Newick trees

#NEXUS

Begin trees;
    Translate
        1 Chlamydopodium_vacuolatum_EF113426,
        2 Protosiphon_sp_FRT2000_JN880462,
        3 Protosiphon_botryoides_UTEX_B99_JN880463,
        4 Protosiphon_botryoides_UTEX_B461_JN880464,
        5 Protosiphon_botryoides_f_parieticola_UTEX_46_JN880465,
        6 Protosiphon_botryoides_UTEX_47_JN880466
        ;
tree 'PAUP_1' = [&U] (1:0.104899,((2:0.009446,(4:0.001635,6:7.29892e-07):0.030410):
0.005612,3:0.007100):0.002552,5:0.001416);
End;

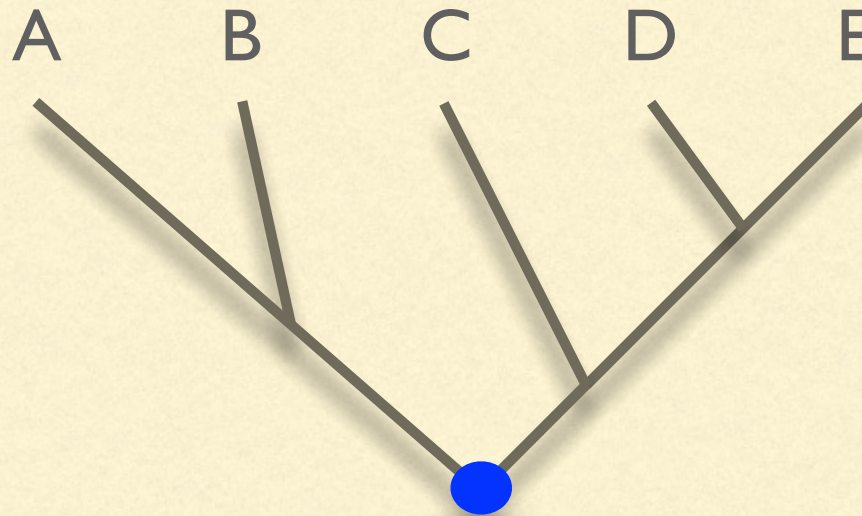https://en.wikipedia.org/wiki/Newick_format

Newick tree descriptions

$$((A,B),(C,(D,E)))$$

Newick tree descriptions
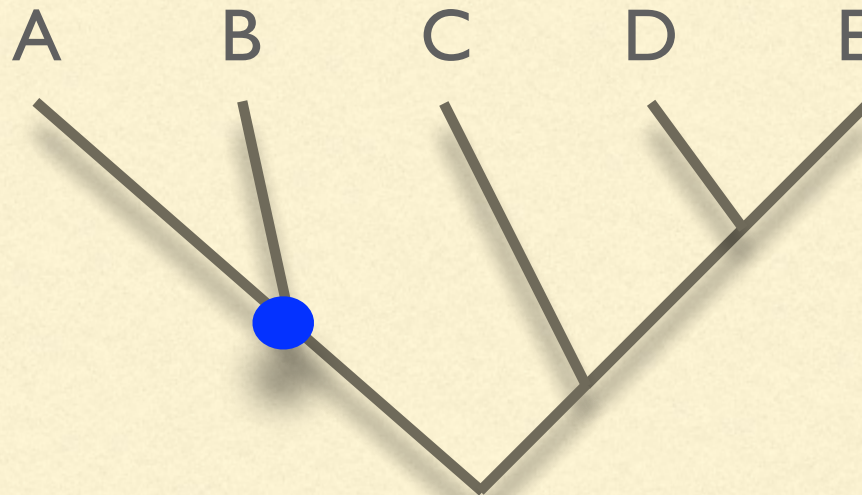

$$((A,B),(C,(D,E)))$$

Newick tree descriptions

$$((A,B),(C,(D,E)))$$

Newick tree descriptions

((A,B),(C,(D,E)))

Newick tree descriptions

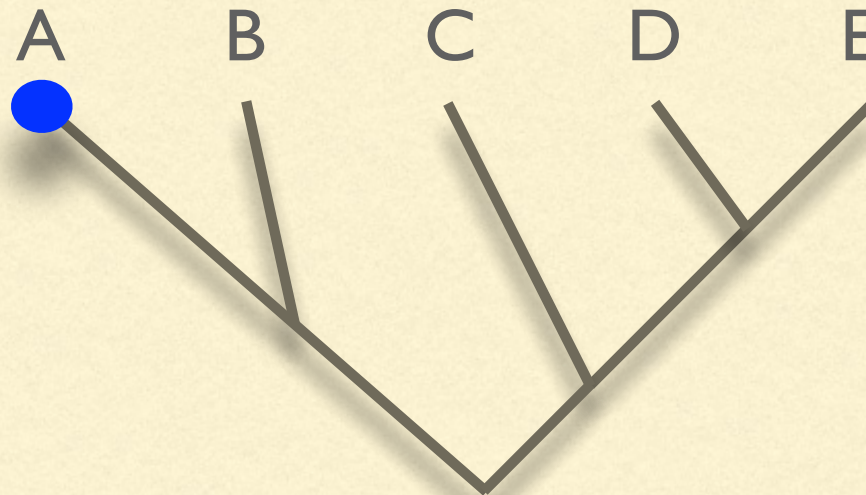$$((A,B),(C,(D,E)))$$

Newick tree descriptions

((A,B),(C,(D,E)))

Newick tree descriptions

$$((A,B),(C,(D,E)))$$

Newick tree descriptions

$$((A,B),(C,(D,E)))$$

Newick tree descriptions

((A,B),(C,(D,E)))

