Modeling phylogenetic change - Wrap up from last time

Breakout Groups

Questions:

1. Explain how you get the likelihood of a linear regression

Ex: X = [3,1,4]; Y = [5,2,9]; m = 1, b = 0, σ^2=1
What's lnL? What does Maximum Likelihood mean?

2. What's the difference between Bayesian & Likelihood methods?

How does it differ in the Breast Cancer/Cowboy hat example and why?
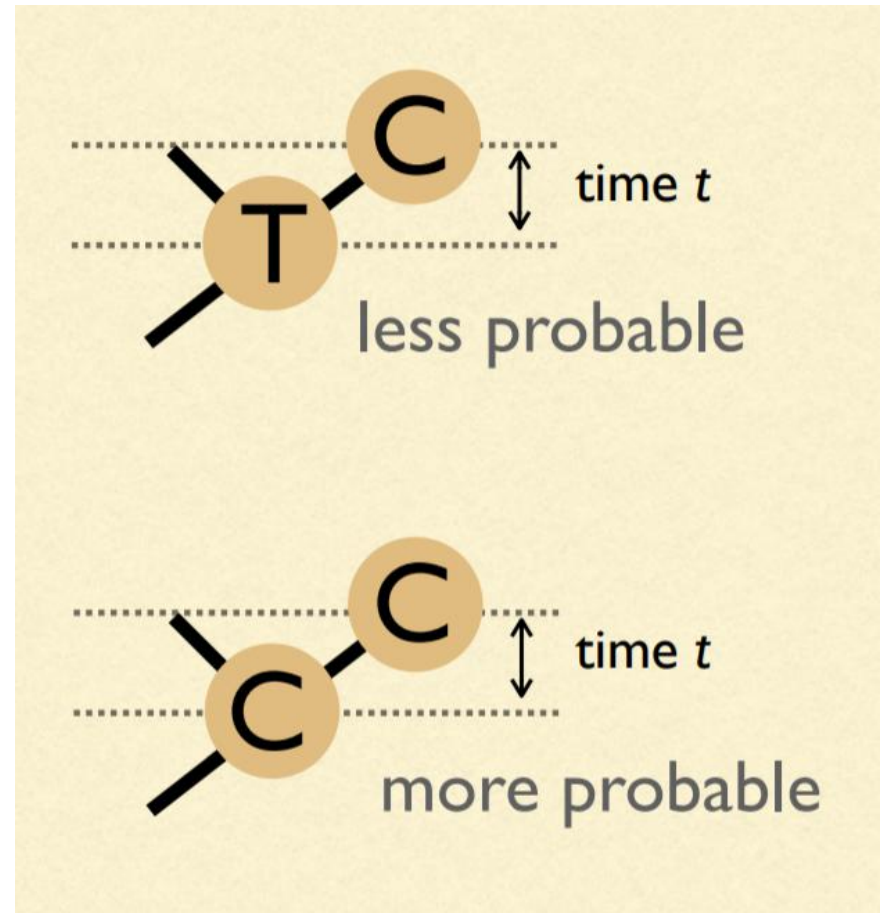
Continuous-Time Markov Models

Markov assumption = probability of change depends only on current state, not how long it has been in that state
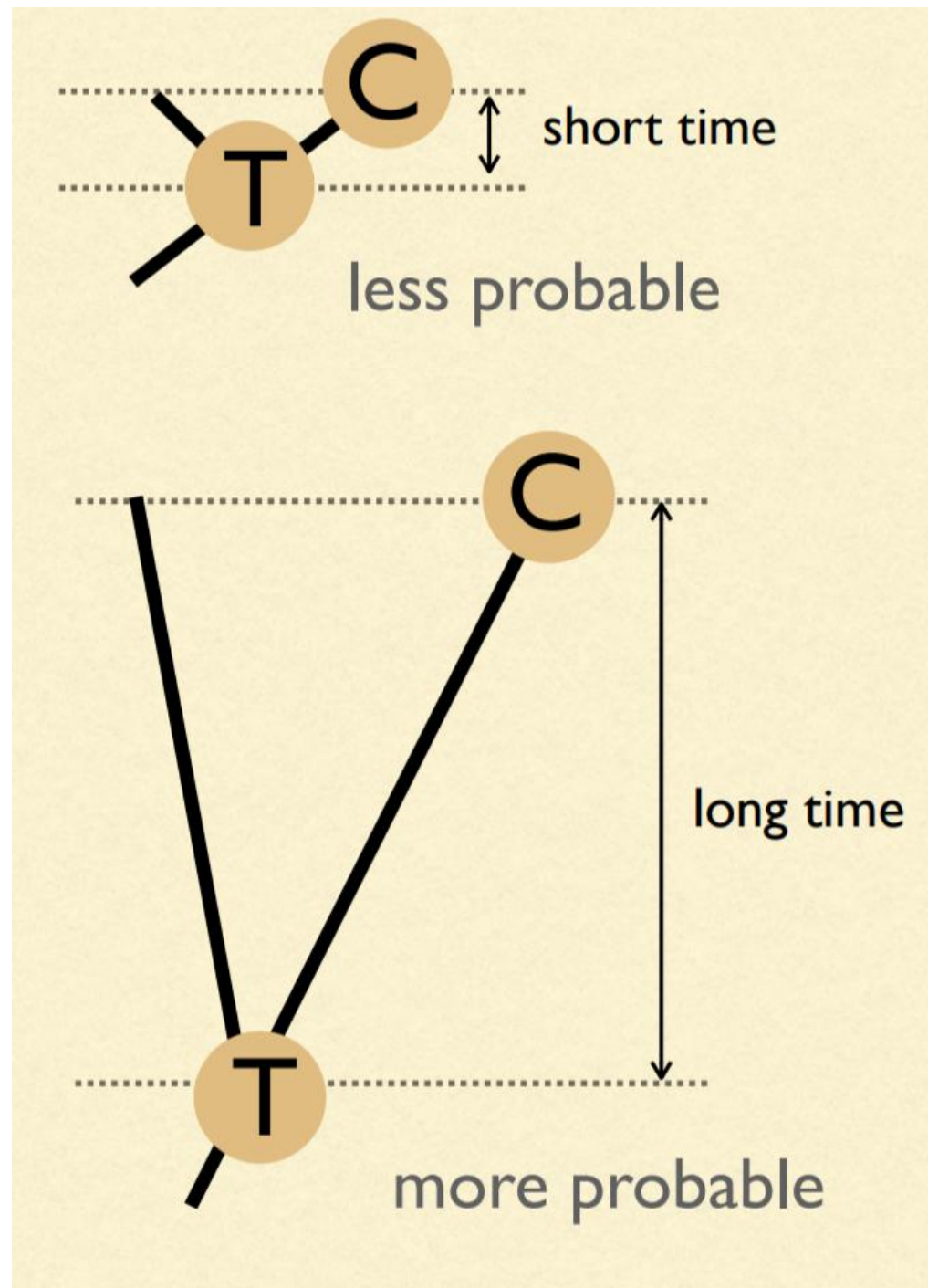
Our model of change depends on time:
We must estimate branch lengths

Units of branch length
will be expected
number of
substitutions per site
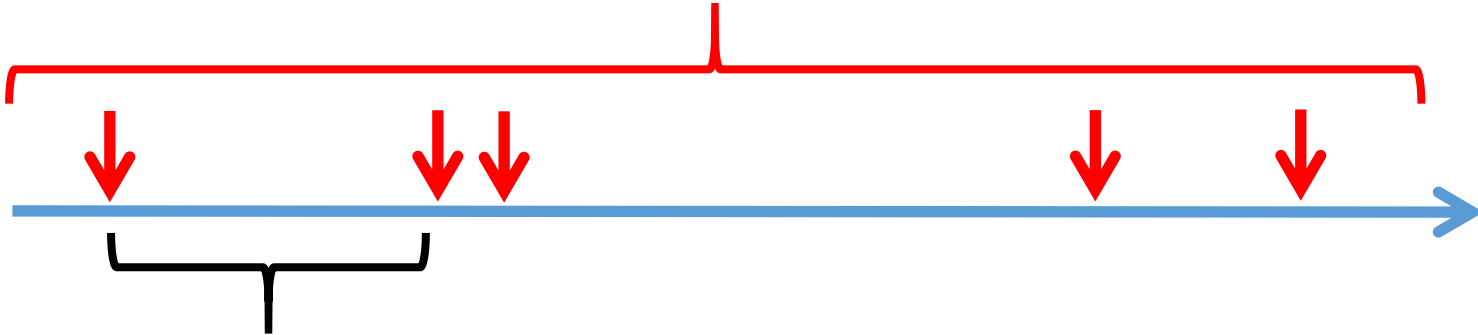
(= rate of substitions x
time)

Probabilities are dependent on time

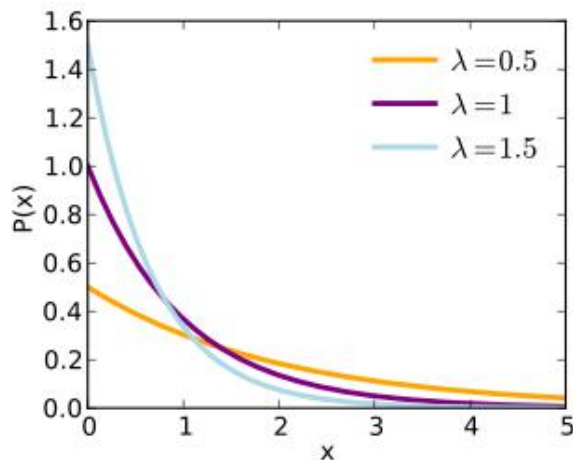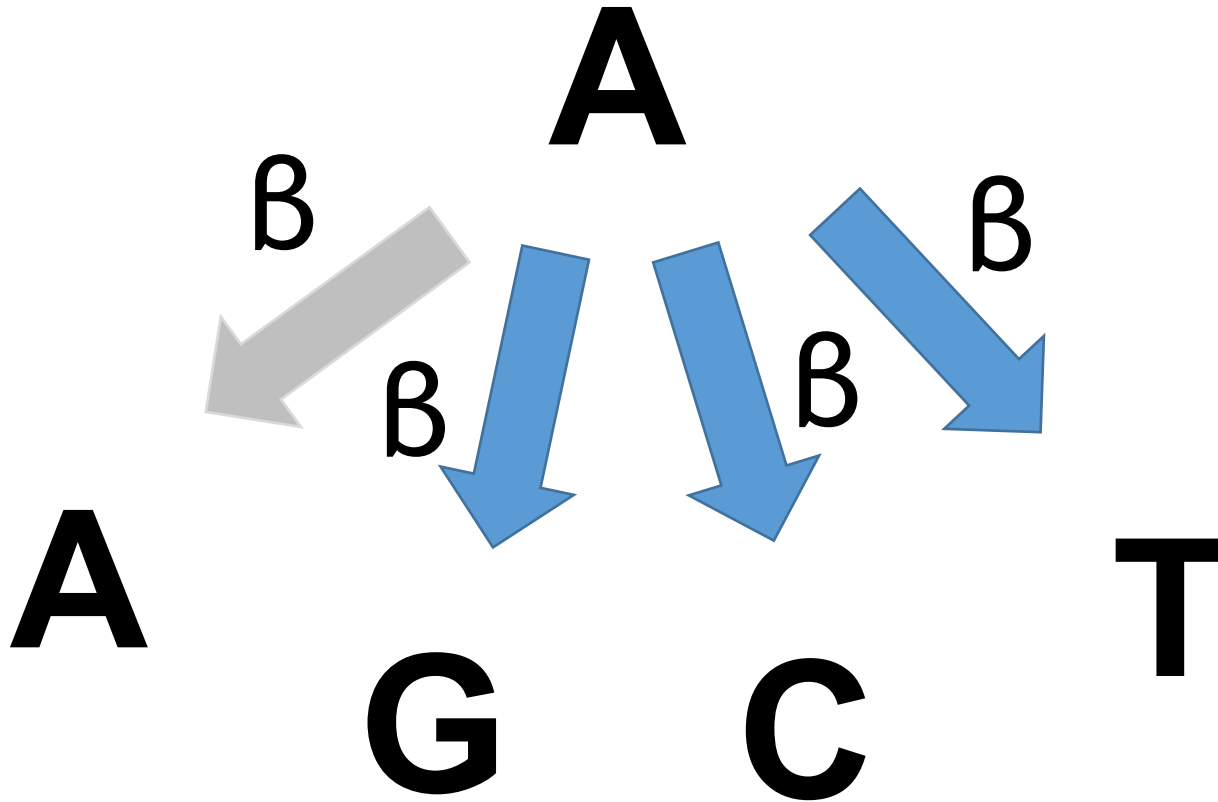# Poisson Process

