assignment snippet: how to start with the root

```
def rootgen(rootnode, basefreq, numsites):
        """
        this function generates numsites nucleotide
        given a set of probabilities
        """
        r = [[0.0,'']]*4
        b = ['A','C','G','T']

        r[0] = [probs[0],b[0]]
        for i in range(len(r)-1):
            r[i+1] = [r[i][0]+basefreq[i+1],b[i+1]]

        for i in range(numsites):
            ran = random.uniform(0.,1.0)
            for ri in r:
                if (ran < ri[0]):
                    rootnode.sequence.append(ri[1])
                    break
```

Jukes-Cantor (JC) allows for a single parameter and has a transition matrix

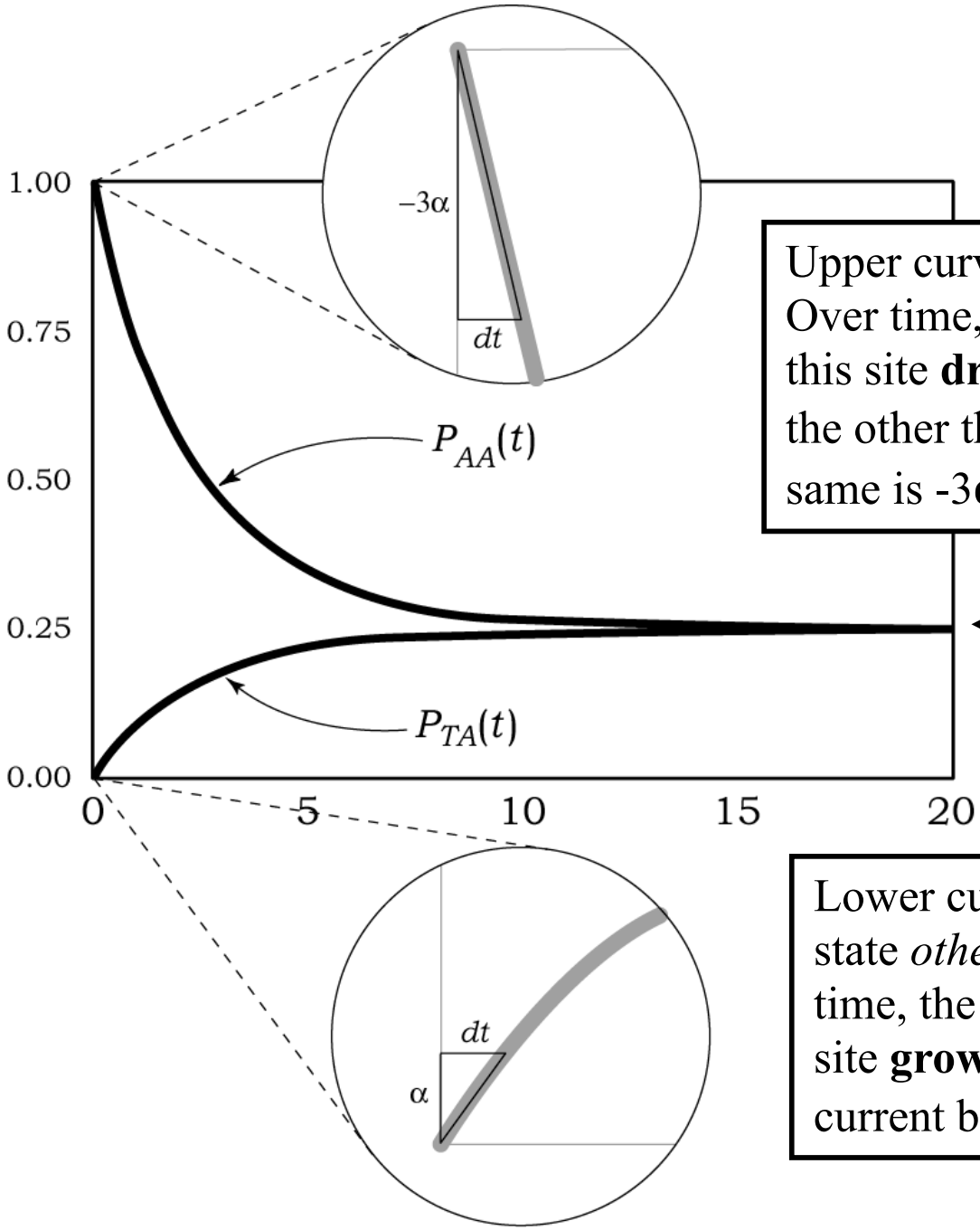
$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

The base frequencies $\pi_A, \pi_C, \pi_G, \pi_T$ are all the same and 0.25. There are only two types of changes possible, either one does not change or one changes. This results in two probabilities:

$$\text{Prob}(t)_{ii} = \frac{1}{4} + \frac{3}{4}e^{-\mu t} \tag{19}$$

$$\text{Prob}(t)_{ij} = \frac{1}{4} - \frac{1}{4}e^{-\mu t} \tag{20}$$

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).

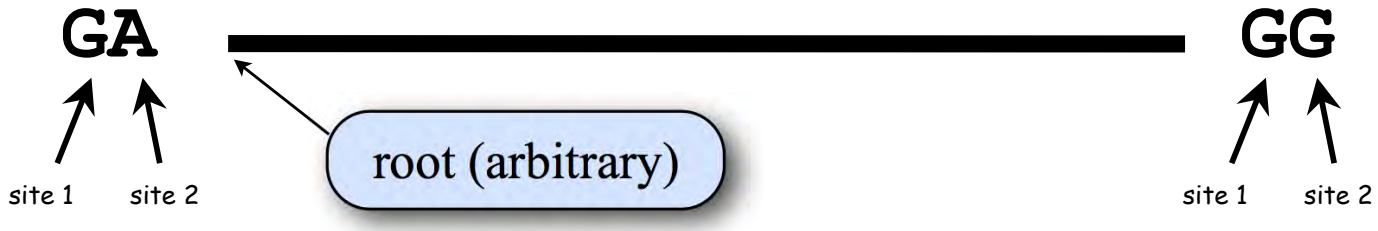
The equilibrium relative frequency of A is 0.25

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.

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, αt) Pr(A) Pr(G|A, αt)

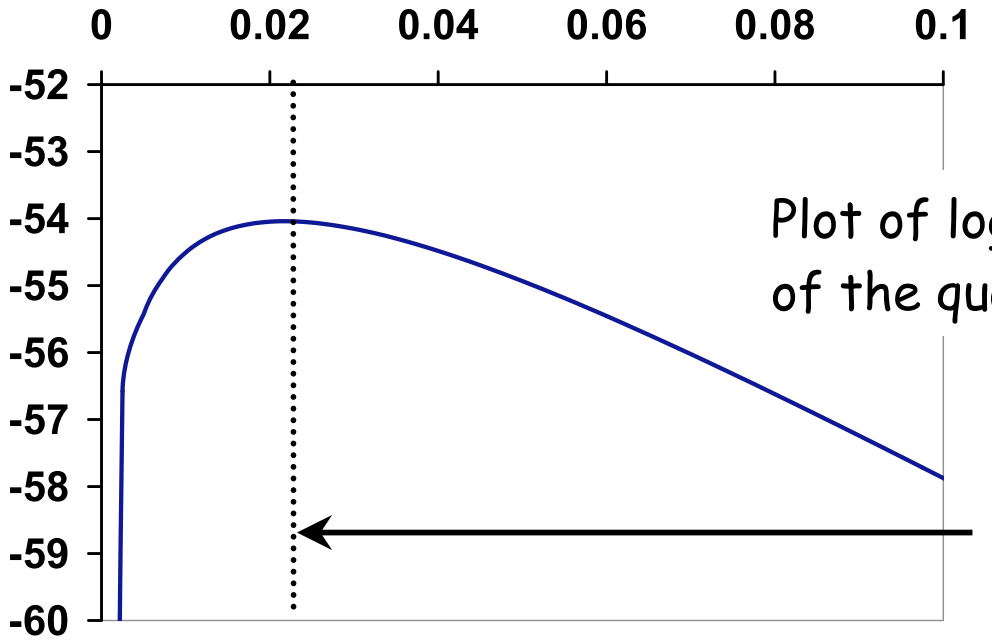
Note that we are NOT assuming independence here

