Evolutionary Inferences from Phylogenies: A Review of Methods

Brian C. O'Meara

Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee 37996; email: bomeara@utk.edu, **http://www.brianomeara.info**

Jukes-Cantor Model (JC69)

Q matrix *(instantaneous rates)*

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} -3\beta & \beta & \beta & \beta \\ \beta & -3\beta & \beta & \beta \\ \beta & \beta & -3\beta & \beta \\ \beta & \beta & \beta & -3\beta \end{bmatrix} \end{array}$$

Transition probabilities:

```
> Q
     [,1] [,2] [,3]
[1,] -0.3  0.1  0.2
[2,]  0.0 -0.3  0.3
[3,]  0.0  0.0  0.0

> P(0)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1

> P(10)
      [,1] [,2] [,3]
[1,] 0.05 0.05 0.90
[2,] 0.00 0.05 0.95
[3,] 0.00 0.00 1.00

> P(100)
     [,1] [,2] [,3]
[1,]    0    0    1
[2,]    0    0    1
[3,]    0    0    1
```

$$P = e^{Qt}$$

Matrix exponentiation

JC69 is our most basic model. We will be able to do amazing things with generalizations of this one model!

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \beta & \beta & \beta \\ \beta & - & \beta & \beta \\ \beta & \beta & - & \beta \\ \beta & \beta & \beta & - \end{bmatrix} \end{array}$$

Back to biology - Are JC69's assumptions realistic?

Assumptions of JC69:

1. All substitutions equally likely
2. Base frequencies equal
3. Every site has equal probability of substitution
4. Process is constant through time
5. Sites are independent of each other
6. Substitution is Markovian (memoryless)
7. All sites have the same evolutionary history

Can we relax any of these assumptions?

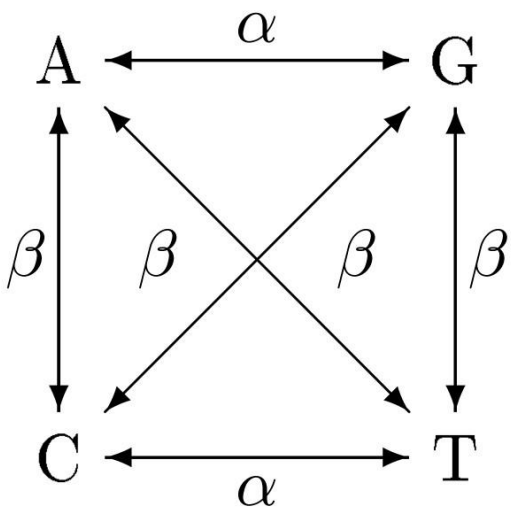
Back to biology - Are JC69's assumptions realistic?

Assumptions of JC69:

1. All substitutions equally likely
2. Base frequencies equal
3. Every site has equal probability of substitution
4. Process is constant through time
5. Sites are independent of each other
6. Substitution is Markovian (memoryless)
7. All sites have the same evolutionary history

Can we relax any of these assumptions?

Kimura 2 Parameter model: K2P



$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{bmatrix} \end{array}$$

Our general CTMC model:

Q matrix

	A	B	C	D	E
A	-	r_{AB}	r_{AC}	r_{AD}	r_{AE}
B	r_{BA}	-	r_{BC}	r_{BD}	r_{BE}
C	r_{CA}	r_{CB}	-	r_{CD}	r_{CE}
D	r_{DA}	r_{DB}	r_{DC}	-	r_{DE}
E	r_{EA}	r_{EB}	r_{EC}	r_{ED}	-

Frequency vector

A	f_A
B	f_B
C	f_C
D	f_D
E	f_E

Differences between statistical phylogenetics and parsimony

Branch lengths in expected # of changes vs. minimum # of changes

Probabilistically incorporate all possible paths, not just the shortest path

Flexibility to modify and compare models

Can easily convert branch lengths to time (w/fossils or other constraints)

All data informative, not just parsimony-informative sites

Representing hypotheses & processes with transition matrices

(Important skill!!)

How many parameters for each?

What do the models suggest about evolution?

A.

	A	G	C	T
A	-	r_{AG}	r_{AC}	r_{AT}
G	r_{AG}	-	r_{GC}	r_{GT}
C	r_{AC}	r_{GC}	-	r_{CT}
T	r_{AT}	r_{GT}	r_{CT}	-

B.

	00	01	11	10
00	-	r_A	0	r_B
01	r_C	-	r_D	0
11	0	r_E	-	r_F
10	r_G	0	r_H	-

C.

	A+	T+	A–	T–
A+	-	r_{AT}	δ	0
T+	r_{TA}	-	0	δ
A–	$k\delta$	0	-	0
T–	0	$k\delta$	0	-

D.

