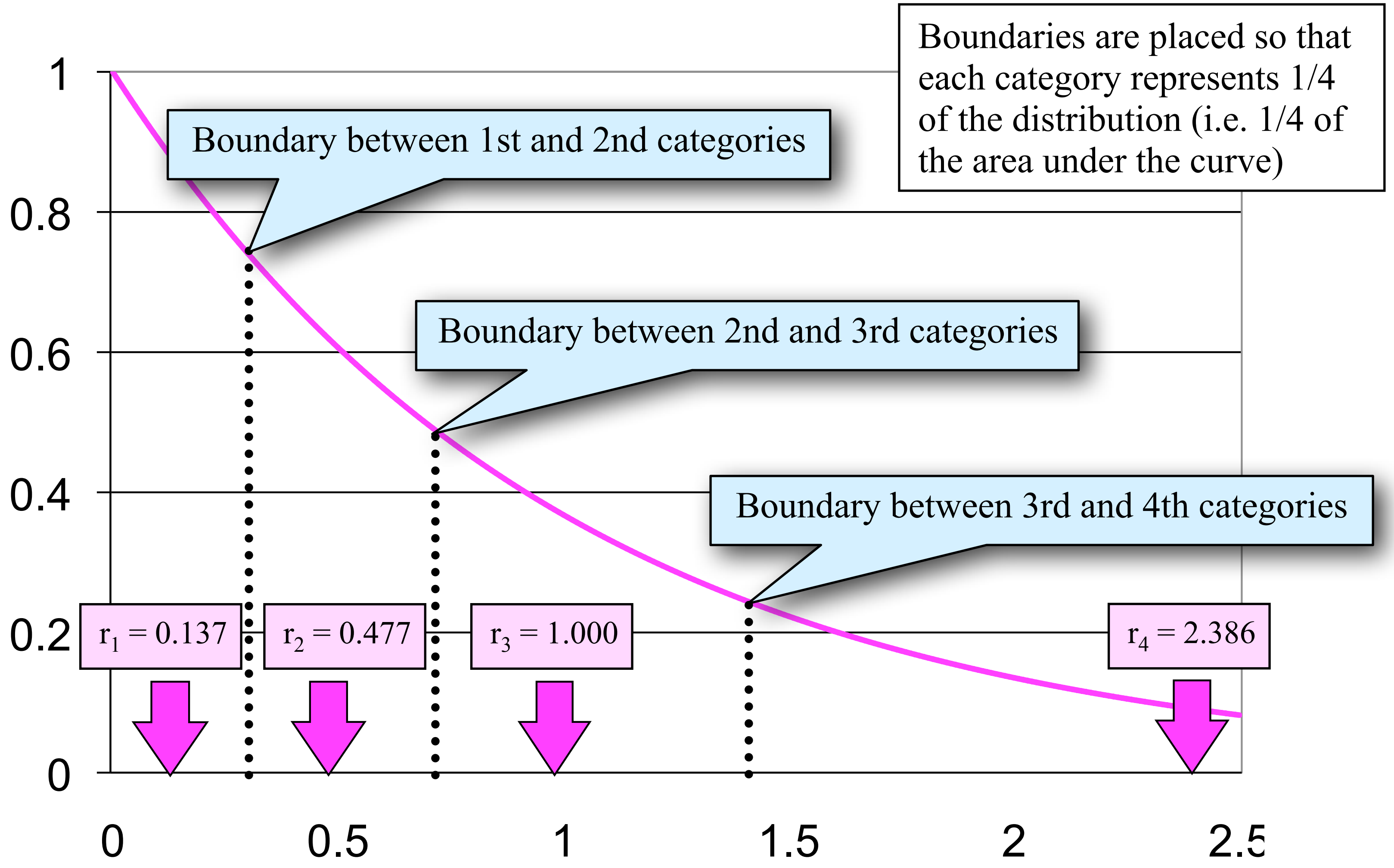
$$L = (\tfrac{1}{4})\,\Pr(D_i|r_1) + (\tfrac{1}{4})\,\Pr(D_i|r_2) + (\tfrac{1}{4})\,\Pr(D_i|r_3) + (\tfrac{1}{4})\,\Pr(D_i|r_4)$$

Relative rates are constrained to a discrete gamma distribution
Number of rate categories can vary (4 used here)

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396-1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306-314.