Maximum likelihood estimation

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

```
gorilla    GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan  GGACTCCTTGAGAAATAAACTGCACACACTGG
```

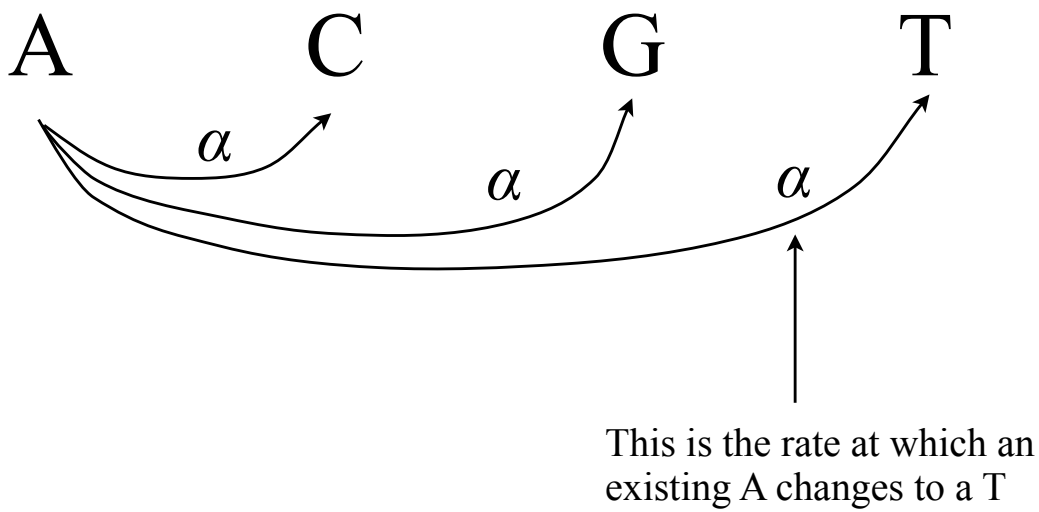
$$L = \left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}+\tfrac{3}{4}e^{-4\alpha t}\right)\right]^{30}\left[\left(\tfrac{1}{4}\right)\ \left(\tfrac{1}{4}-\tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$



Plot of log-likelihood as a function of the quantity αt

Maximum likelihood estimate (MLE) of αt is 0.021753

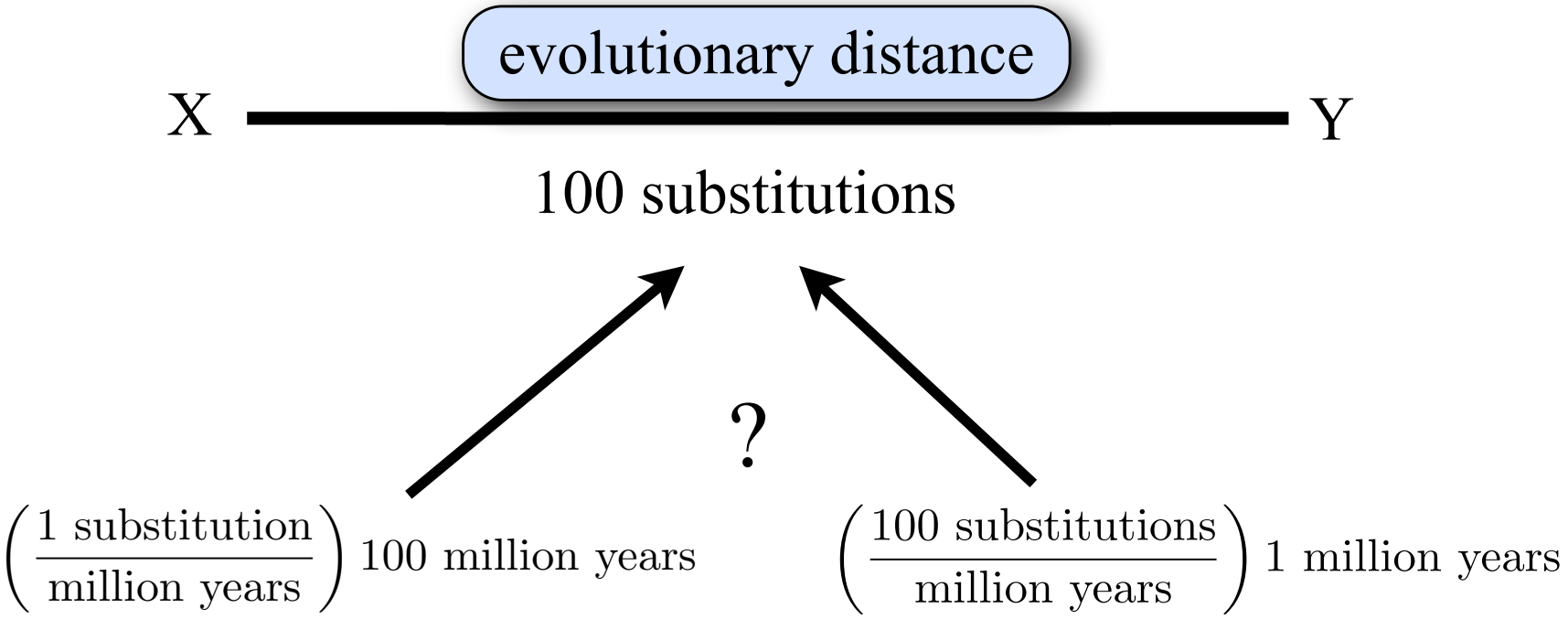
number of substitutions = rate × time



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

$$v = 3\alpha t$$

Rate and time are confounded



On Tuesday, Tracy Heath will introduce models that allow separate estimation of rates and times, but without extra information/constraints, sequence data allow only estimation of the **number** of substitutions.

Evolutionary distances for several common models

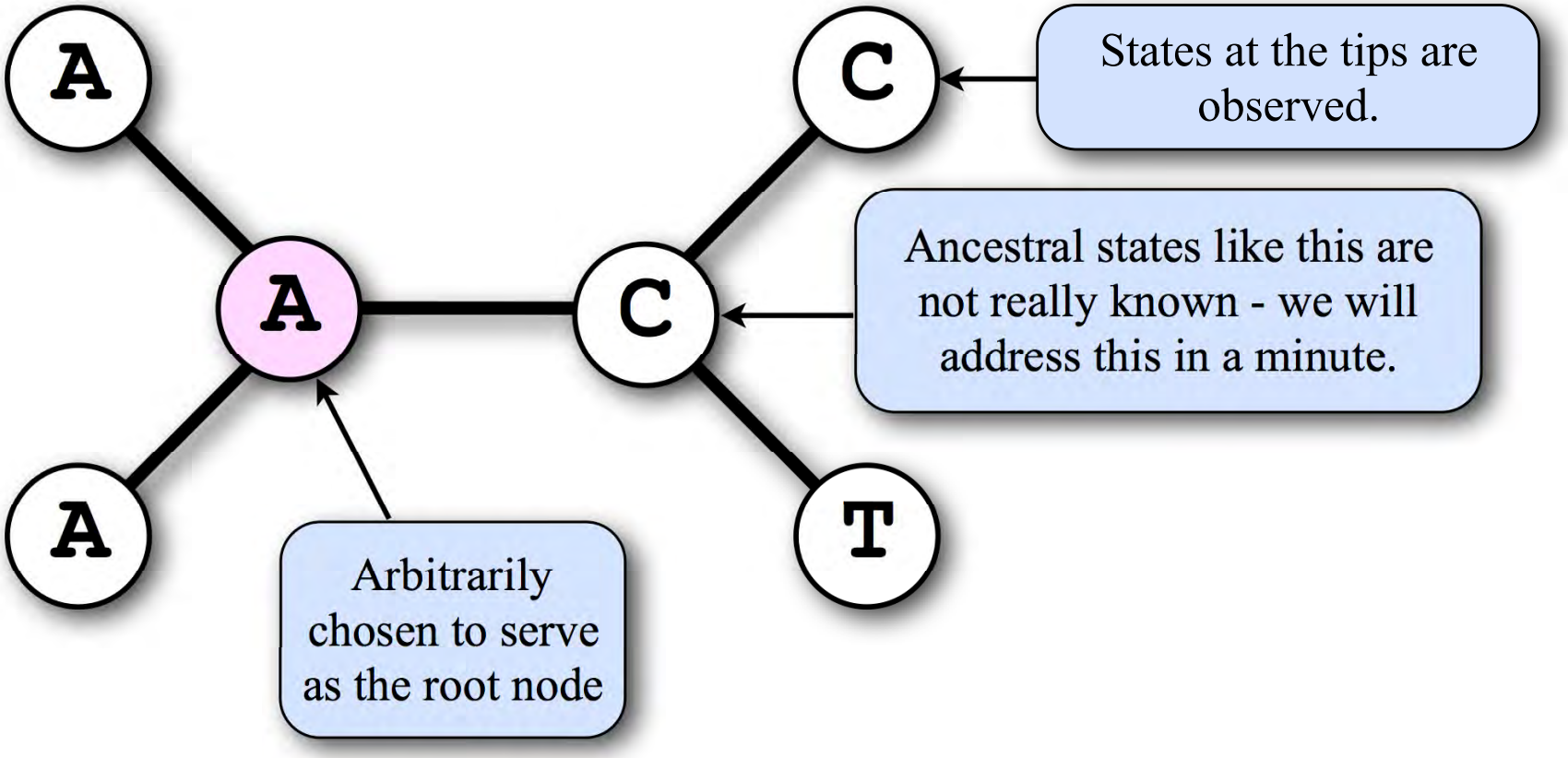
Model	Expected no. substitutions: $v = \{r\}t$
JC69	$v = \{3\alpha\}\, t$
F81	$v = \{2\mu(\pi_R\pi_Y + \pi_A\pi_G + \pi_C\pi_T)\}\, t$
K80	$v = \{\beta(\kappa + 2)\}\, t$
HKY85	$v = \{2\mu\,[\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)]\}\, t$