	0	1	2	3
0	-	r_{01}	0	0
1	r_{10}	-	r_{12}	0
2	0	r_{21}	-	r_{23}
3	0	0	r_{32}	-

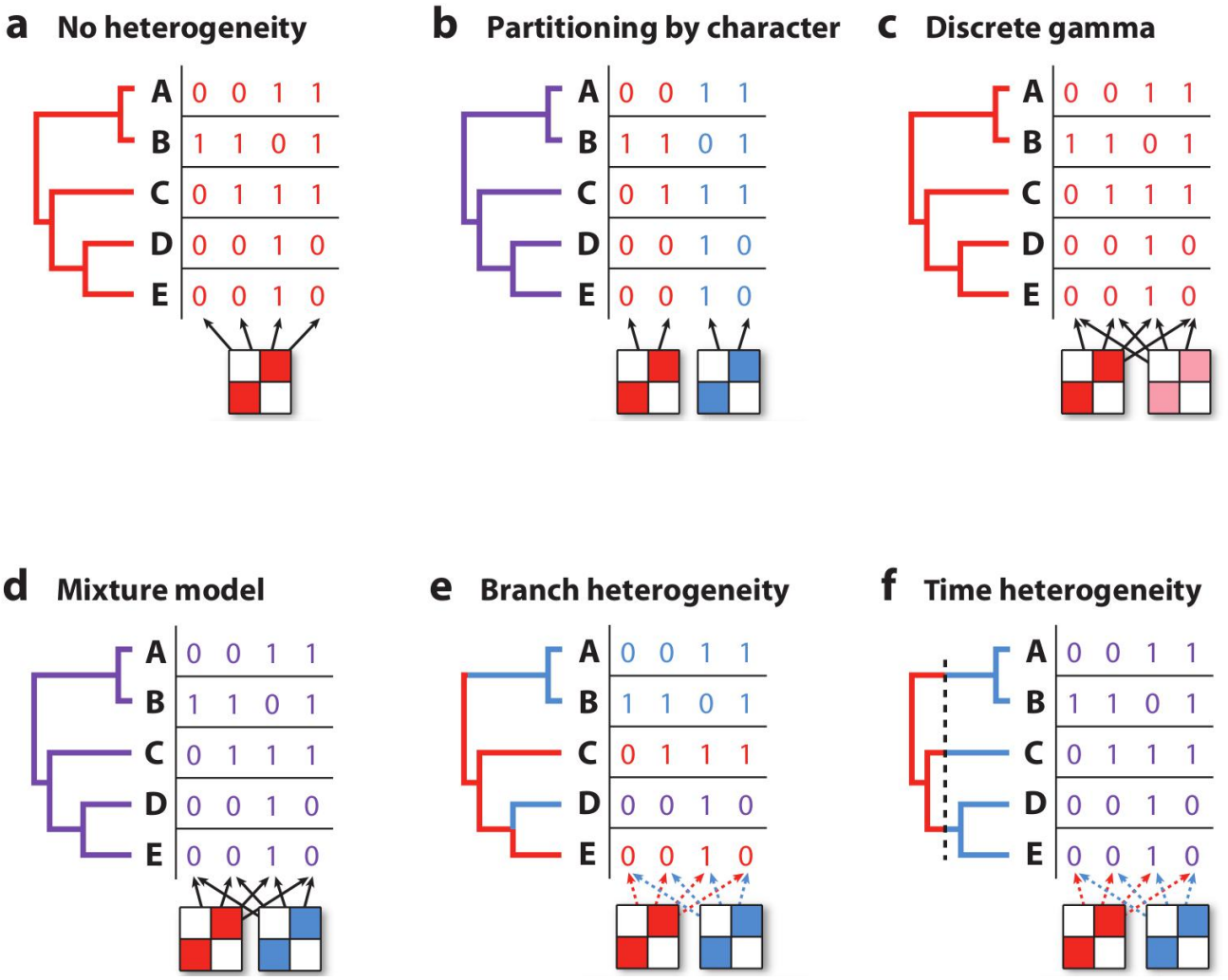
Back to biology - Are JC69's assumptions realistic?

Assumptions of JC69:

1. All substitutions equally likely
2. Base frequencies equal
3. Every site has equal probability of substitution
4. Process is constant through time
5. Sites are independent of each other
6. Substitution is Markovian (memoryless)
7. All sites have the same evolutionary history

Can we relax any of these assumptions?

Dealing with among-site and time-heterogeneity



Molecular Phylogenetics and Evolution 129 (2018) 96–105

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Multi-locus phylogeny of the tribe Tragelaphini (Mammalia, Bovidae) and species delimitation in bushbuck: Evidence for chromosomal speciation mediated by interspecific hybridization

Alexandre Hassanin[a,*], Marlys L. Houck[b], Didier Tshikung[c], Blaise Kadjo[d], Heidi Davis[b], Anne Ropiquet[e]

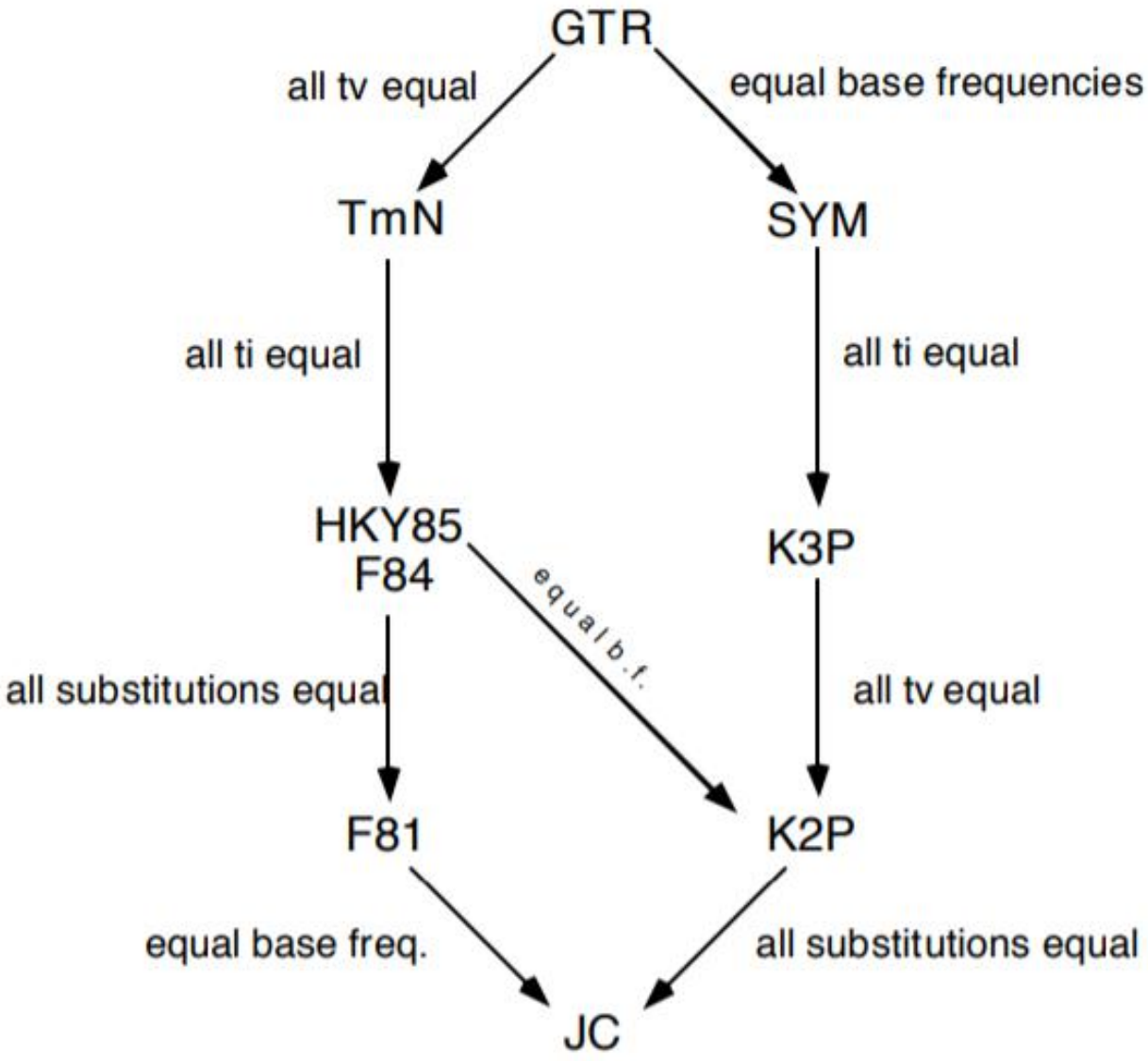
[a] Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE; 57 rue Cuvier, CP 51, 75005 Paris, France
[b] San Diego Zoo Institute for Conservation Research; 15600 San Pasqual Valley Road, Escondido, CA 92027, USA
[c] Faculté de médicine vétérinaire; Université de Lubumbashi, 243 BP 1825, The Democratic Republic of the Congo
[d] Université Félix-Houphouët-Boigny, UFR Biosciences; 22 BP 582, Abidjan 22, Cote d'Ivoire
[e] Middlesex University, Department of Natural Sciences, Faculty of Science and Technology, The Burroughs, Hendon, London NW4 4BT, United Kingdom