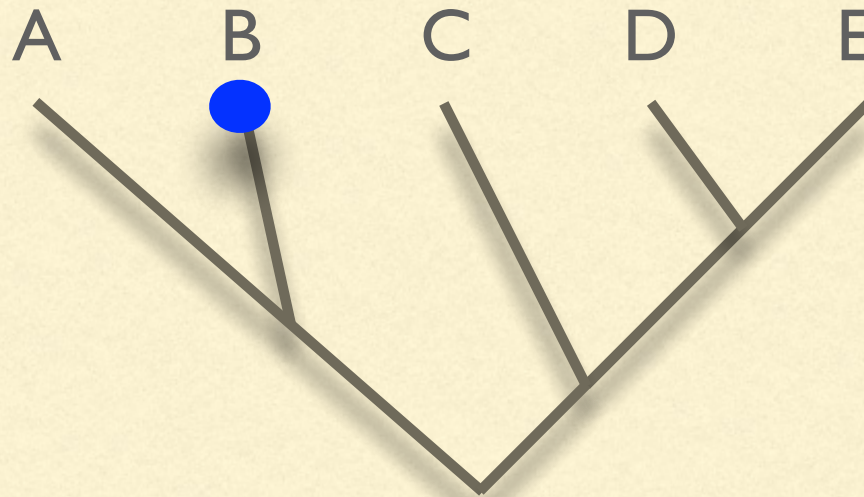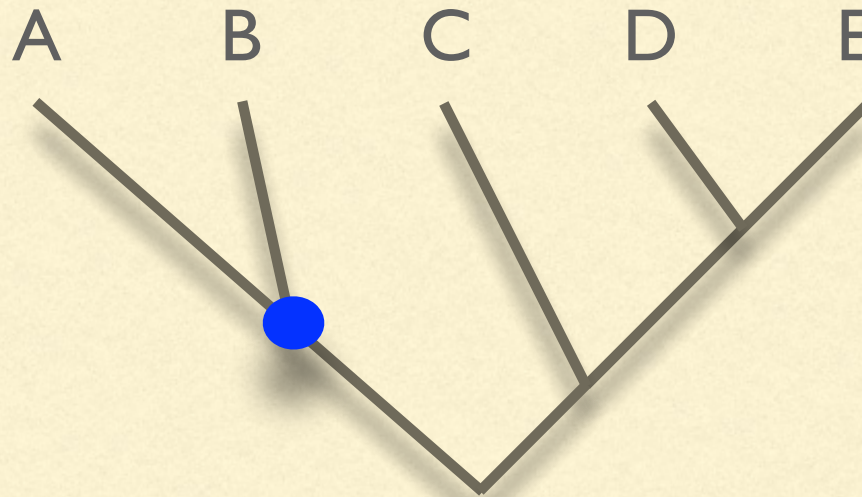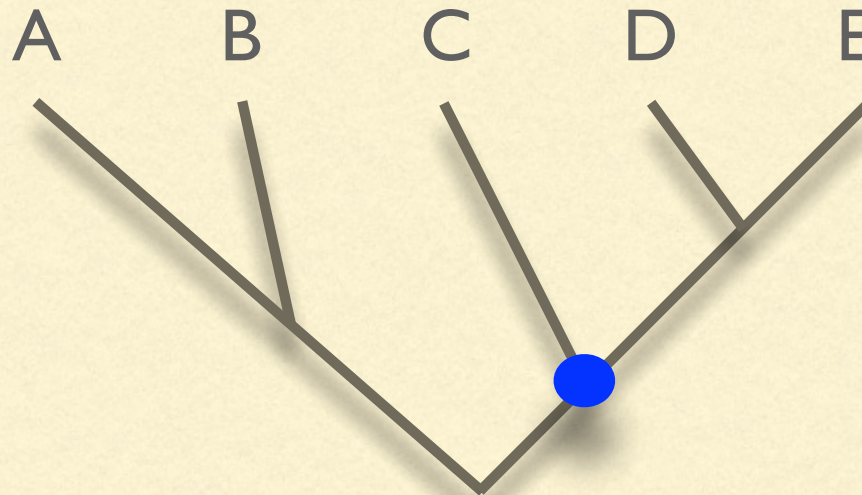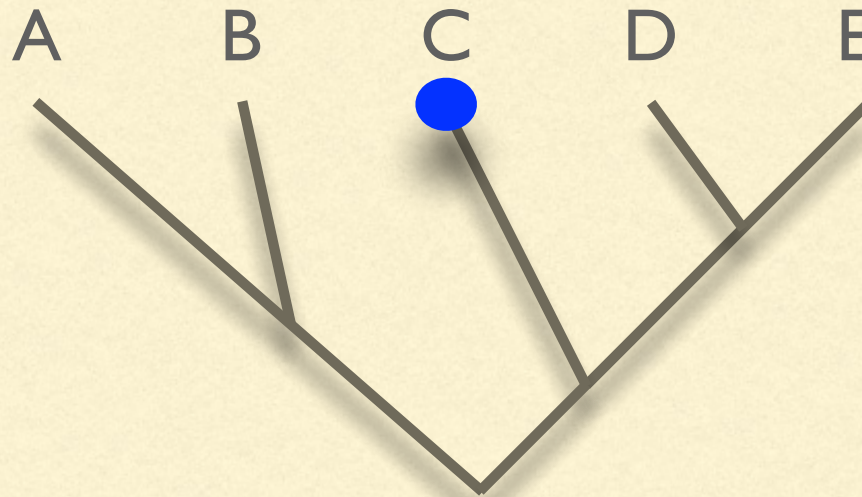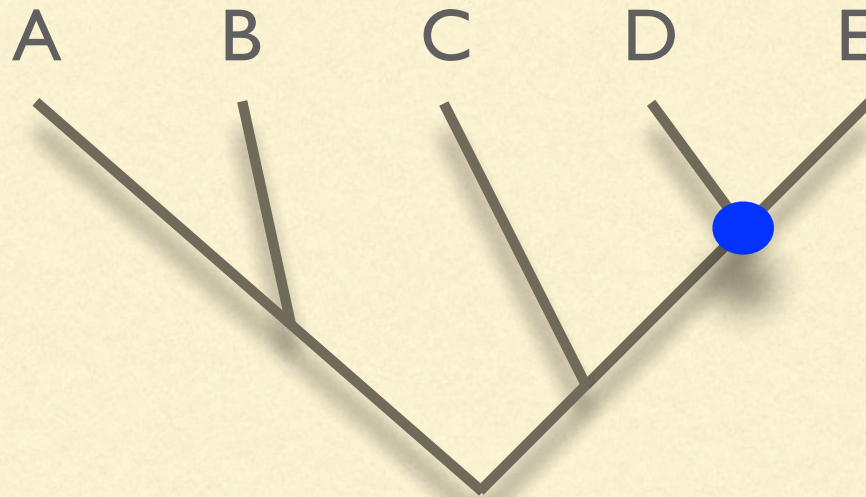# Newick tree descriptions
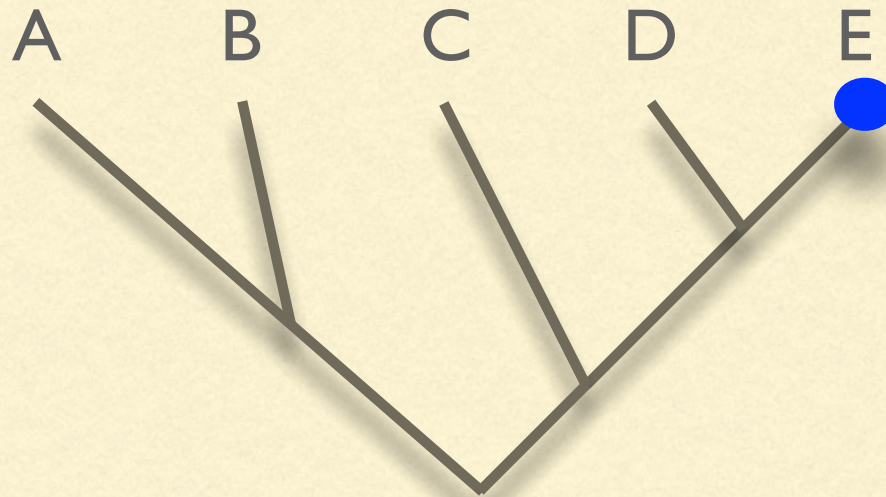
((A,B),(C,(D,E)))

Newick tree descriptions

((A,B),(C,(D,E)))