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

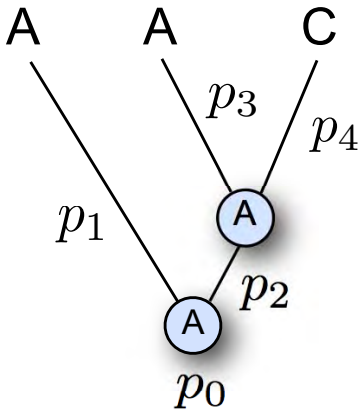
One substitution model parameter is always determined from the edge length; the others are usually global (i.e. same value applies to all edges).

Likelihood of an unrooted tree

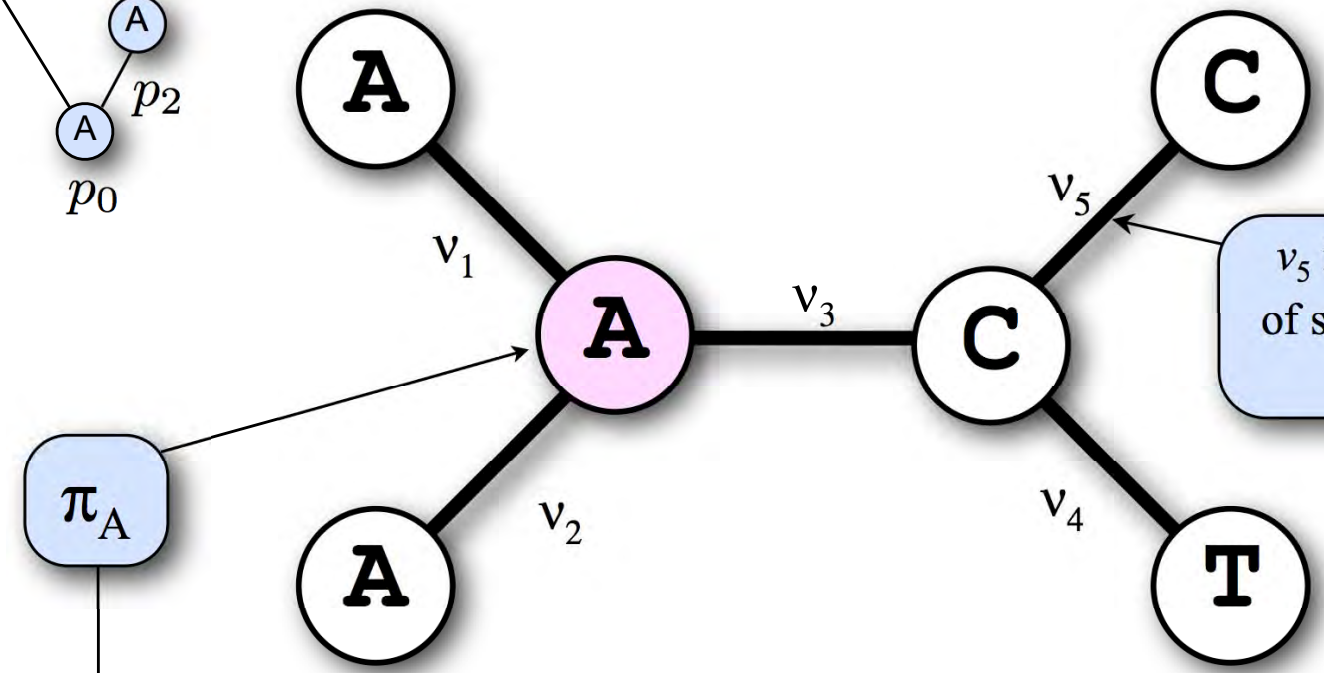
(data shown for only one site)



From slide 6



Likelihood for site *k*



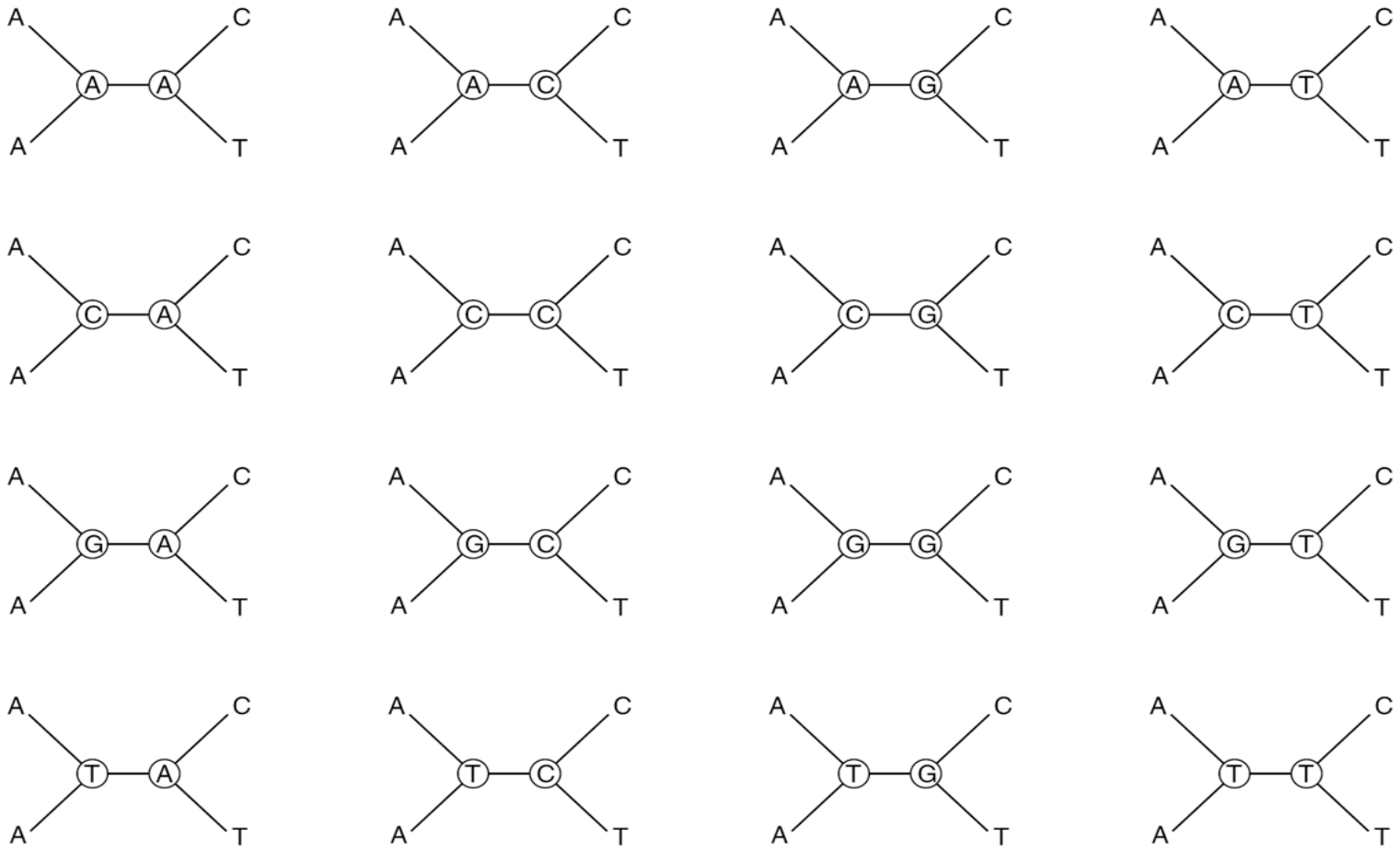
v_5 is the expected number of substitutions for just this one branch