2.2. Phylogenetic analyses

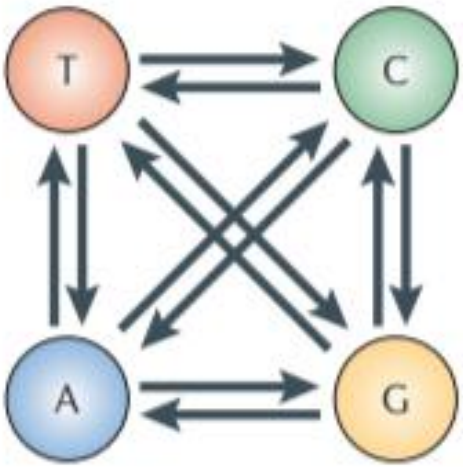
The 22 new mtDNA sequences (cytochrome *b* + control region) of bushbuck generated for this study were aligned on AliView 1.22 (Larsson, 2014) with those published in Moodley and Bruford (2007) and Moodley et al. (2009) and those extracted from the complete mitochondrial genomes published in Hassanin et al. (2012). Our final alignment contains 1901 characters for 206 sequences (available upon request to corresponding author) which includes 183 bushbuck and 23 outgroup taxa of the tribes Caprini, Boselaphini, Bovini, and Tragelaphini. A Bayesian tree was constructed as detailed below for other alignments using the resources available from the CIPRES Science Gateway (Miller et al., 2010).

The phylogeny of the tribe Tragelaphini was investigated by analyzing 17 independent genes (the mtDNA fragment and 16 nuclear genes) on a reduced sample of 30 taxa including seven bushbuck. For each gene, DNA sequences were aligned with AliView 1.22 (Larsson, 2014). The best-fitting models of sequence evolution were found under jModelTest 2.1.7 (Darriba et al., 2012). Using the Akaike Information Criterion (AIC), we selected the K80 + G model for LF (342 nt) and TG (786 nt), the HKY + I model for TUFM (860 nt), the HKY + G model for CSN3 (418 nt), PRKCI (525 nt), SPTBN1 (599 nt) and ZFYVE27 (701 nt), the GTR model for EXOSC9 (993 nt), HDAC2 (722 nt), and PABPN1 (896 nt), the GTR + I model for DIS3 (851 nt), FGB (695 nt), LALBA (494 nt), and RIOK3 (722 nt), the GTR + G model for CCAR1 (1031 nt) and CHPF2 (890 nt), and the GTR + I + G model for the mtDNA fragment (1901 nt) and the concatenation of the 16 nuclear markers (nuDNA, 11,525 nt).

Figure 11 from Swofford et al. (1996).

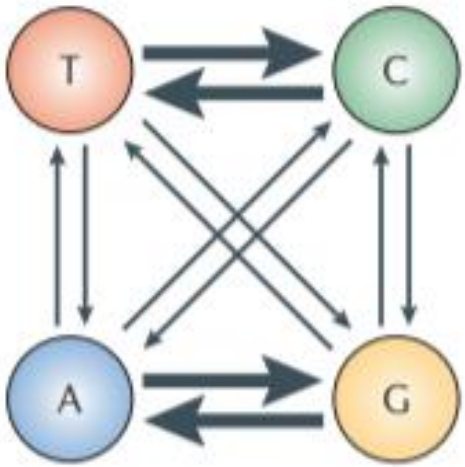


JC69



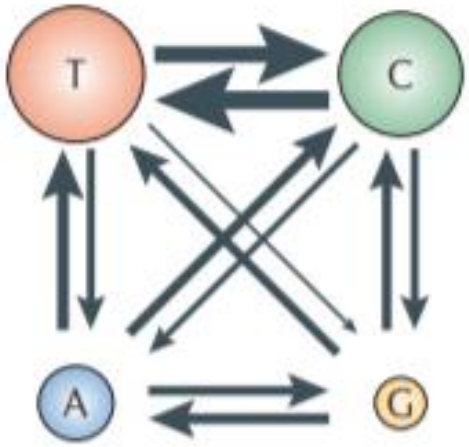
$$\begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha. \end{pmatrix}$$

K80



$$\begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix}$$

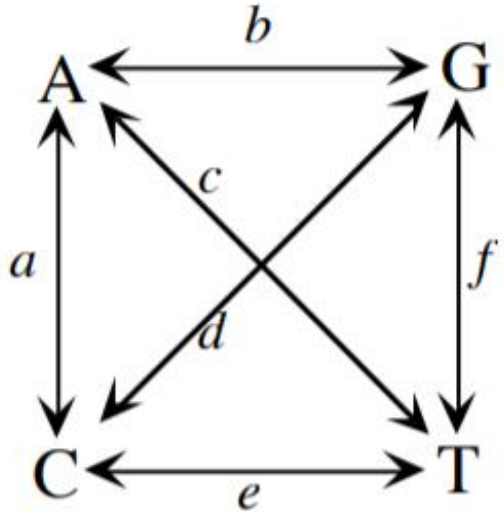
HKY85



$$\begin{pmatrix} -\mu(\kappa\pi_G + \pi_Y) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T + \pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_R) \end{pmatrix}$$

General Time-Reversible Model

$$\mathbf{Q} = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G - e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{bmatrix}$$



