Phylogenetic likelihood and models

Emily Jane McTavish

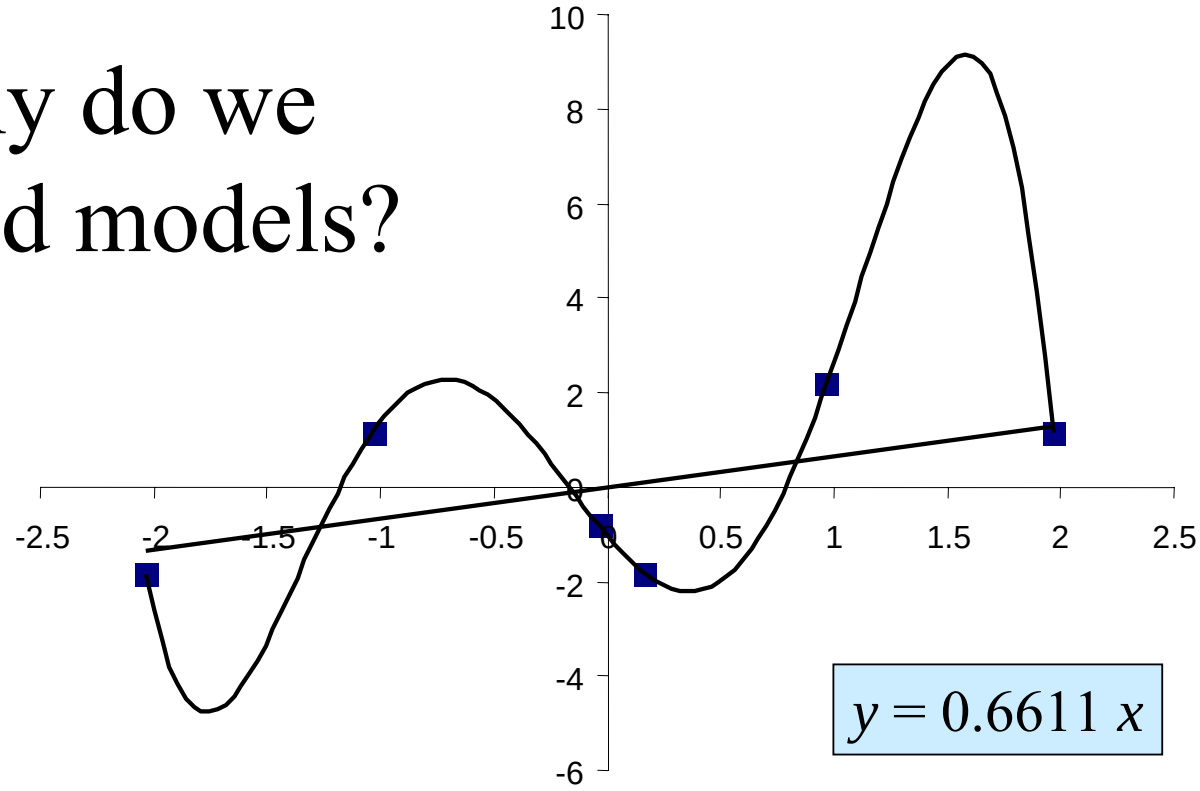
Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Mark Holder and Paul Lewis for slides)

Substitution Models

$$y = -1.5972\,x^5 + 23.167\,x^4 - 126.18\,x^3 + 319.17\,x^2 - 369.22\,x + 155.67$$

Why do we need models?



$$y = 0.6611\,x$$

Models

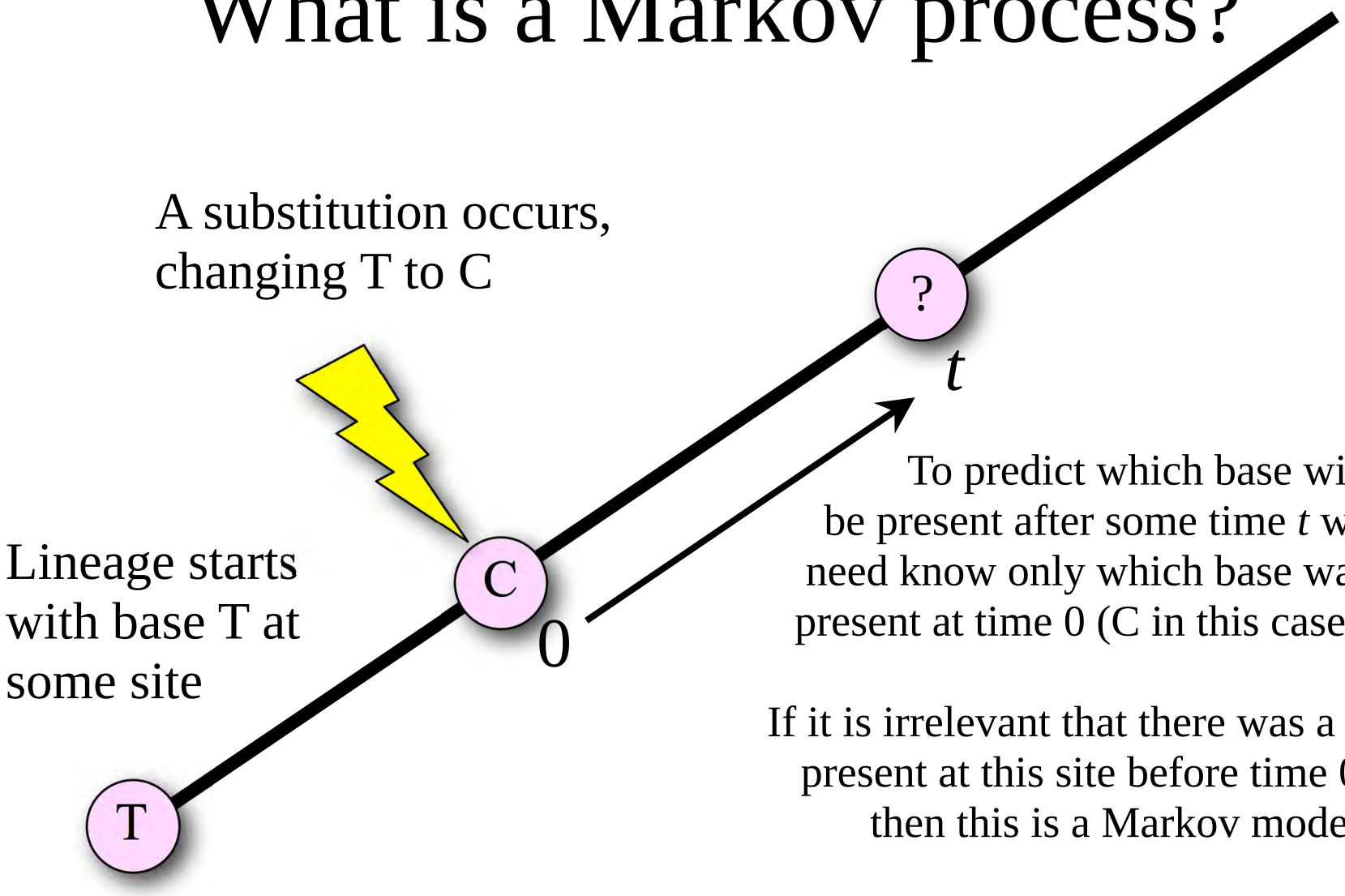
- Models help us intelligently **interpolate between our observations** for purposes of **making predictions**
- **Adding parameters** to a model generally increases its fit to the data
- **Underparameterized** models lead to poor fit to observed data points
- **Overparameterized** models lead to poor prediction of future observations
- Criteria for choosing models include likelihood ratio tests, AIC, BIC, Bayes Factors, etc.
  - all provide a way to choose a model that is neither underparameterized nor overparameterized

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 ($r_{A \leftrightarrow C} = r_{A \leftrightarrow G} = r_{A \leftrightarrow T} = r_{C \leftrightarrow G} = r_{C \leftrightarrow T} = r_{G \leftrightarrow T} = \alpha$)
- Usually described as a **1-parameter** model (the parameter being the edge length)
  - Remember, however, that each edge 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.