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

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

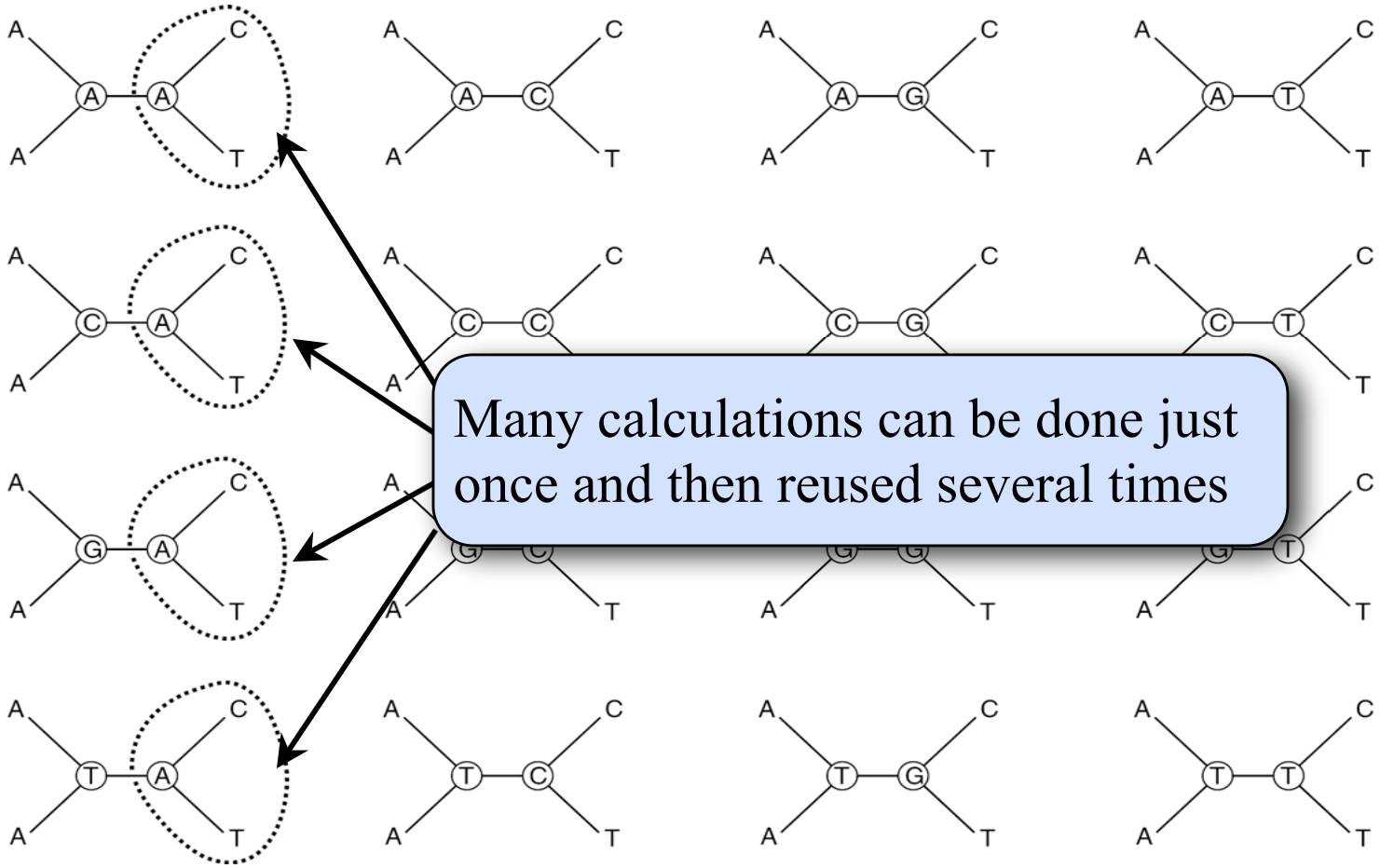
Note use of the AND probability rule

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)



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

Algorithm 1 Likelihood downpass algorithm

for all i **do**

$h_i^{(q)} \leftarrow 0$

for all j **do**

$h_i^{(q)} \leftarrow h_i^{(q)} + p_{ij}^{(q)} g_j^{(q)}$

end for

end for

for all i **do**

$h_i^{(r)} \leftarrow 0$

for all j **do**

$h_i^{(r)} \leftarrow h_i^{(r)} + p_{ij}^{(r)} g_j^{(r)}$

end for

end for

for all i **do**

$g_i^{(p)} \leftarrow h_i^{(q)} h_i^{(r)}$

end for

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
```

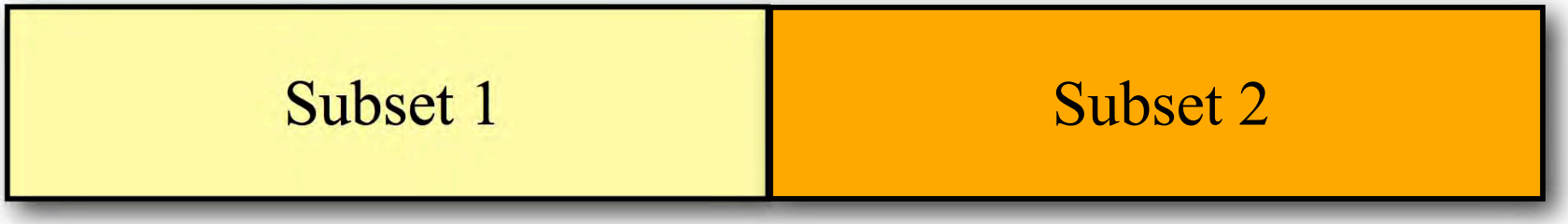
```
Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAACTAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA
.....A..T........A.........G..T..G.........A.........A..A.........T.....G.....A.........T..T..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



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:

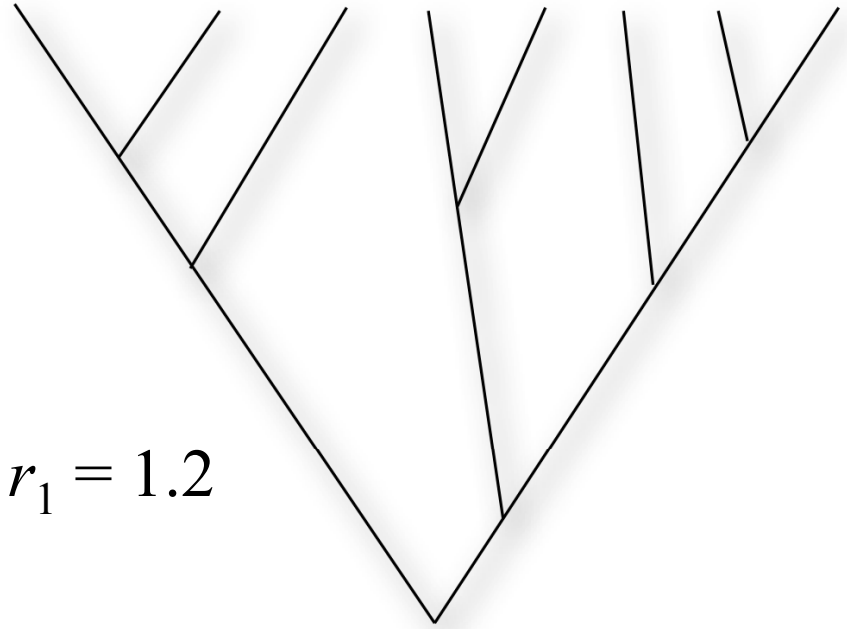
$$\frac{r_1 + r_2}{2} = 1$$

More generally:

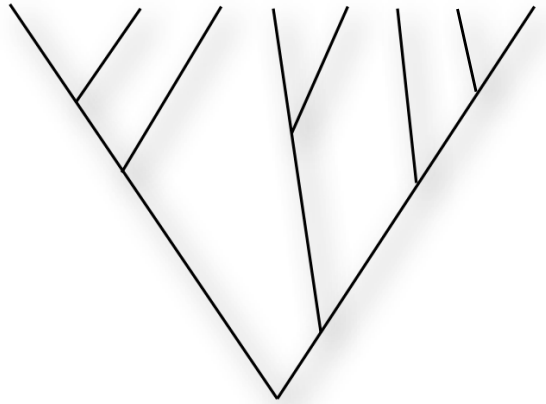
$$r_1\ p(r_1) + r_2\ p(r_2) = 1$$

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**:

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

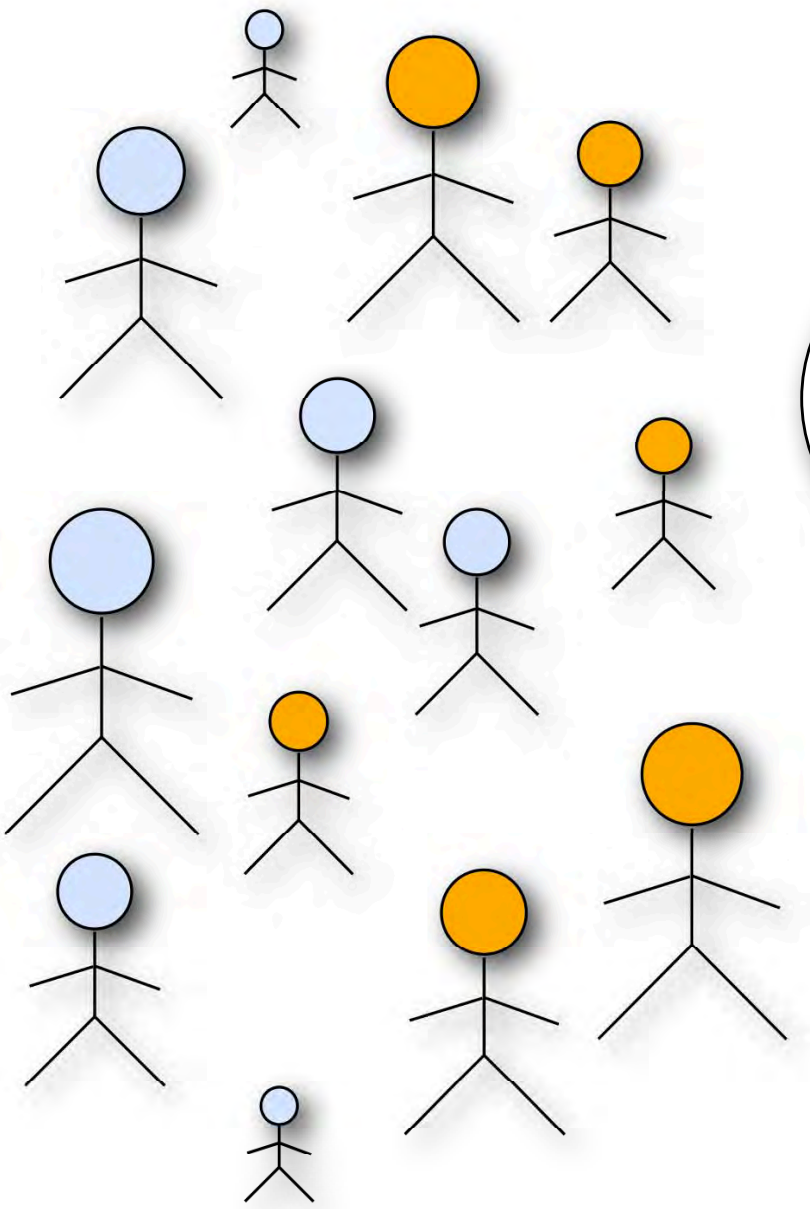
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4r_1\alpha t}$$

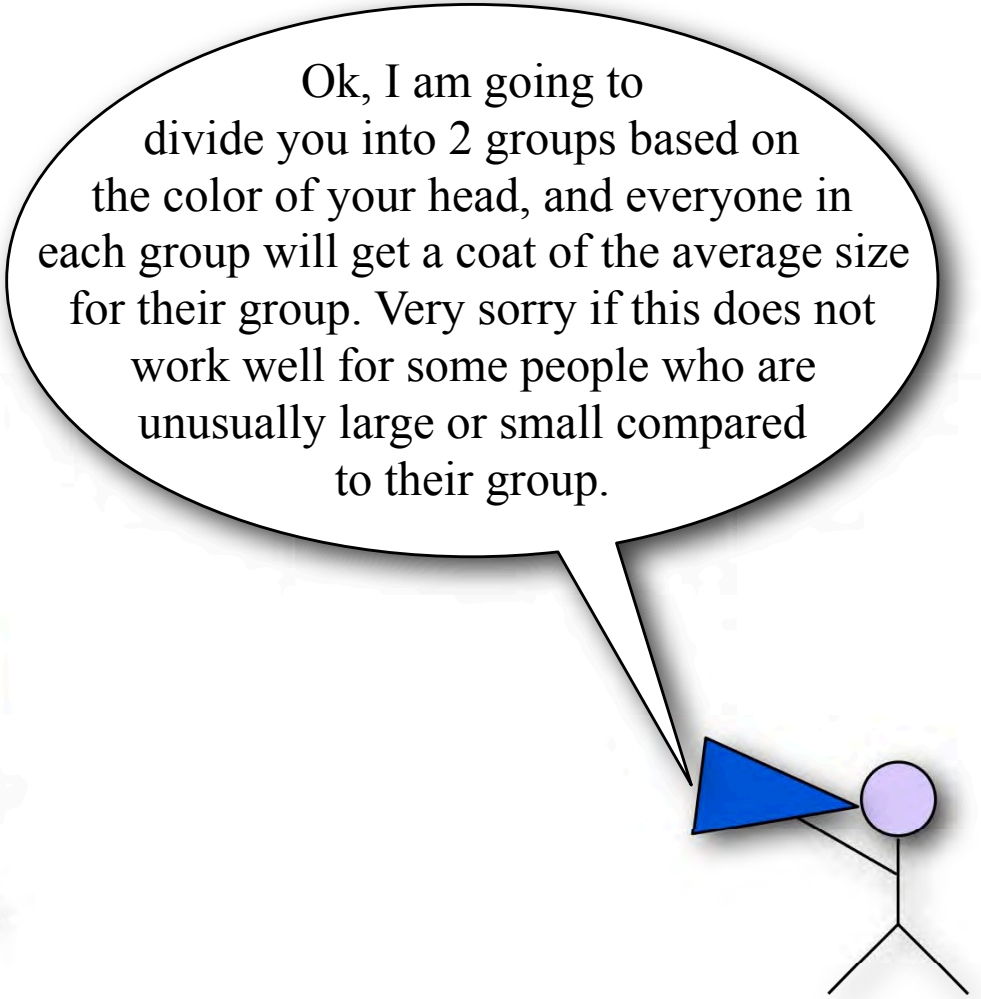
JC69 transition probabilities that would be used for sites in **gene 2**:

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

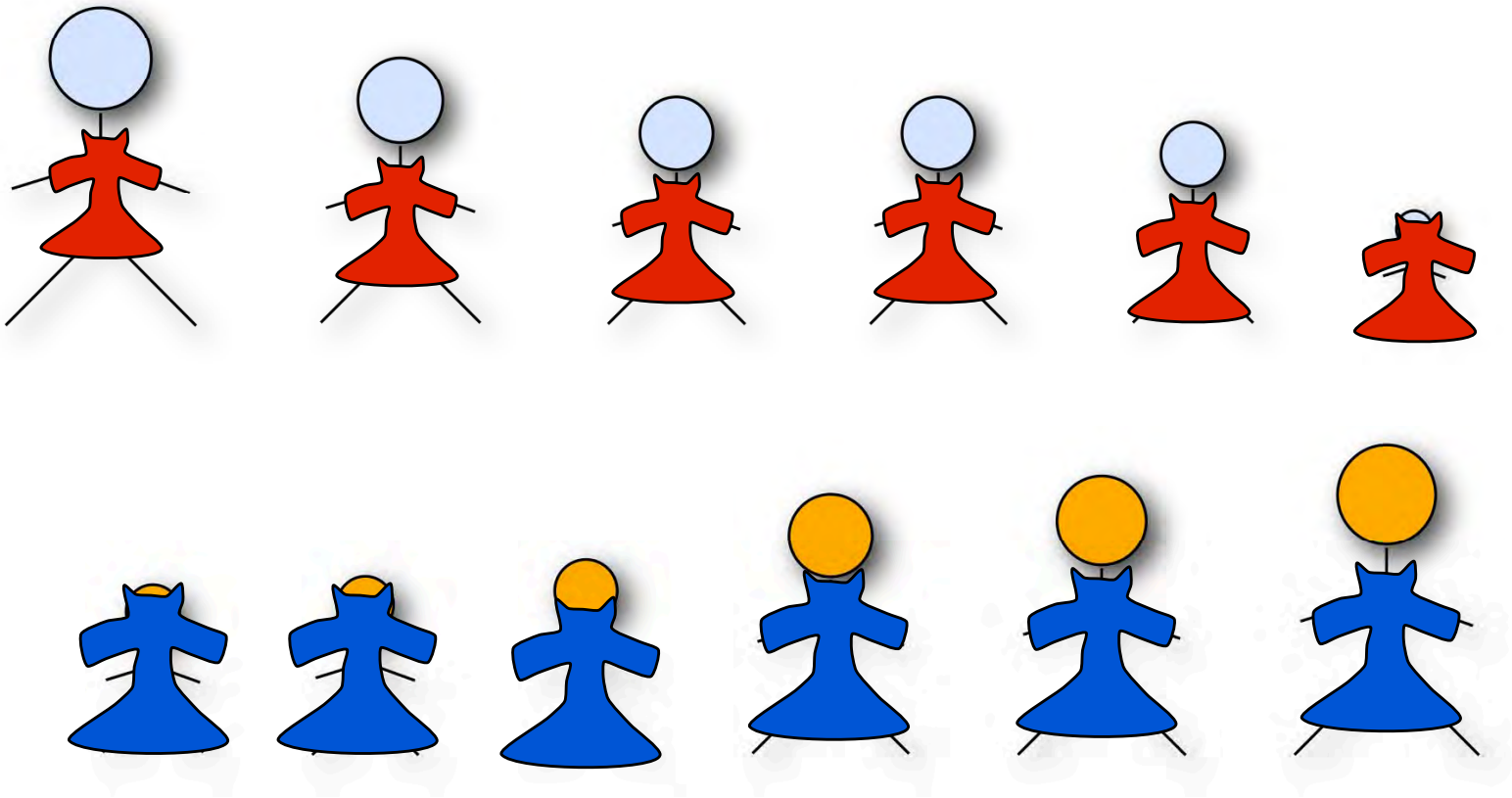
$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4r_2\alpha t}$$

Site-specific Approach





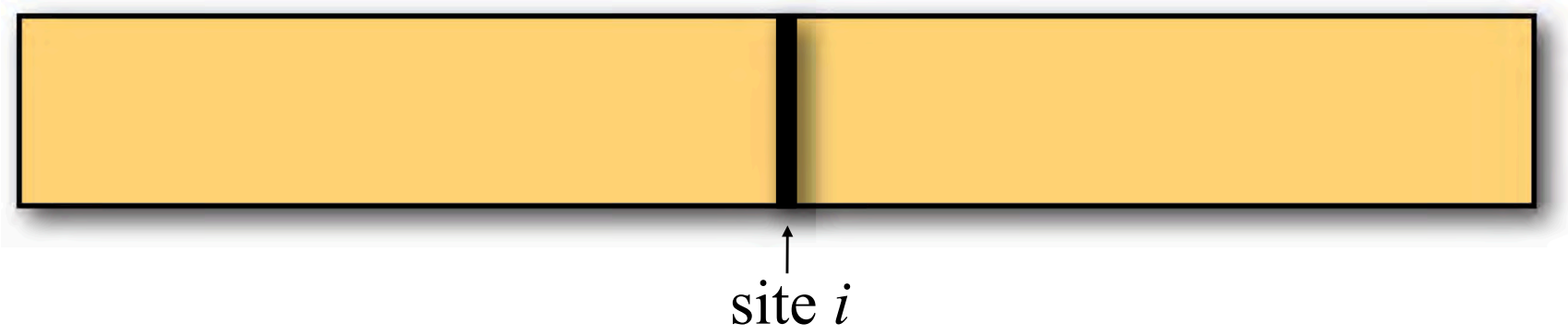
Site-specific Approach



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

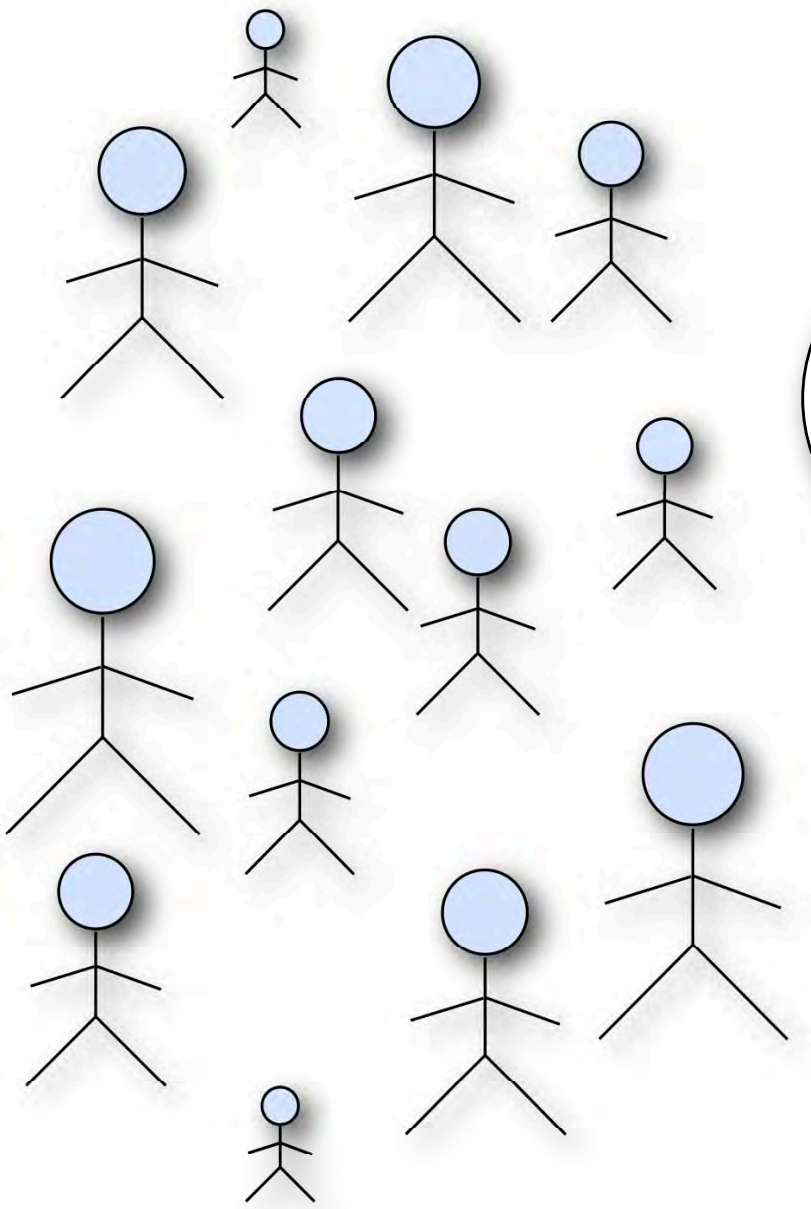
All relative rates applied to every site