Expected number of events ~
Poisson (λt)

Interinterval time ~ Exponential (λ)
Mean = 1/λ

A (very) simple phylogeny...

time

A →————————————→ **A**

P_{AA} =

Probability nothing happened +

Probability something happened, but that the last thing that happened ended in an A

A (very) simple phylogeny...

time

A $\longrightarrow$ **A**

$P_{AA} = \quad (e^{-\mu t}) \quad + \quad (1 - e^{-\mu t})(1/4)$

Probability something doesn't happen

Probability at least one thing happens

Probability that the last thing that happened ends in an A

$(\pi_A = \pi_G = \pi_C = \pi_T = 1/4)$

A (very) simple phylogeny...

time

A → **G**

$$P_{AG} = (1 - e^{-\mu t})(1/4)$$

Probability at least one thing happens

Probability that the last thing that happened ends in an G

A (very) simple phylogeny...

time

A $\longrightarrow$ C

P_{AC} = $(1 - e^{-\mu t})(1/4)$

Probability at least one thing happens

Probability that the last thing that happened ends in an C

A (very) simple phylogeny...

time

A $\longrightarrow$ T

P_{AT} = $(1 - e^{-\mu t})(1/4)$

Probability at least one thing happens

Probability that the last thing that happened ends in an T

One last bit...substitutions vs. "events"

$$v = (3/4)\mu t = 3\beta t$$

$$4v/3 = \mu t$$

Only 3 out of 4 events results in a substitution. Thus, we can define the substitution rate v.

$$P_{AA} = (e^{-\mu t}) + (1 - e^{-\mu t})(1/4)$$

$$P_{AG} = (1 - e^{-\mu t})(1/4)$$

$$P_{AC} = (1 - e^{-\mu t})(1/4)$$

$$P_{AT} = (1 - e^{-\mu t})(1/4)$$

$$P_{AA} = (1/4) + (3/4)(e^{-4v/3})$$

$$P_{AG} = (1/4) - (1/4)(e^{-4v/3})$$
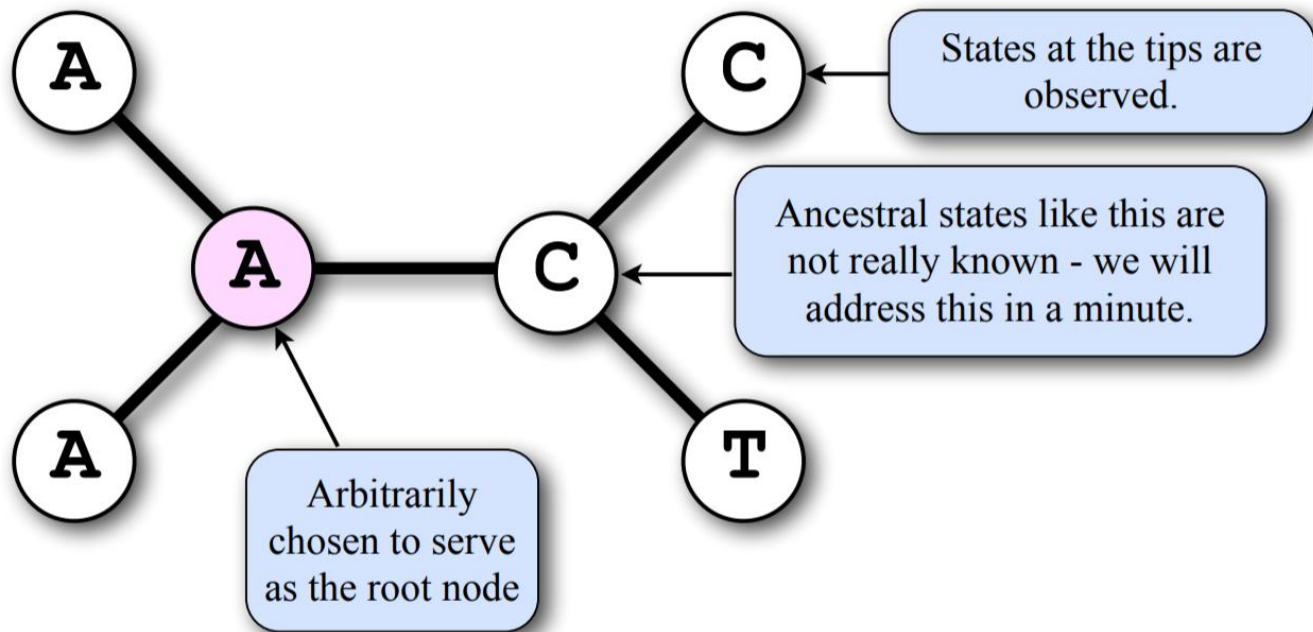
$$P_{AC} = (1/4) - (1/4)(e^{-4v/3})$$

$$P_{AT} = (1/4) - (1/4)(e^{-4v/3})$$

Sanity check:
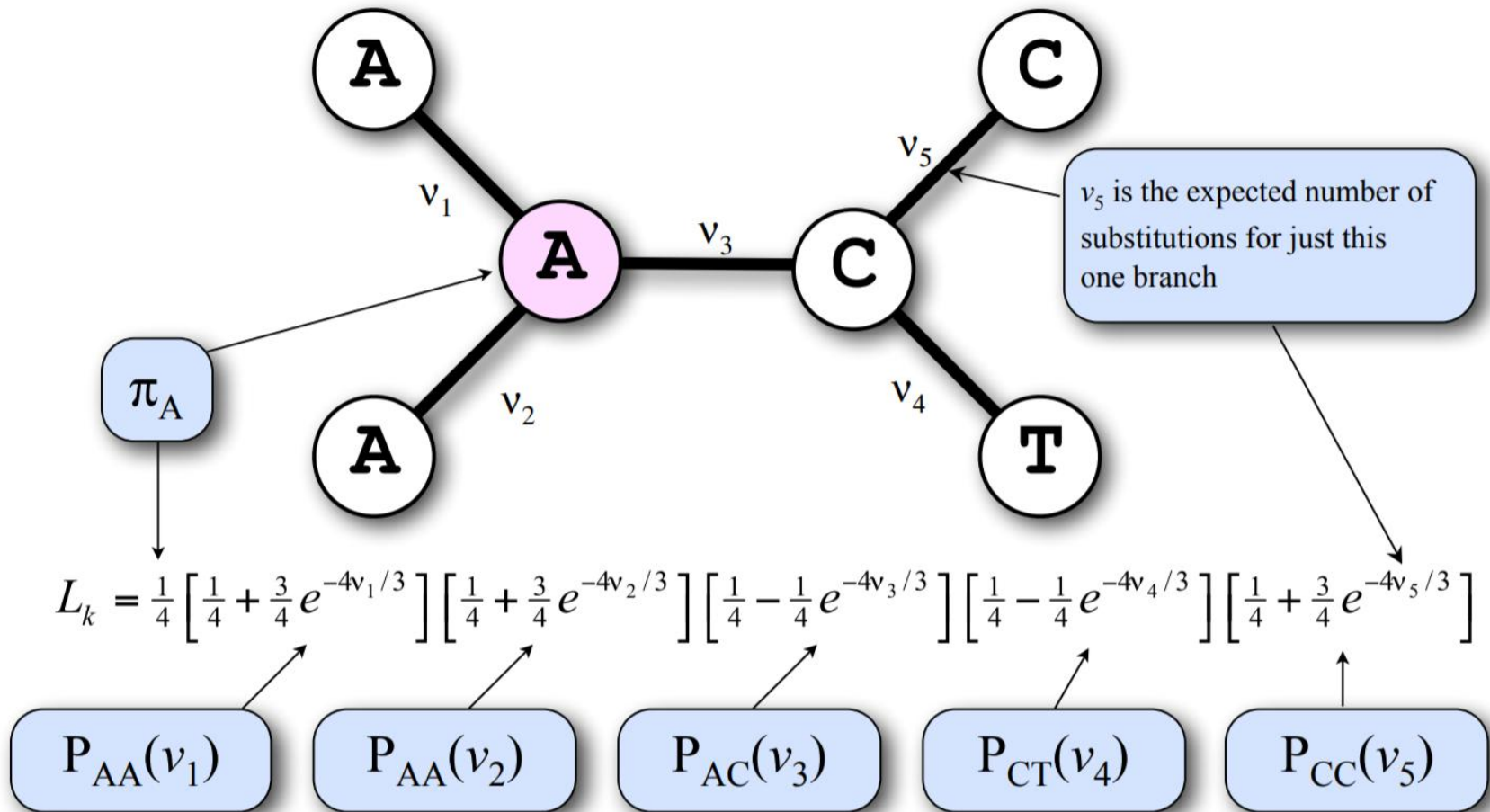Do they all add to 1?

Likelihood of an unrooted tree

(data shown for only one site)

States at the tips are observed.