Newick tree descriptions
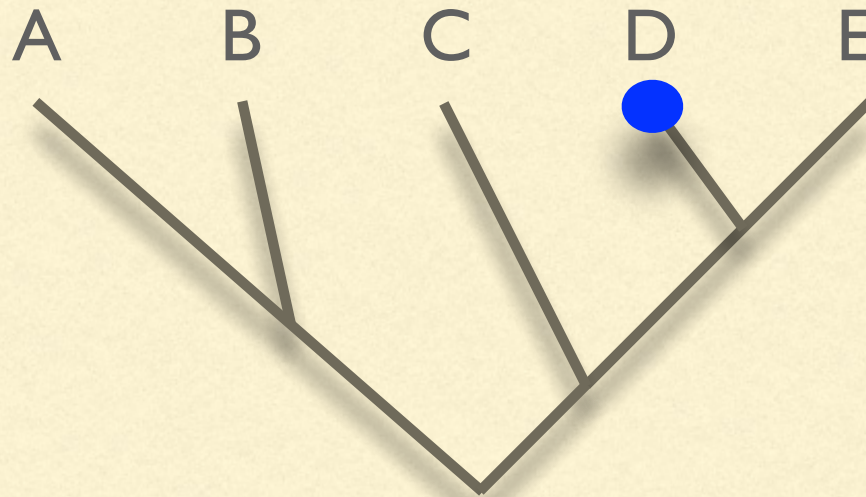

((A,B),(C,(D,E)))

Newick tree descriptions

$$((A,B),(C,(D,E)))$$

Newick tree descriptions

$$((A:.12,B:.12):.13,(C:.2,(D:.06,E:.06):.14):.05)$$

edge lengths follow colon after node name (if present)

Hasegawa, Kishino, and Yano (1985)

HKY85 model

Parameters: $\mu, \boxed{\kappa, \pi_A, \pi_C, \pi_G}$

these are global parameters (apply to all edge lengths)

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

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

Tavaré (1986)

GTR model

Parameters: $a, b, c, d, e, \mu, \pi_A, \pi_C, \pi_G$

$$
\begin{array}{c}
 \quad\;\; A \qquad\quad C \qquad\quad G \qquad\quad T \\
\begin{array}{c} A \\[2.5em] C \\[2.5em] G \\[2.5em] T \end{array}
\left[
\begin{array}{cccc}
- & \pi_C\mu\,\boxed{a} & \pi_G\mu\,\boxed{b} & \pi_T\mu\,\boxed{c} \\[1.5em]
\pi_A\mu\,a & - & \pi_G\mu\,\boxed{d} & \pi_T\mu\,\boxed{e} \\[1.5em]
\pi_A\mu\,b & \pi_C\mu\,d & - & \pi_T\mu\,\boxed{f} \\[1.5em]
\pi_A\mu\,c & \pi_C\mu\,e & \pi_G\mu\,f & -
\end{array}
\right]
\end{array}
$$

exchangeability parameters are circled

GTR = General Time Reversible
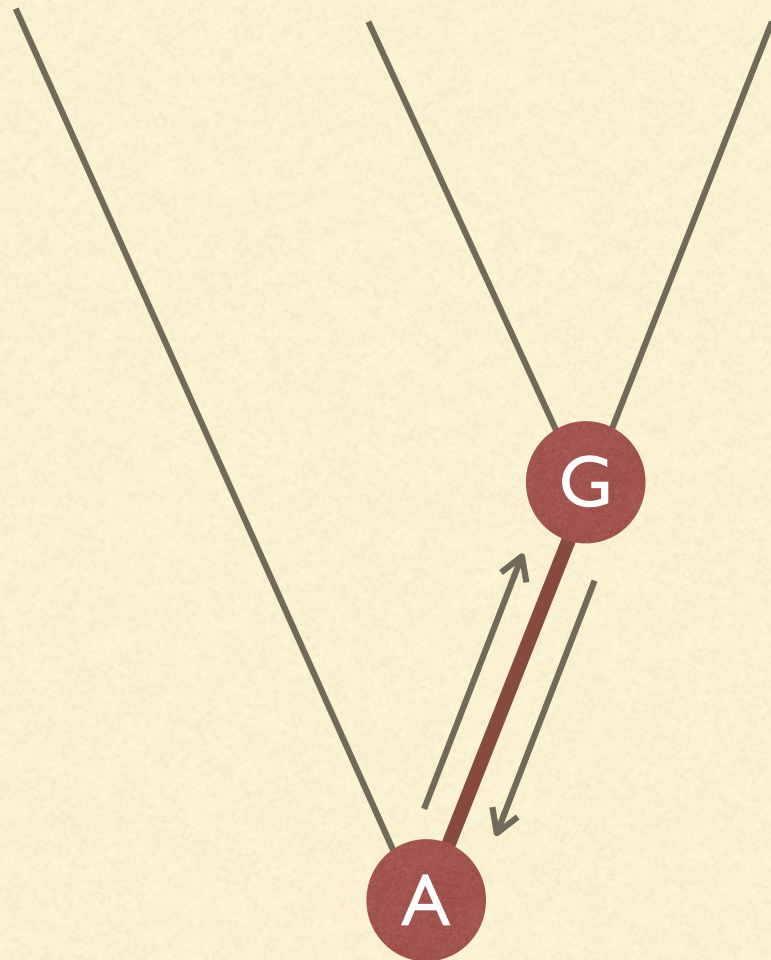
Time reversibility means...