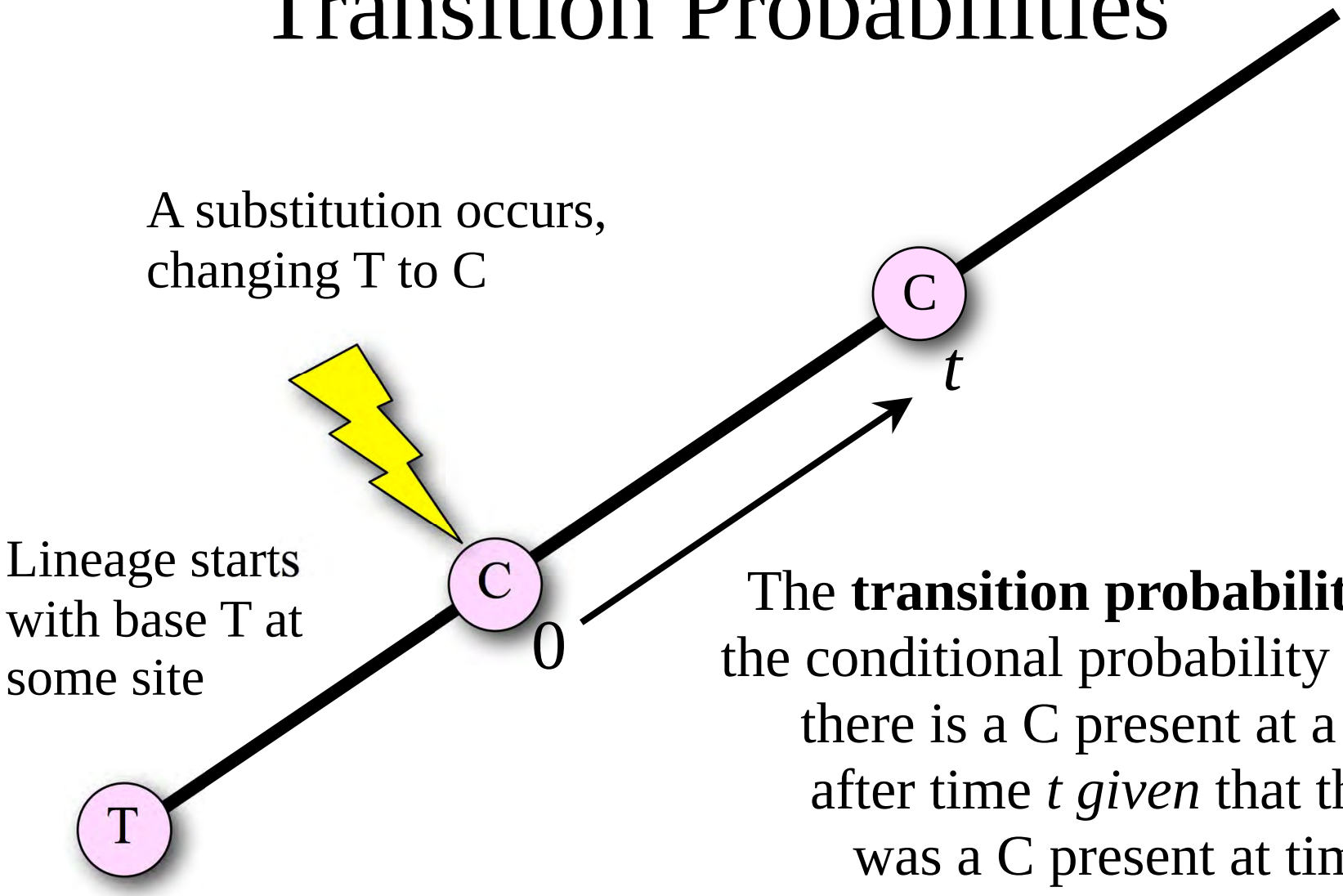
What is a Markov process?



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



The **transition probability** is the conditional probability that there is a C present at a site after time *t given* that there was a C present at time 0

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:

$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\alpha t}\right) = \Pr(G|T, \alpha t)$$

The JC69 model has only one unknown quantity: αt

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

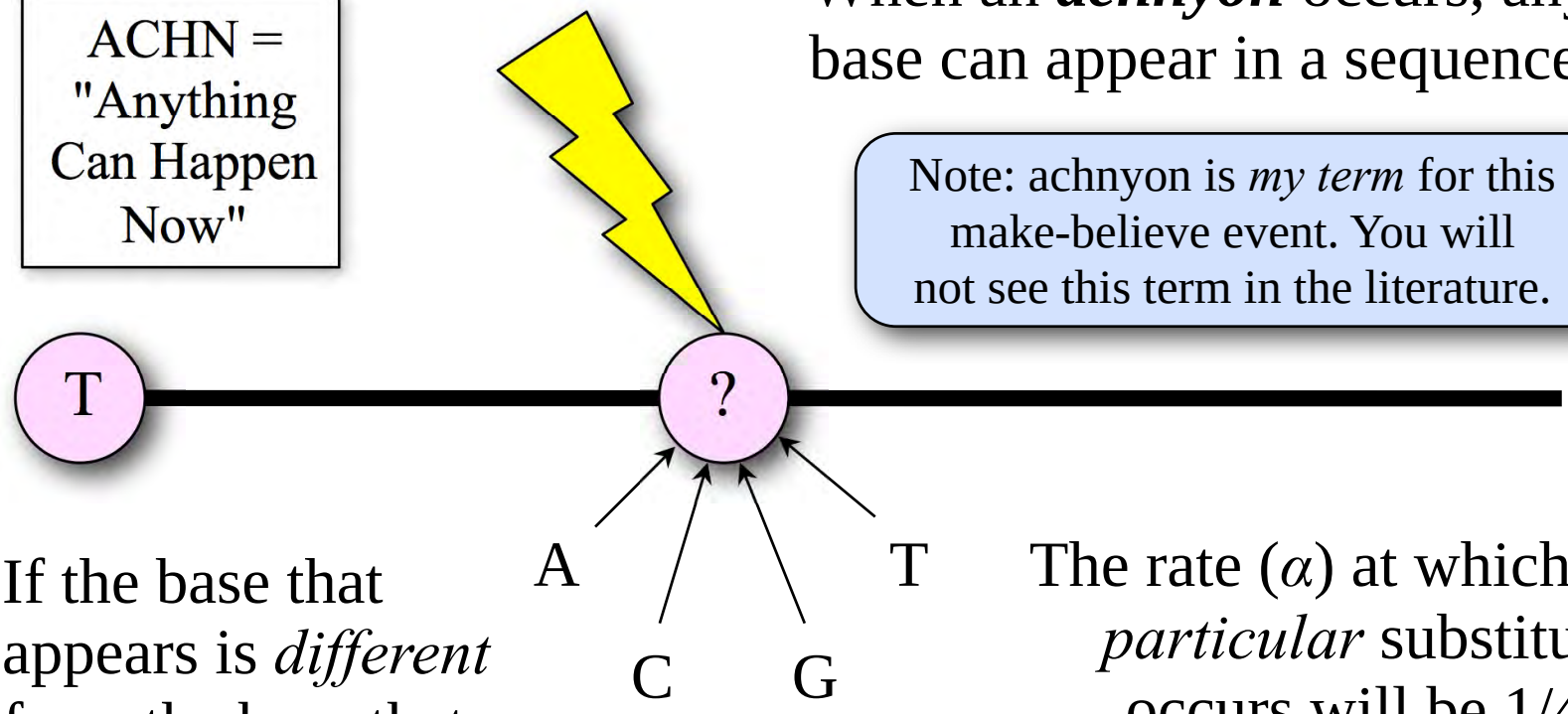
Where does a transition probability formula such as this come from?

"ACHNyons" vs. substitutions

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

ACHN = "Anything Can Happen Now"

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 $\mu = 4\alpha$)

The Poisson distribution

Probability distribution on the number of events when:

1. events are assumed to be independent,
2. the *rate* of events some constant, μ, and
3. the process continues for some duration of time, t.

The expectation of the number of events is $\nu = \mu t$.

Note that ν can be any non-negative number, but the Poisson is a discrete distribution – it gives the probabilities of the number of events (and this number will always be a non-negative integer).

Poisson distibution can be used to explain statistical regularities of rare events

$P\left(\text{k events in interval}\right) = \frac{e^{-\nu}\nu^{k}}{k!}$

- ν is the average number of events per interval (rate times time)
- e is the number 2.71828... (Euler's number) the base of the natural logarithms
- k takes values 0, 1, 2, ...
- $k! = k * (k-1) * (k-2) * \ldots * 2 * 1$ is the factorial of k.

from wikipedia

$$P\left(\text{k events in interval}\right) = \frac{e^{-\nu}\nu^{k}}{k!}$$
$$P\left(0 \text{ events}\right) = \frac{e^{-\nu}\nu^{0}}{0!} = e^{-\nu} = e^{-\mu t}$$
$$P\left(>= 1 \text{ events}\right) = 1 - e^{-\mu t}$$