General Time-Reversible Model

$$\mathbf{R} = \begin{matrix} --- & \mu a & \mu b & \mu c \\ \mu a & --- & \mu d & \mu e \\ \mu b & \mu d & --- & \mu f \\ \mu c & \mu e & \mu f & --- \end{matrix}$$

and

$$\mathbf{\Pi} = \begin{matrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{matrix}$$

Among-site rate variation

GENE NAME: ATP1a1

Human	GAT	AAC	TCC	TCG	CTC	ACT	GGT	GAA	TCA	GAA	CCC	CAG	ACT	AGG	TCT	CCA	GAT	TTC	ACA	AAT	GAA	AAC	CCC	CTG	GAG	ACG
Human 2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Mouse	...	...	...	..A	...	...	...	...	...	...	...	...	...	...	C..	..C	..G	...	...	...	..C	..G	...	...	T..	...	..A
Rat	...	...	...	..A	...	...	...	...	...	...	...	...	...	...	C..	..C	..G	...	...	...	..C	..G	...	...	T..	...	..A
Rabbit	...	...	...	..A	...	...	...	...	...	...	...	...	...	..C	...	...	..G	...	..T	...	...	...	...	..G	...	...	...
Wolf	...	...	...	...	...	...	...	...	...	..G	...	...	...	..A	...	...	..C	...	...	...	...	...	...	...	...	..A	...
Pig	..C	...	...	..C	...	...	...	...	...	...	..G	...	..C	...	...	...	...	...	...	..C	...	..G	...	...	...	...	..T
Chicken	...	...	...	...	..T	...	...	..G	...	..G	..T	...	..C	...	...	...	...	...	T.C	...	..G	...	..A	...	..A	..T
African Clawed Frog	...	...	..T	..C	...	...	...	..G	..C	..G	...	...	..C	C.C	...	..T	..C	...	..C	...	..G	...	..G	...	...	..T
Zebrafish	..C	...	...	..C	...	...	..A	...	..T	...	...	...	...	C.T	A..	..T	..C	..T	T.C	...	..T	...	..A	...	...	..C

Human	Asp	Asn	Ser	Ser	Leu	Thr	Gly	Glu	Ser	Glu	Pro	Gln	Thr	Arg	Ser	Pro	Asp	Phe	Thr	Asn	Glu	Asn	Pro	Leu	Glu	Thr
Human 2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Mouse	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Rat	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Rabbit	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Wolf	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Pig	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Chicken	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Ser	.	.	.	.	.	.	.
African Clawed Frog	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Zebrafish	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Thr	.	.	.	Ser	.	Asp	.	.	.	.	.

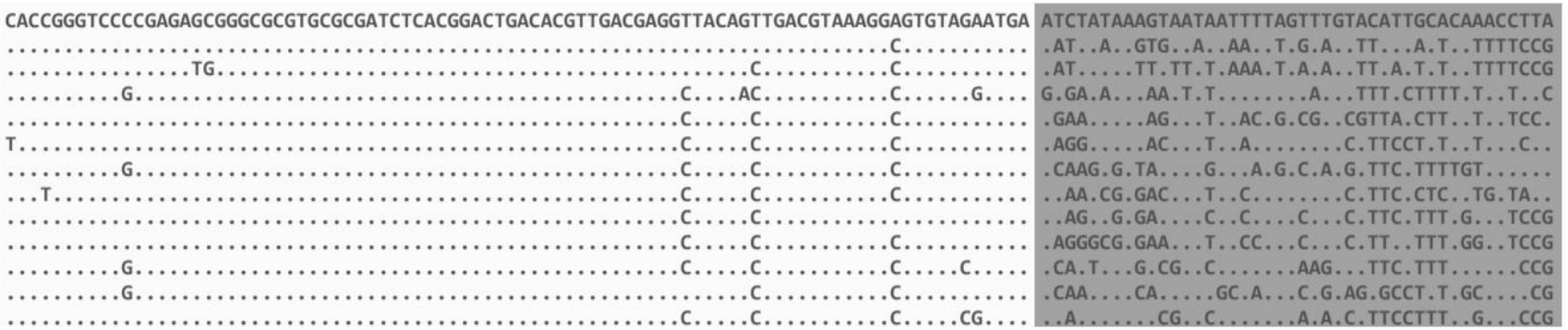
GENE NAME: FOXp2

Mouse	ACC	ACG	TCC	AAA	GCA	TCA	CCA	CCC	ATC	ACA	CAT	CAT	TCC	ATA	GTG	AAC	GGA	CAG	TCT	TCA	GTT	CTG	AAT	GCA	AGG	CGG
Macaque	...	..T	...	...	...	...	...	..A	..A	..T	...	...	...	...	...	..T	...	...	...	...	...	..A	...	...	..A	..A
Baboon	...	..T	...	...	...	...	...	..A	..A	..T	...	...	...	...	...	..T	...	...	...	...	...	..A	...	...	..A	..A
Orangutan	...	..T	...	...	..G	...	...	..A	..A	..T	...	...	...	..C	...	..T	...	...	...	...	...	..A	...	...	..A	..A
Gorilla	...	..T	...	...	...	...	...	..A	..A	..T	...	...	...	...	...	..T	...	...	...	...	...	..A	...	...	..A	..A
Bonobo	...	..T	...	...	..G	...	...	..A	..A	..T	...	...	...	..C	...	..T	...	...	...	...	...	..A	...	...	..A	..A
Chimpanzee	...	..T	...	...	..G	...	...	..A	..A	..T	...	...	...	..C	...	..T	...	...	...	...	...	..A	...	...	..A	..A
Human	.A.	..T	...	...	...	...	...	..A	..A	..T	...	...	...	...	...	..T	...	...	...	...	...	..A	.G.	...	..A	..A

Mouse	Ser	Thr	Thr	Ser	Lys	Ala	Ser	Pro	Pro	Ile	Thr	His	His	Ser	Ile	Val	Asn	Gly	Gln	Ser	Ser	Val	Leu	Asn	Ala	Arg
Macaque	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Baboon	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Orangutan	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Gorilla	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Bonobo	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Chimpanzee	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human	Asn	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Ser	.	.	.