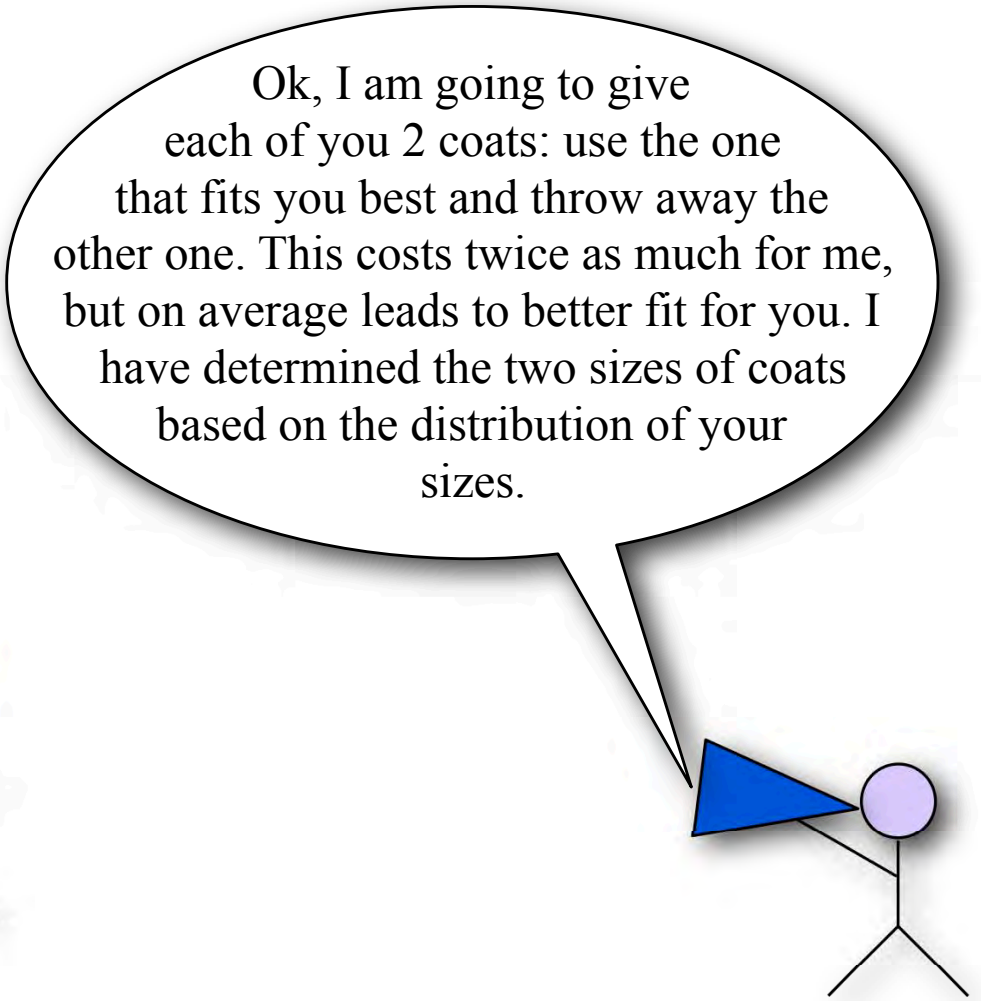


$$L_i = \Pr(D_i|r_1)\Pr(r_1) + \Pr(D_i|r_2)\Pr(r_2)$$

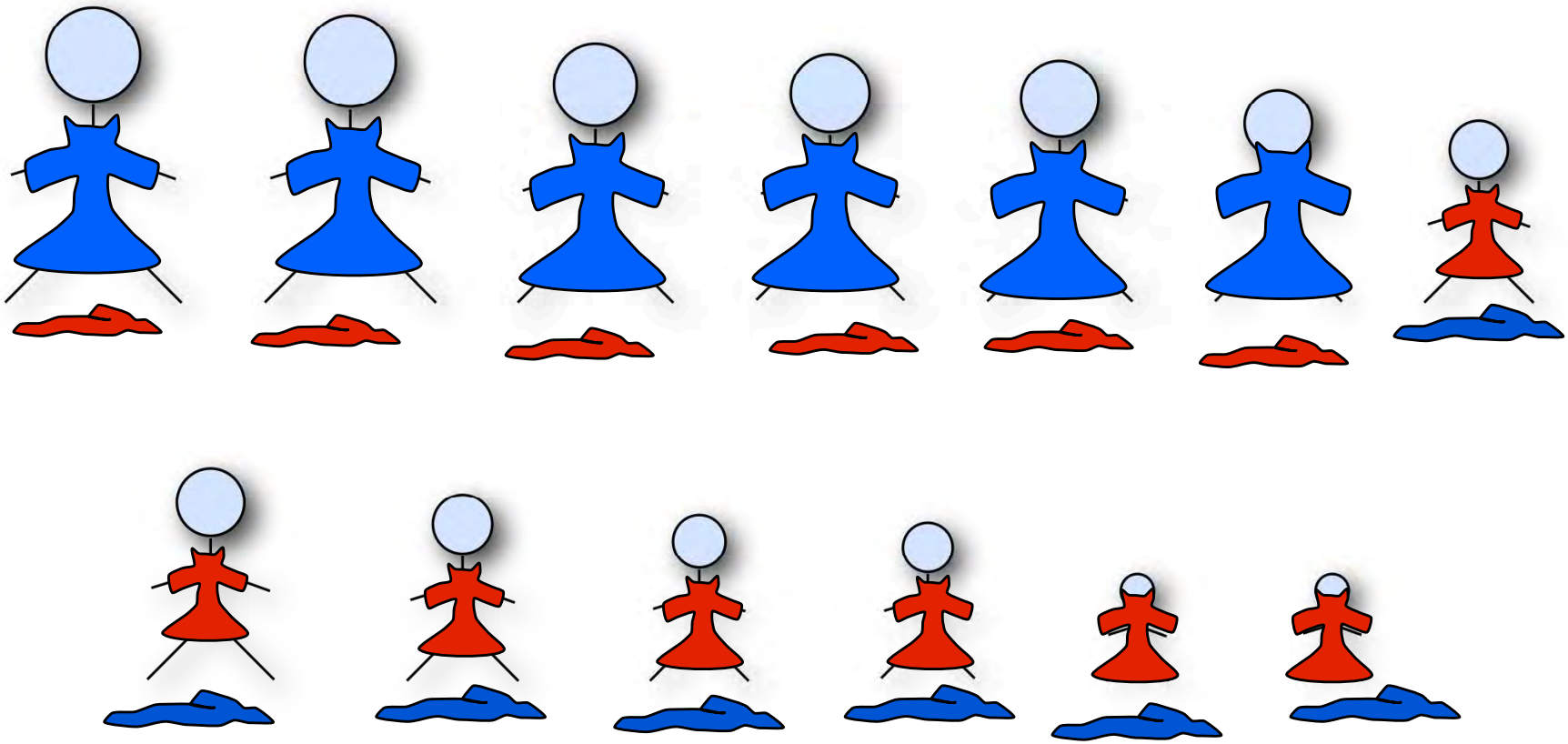
Common examples
- Invariable sites (I) model
- Discrete Gamma (G) model