Relative rates in 4-category case



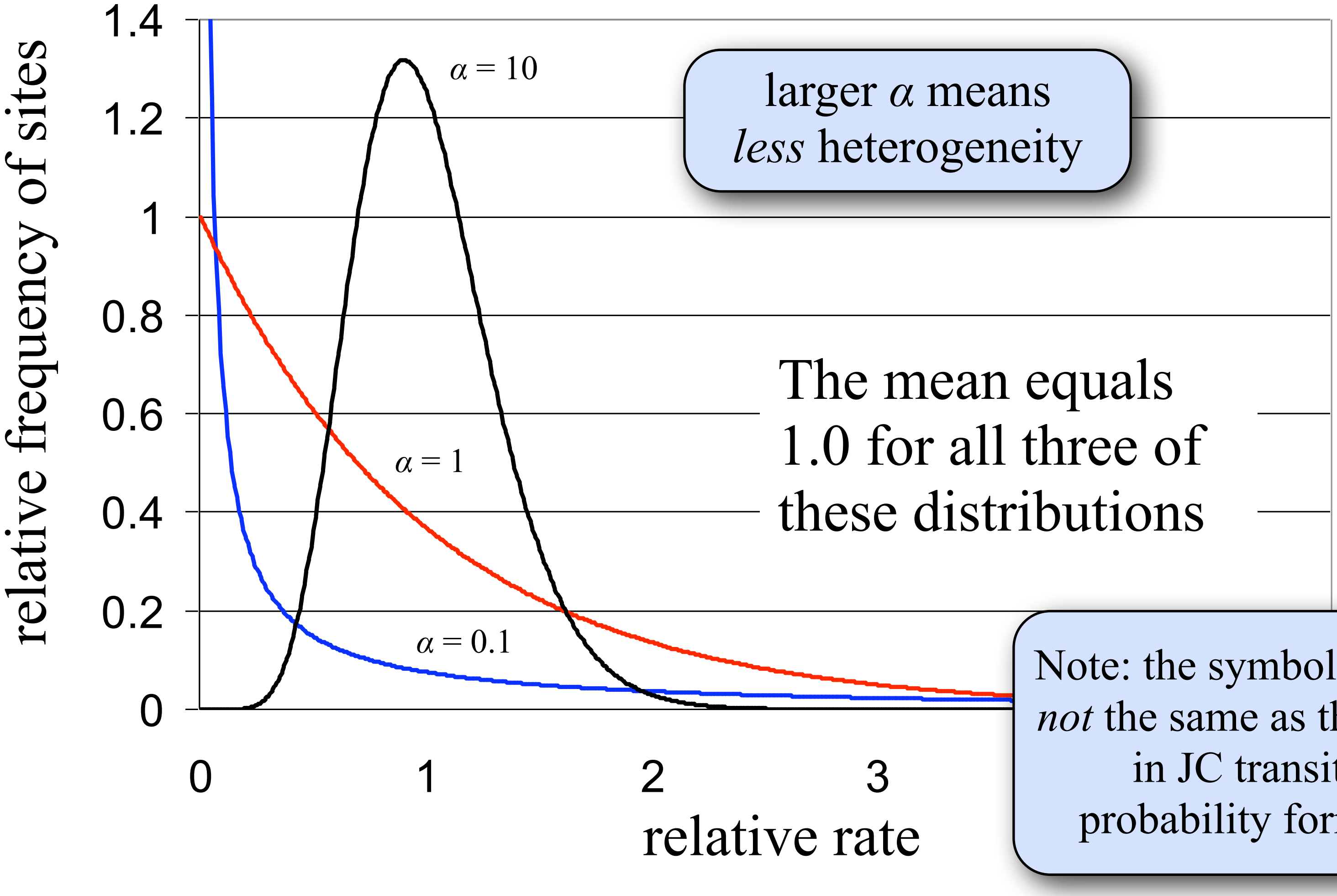
Boundary between 1st and 2nd categories

Boundary between 2nd and 3rd categories

Boundary between 3rd and 4th categories

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

Gamma distributions



larger α means *less* heterogeneity

The mean equals 1.0 for all three of these distributions

Note: the symbol α here is *not* the same as the α used in JC transition probability formulas!