Site-specific rates

Each defined subset (e.g. 1st+2nd pos. versus 3rd pos.) has its own relative rate



CACCGGGTCCCCGAGAGCGGGCGCGTGCGCGATCTCACGGACTGACACGTTGACGAGGTTACAGTTGACGTAAAGGAGTGTAGAATGA ATCTATAAAGTAATAATTTTAGTTTGTACATTGCACAAACCTTA
.....C..... .AT..A..GTG..A..AA..T.G.A..TT...A.T..TTTTCCG
.....TG.....C.....C..... .AT.....TT.TT.T.AAA.T.A.A..TT.A.T.T..TTTTCCG
.....G.....C....AC.....C.....G.... G.GA.A...AA.T.T........A...TTT.CTTTT.T..T..C
.....C.....C.....C..... .GAA.....AG...T..AC.G.CG..CGTTA.CTT..T..TCC.
T.....C.....C.....C..... .AGG.....AC...T..A........C.TTCCT.T..T...C..
.....G.....C.....C.....C..... .CAAG.G.TA....G...A.G.C.A.G.TTC.TTTTGT......
...T.....C.....C.....C..... ..AA.CG.GAC...T..C........C.TTC.CTC..TG.TA..
.....C.....C..... ..AG..G.GA....C..C....C...C.TTC.TTT.G...TCCG
.....C.....C.....C..... .AGGGCG.GAA...T..CC...C...C.TT..TTT.GG..TCCG
.....G.....C.....C.....C.....C..... .CA.T...G.CG..C.......AAG...TTC.TTT......CCG
.....G.....C.....C..... .CAA....CA.....GC.A...C.G.AG.GCCT.T.GC....CG
.....C.....C.....C.....CG.... ..A.......CG..C.......A.A.C.TTCCTTT..G...CCG

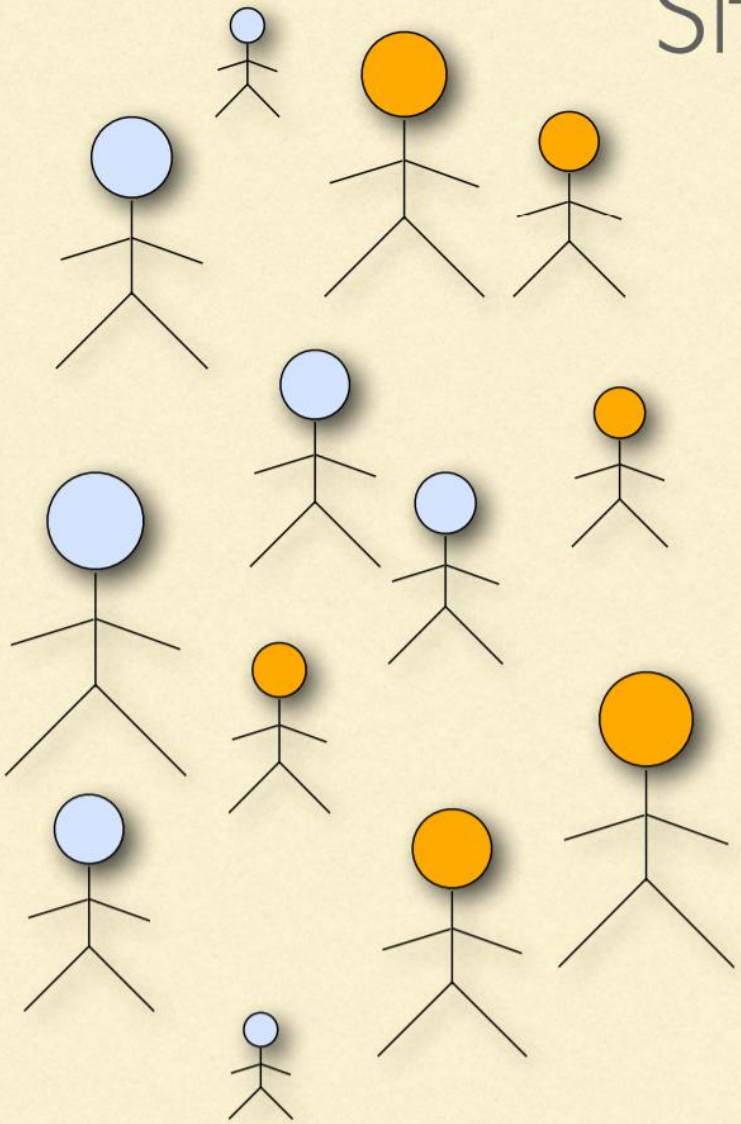
r_1 applies to subset 1

1st+2nd codon positions
(sites 1 - 88)

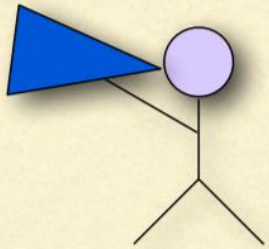
r_2 applies to subset 2

3rd codon positions
(sites 89-132)