Mixture Model Approach





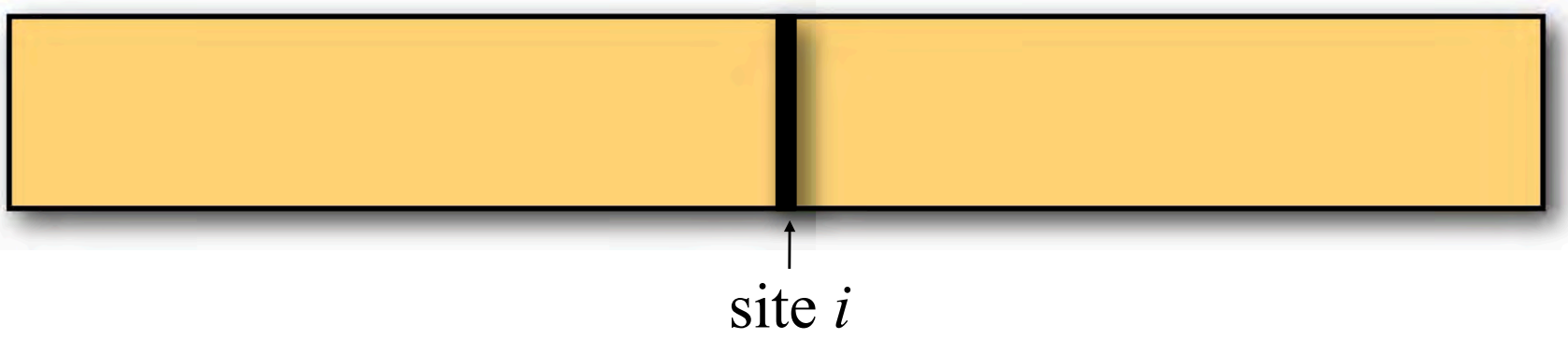
Mixture Model Approach



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)



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

$$r_1 = 0.0$$

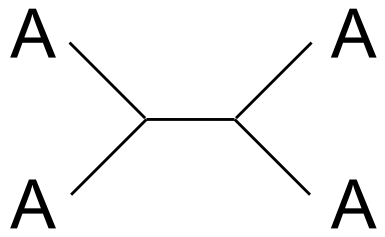
$$r_2 = \frac{1}{1 - p_{\text{invar}}}$$

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.

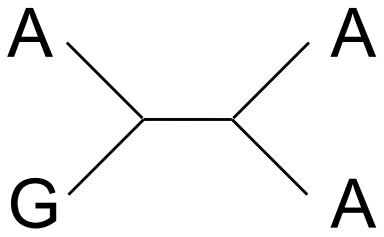
Invariable sites model

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



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

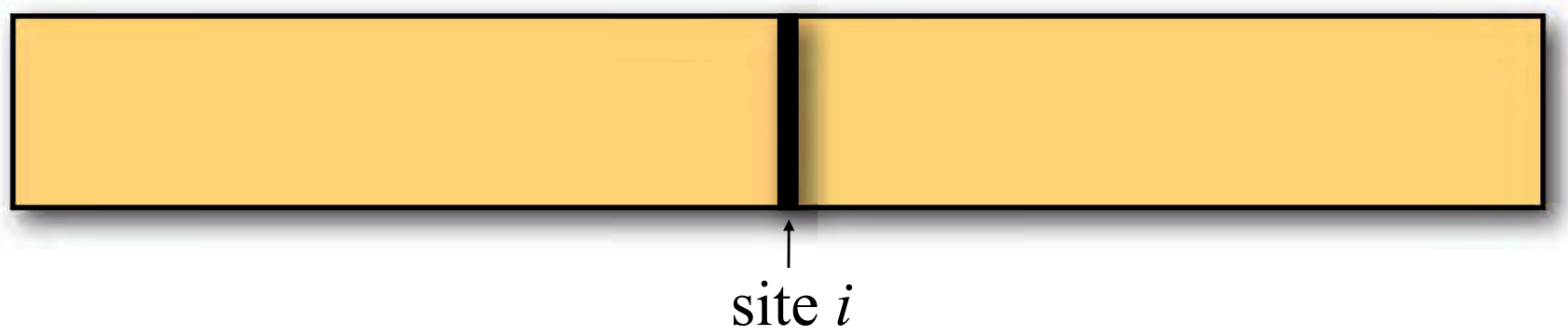
If site *i* is a *variable* site, there is no way to explain the data with a zero rate, so the first term is zero:



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

Discrete Gamma Model

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



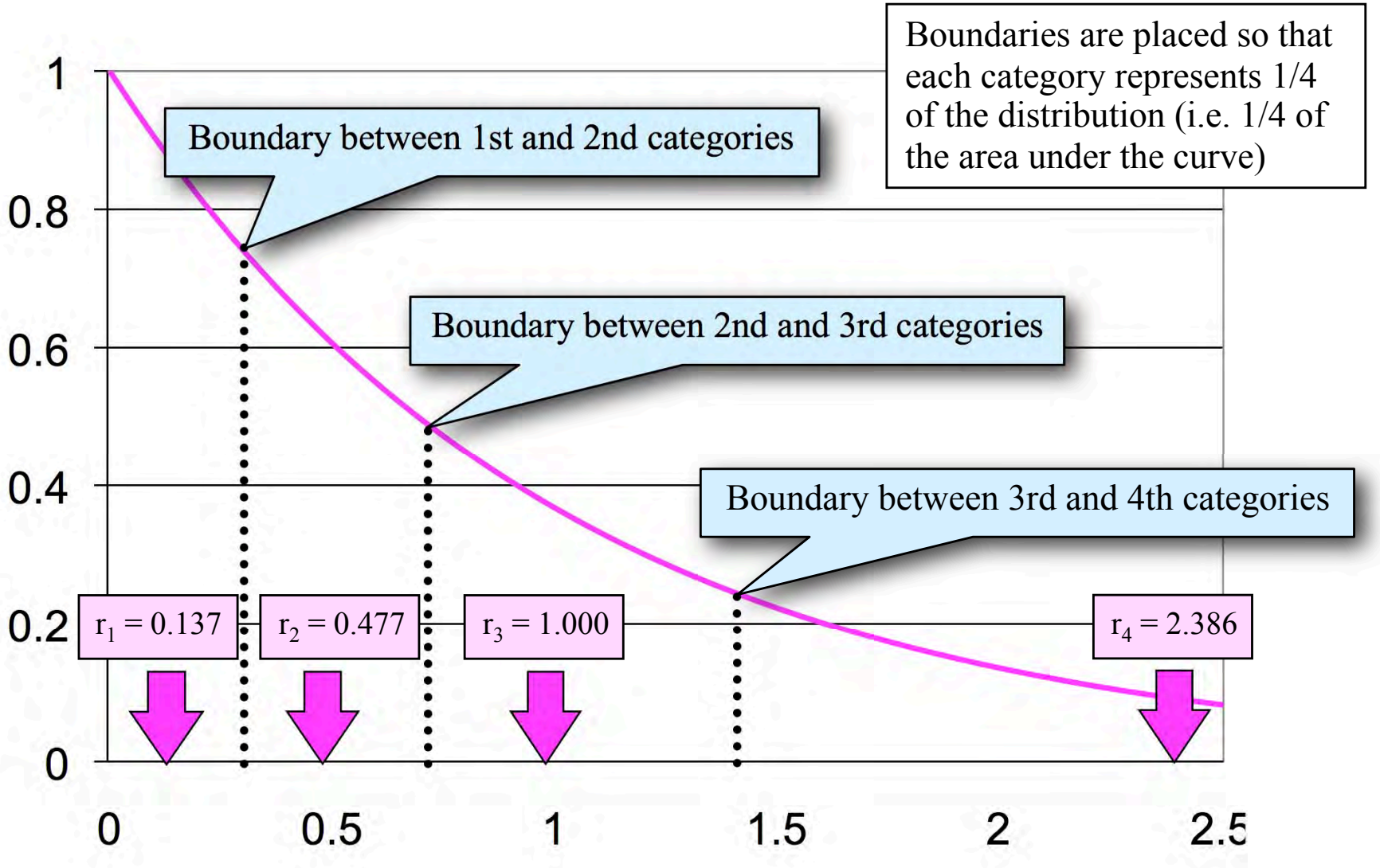
$$L = \left(\tfrac{1}{4}\right)\Pr(D_i|r_1) + \left(\tfrac{1}{4}\right)\Pr(D_i|r_2) + \left(\tfrac{1}{4}\right)\Pr(D_i|r_3) + \left(\tfrac{1}{4}\right)\Pr(D_i|r_4)$$

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