Codon models

The Genetic Code

First 12 nucleotides at the 5' end of the *rbc*L gene in corn:

```
5'-ATG|TCA|CCA|CAA-3'  coding strand
3'-TAC|AGT|GGT|GTT-5'  template strand
```

} DNA double helix

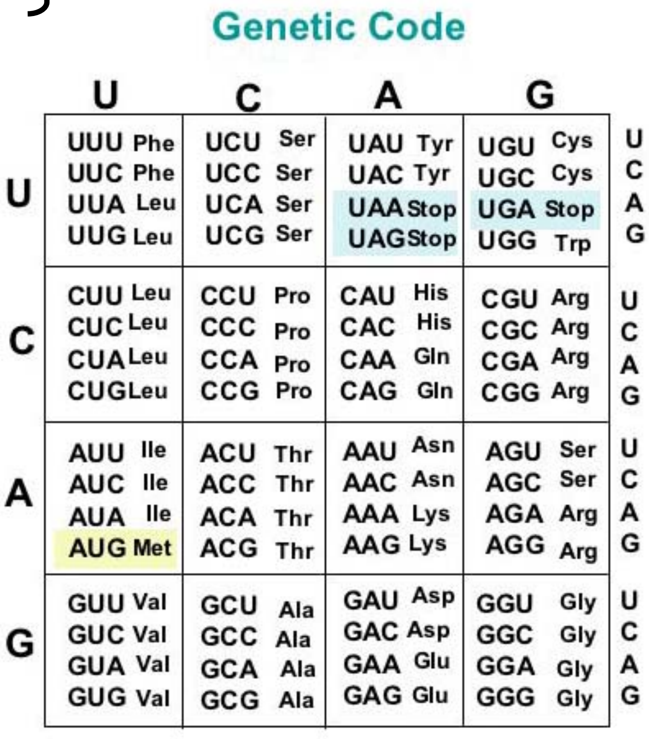
transcription

```
5'-AUG|UCA|CCA|CAA-3'  mRNA
```

translation

```
 N-Met|Ser|Pro|Gln-C  polypeptide
```

Codon models treat codons as the independent units, not individual nucleotide sites.



Genetic Code

	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

http://www.langara.bc.ca/biology/mario/Assets/Geneticode.jpg

First codon models

- Muse and Gaut model (MG94) is simplest

  α = synonymous substitution rate

  β = nonsynonymous substitution rate

  π_A, π_C, π_G, π_T = base frequencies

- Goldman and Yang model (GY94) similar
  – accounts for synon./nonsynon. *and* trs/trv bias *and* amino acid properties (later simplified, see Yang et

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715-724.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11:725-736.

Yang, Z., Nielsen, R., and Hasegawa, M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Molecular Biology and Evolution 15:1600-1611.