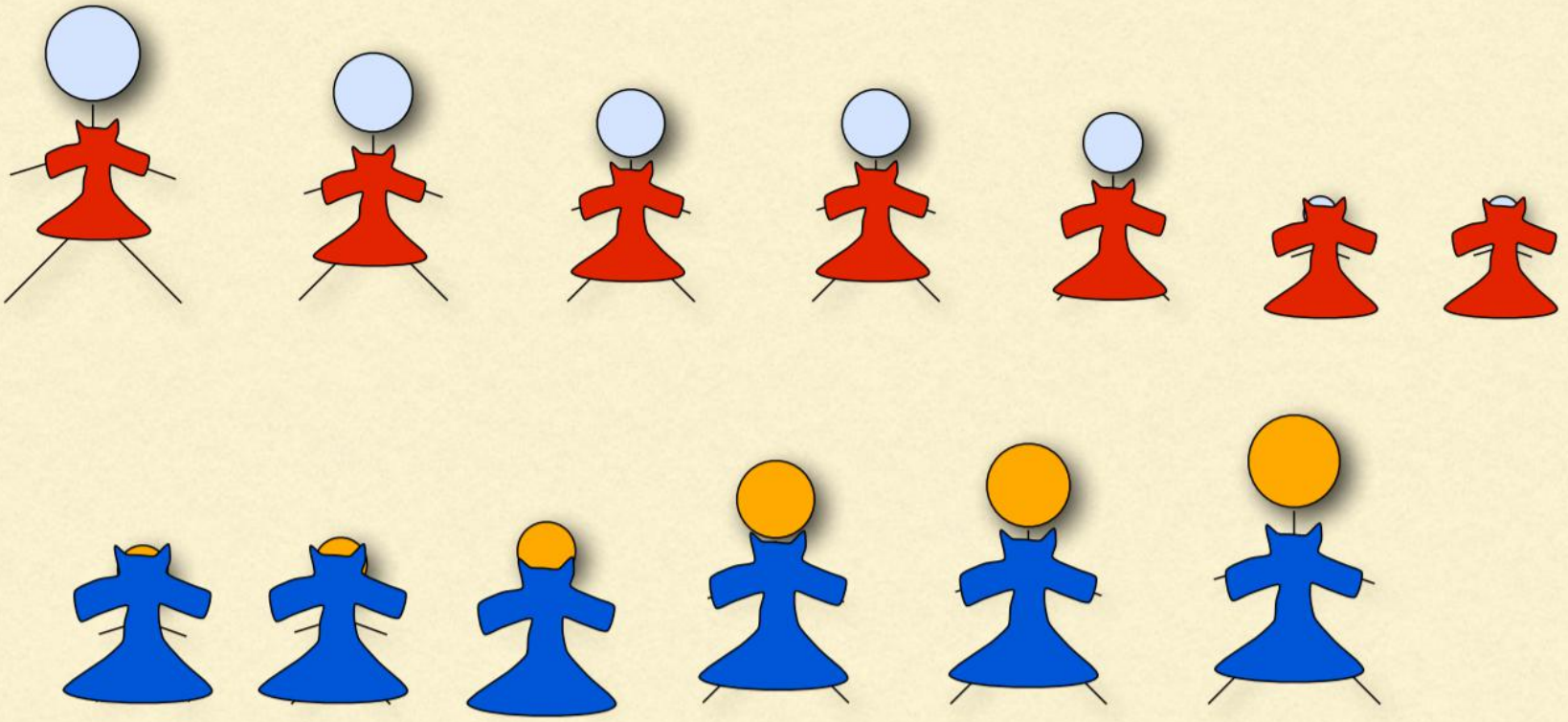
Site-specific approach



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.

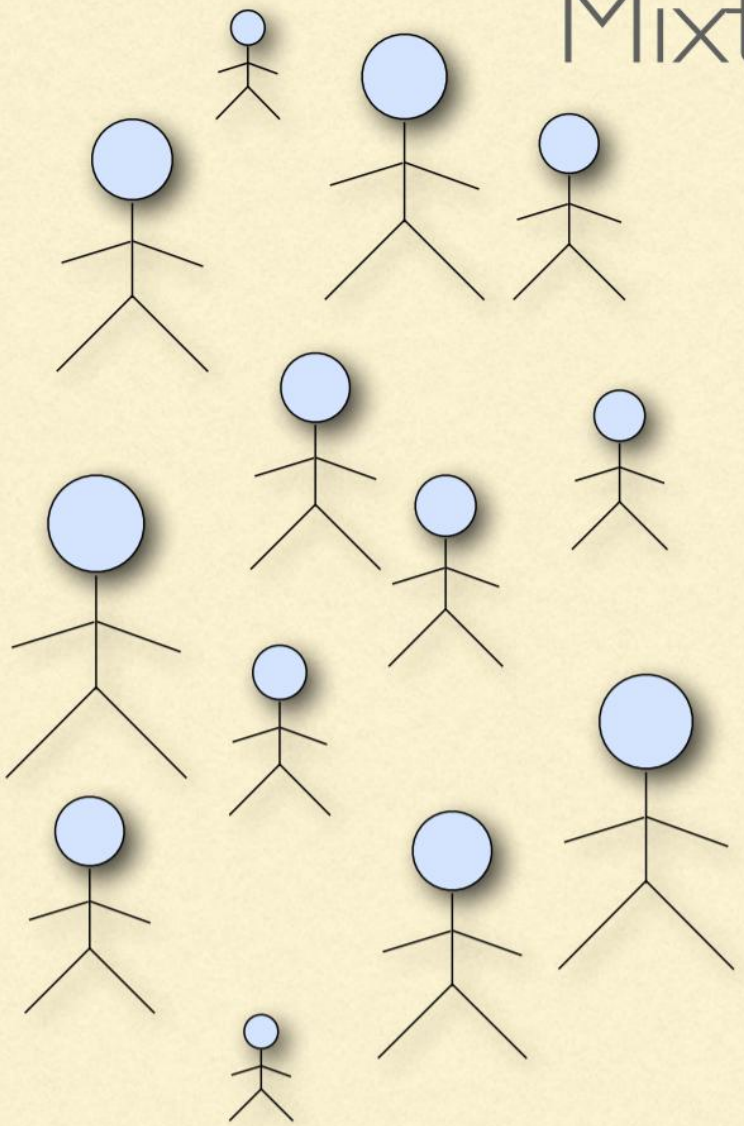


Site-specific approach

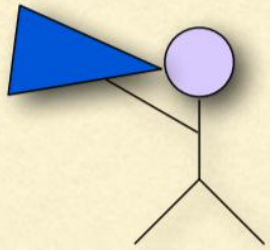


Good: costs less: need to buy just one coat for every person
Bad: every person in a group has to wear the same size coat