$$\mathrm{Pr}(A)\,\mathrm{Pr}(G|A, v) = \mathrm{Pr}(G)\,\mathrm{Pr}(A|G, v)$$

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

GTR family

GTR: $a, b, c, d, e, \mu, \pi_A, \pi_C, \pi_G$

$a=c=d=1, b=e=\kappa$

HKY85: $\mu, \kappa, \pi_A, \pi_C, \pi_G$

$\mu=4\beta, \pi_A=\pi_C=\pi_G=1/4$

$\kappa=1$

K80: κ, β

F81: μ, π_A, π_C, π_G

$\kappa=1$

$\mu=4\beta, \pi_A=\pi_C=\pi_G=1/4$

JC69: β
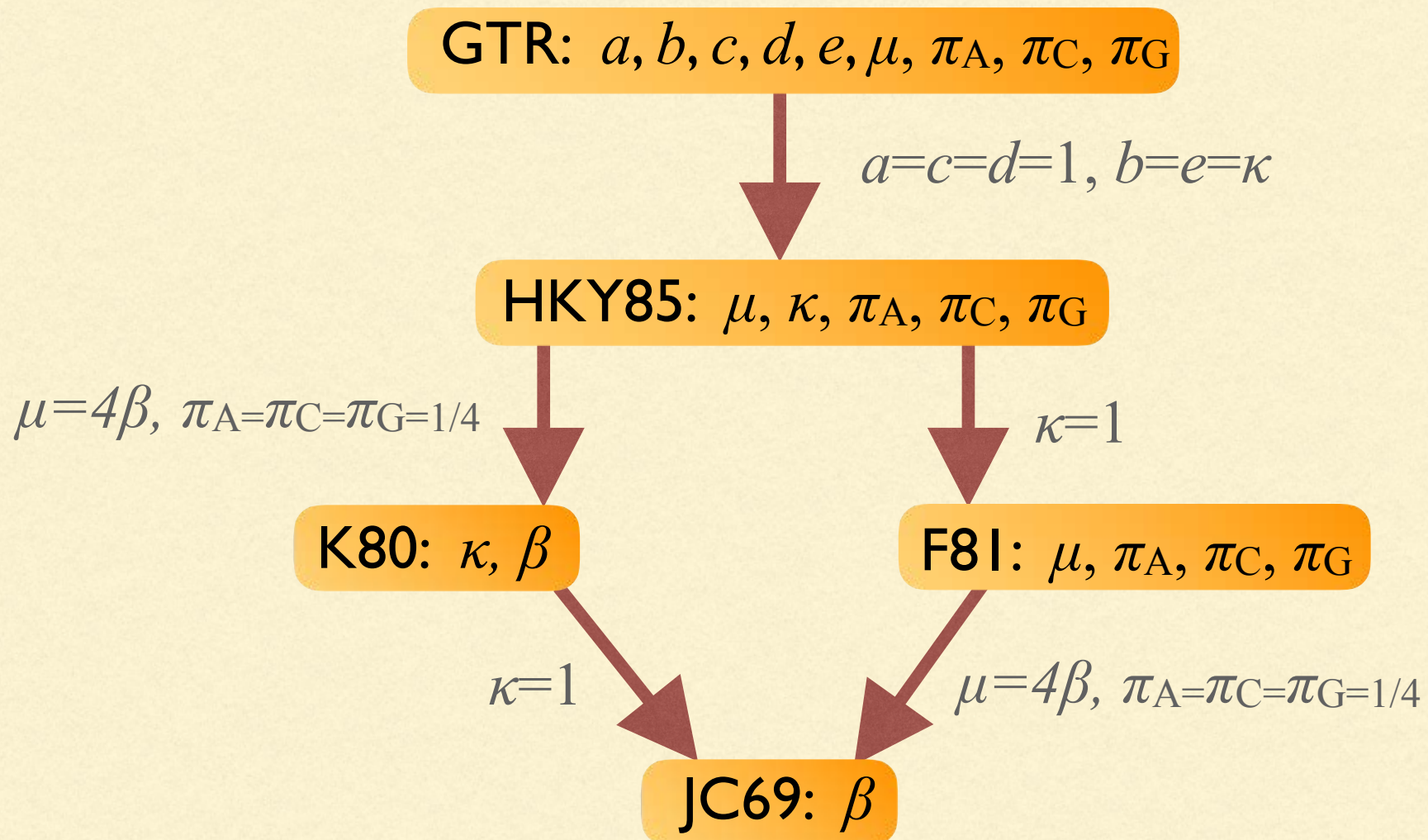
Rate heterogeneity

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...T
Osmunda      (fern)                             ........TC...G...C..........C.T...G.G..C..G.......T.....G..A......C..AA...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..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........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.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..........T....C..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...C
.....A..G.........G..A..G.A..........C....C........C..C.T...C.T..C...C.T..T....G.........T.C..C
.....A..G..B..C.G..C..G..A....A.....T....C.C.......C..C........C.T...C.T..C...C.T..T....G..GC.....T.C..C
....C..A.......TG......G...C..G.......C...............A..A..G.......T...C.T..C...C.T..T..........C.......C.C..C.G
....C..A..A..G.......C..A...........G..C..A.......C...G....A....G.G..C..CC.T....T......G..CC......C..C.G
......C..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

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

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

$r_1 = 0.12255$

mean relative rate:

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

$r_2 = 2.75490$
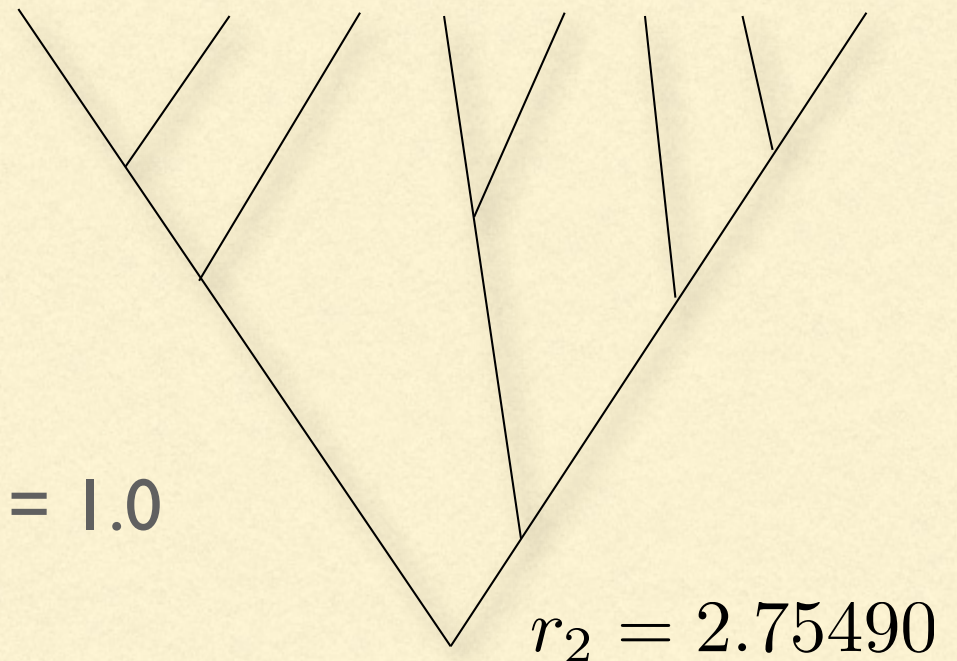
Site-specific rates

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

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

C ———————— C
identity

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

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**:

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

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

Site-specific approach
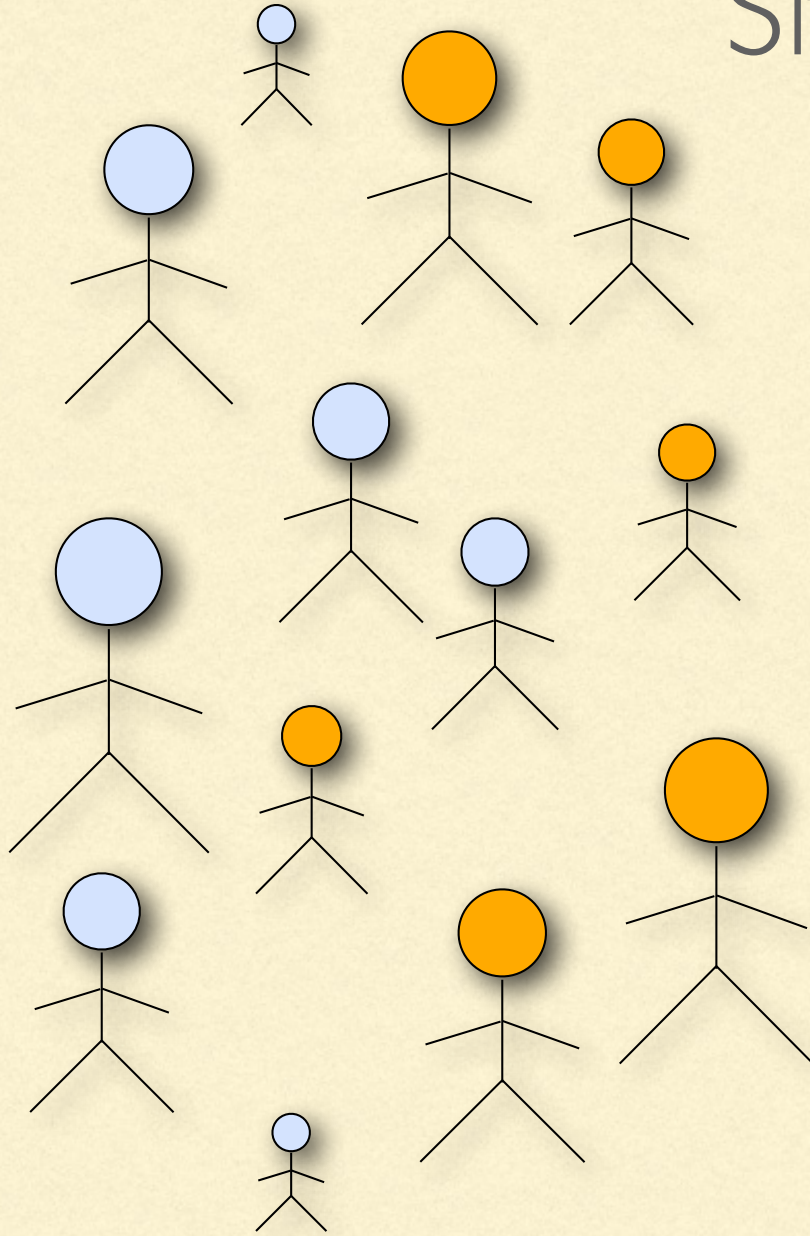
OK, I am going to divide you into 2 groups based on the color of your head, and everyone in each group will get a coat of the average size for their group. Very sorry if this does not work well for some people who are unusually large or small compared to their group.
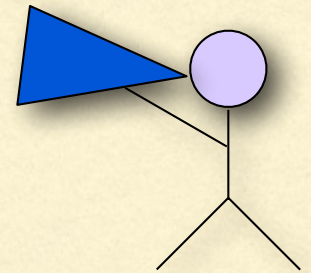
Site-specific approach
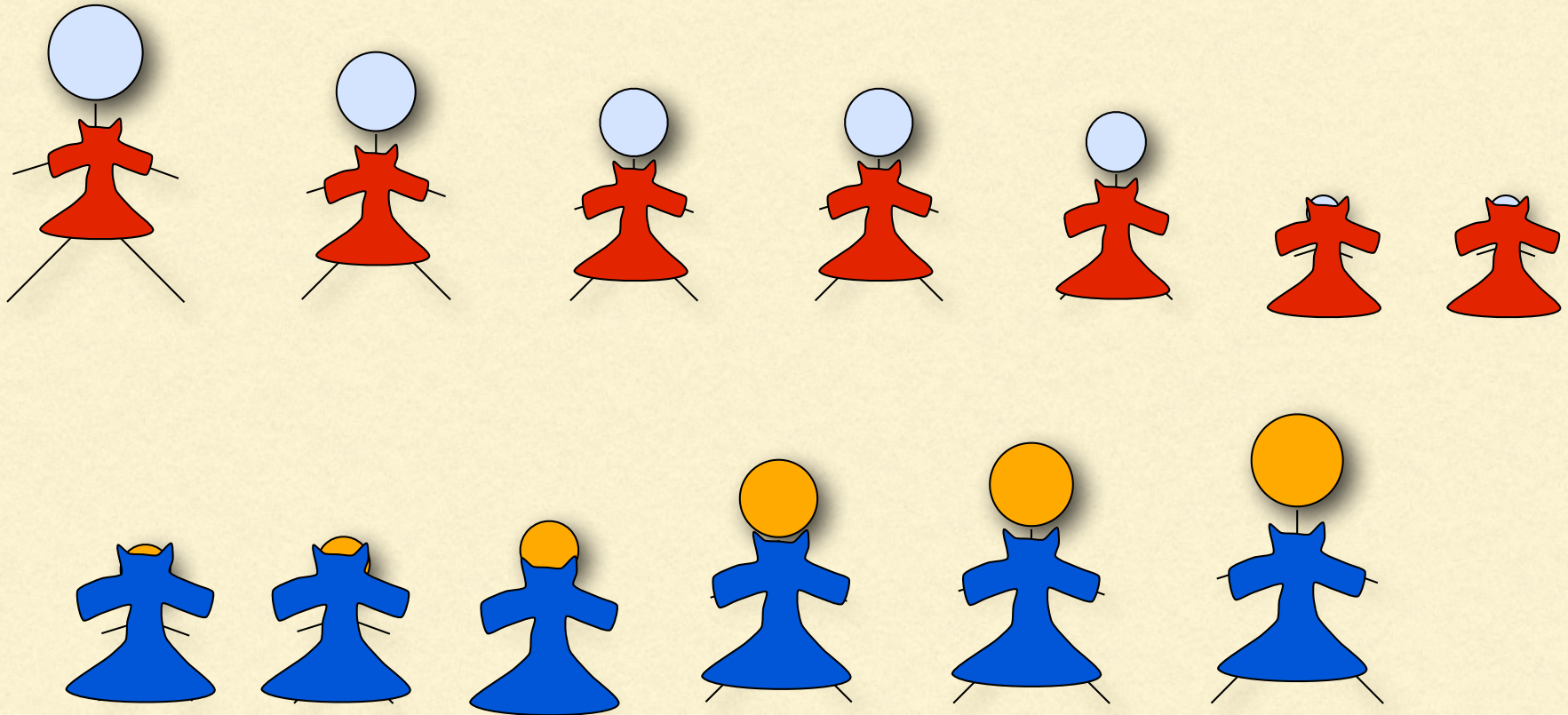
Good: costs less: need to buy just one coat for every person
Bad: every person in a group has to wear the same size coat