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

Arbitrarily chosen to serve as the root node

Likelihood for site k

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

π_A

$$L_k = \frac{1}{4}\left[\frac{1}{4} + \frac{3}{4}e^{-4v_1/3}\right]\left[\frac{1}{4} + \frac{3}{4}e^{-4v_2/3}\right]\left[\frac{1}{4} - \frac{1}{4}e^{-4v_3/3}\right]\left[\frac{1}{4} - \frac{1}{4}e^{-4v_4/3}\right]\left[\frac{1}{4} + \frac{3}{4}e^{-4v_5/3}\right]$$

$P_{AA}(v_1)$ $P_{AA}(v_2)$ $P_{AC}(v_3)$ $P_{CT}(v_4)$ $P_{CC}(v_5)$
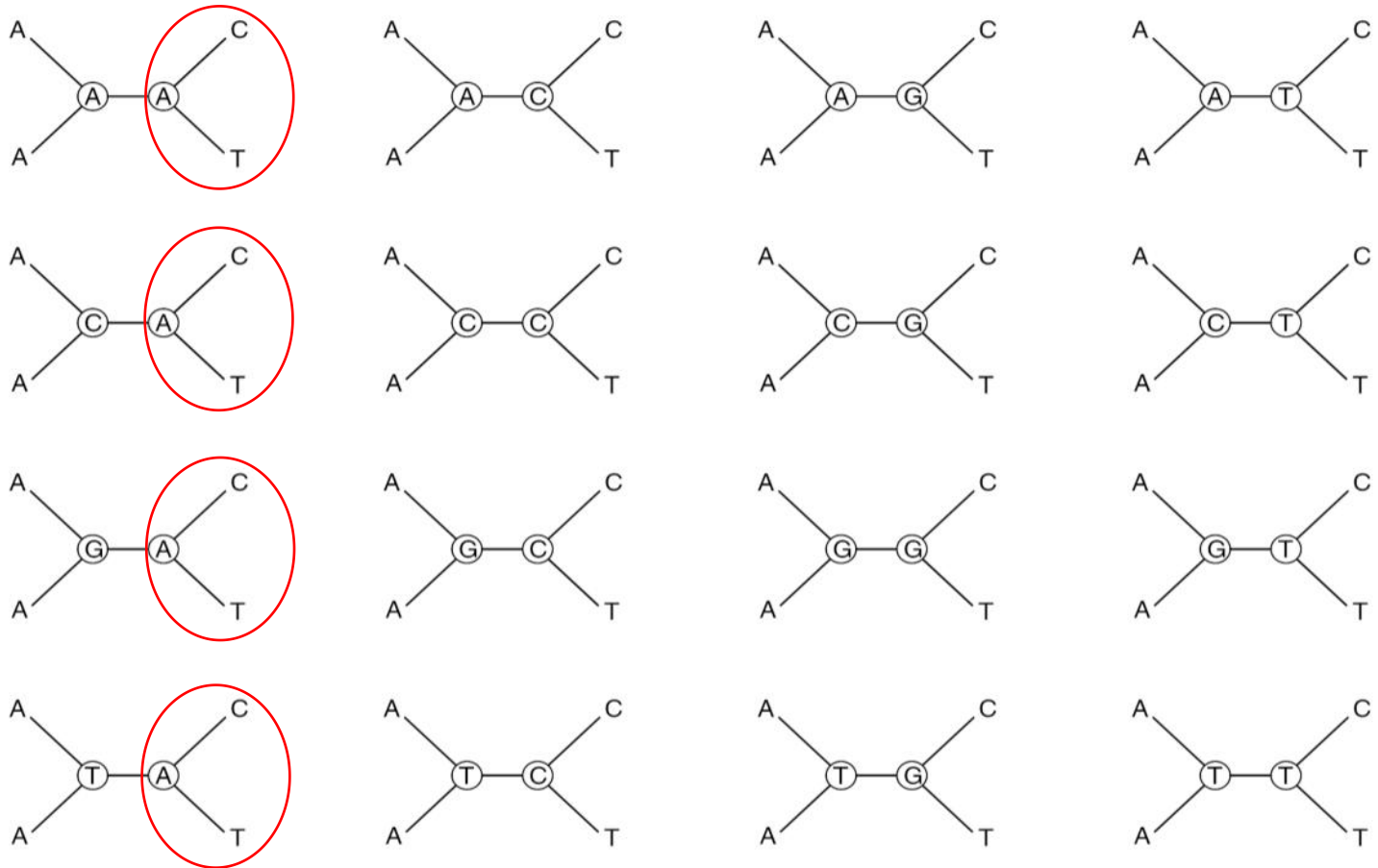
Note use of the AND probability rule

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them
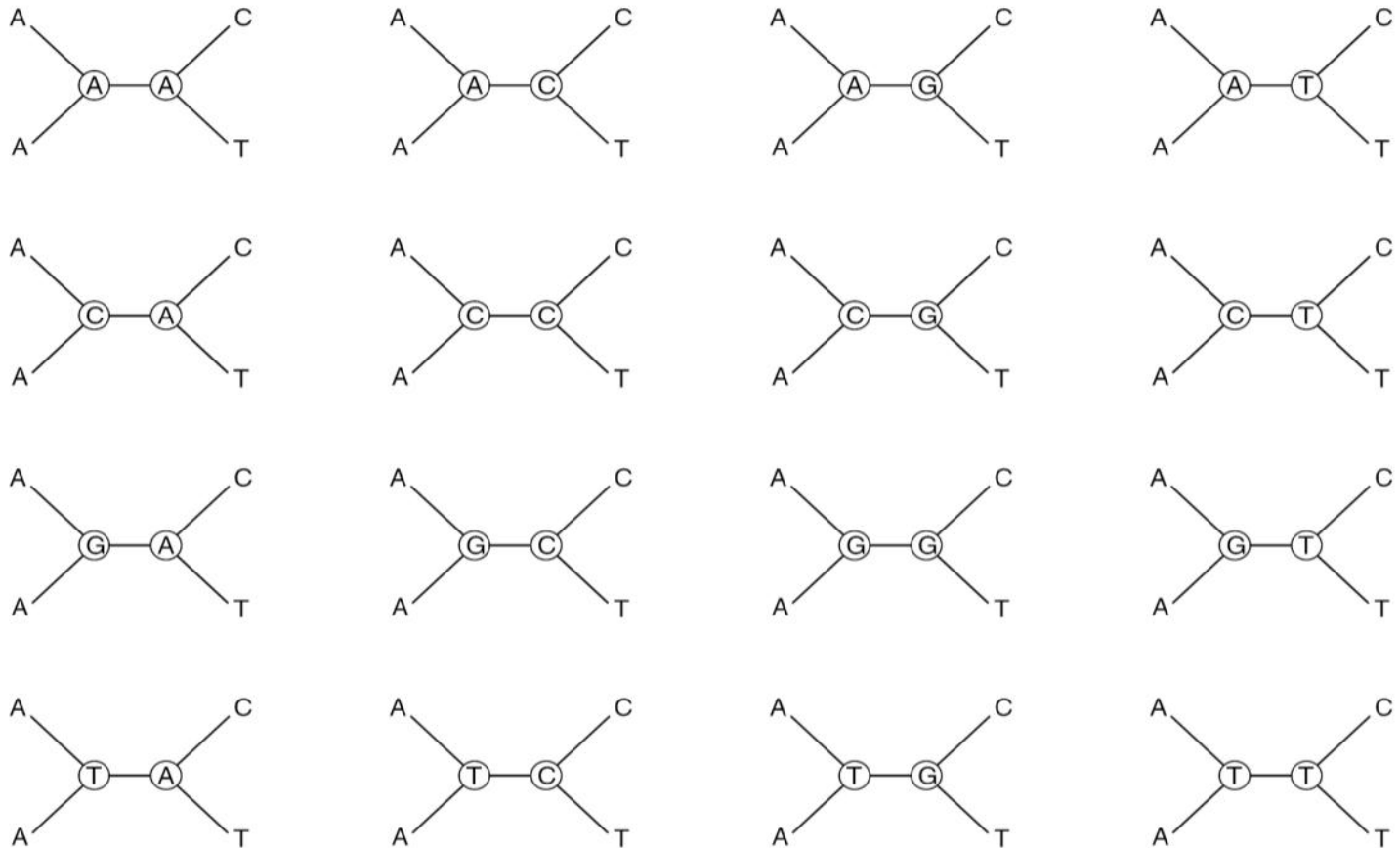
Note use of the OR probability rule