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) = $1 - e^{-\mu t}$

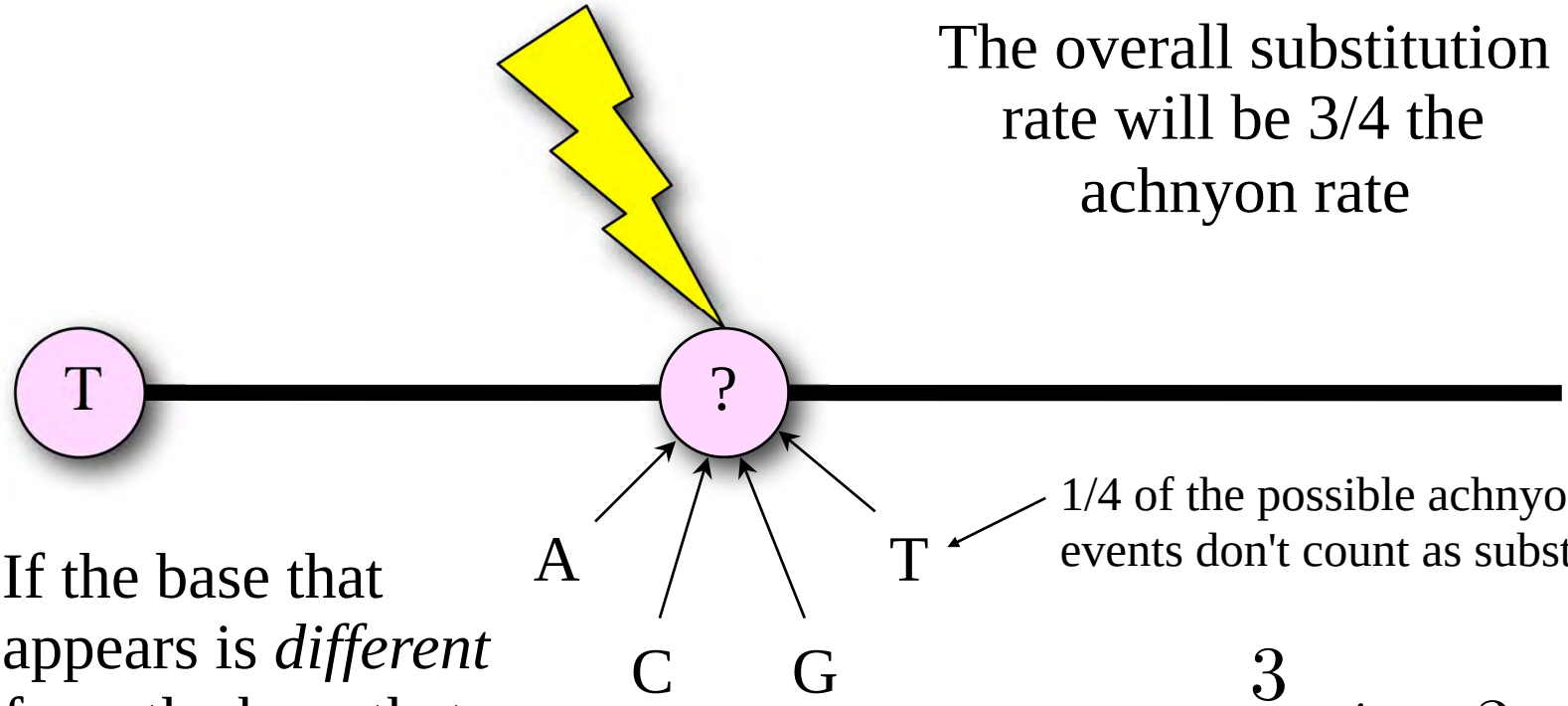
Pr(last achnyon results in base G) = $\frac{1}{4}$

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):

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

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

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

$$1 - e^{-4\alpha t}$$

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

Doh! Something must be wrong here...

Transition Probabilities: Remarks

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

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

Equilibrium frequencies

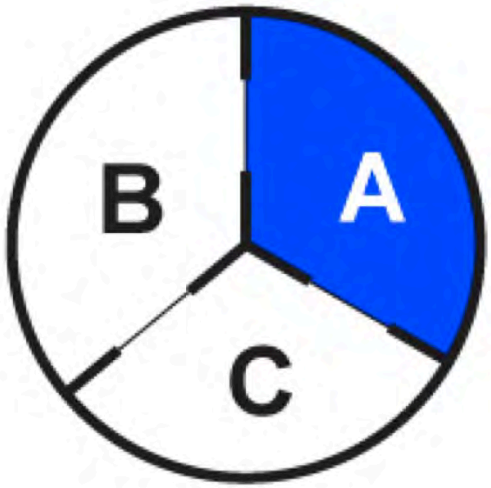
- The JC69 model assumes that the frequencies of the four bases (A, C, G, T) are equal
- The equilibrium relative frequency of each base is thus 0.25
- Why are they called *equilibrium* frequencies?

Equilibrium Frequencies

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

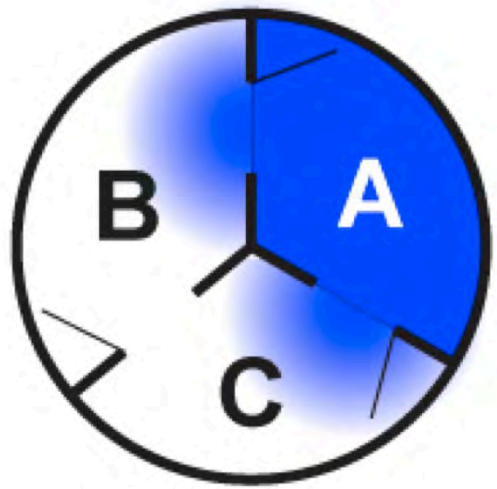
The doors to the other rooms are closed, so the perfume has, thus far, not been able to spread.

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.

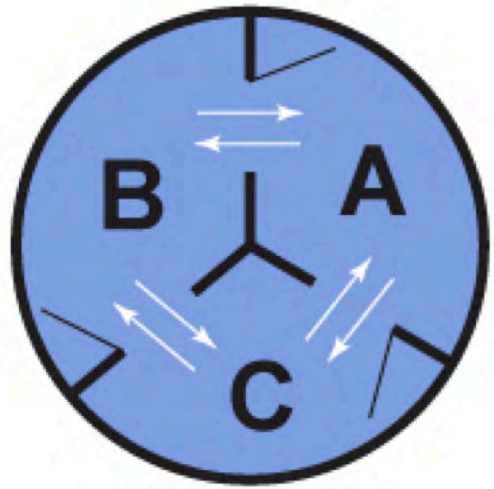


Molecules of perfume go both ways through open doors, but more pass one way than another, leading to a net flow from room A to rooms B and C.

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

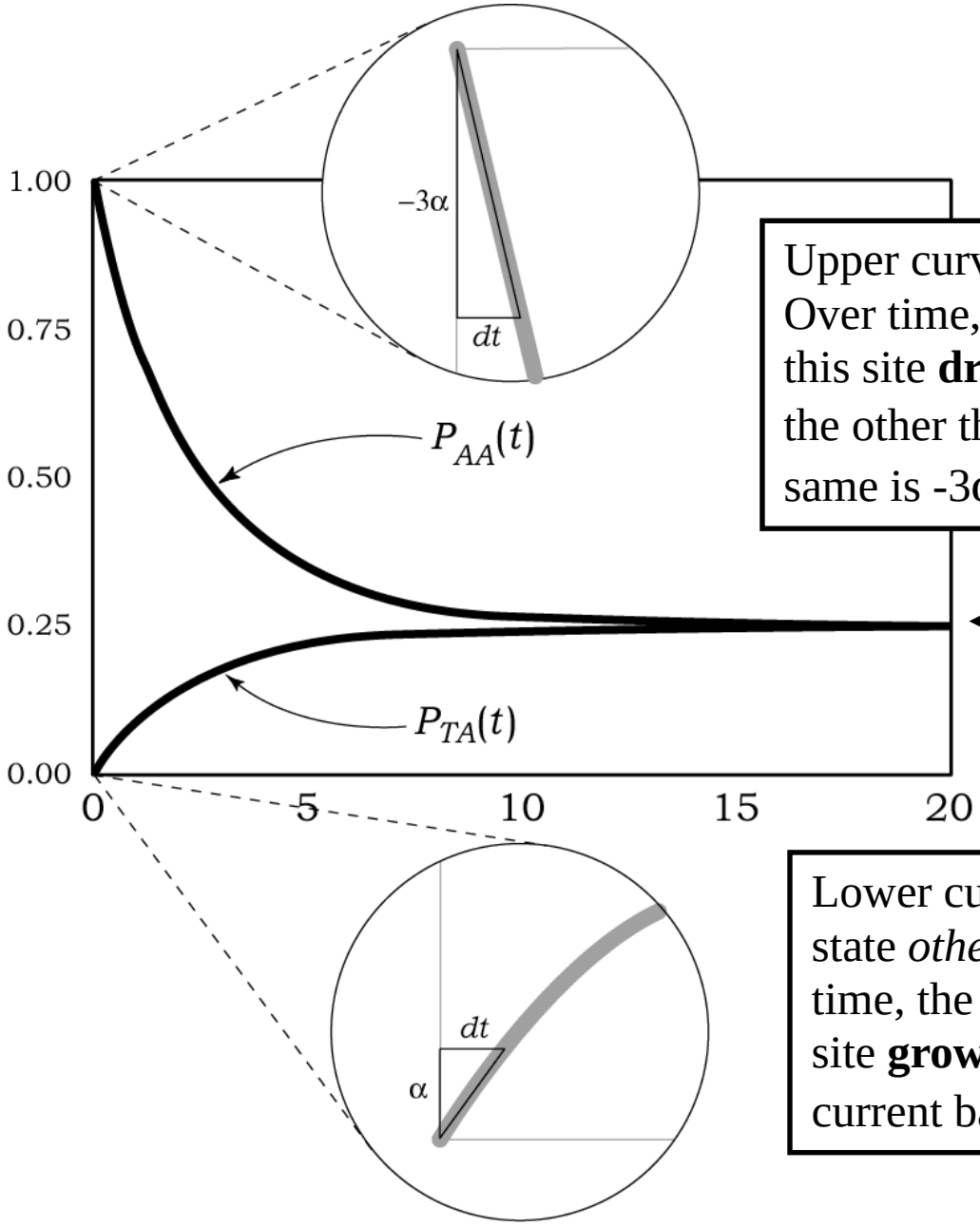
Eventually, all 3 rooms have essentially the same concentration of perfume.