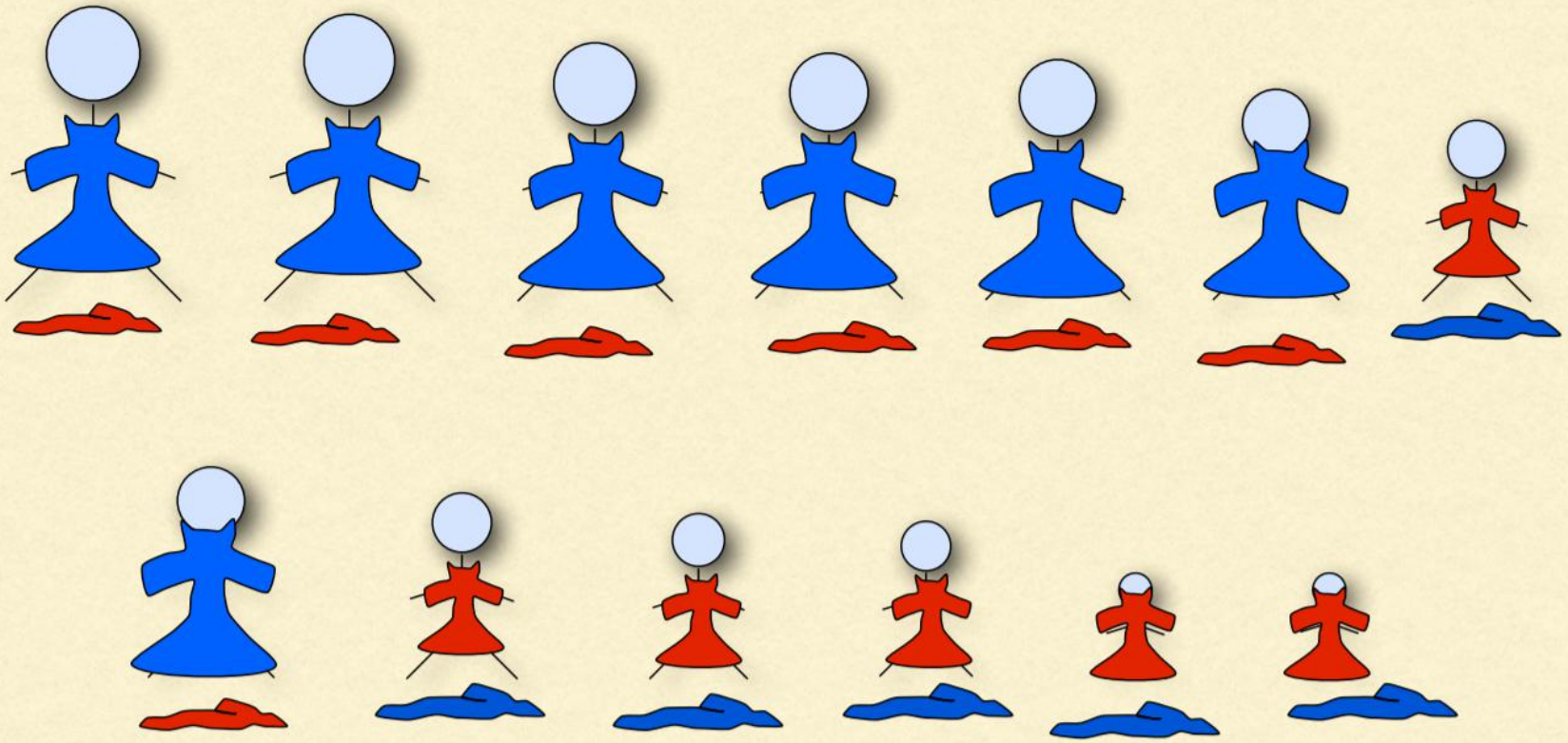
Mixture model approach



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



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

Every site has a probability, P_{invar} of being *invariant*

P_{invar} estimated from the data

Every site gets its likelihood estimated in two parts: If it were invariant & variable (as normal)

Often designated "+I" (e.g. HKY85 + I)

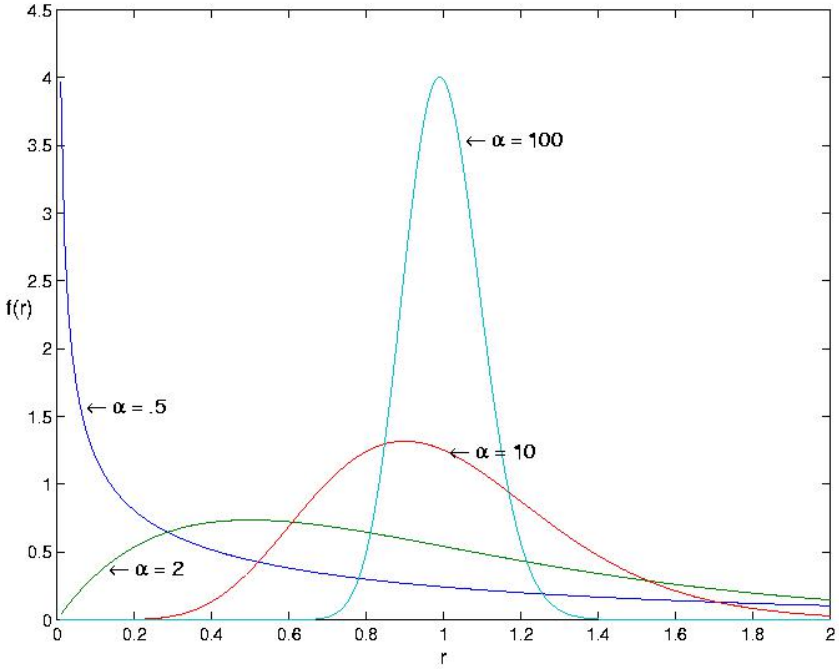
Gamma-distributed rates

Rate categories set by a discrete gamma distribution (e.g. 4 rate categories)