Pruning algorithm

Note use of the OR probability rule

Pruning algorithm

Note use of the OR probability rule

Pruning algorithm

Note use of the OR probability rule

Pruning algorithm

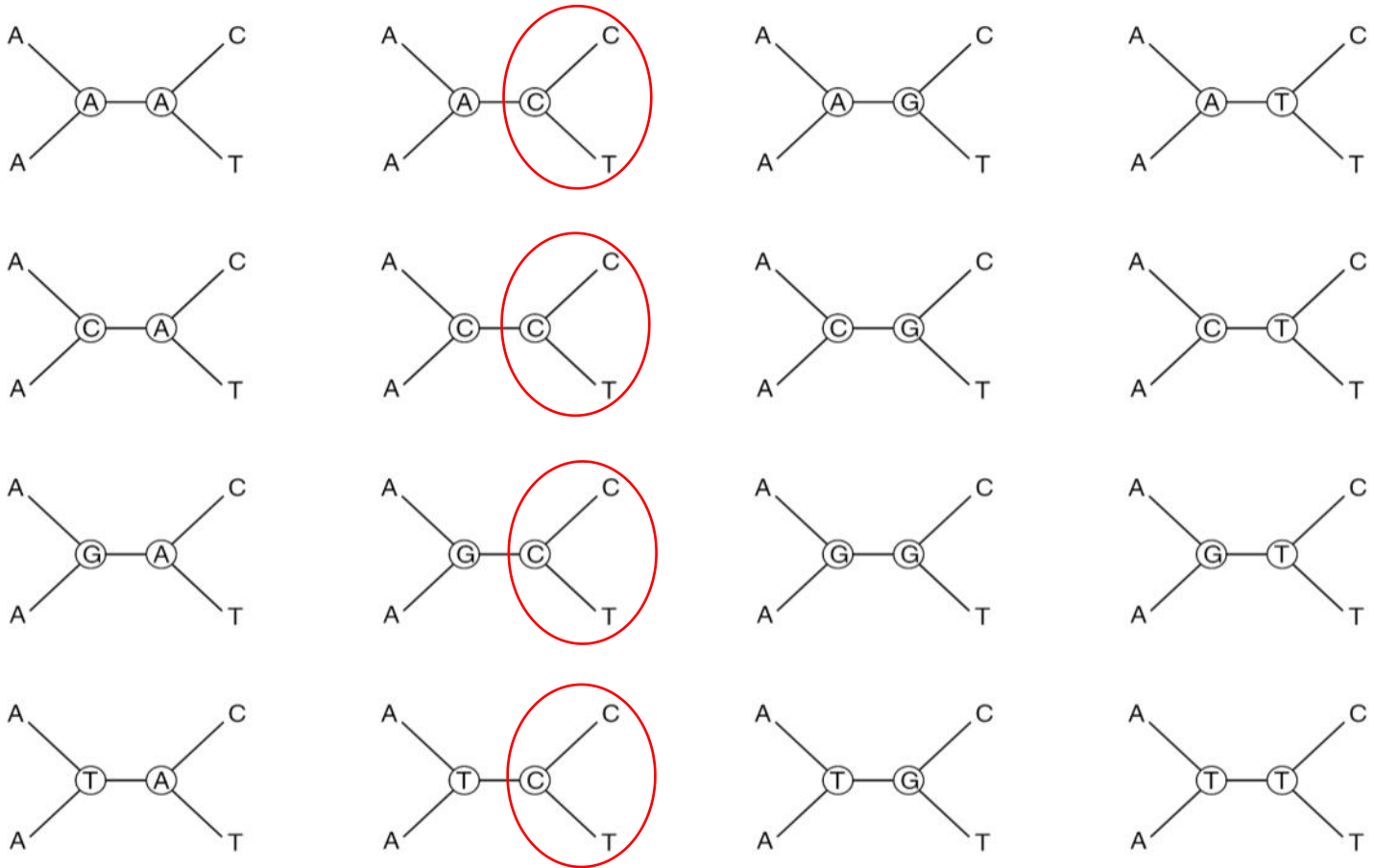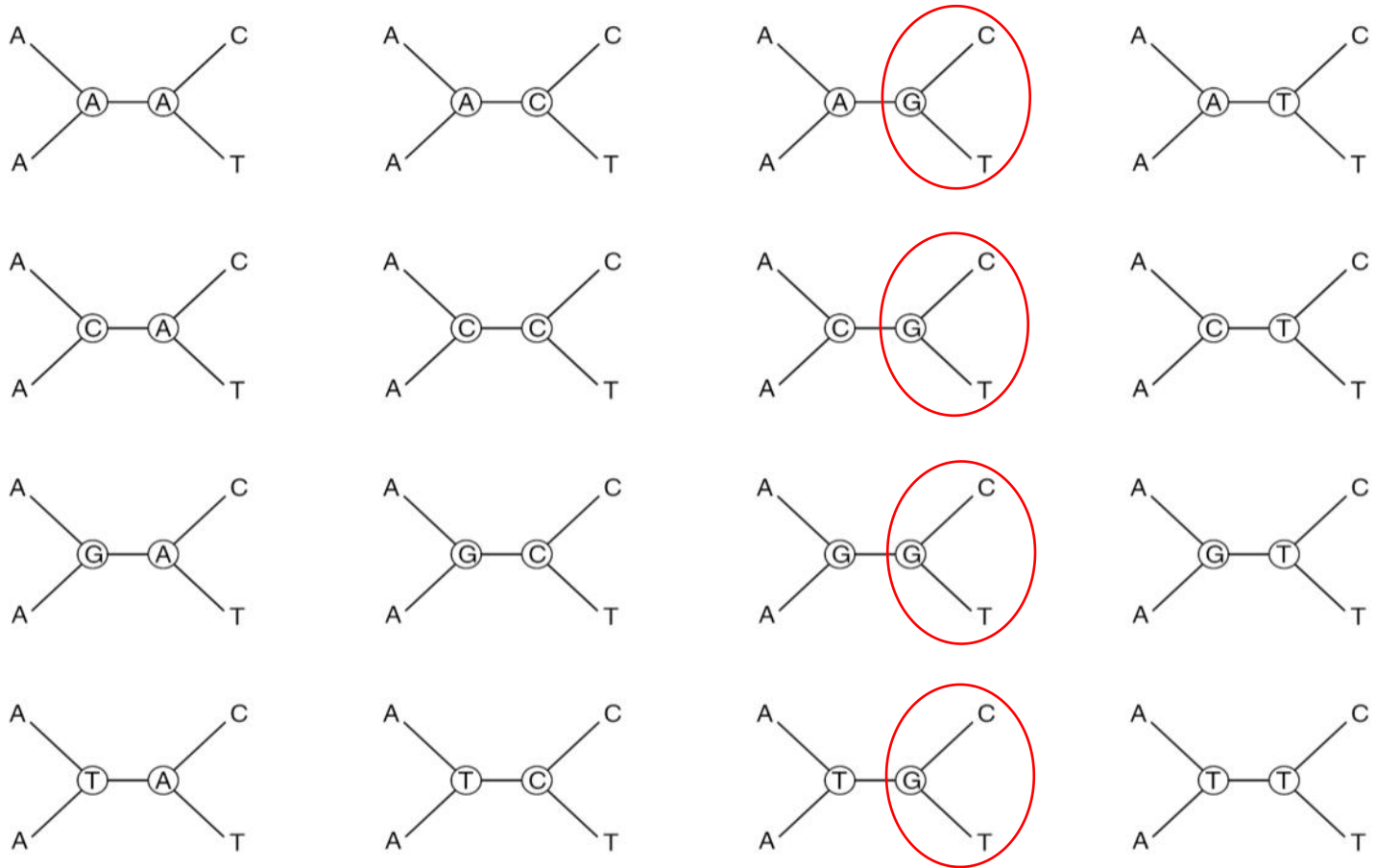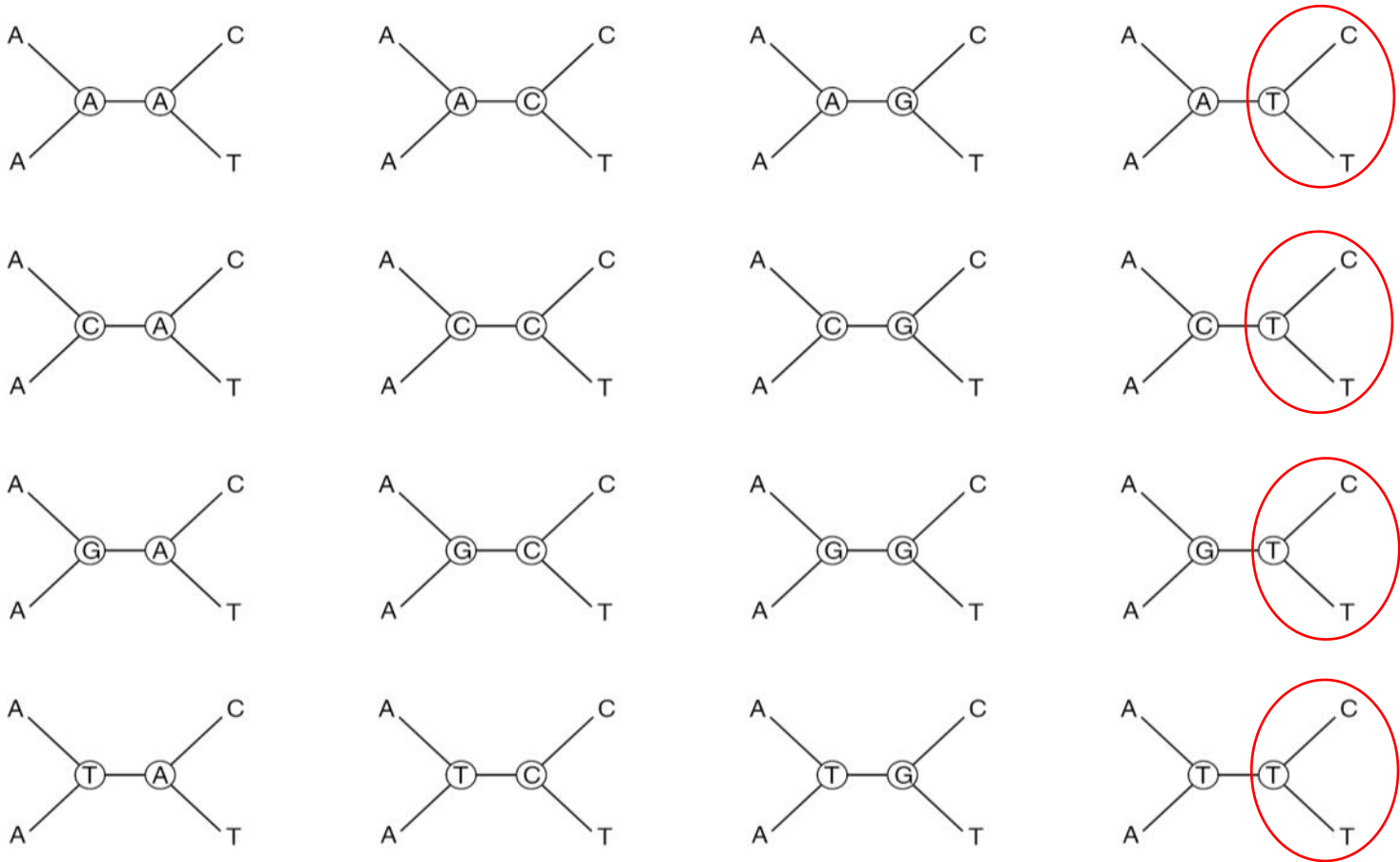

Note use of the OR probability rule

Pruning algorithm

Note use of the OR probability rule
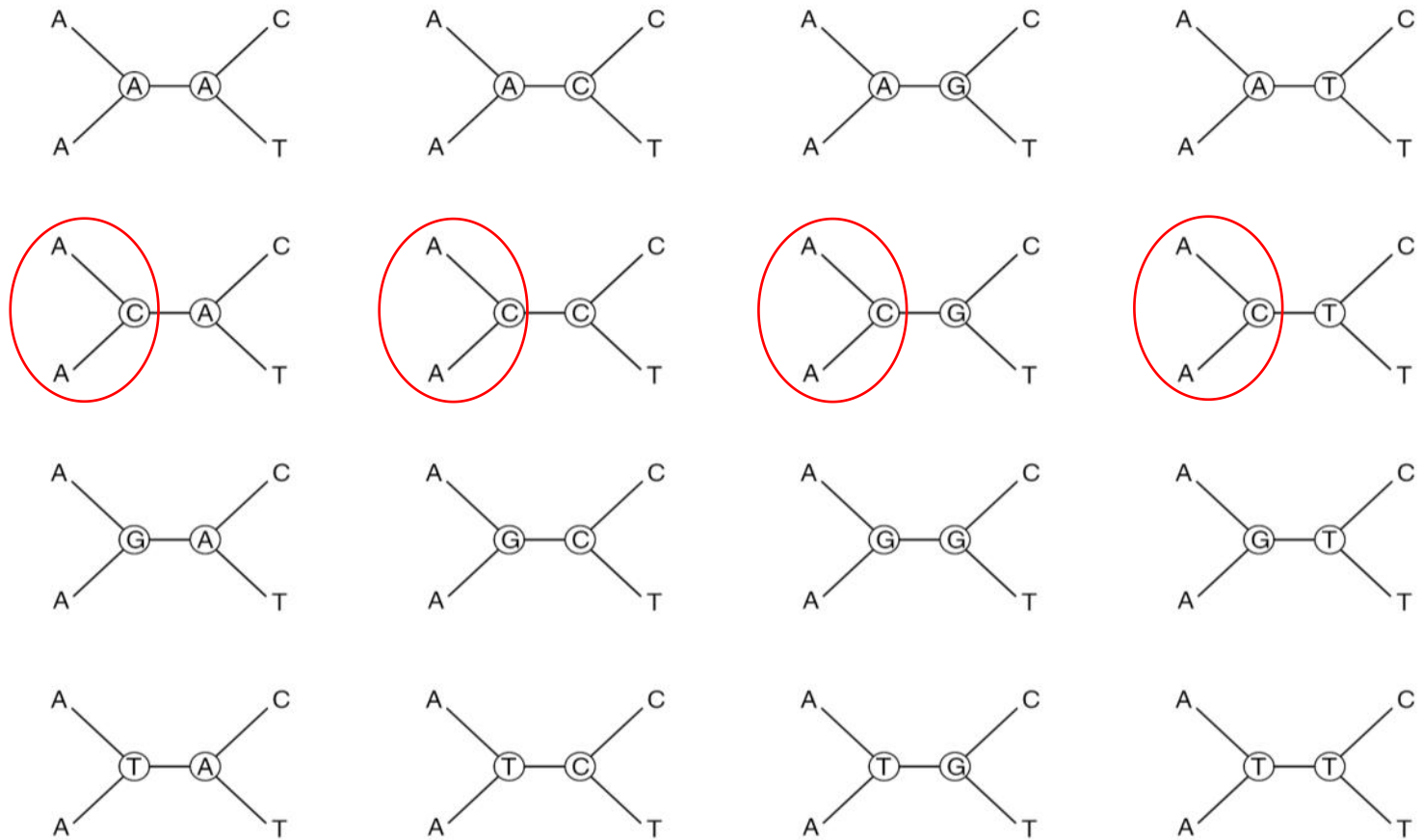
Pruning algorithm