Molecules still move through open 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

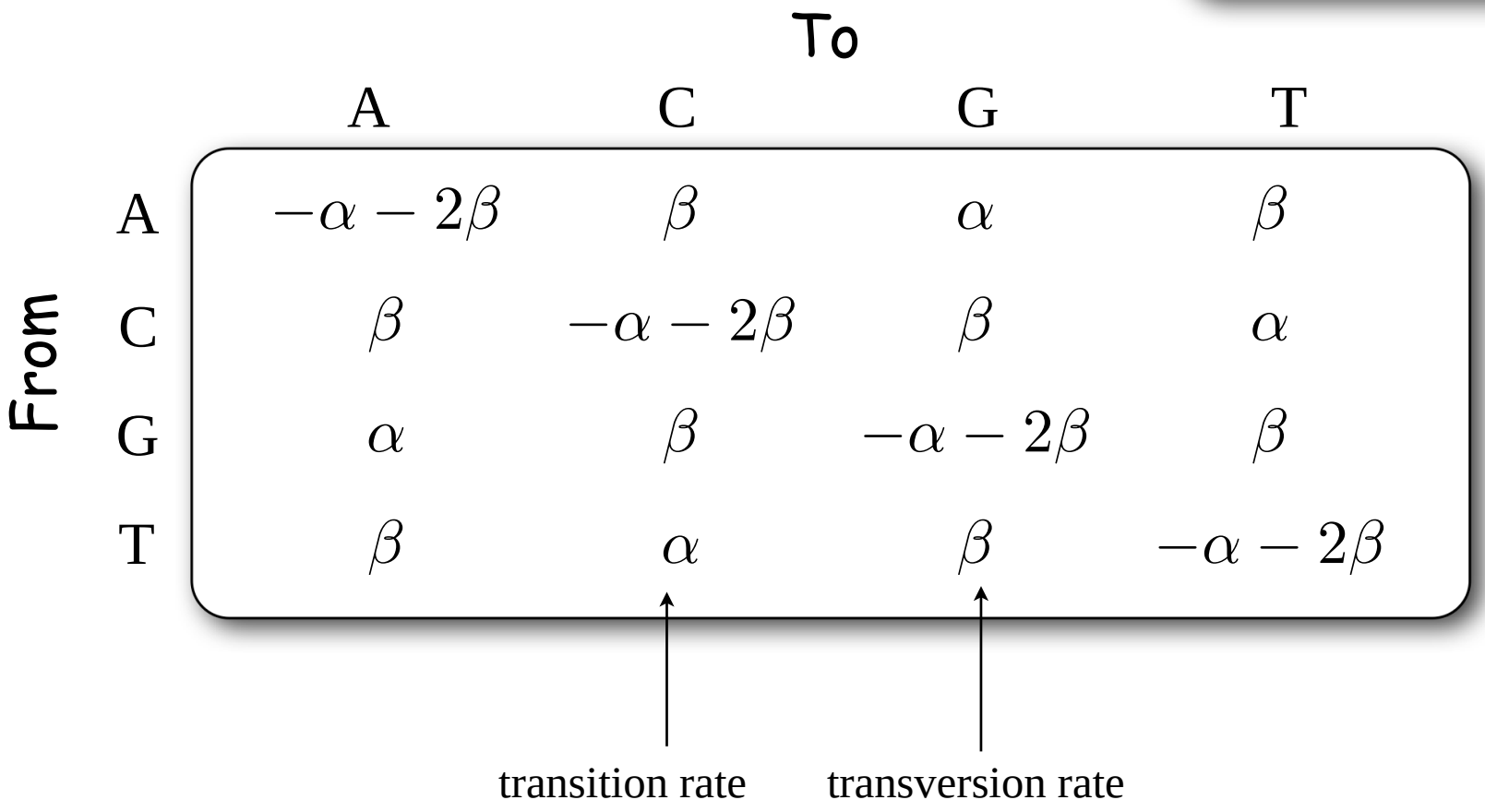
1 parameter: α

From \ To	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

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



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

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

All I've done is re-parameterize the rate matrix, letting κ equal the *transition/transversion rate ratio*

$$\kappa = \frac{\alpha}{\beta}$$

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

Transition/transversion ratio (tratio) versus Transition/transversion *rate* ratio (kappa)

Cobbler analogy:

- 4 cobblers in a factory make loafers
- 8 cobblers in the factory make work boots
- all cobblers produce the same number of shoes per unit time, regardless of shoe type
- what is the loafer/boot *rate ratio* and how does that compare to the loafer/boot *ratio*?

The loafer/boot *rate ratio* is 1.0 because each cobbler cranks out shoes at the same rate.

The loafer/boot *ratio*, however, is 0.5 because there are twice as many cobblers making boots as there are cobblers making loafers.

There are 8 possible transversion-type substitutions and only 4 possible transition-type substitutions: the transition/transversion ratio is thus 0.5 when the transition/transversion rate ratio is 1.

Image from http://www.gaseenterprise.com/07130%20Shoemaker.jpg

F81 rate matrix

4 parameters:
μ
π_A
π_C
π_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	$\pi_A\mu$	$\pi_C\mu$	$\pi_G\mu$	$-\mu(1-\pi_T)$

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

	A	C	G	T
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	$\pi_A\beta$	$\pi_C\beta\kappa$	$\pi_G\beta$	—

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

	A	C	G	T
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	$\pi_A c\mu$	$\pi_C e\mu$	$\pi_G f\mu$	$-$

9 parameters:

π_A
π_C
π_G
a
b
c
d
e
μ

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
```

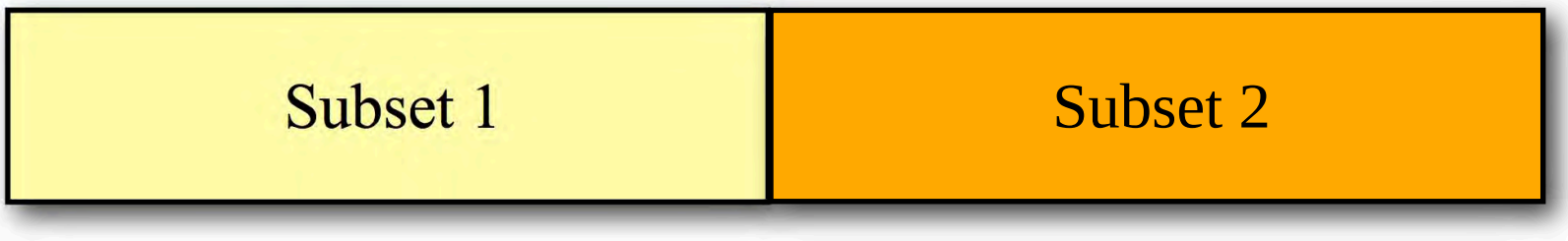
```
Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAACTAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA
.....A..T........A........G..T..G........A........A..A........T.....G.....A.........T..T...........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.....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.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



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

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

Relative rates have mean 1:

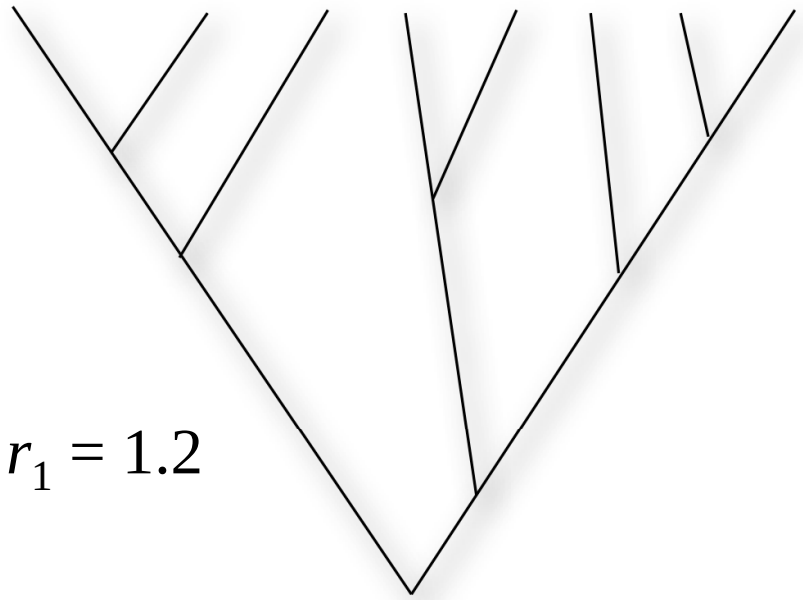
$$\frac{r_1 + r_2}{2} = 1$$

More generally:

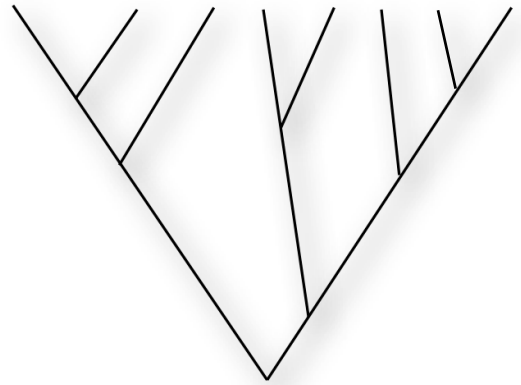
$$r_1\ p(r_1) + r_2\ p(r_2) = 1$$

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

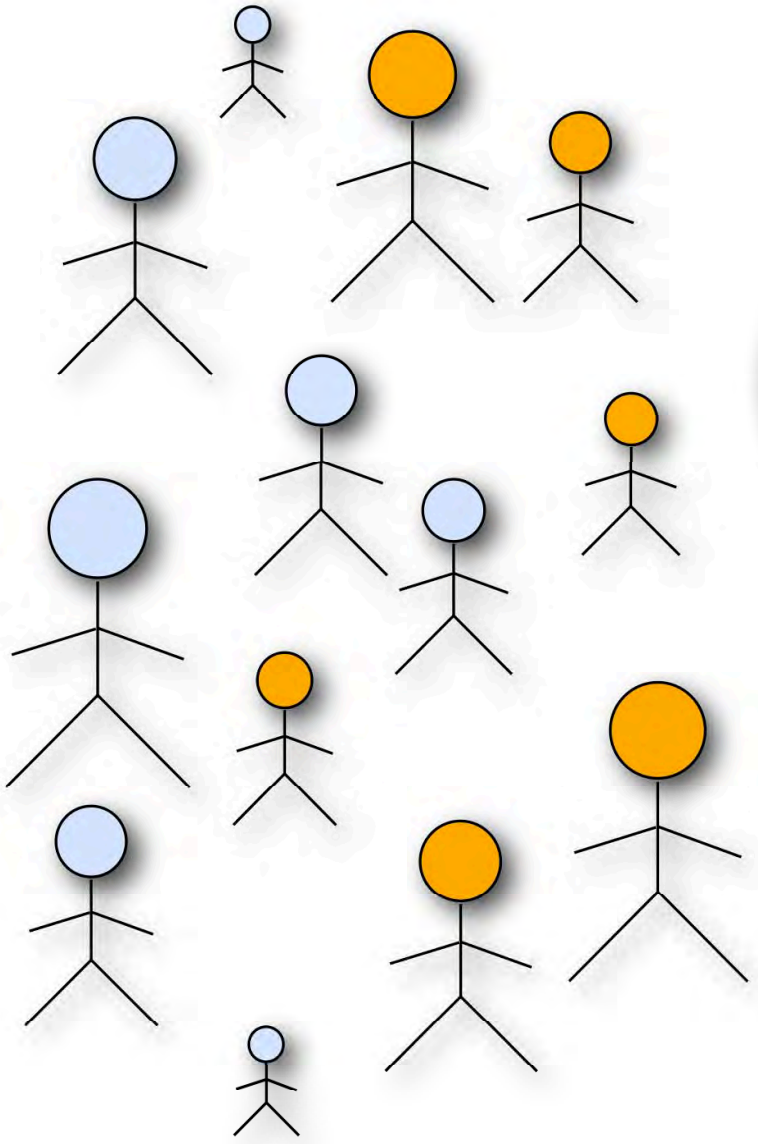
JC69 transition probabilities that would be used for sites in **gene 1**:

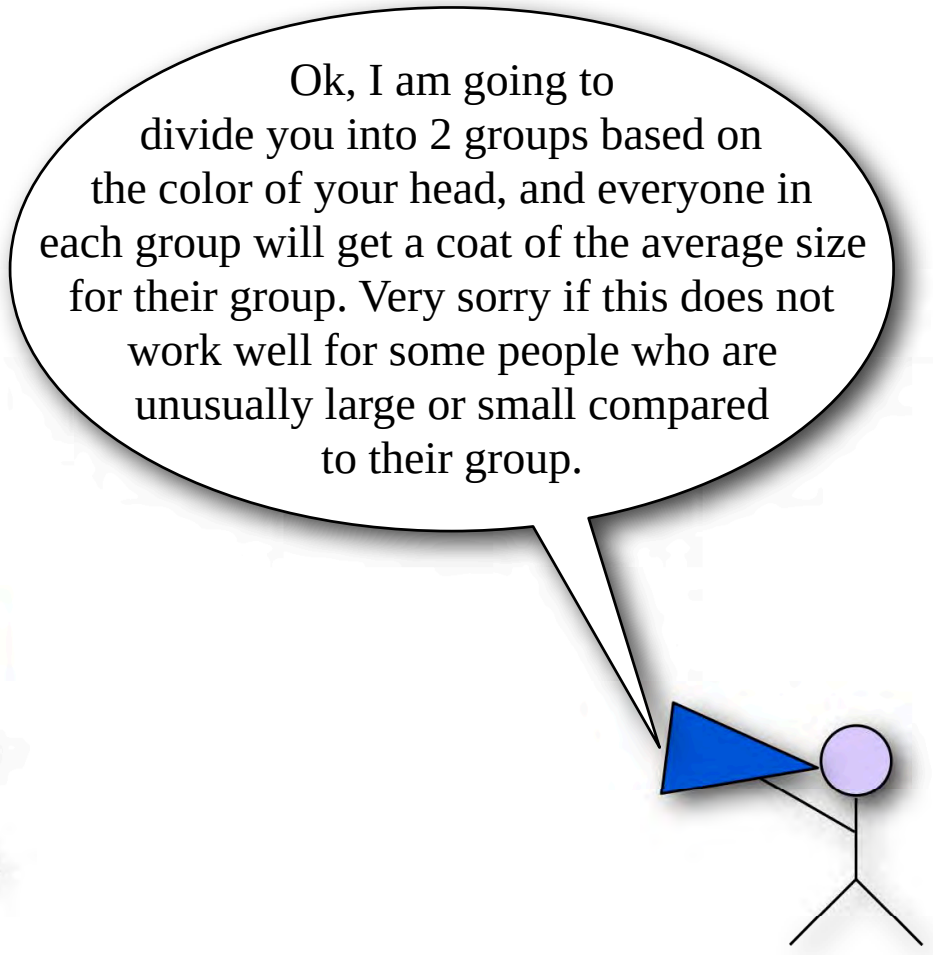
$$
\begin{aligned}
P_{ii}(t) &= \frac{1}{4} + \frac{3}{4}e^{-4r_1\alpha t} \\
P_{ij}(t) &= \frac{1}{4} - \frac{1}{4}e^{-4r_1\alpha t}
\end{aligned}
$$