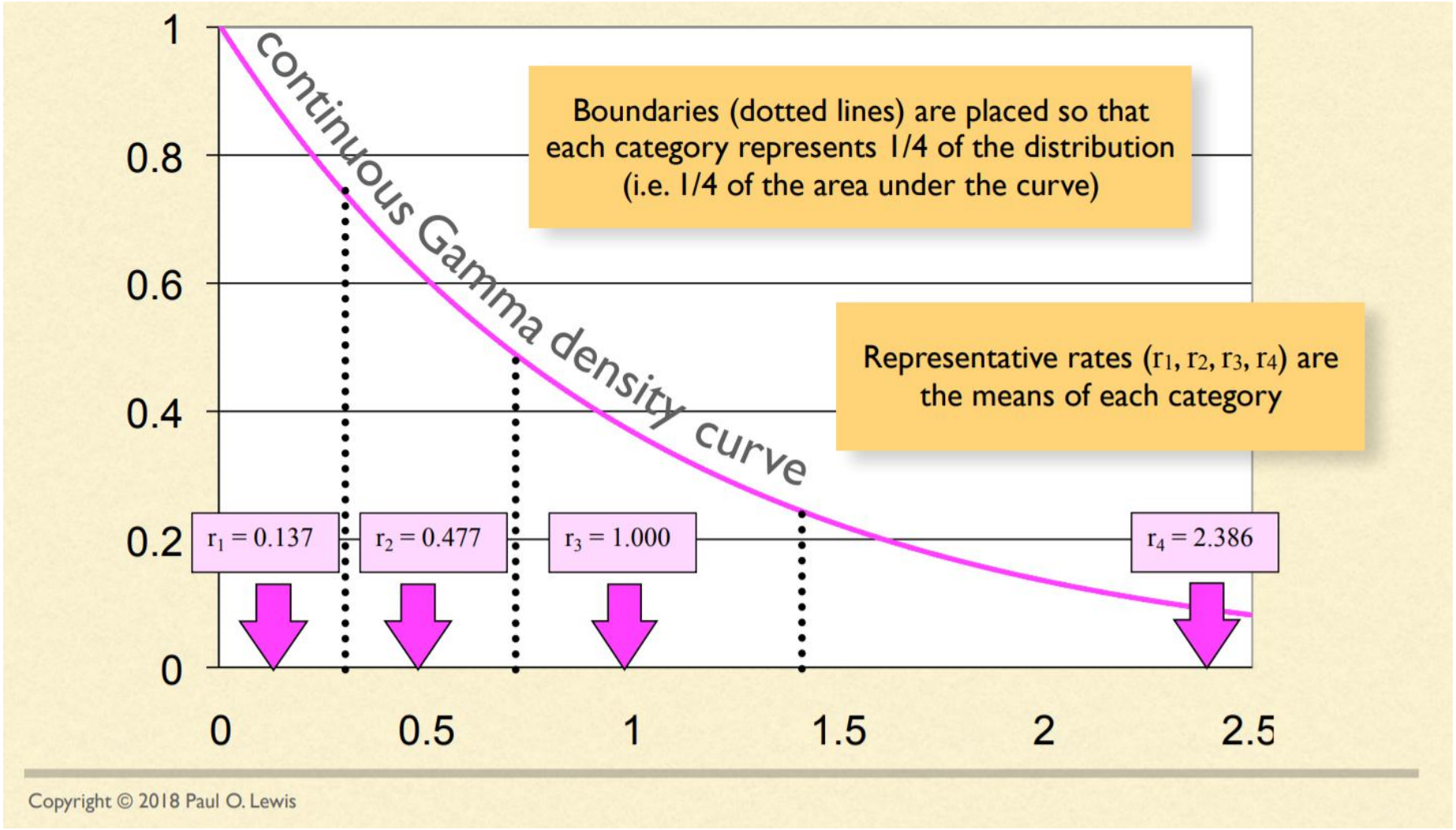
Estimate *α*

Site likelihoods calculated for each rate category and summed

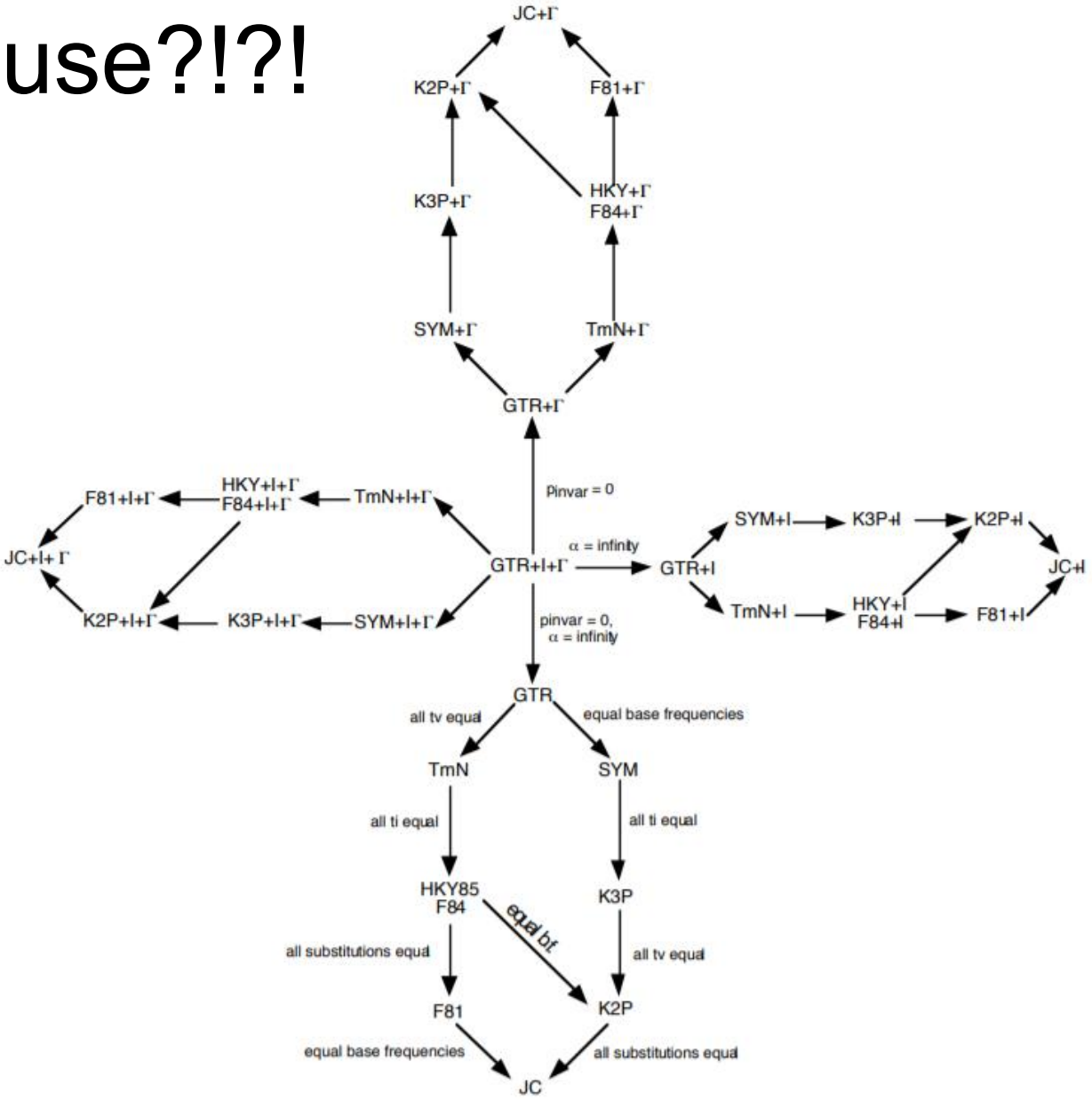
" + G " or “ + Γ ”
(e.g. HKY85 + I + G)



Discretizing the Gamma distribution



OK...but how do we decide which model to use?!?!



Model testing

Ideally, fit all models while simultaneously estimating tree

More practical:

Start with a not-too-bad tree, fit all the possible models, rank them by some criterion, pick the best one and use in a full analysis

Criteria: Likelihood ratio test, Akaike Information Criterion, Bayesian Information Criterion, Decision Theory, Bayesian model selection, Bayesian hypothesis testing or model averaging