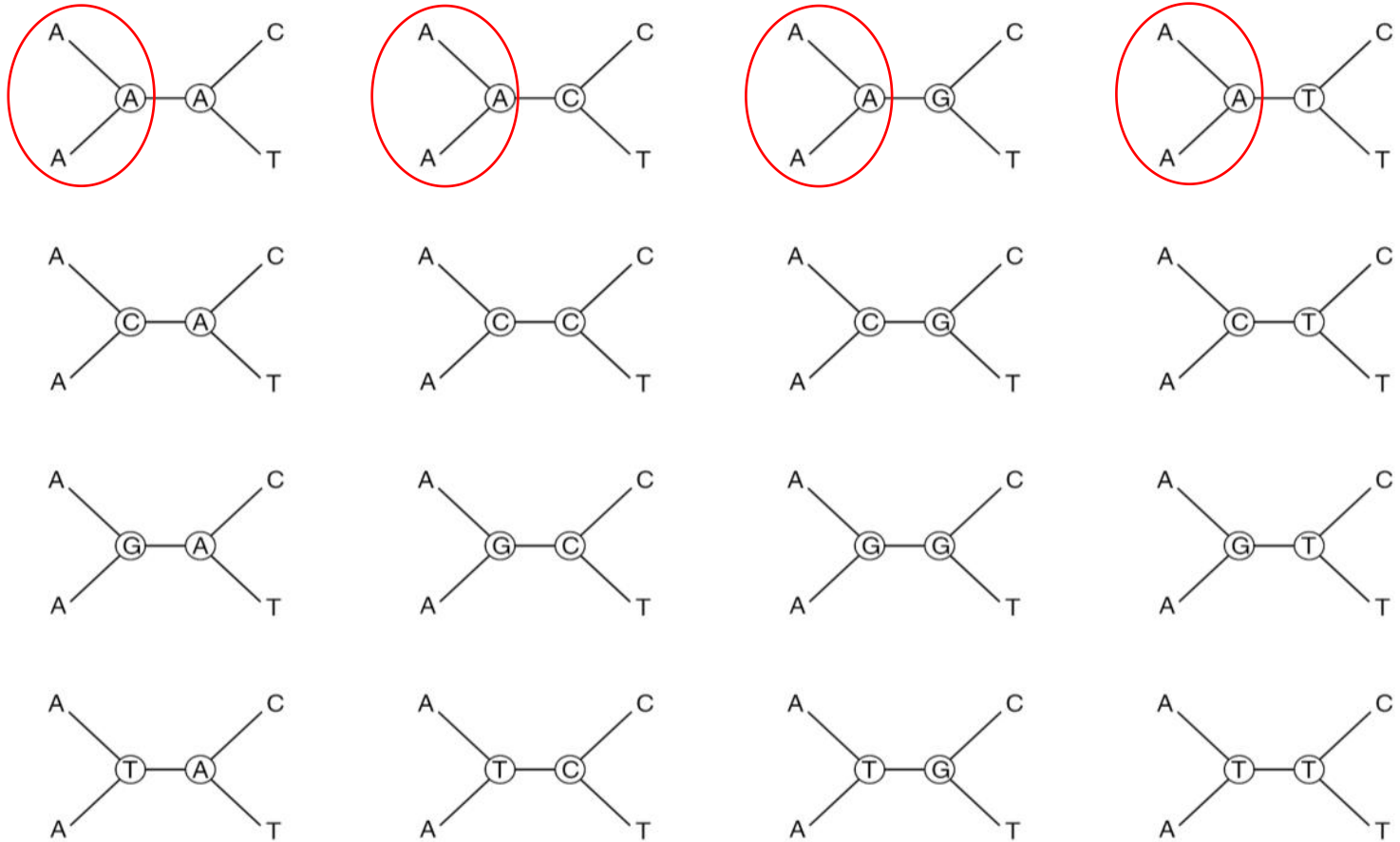Note use of the OR probability rule
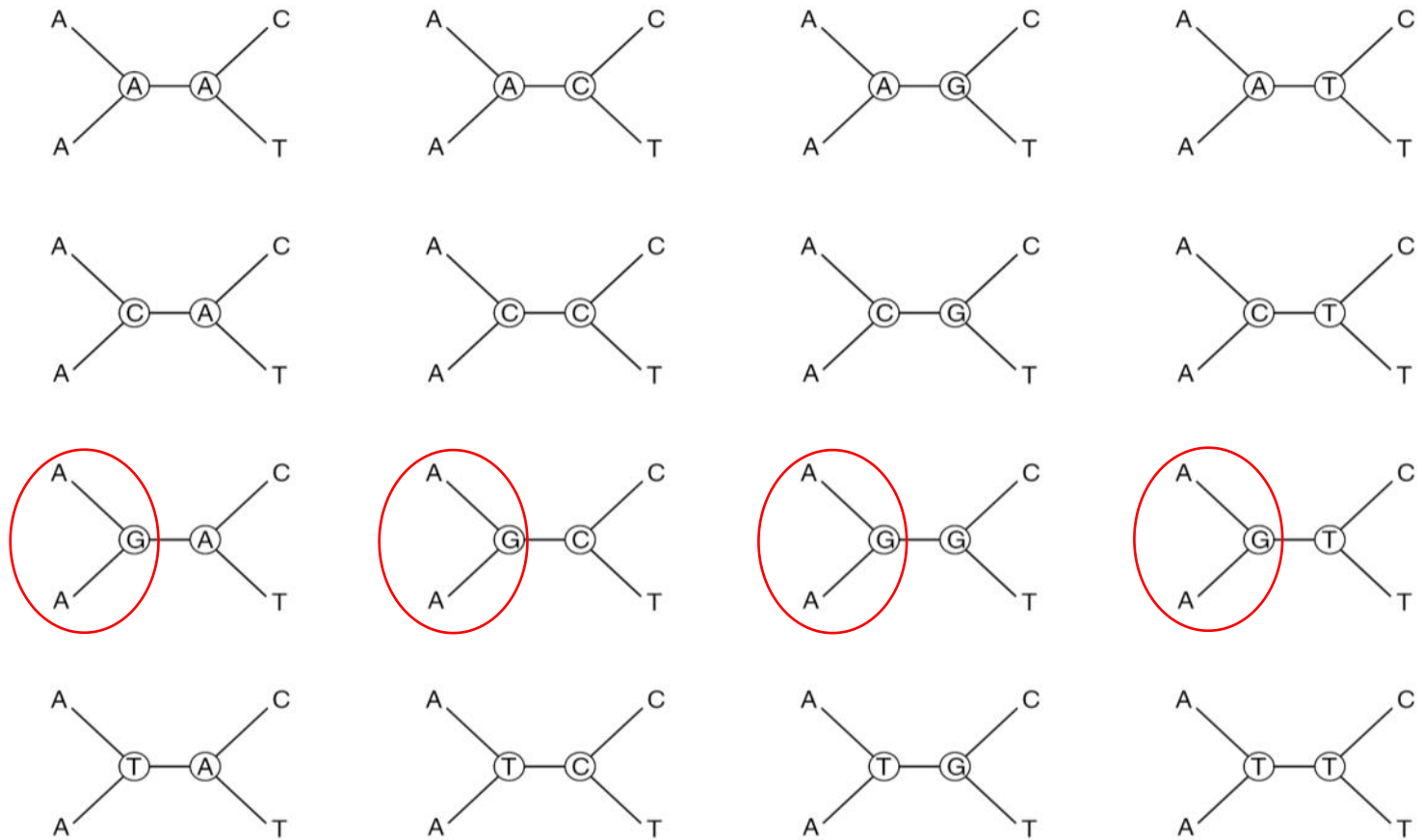
Pruning algorithm

Note use of the OR probability rule

Pruning algorithm
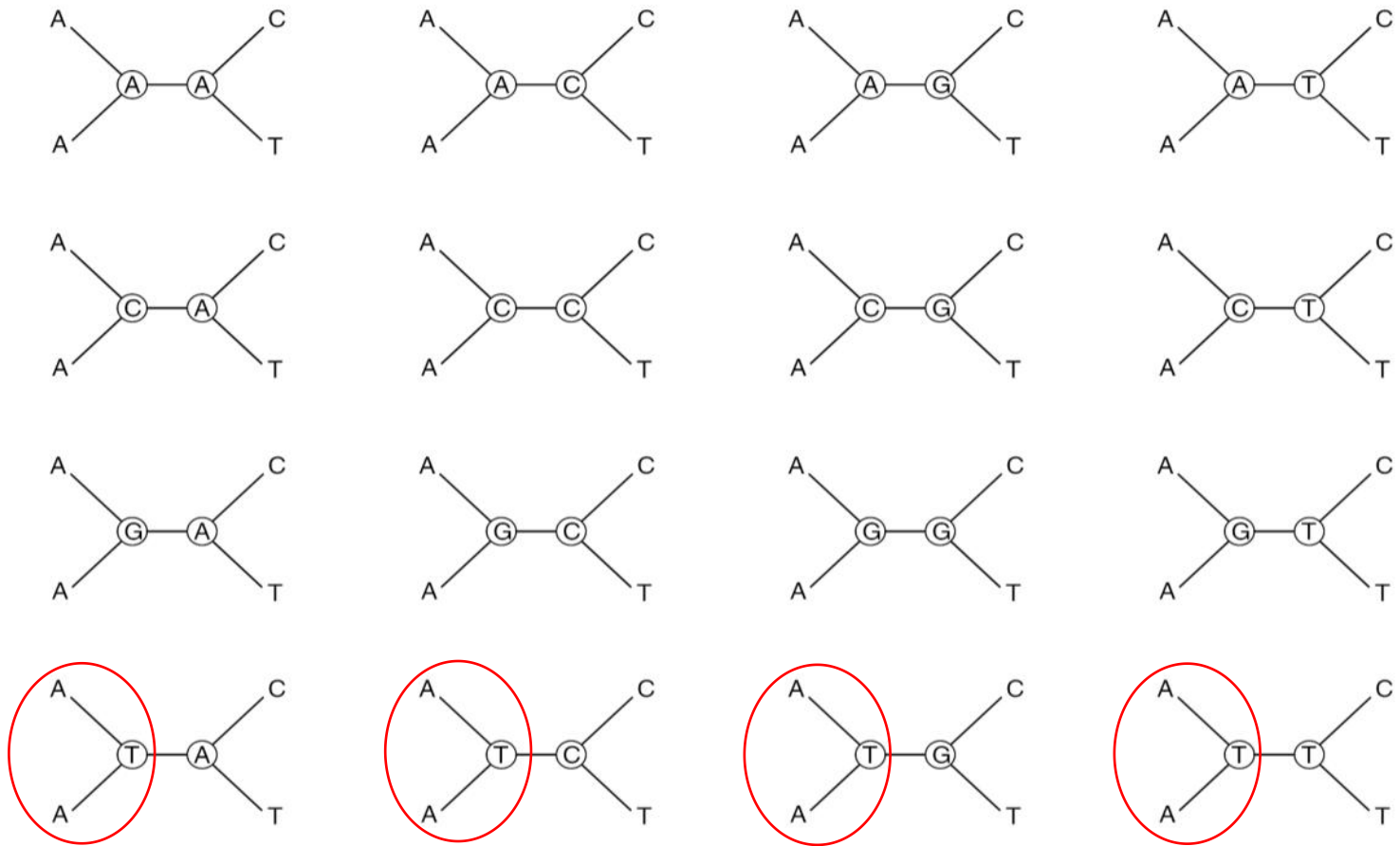
Note use of the OR probability rule