Mixture models

All k relative rates applied to every site

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

Mixture model approach
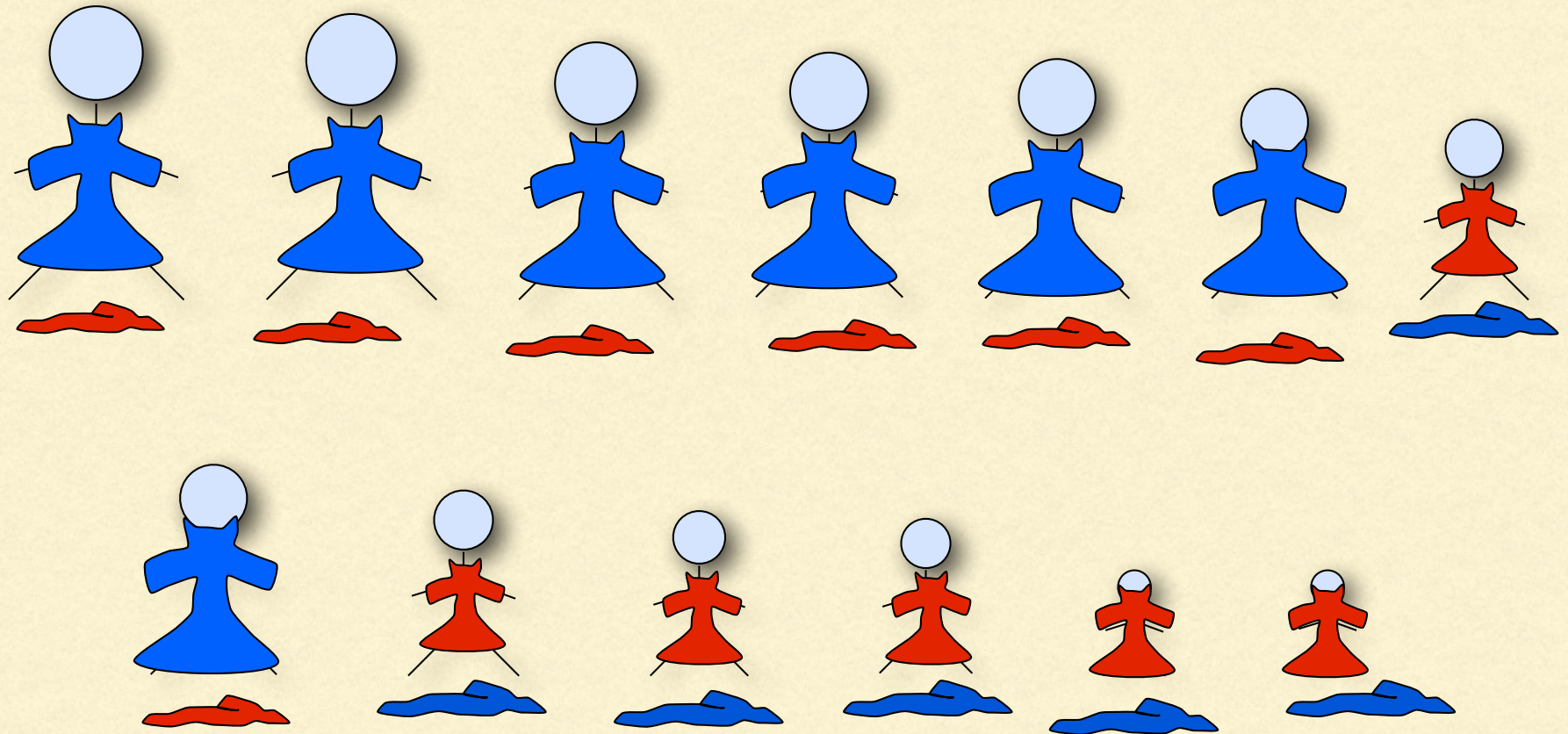
Good: every person experiences better fit because they can choose the size coat that fits best

Bad: costs more because two coats much be provided for each person

Invariable sites model (Reeves 1992)

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

$$L_i = p(\mathbf{y}_i|r_1)p_{\text{invar}} + p(\mathbf{y}_i|r_2)(1 - p_{\text{invar}})$$

$r_1 = 0.0$

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

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

$$\bar{r} = p_{\text{invar}}(0.0) + (1 - p_{\text{invar}})\left(\dfrac{1}{1 - p_{\text{invar}}}\right) = 1.0$$

Invariable sites model

If site *i* is a *constant* site, both terms will contribute to the site likelihood:
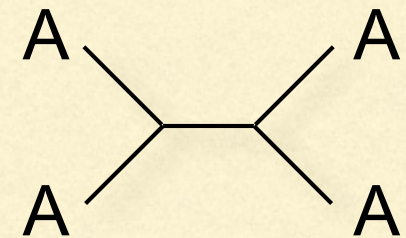
$$L_i = p(\mathbf{y}_i|r_1)p_{\text{invar}} + p(\mathbf{y}_i|r_2)(1 - p_{\text{invar}})$$

If site *i* is a *variable* site, there is no way to explain the data with a zero rate, so the likelihood in the first term equals zero:
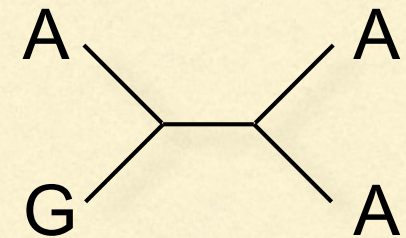
$$L_i = p(\mathbf{y}_i|r_1)p_{\text{invar}} + p(\mathbf{y}_i|r_2)(1 - p_{\text{invar}})$$

Discrete Gamma model (Yang 1994)