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

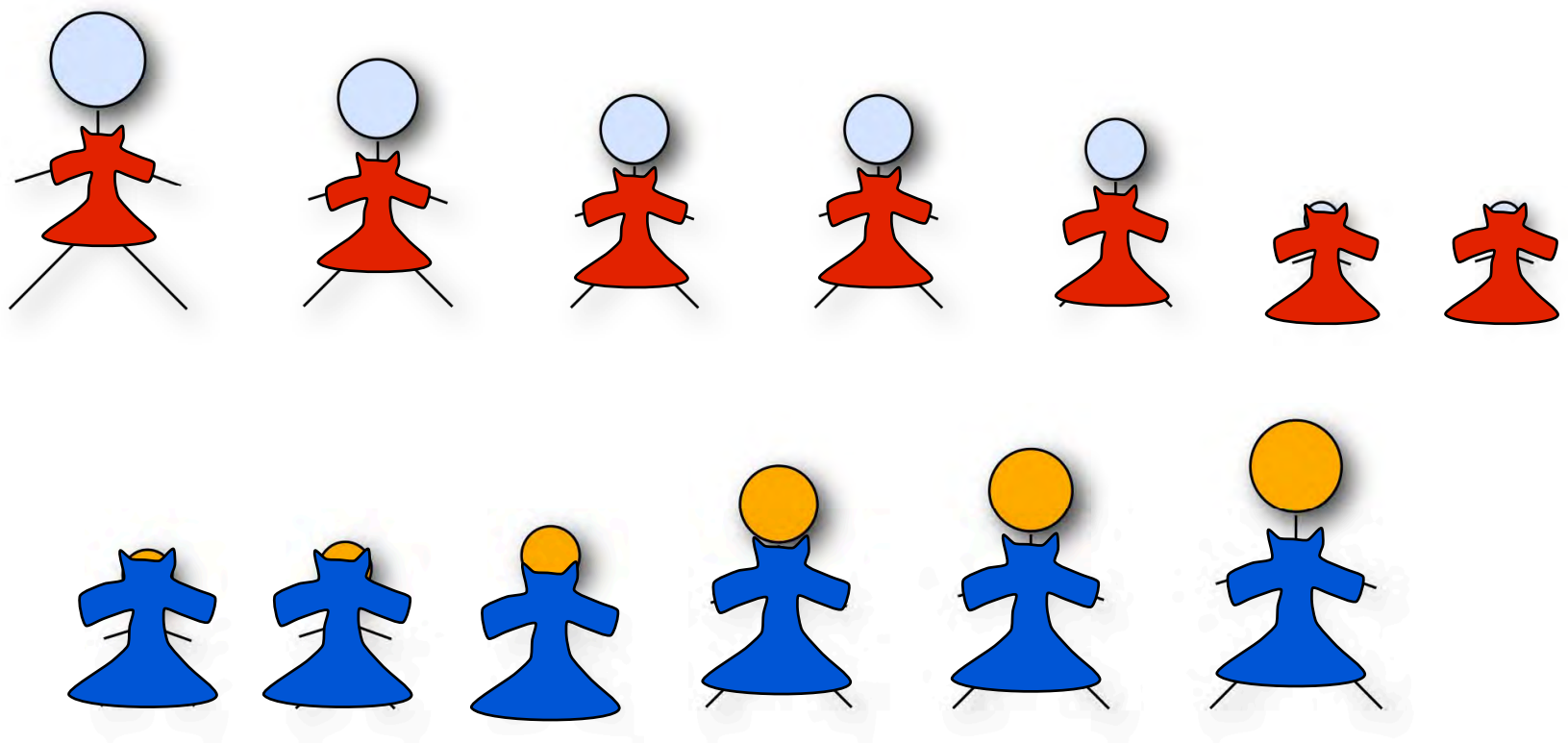
$$
\begin{aligned}
P_{ii}(t) &= \frac{1}{4} + \frac{3}{4}e^{-4r_2\alpha t} \\
P_{ij}(t) &= \frac{1}{4} - \frac{1}{4}e^{-4r_2\alpha t}
\end{aligned}
$$

Site-specific Approach





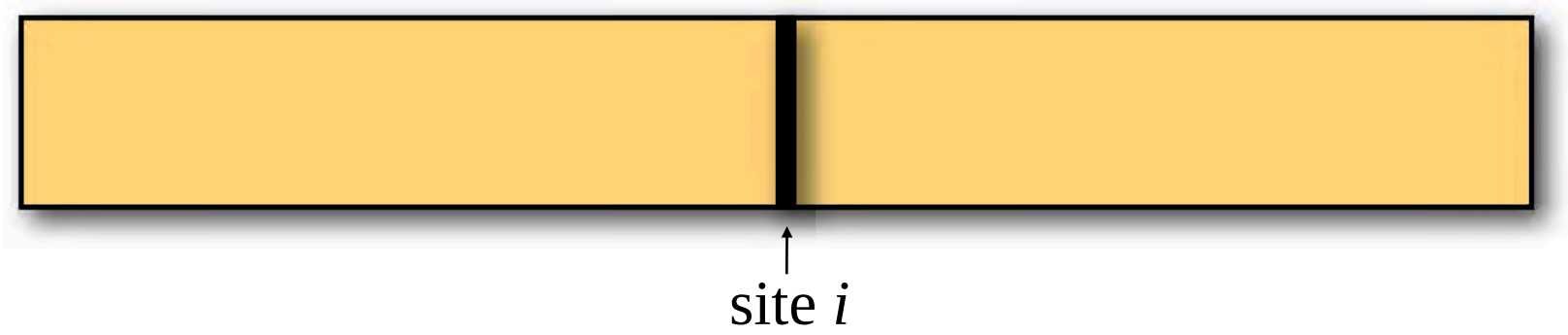
Site-specific Approach



Good: costs less: need to buy just one coat for every person
Bad: 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

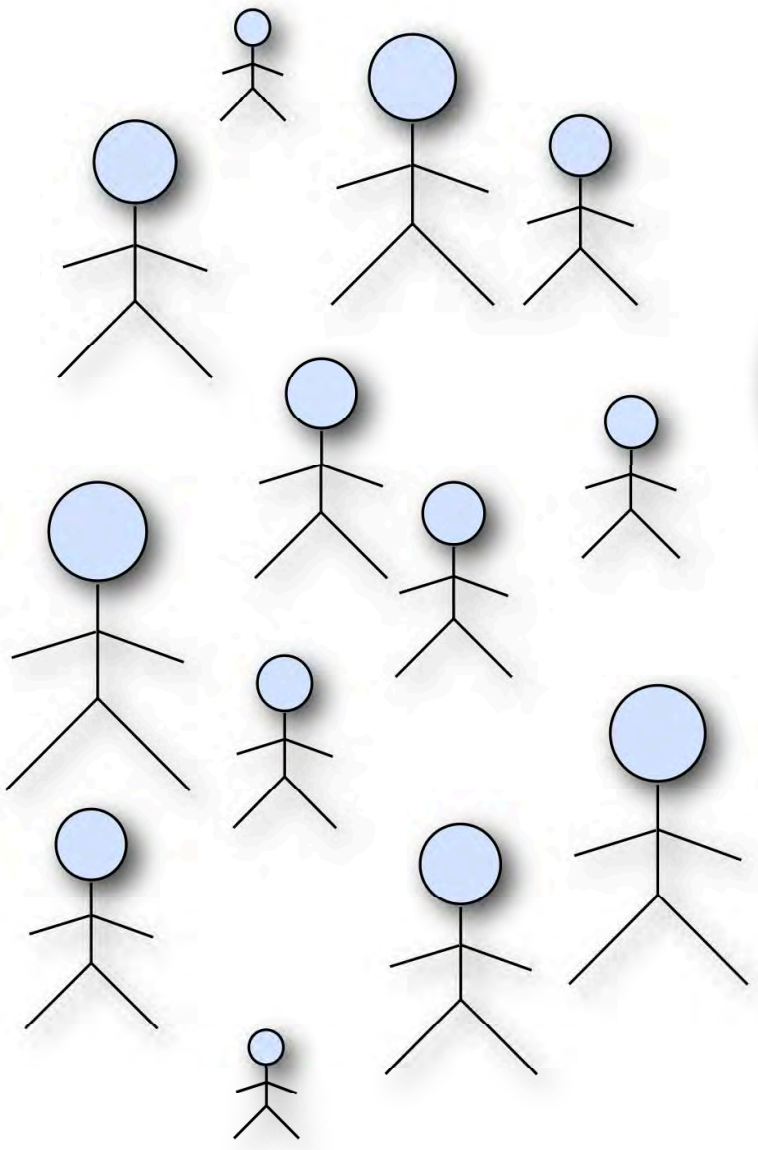
All relative rates applied to every site