Species	A	B	C	D	E	F
Character state at tips	0	1	0	2	2	1

Species	A	B	C	D	E	F
Character state at tips	0	1	0	2	2	1

Species	A	B	C	D	E	F
Character state at tips	0	1	0	2	2	1
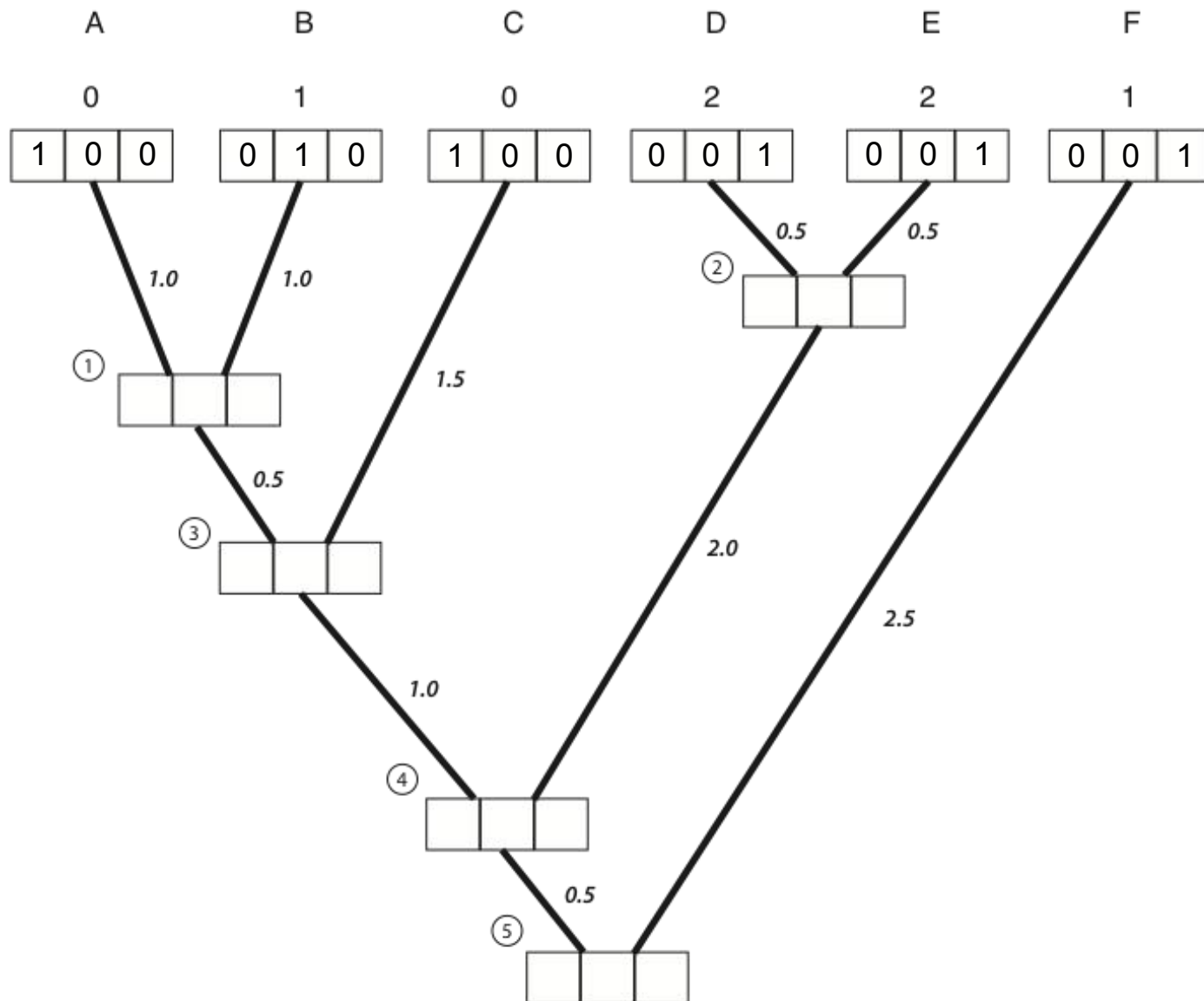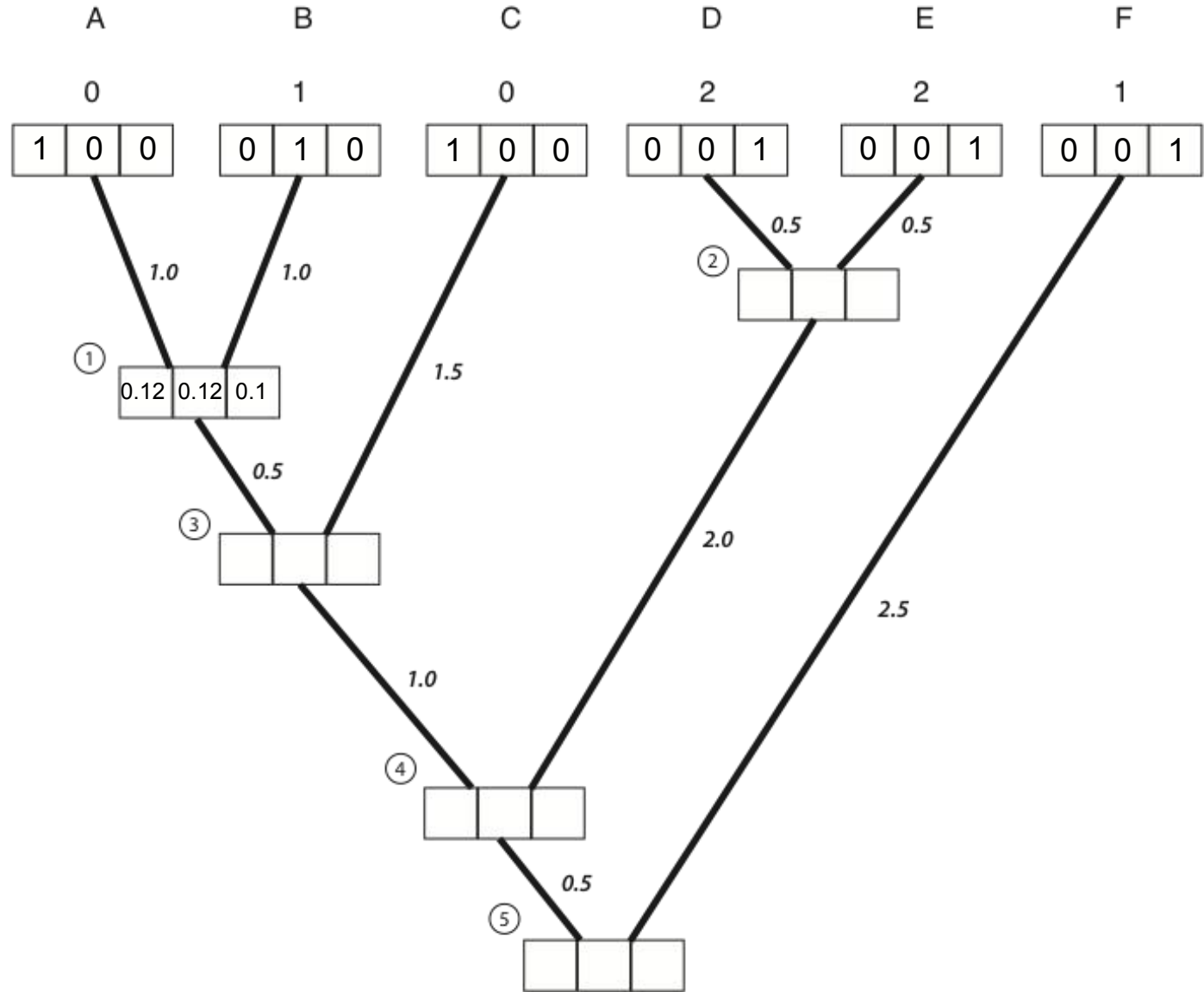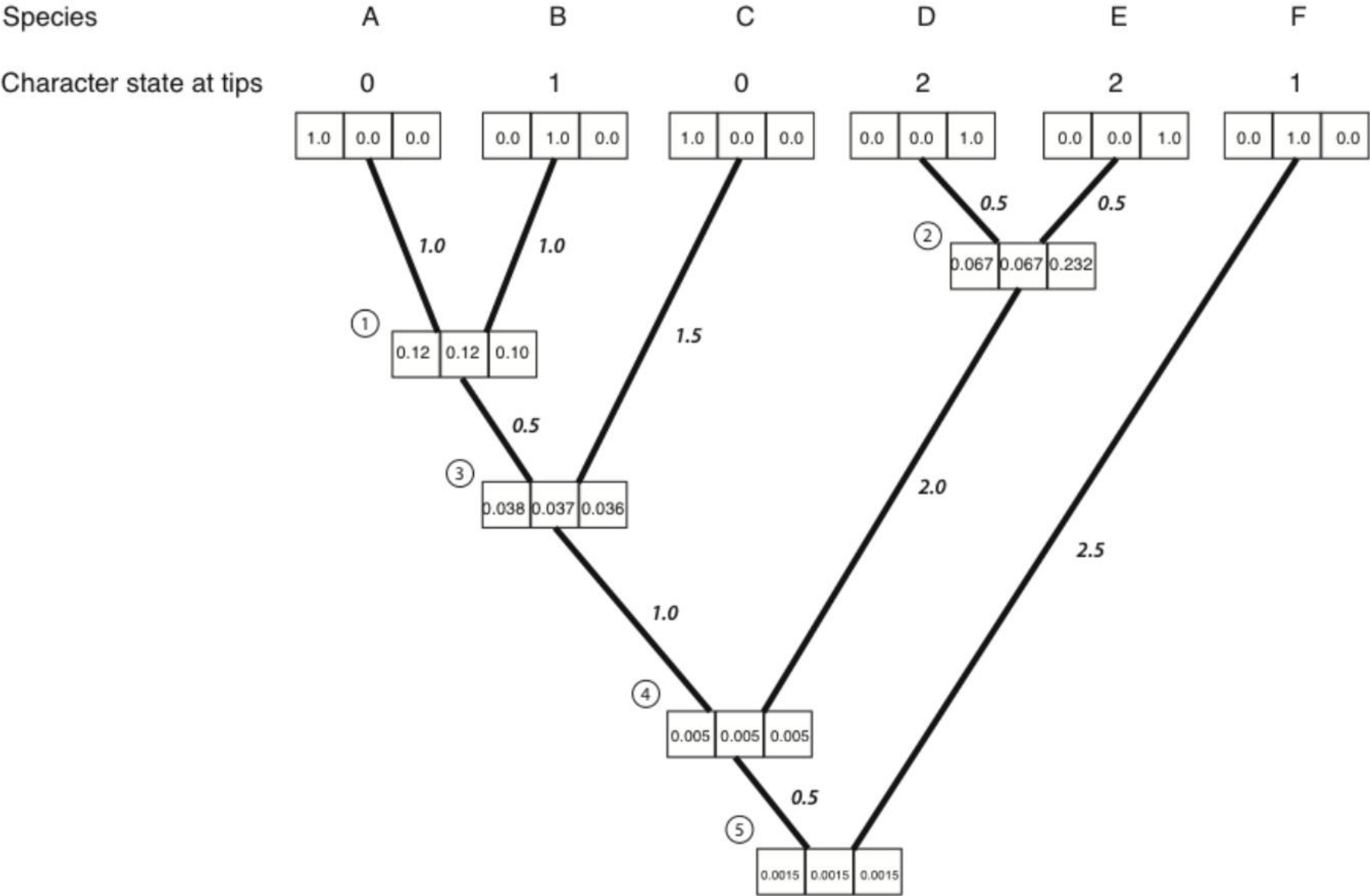
Returning to our original problem:
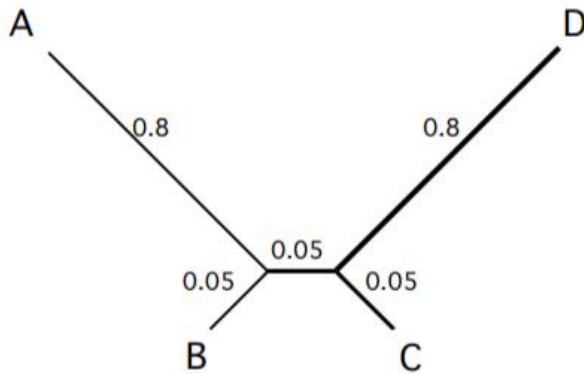
True Tree
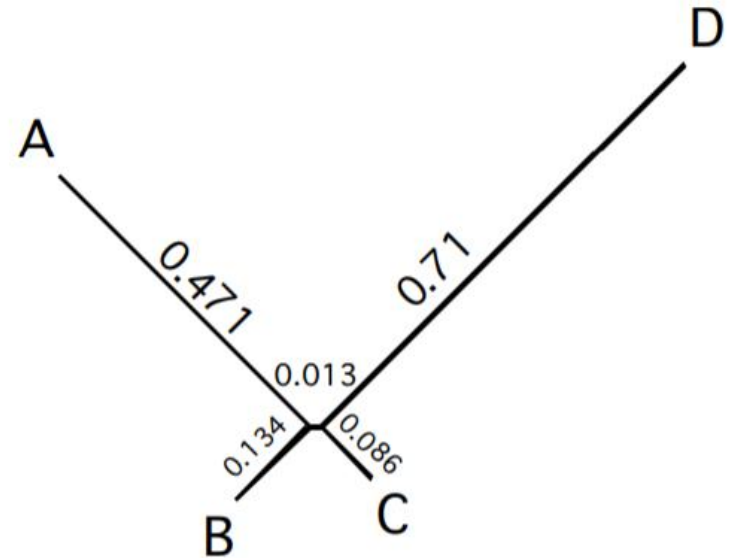
ML Tree