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

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

continuous Gamma density curve

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

$r_1 = 0.137$

$r_2 = 0.477$

$r_3 = 1.000$

$r_4 = 2.386$

Gamma distributions

larger shape means less heterogeneity
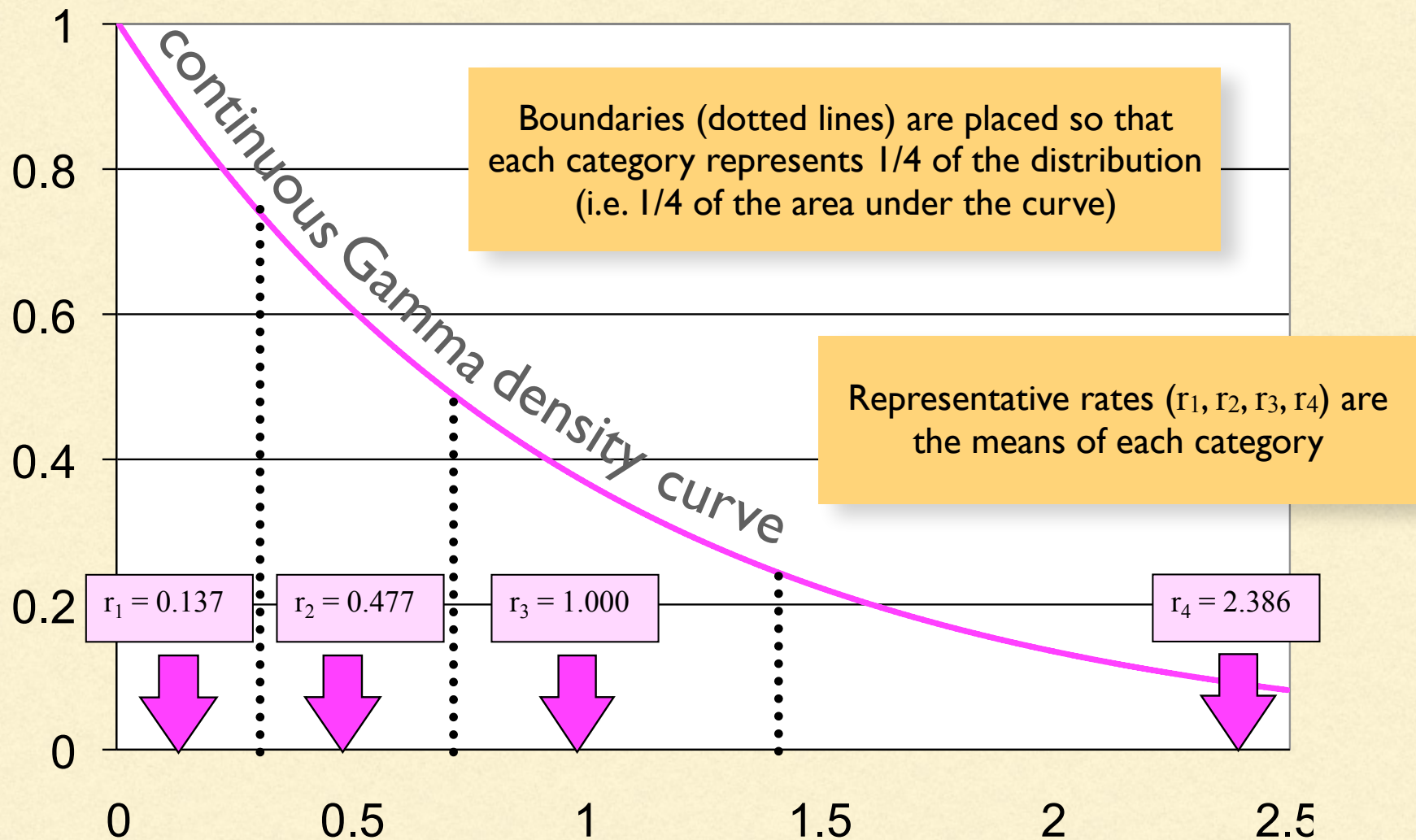
The mean relative rate equals 1.0 for all three of these distributions
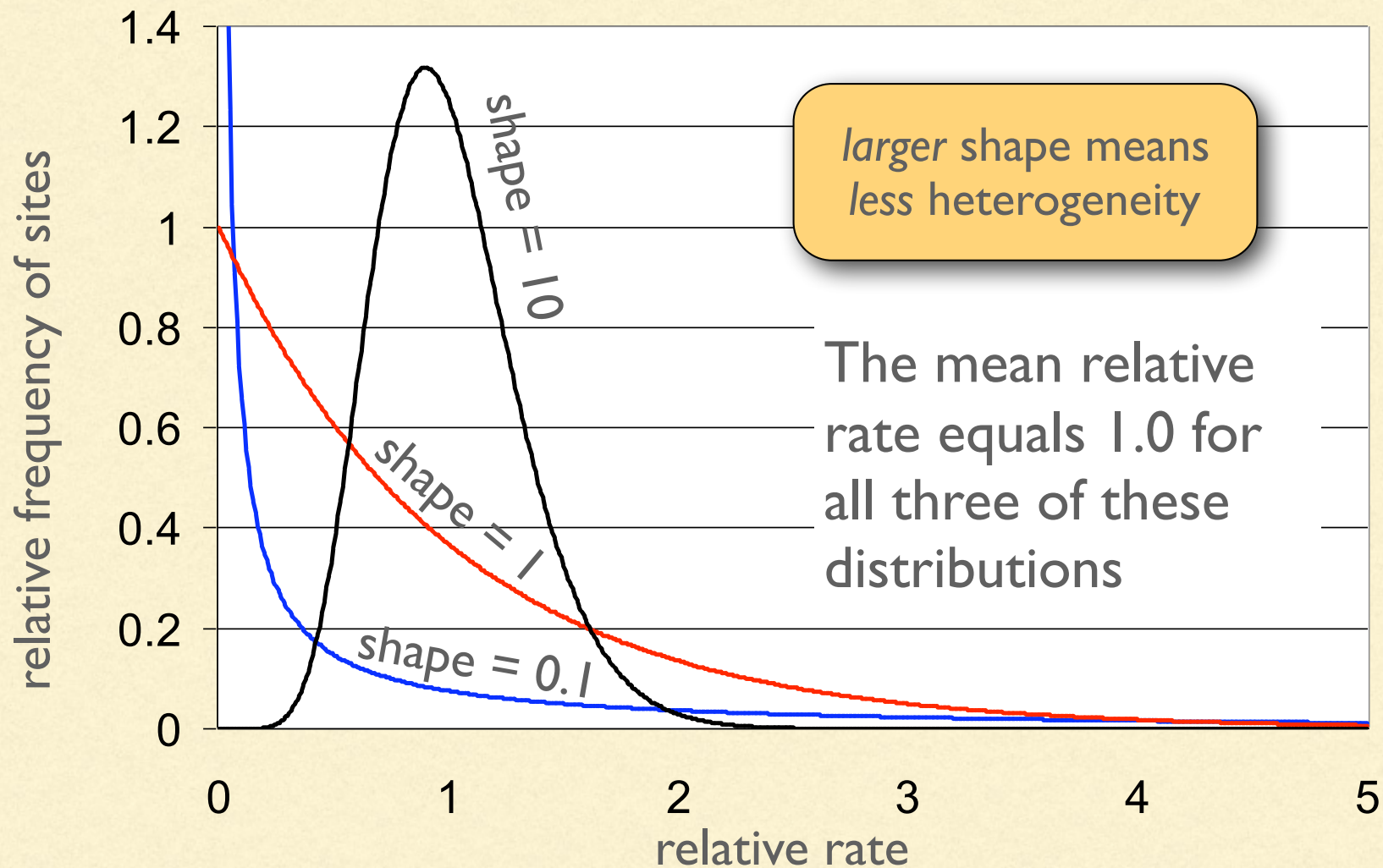
Codon models