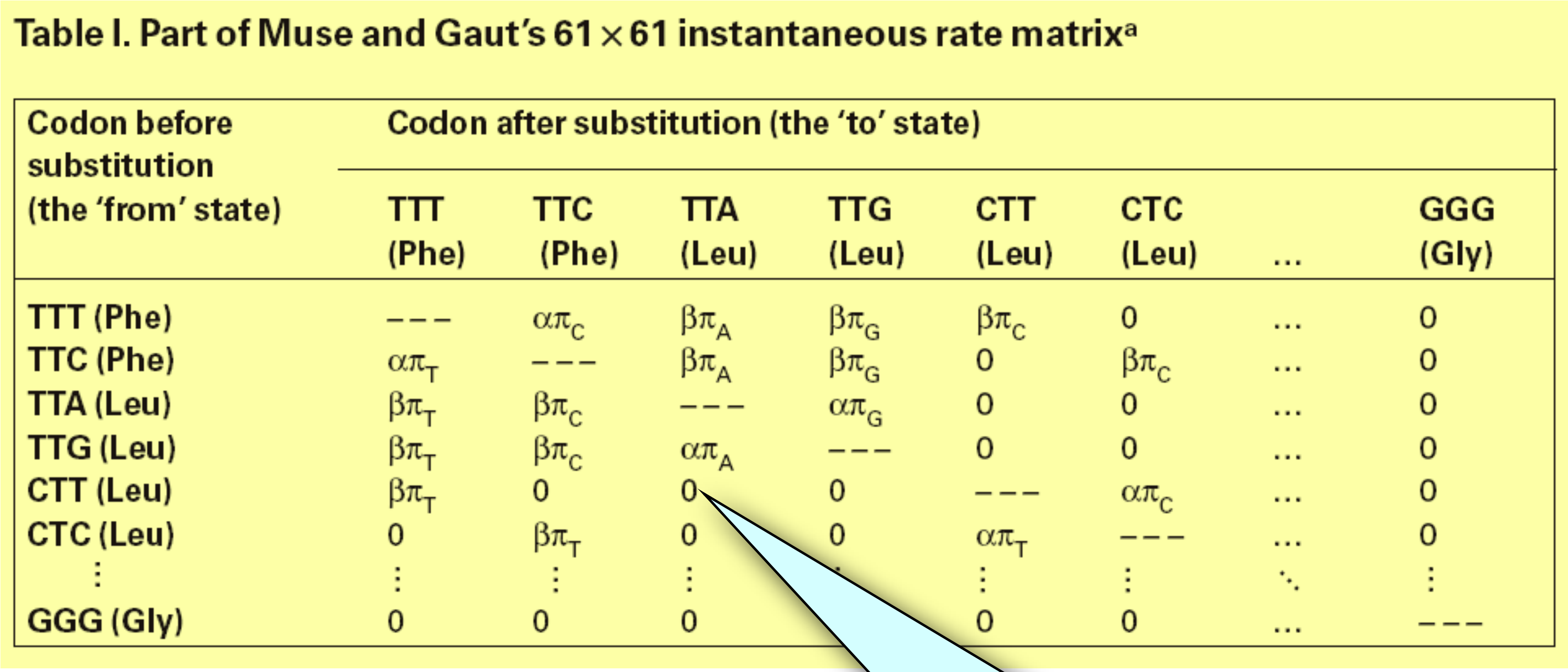


Table I. Part of Muse and Gaut's 61 × 61 instantaneous rate matrix[a]

Codon before substitution (the 'from' state)	Codon after substitution (the 'to' state)							
	TTT (Phe)	TTC (Phe)	TTA (Leu)	TTG (Leu)	CTT (Leu)	CTC (Leu)	...	GGG (Gly)
TTT (Phe)	---	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$	$\beta\pi_C$	0	...	0
TTC (Phe)	$\alpha\pi_T$	---	$\beta\pi_A$	$\beta\pi_G$	0	$\beta\pi_C$	...	0
TTA (Leu)	$\beta\pi_T$	$\beta\pi_C$	---	$\alpha\pi_G$	0	0	...	0
TTG (Leu)	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	---	0	0	...	0
CTT (Leu)	$\beta\pi_T$	0	0	0	---	$\alpha\pi_C$	...	0
CTC (Leu)	0	$\beta\pi_T$	0	0	$\alpha\pi_T$	---	...	0
⋮	⋮	⋮	⋮		⋮	⋮	⋱	⋮
GGG (Gly)	0	0	0		0	0	...	---

Instantaneous rate is 0.0 if two or more nucleotides must change during the codon transition

Note that it is still easy for the change CTT → TTA to occur, it just requires more than one instant of time

Table 1 from: Lewis, P. O. 2001. Phylogenetic systematics turns over a new leaf. Trends in Ecology and Evolution 16:30-37.

Interpreting codon model results

$\omega = \beta/\alpha$ is the nonsynonymous/synonymous rate ratio

omega	mode of selection	example(s)
$\omega < 1$	stabilizing selection (nucleotide substitutions rarely change the amino acid)	functional protein coding genes
$\omega = 1$	neutral evolution (synonymous and nonsynonymous substitutions occur at the same rate)	pseudogenes
$\omega > 1$	positive selection (nucleotide substitutions often change the amino acid)	envelope proteins in viruses under active positive selection