Jukes-Cantor Model (JC69)

Q matrix *(instantaneous rates)*

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

Jukes-Cantor Model (JC69)

Transition probabilities:

$$P = e^{Qt}$$

Matrix exponentiation

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

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

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

$$
Q = \begin{array}{c c} & \begin{array}{c c c c} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \left[\begin{array}{c c c c} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{array} \right] \end{array}
$$
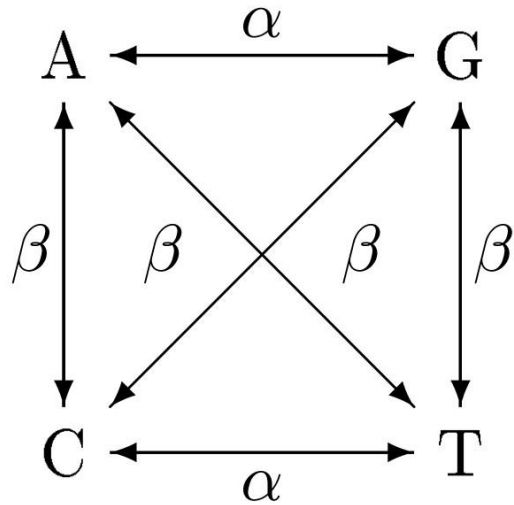
Our general CTMC model:

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

We get branch lengths in expected number of changes rather than minimum # of changes

We expect and probabilistically incorporate all possible paths to the data, not just the shortest path

We have the flexbility to modify and compare models

We have a straight-forward way to convert branch lengths to time (with fossils or other constraints)

We use ALL OF THE DATA, not just parsimony-informative sites

Example: Parsimony says node 2 can be either A, C, or G with equal number of steps.
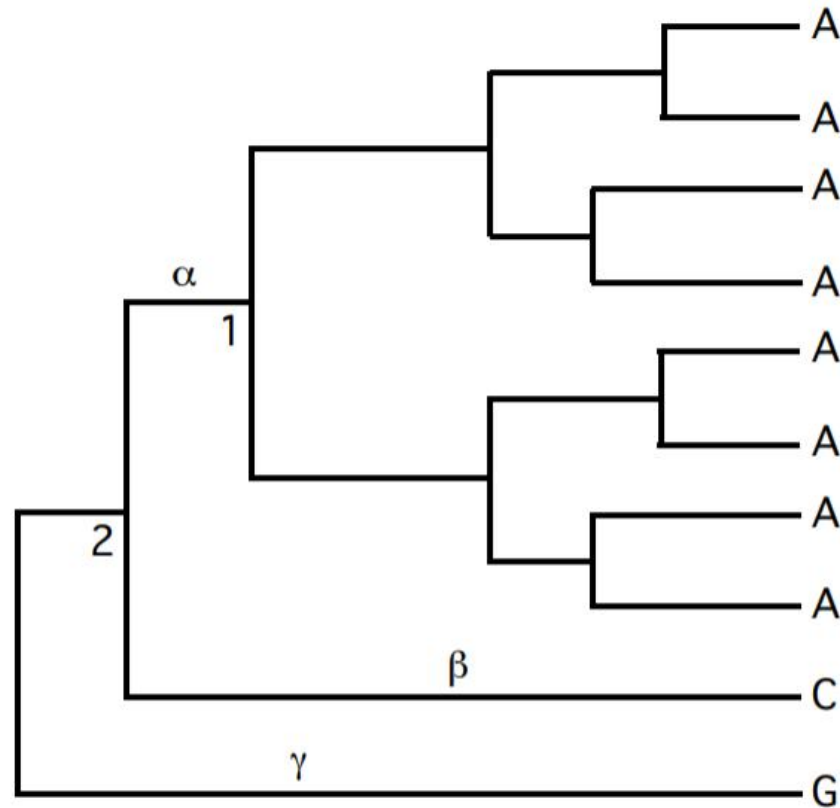
Example: What evidence is here that says node 2 is more likely A, than C and G?

How many parameters for each?

What do the models suggest about evolution?

a **GTR nucleotides**

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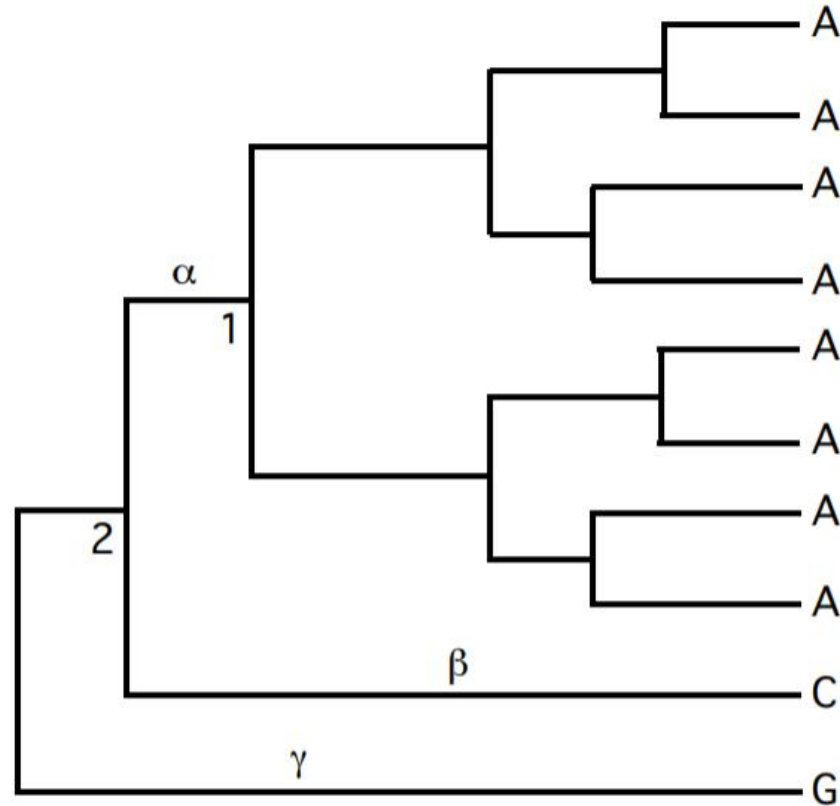
	00	01	11	10
00	-	r_A	0	r_B
01	r_C	-	r_D	0
11	0	r_E	-	r_F
10	r_G	0	r_H	-

c **Covarion**

	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 **Ordered transitions**

	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}	-

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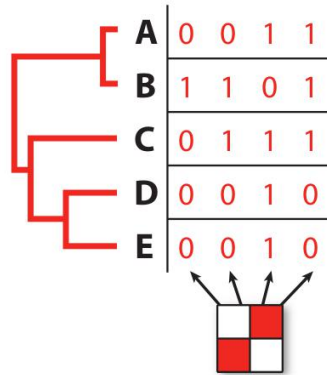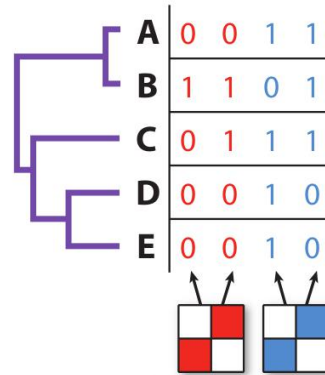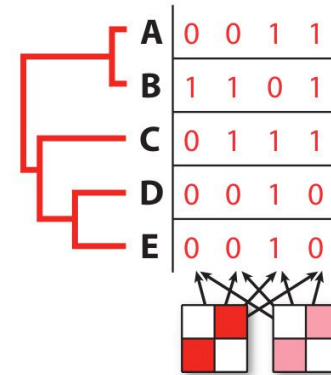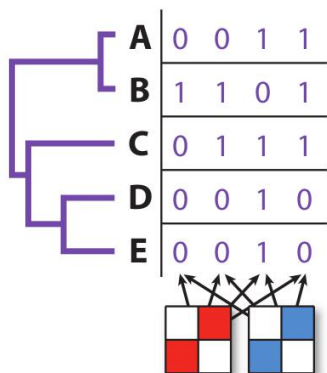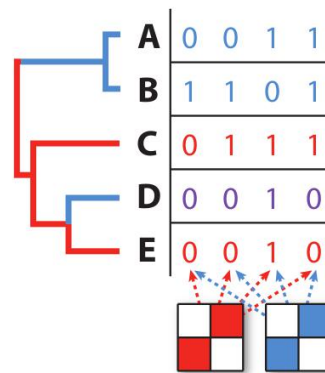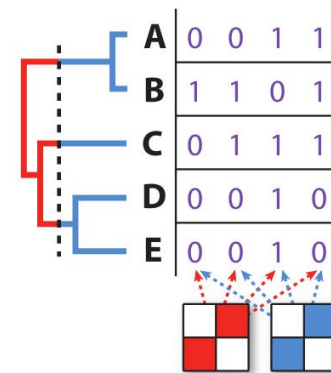
Multi-locus phylogeny of the tribe Tragelaphini (Mammalia, Bovidae) and species delimitation in bushbuck: Evidence for chromosomal speciation mediated by interspecific hybridization