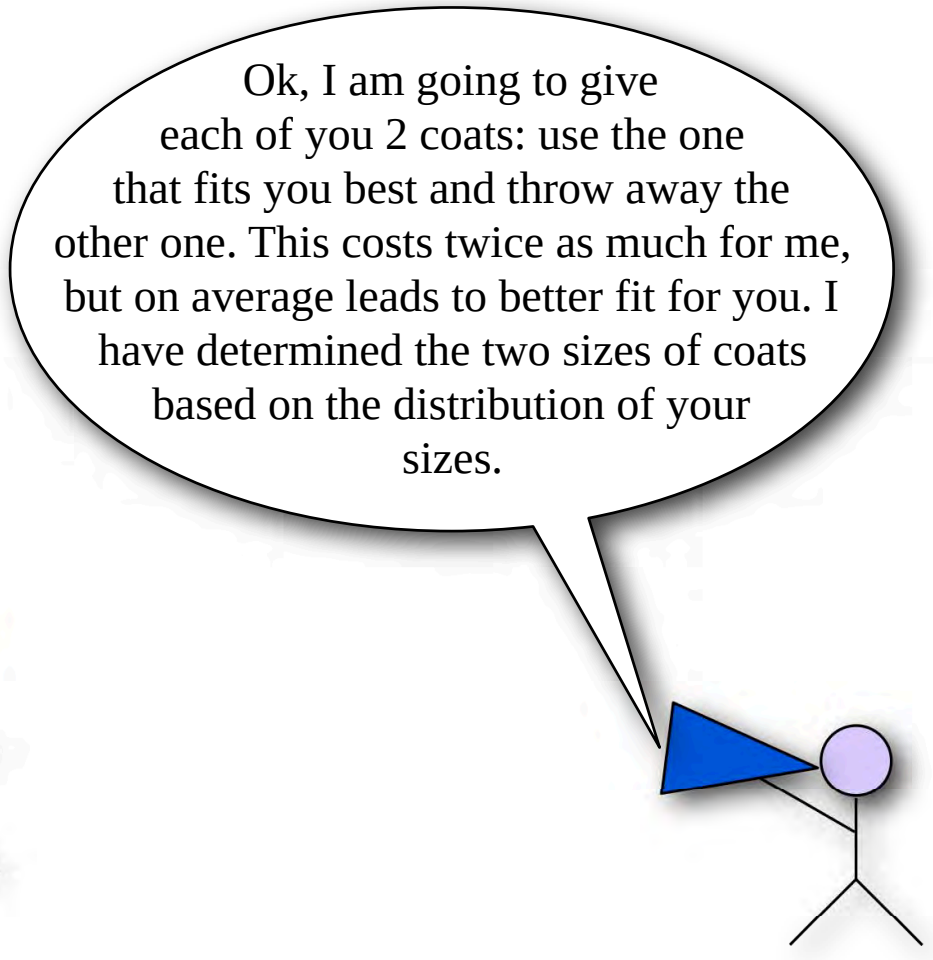


$$L_i = \Pr(D_i|r_1)\Pr(r_1) + \Pr(D_i|r_2)\Pr(r_2)$$

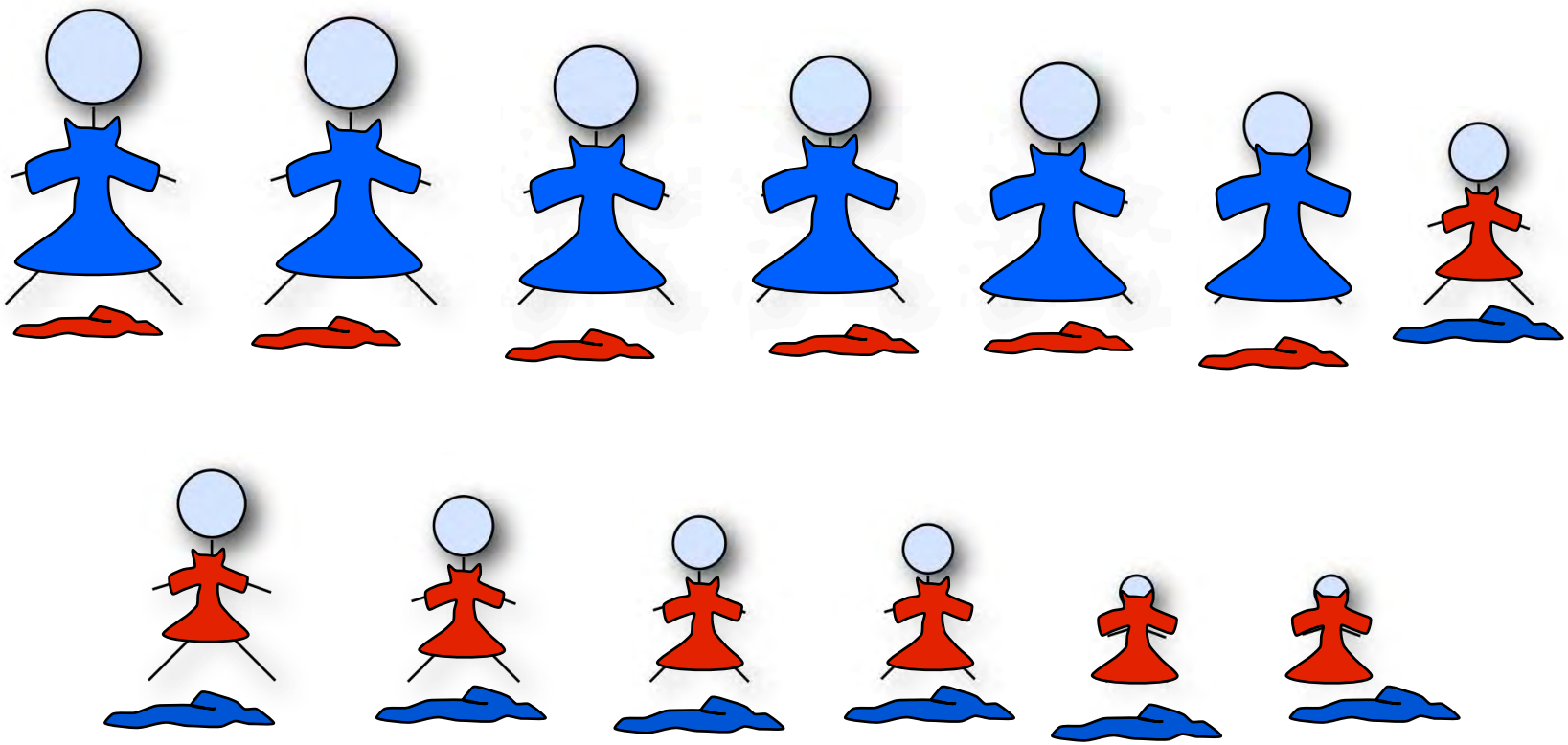
Common examples
- Invariable sites (I) model
- Discrete Gamma (G) model

Mixture Model Approach





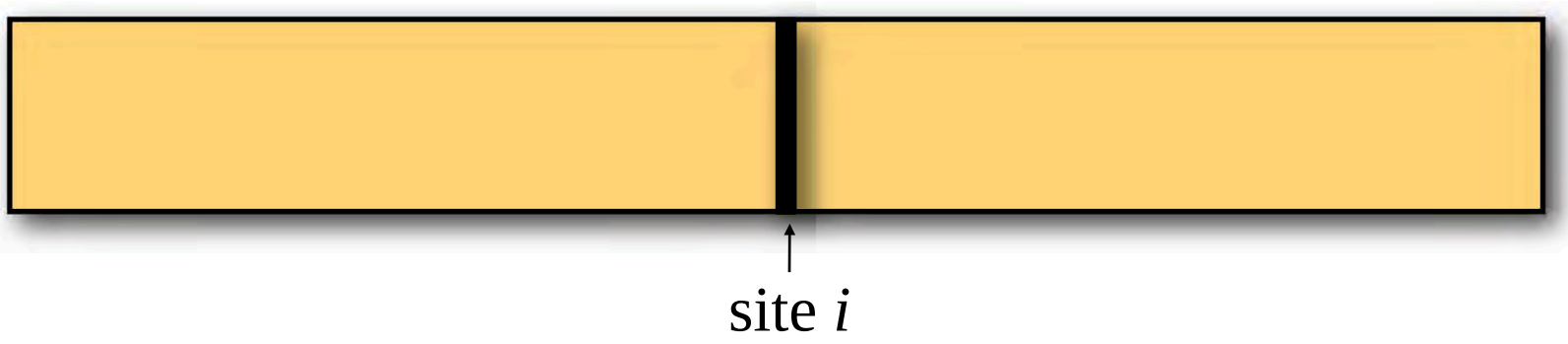
Mixture Model Approach



Good: every person experiences better fit because they can choose the size coat that fits best
Bad: 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)



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

$$r_1 = 0.0$$

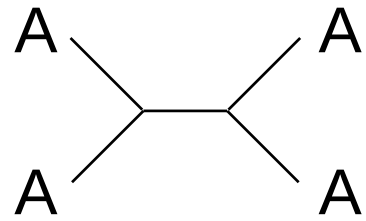
$$r_2 = \frac{1}{1 - p_{\text{invar}}}$$

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.

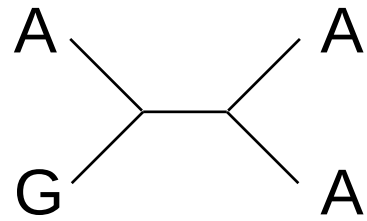
Invariable sites model

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



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

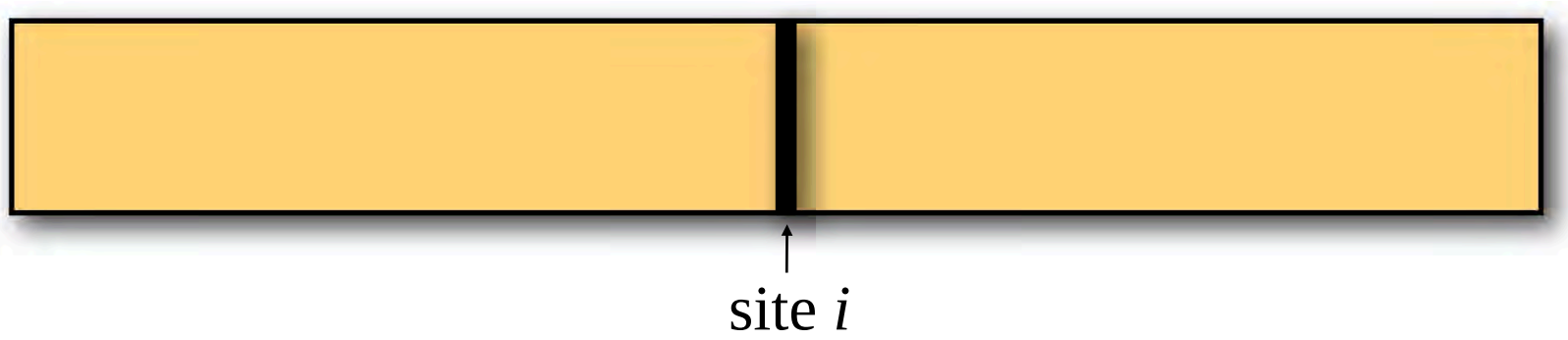
If site *i* is a *variable* site, there is no way to explain the data with a zero rate, so the first term is zero:



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

Discrete Gamma Model

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



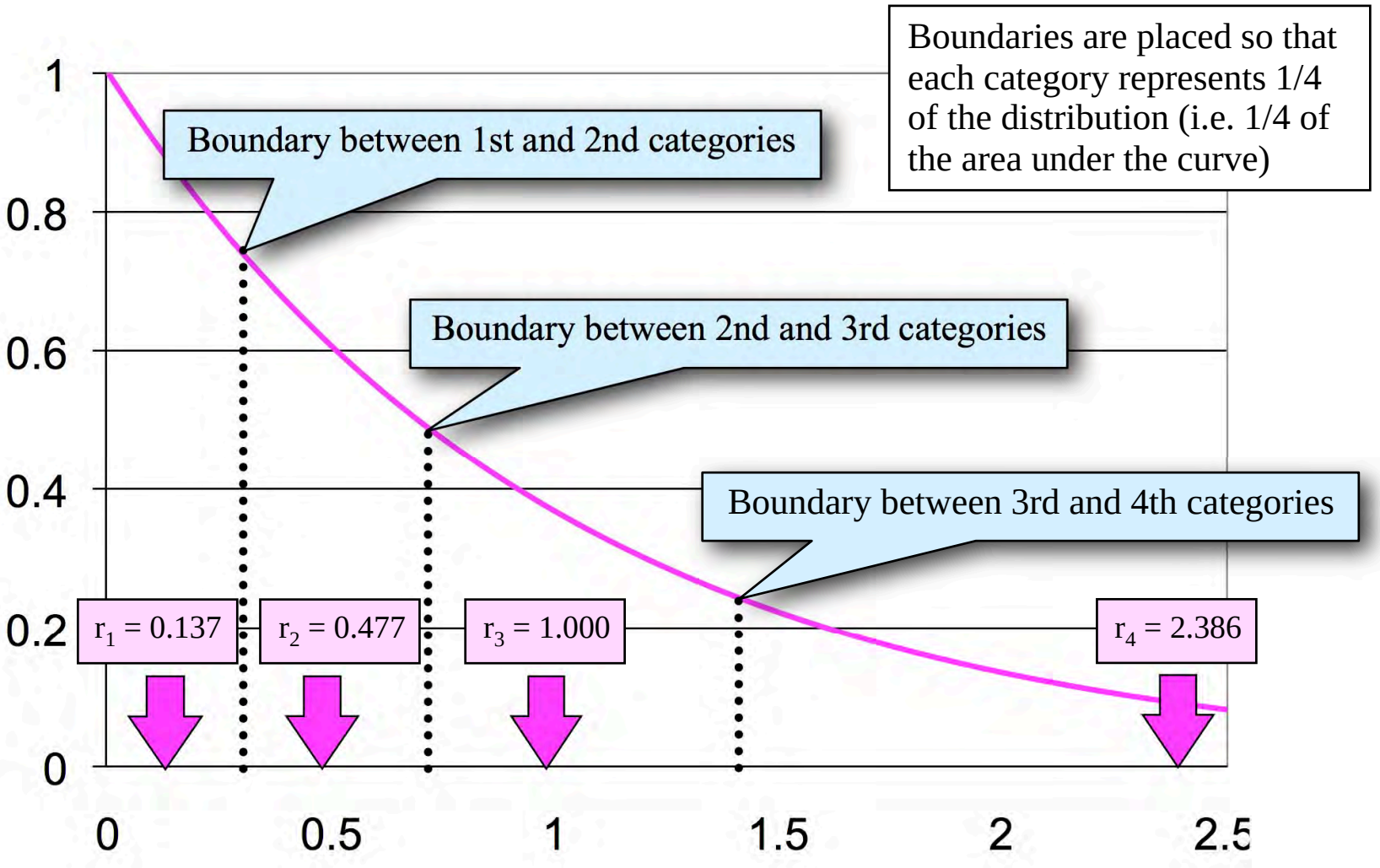
$$L = (\tfrac{1}{4})\Pr(D_i|r_1) + (\tfrac{1}{4})\Pr(D_i|r_2) + (\tfrac{1}{4})\Pr(D_i|r_3) + (\tfrac{1}{4})\Pr(D_i|r_4)$$

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

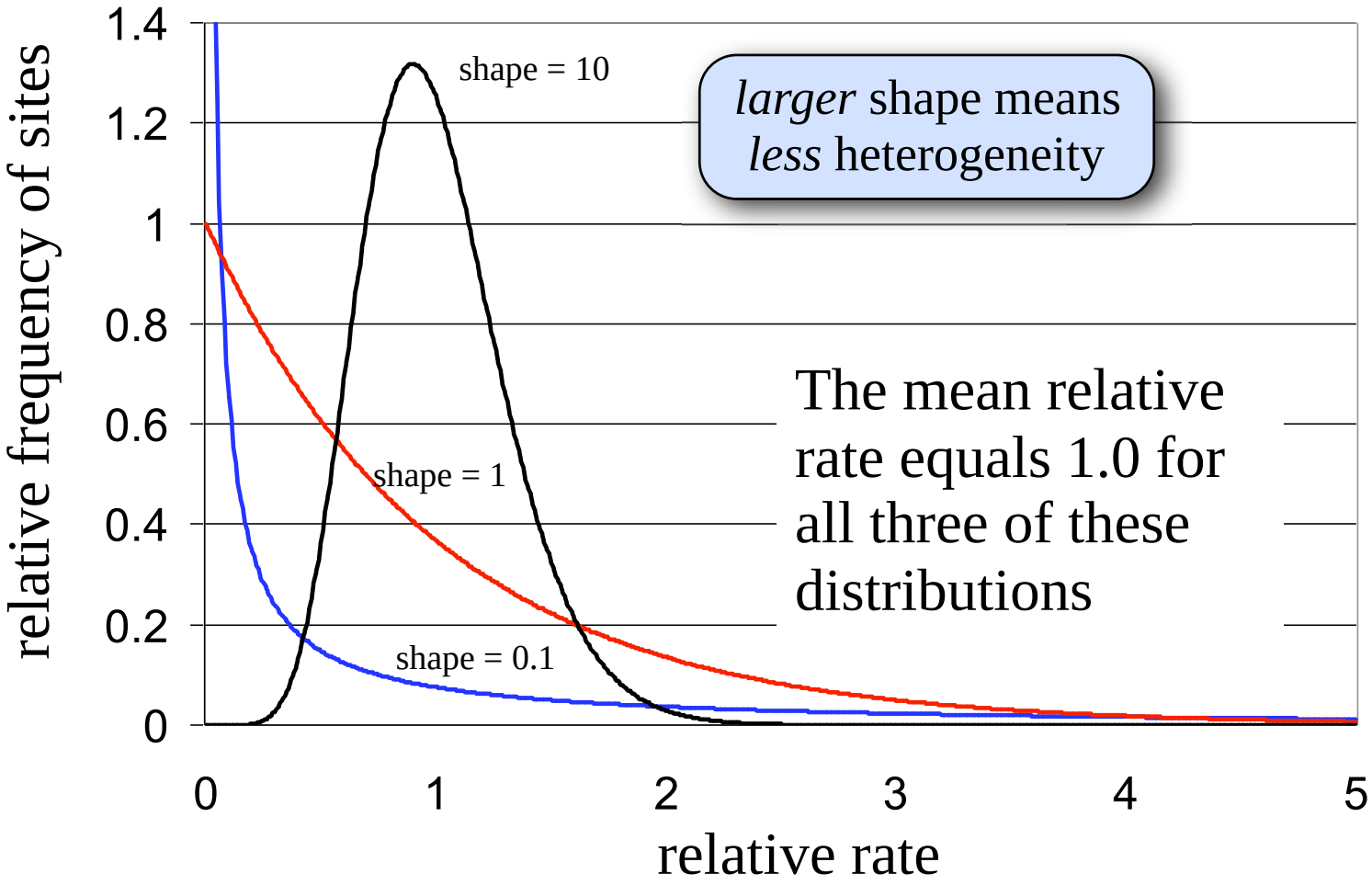
Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396-1401.

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



Gamma distributions



lqtree tutorial