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'  DNA
                       template strand
```

transcription

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

translation

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

Codon Table

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

Muse & Gaut (1994); Goldman & Yang(1994)

	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)			—	$\alpha\,\pi_G$	0	0	...	0
TTG (Leu)		π_A	—		0	0	...	0
CTT (Leu)				0	—	$\alpha\,\pi_C$	...	0
CTC (Leu)				0	$\alpha\,\pi_T$	—	...	0
⋮			⋮	⋮	⋮			
GGG (Gly)	0	0	0	0	0			

synon. subst.
Phe →Phe
T→C

nonsynon. subst.
Phe→Leu
C→A

rate = 0 if more than one nucleotide change is required

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

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

JC69 revisited

$$v = 3\beta t$$

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

P matrix (transition probabilities)

$$
\left[
\begin{array}{cccc}
\frac{1}{4} + \frac{3}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} + \frac{3}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} + \frac{3}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} \\
\frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \frac{1}{4} + \frac{3}{4}e^{-4\beta t}
\end{array}
\right]
$$

2-state version

Q matrix (instantaneous rates)

$$
\begin{array}{cc}
 & \begin{array}{cc} 0 & 1 \end{array} \\
\begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} -\beta & \beta \\ \beta & -\beta \end{array} \right]
\end{array}
\qquad v = \beta t
$$

P matrix (transition probabilities)

$$
\left[\begin{array}{cc} \frac{1}{2} + \frac{1}{2}e^{-2\beta t} & \frac{1}{2} - \frac{1}{2}e^{-2\beta t} \\[2mm] \frac{1}{2} - \frac{1}{2}e^{-2\beta t} & \frac{1}{2} + \frac{1}{2}e^{-2\beta t} \end{array} \right]
$$

2-state version

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

$$\mathbf{Q}$$

$$\begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} -\beta & \beta \\ \beta & -\beta \end{array} \right] \end{array}$$

2-state version

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

$$\mathbf{Q}t$$

$$\begin{array}{c}\quad 0 \qquad 1 \\ \begin{array}{c}0\\1\end{array}\left[\begin{array}{cc} -\beta t & \beta t \\ \beta t & -\beta t \end{array}\right]\end{array}$$

2-state version

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

$$\mathbf{Q}t$$

$$
\begin{array}{c c}
 & \begin{array}{cc} 0 & 1 \end{array} \\
\begin{array}{c} 0 \\ 1 \end{array} &
\begin{bmatrix} -\beta t & \beta t \\ \beta t & -\beta t \end{bmatrix}
\end{array}
$$

$$
\begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}
\begin{bmatrix} 0 & 0 \\ 0 & -2\beta t \end{bmatrix}
\begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}
$$

eigenvector   eigenvalue inverse eigenvector
matrix   matrix   matrix

2-state version

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

$$e^{\mathbf{Q}t}$$

$$
\begin{bmatrix}
\frac{1}{2} + \frac{1}{2}e^{-2\beta t} & \frac{1}{2} - \frac{1}{2}e^{-2\beta t} \\[2ex]
\frac{1}{2} - \frac{1}{2}e^{-2\beta t} & \frac{1}{2} + \frac{1}{2}e^{-2\beta t}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\[2ex]
\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}
\end{bmatrix}
\begin{bmatrix}
e^{0} & 0 \\[2ex]
0 & e^{-2\beta t}
\end{bmatrix}
\begin{bmatrix}
\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\[2ex]
-\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}
\end{bmatrix}
$$

Empirical amino acid models

A different path from Q to P

Once freed from having to derive formulas for transition probabilities, we can use a great variety of Q matrices.

Dayhoff, Schwartz and Orcutt (1978; DSO78) identified 1572 accepted point mutations using closely-related sequences (<15% pairwise divergence), producing this matrix.

Kosiol and Goldman (2005) discussed ways of estimating a Q matrix from these numbers.

		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr
A	Ala																			
R	Arg	30																		
N	Asn	109	17																	
D	Asp	154	0	532																
C	Cys	33	10	0	0															
Q	Gln	93	120	50	76	0														
E	Glu	266	0	94	831	0	422													
G	Gly	579	10	156	162	10	30	112												
H	His	21	103	226	43	10	243	23	10											
I	Ile	66	30	36	13	17	8	35	0	3										
L	Leu	95	17	37	0	0	75	15	17	40	253									
K	Lys	57	477	322	85	0	147	104	60	23	43	39								
M	Met	29	17	0	0	0	20	7	7	0	57	207	90							
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17						
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7					
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269				
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696			
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0		
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17

Figure 80. Numbers of accepted point mutations (X 10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

The elements of Q

The Q matrix is often presented in the following form, factored into a symmetric matrix exchangeabilities and a set of state frequencies.

```
Ala
Arg   0.267828
Asn   0.984474   0.327059
Asp   1.199805   0.000000   8.931515
Cys   0.360016   0.232374   0.000000   0.000000
Gln   0.887753   2.439939   1.028509   1.348551   0.000000
Glu   1.961167   0.000000   1.493409  11.388659   0.000000   7.086022
Gly   2.386111   0.087791   1.385352   1.240981   0.107278   0.281581   0.811907
His   0.228116   2.383148   5.290024   0.868241   0.282729   6.011613   0.439469 ...
Ile   0.653416   0.632629   0.768024   0.239248   0.438074   0.180393   0.609526 ...
Leu   0.406431   0.154924   0.341113   0.000000   0.000000   0.730772   0.112880 ...
Lys   0.258635   4.610124   3.148371   0.716913   0.000000   1.519078   0.830078 ...
Met   0.717840   0.896321   0.000000   0.000000   0.000000   1.127499   0.304803 ...
Phe   0.183641   0.136906   0.138503   0.000000   0.000000   0.000000   0.000000 ...
Pro   2.485920   1.028313   0.419244   0.133940   0.187550   1.526188   0.507003 ...
Ser   4.051870   1.531590   4.885892   0.956097   1.598356   0.561828   0.793999 ...
Thr   3.680365   0.265745   2.271697   0.660930   0.162366   0.525651   0.340156 ...
Trp   0.000000   2.001375   0.224968   0.000000   0.000000   0.000000   0.000000 ...
Tyr   0.244139   0.078012   0.946940   0.000000   0.953164   0.000000   0.214717 ...
Val   2.059564   0.240368   0.158067   0.178316   0.484678   0.346983   0.367250 ...
         Ala        Arg        Asn        Asp        Cys        Gln        Glu ...
```

exchangeabilities

(only values below diagonal shown)

```
Freq 0.087127   0.040904   0.040432   0.046872   0.033474   0.038255   0.049530 ...
```

frequencies

GTR revisited

$$\mu \begin{bmatrix} - & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & - & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & - & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & - \end{bmatrix}$$

The off-diagonal elements of the GTR Q matrix can similarly be obtained by multiplying a symmetric exchangeability matrix and a diagonal matrix of frequencies.

$$\mu \begin{bmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{bmatrix} \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix}$$

exchangeabilities   frequencies

What does all this accomplish?

- An empirical Q matrix can be constructed from many closely-related pairwise comparisons

- A Q matrix can be extrapolated to any desired value of t using P=exp(Qt) to generate a P matrix

- Models generic features of protein evolution; Q matrix does not necessarily reflect your particular sequences

- Frequencies can be swapped with more appropriate set (locally estimated)

Successive improvements

- JTT model (Jones et al. 1992)

    Based on a much larger protein database

- WAG model (Whelan & Goldman 2001)

    Avoids need to use closely-related sequence pairs by obtaining ML estimate of Q matrix

- LG model (Le & Gascuel 2008)

    Add rate heterogeneity to ML estimation of Q matrix

Literature cited

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. 1978. A model of evolutionary change in proteins. Chapter 22 in *Atlas of protein sequence and structure*, vol. 5, suppl. 3. M.O. Dayhoff (ed.), pp. 345-352, Natl. Biomed. Res. Found., Washington, DC

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

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.

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.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

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.

Kosiol C., and Goldman N. 2005. Different versions of the Dayhoff rate matrix. Molecular Biology and Evolution. 22:193–199.

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.

Literature Cited

Le, S.Q., and Gascuel, O. 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution. 25:1307–1320.

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.

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.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17:57-86.

Whelan, S., and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution. 18:691–699.

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

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.