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 Ha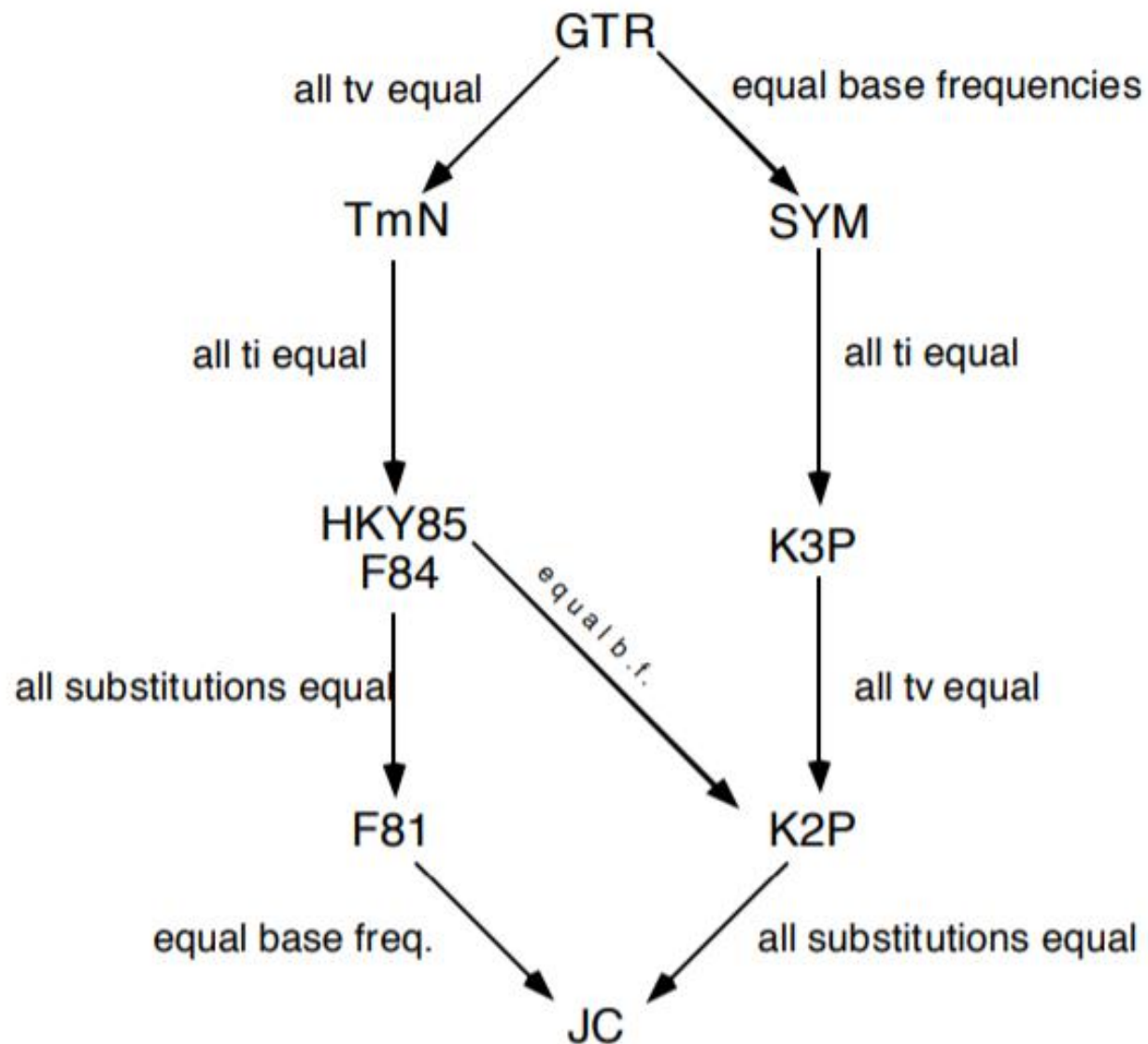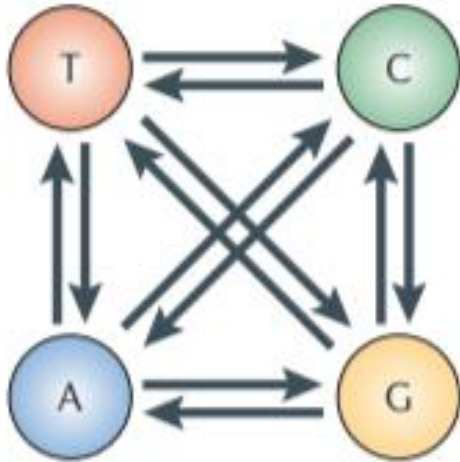ssanin et al. (2012). Our final alignment contains 1901 characters for 206 sequences (available upon request to corresponding author) which includes 183 bushbuck and 23 outgroup taxa of the tribes Caprini, Boselaphini, Bovini, and Tragelaphini. A Bayesian tree was constructed as detailed below for other 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
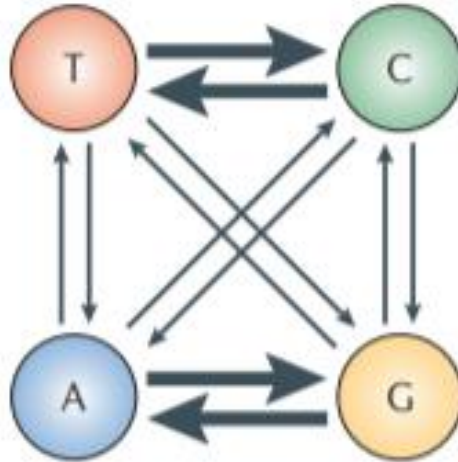
$$\begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix}$$

HKY85
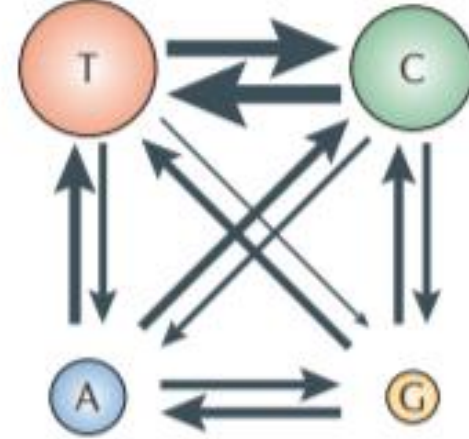
$$\begin{pmatrix} -\mu(\kappa\pi_G + \pi_Y) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T + \pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_R) \end{pmatrix}$$

General Time-Reversible Model

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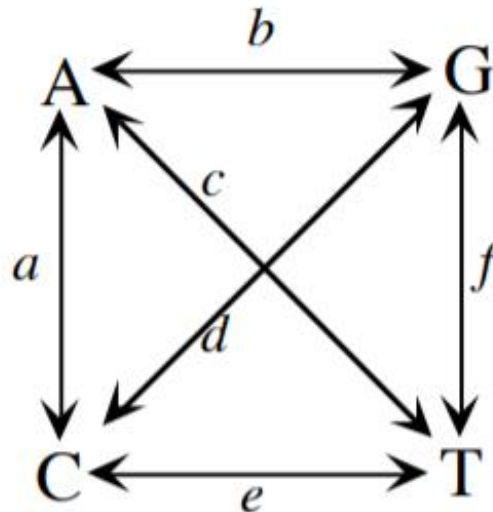
General Time-Reversible Model

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

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


r_1 applies to subset 1
1st+2nd codon positions
(sites 1 - 88)

r_2 applies to subset 2
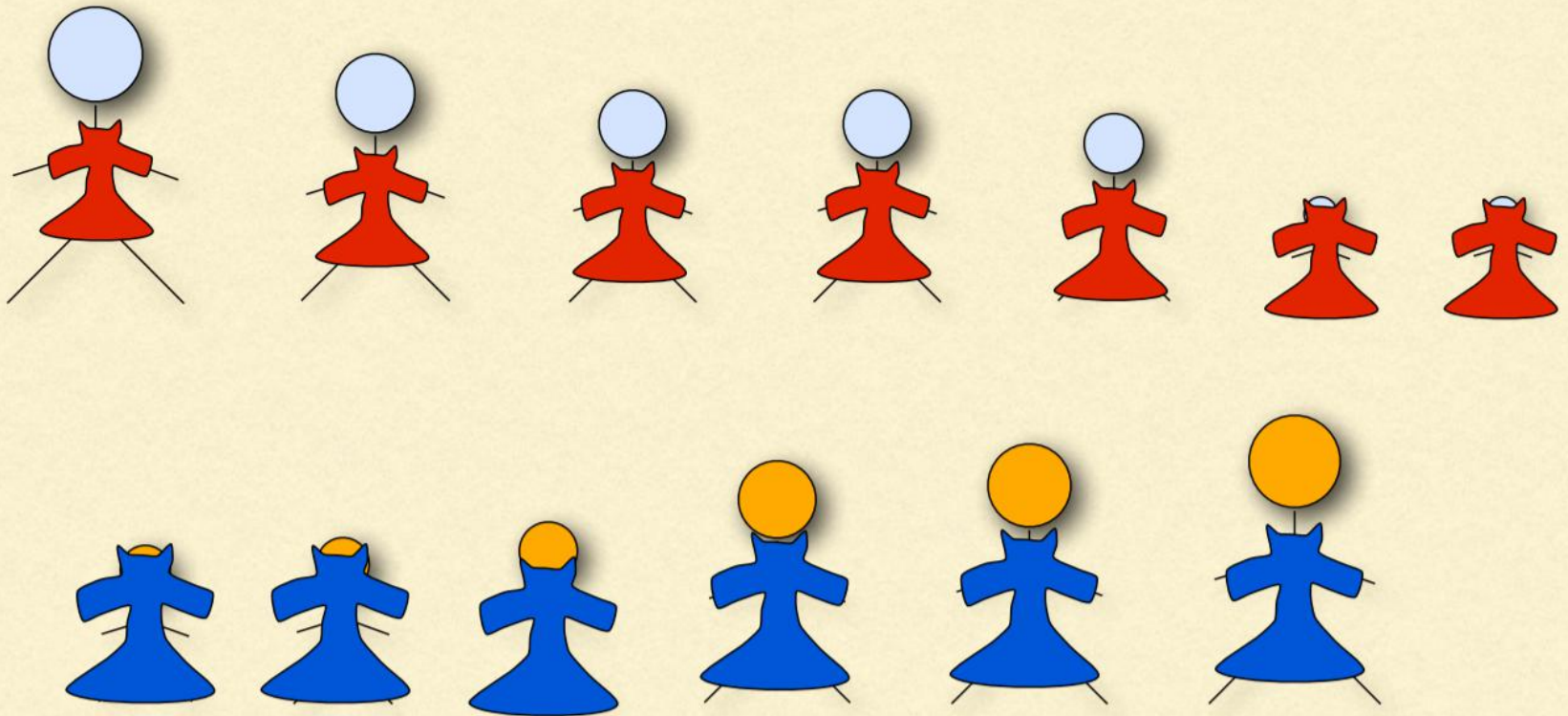3rd codon positions
(sites 89-132)

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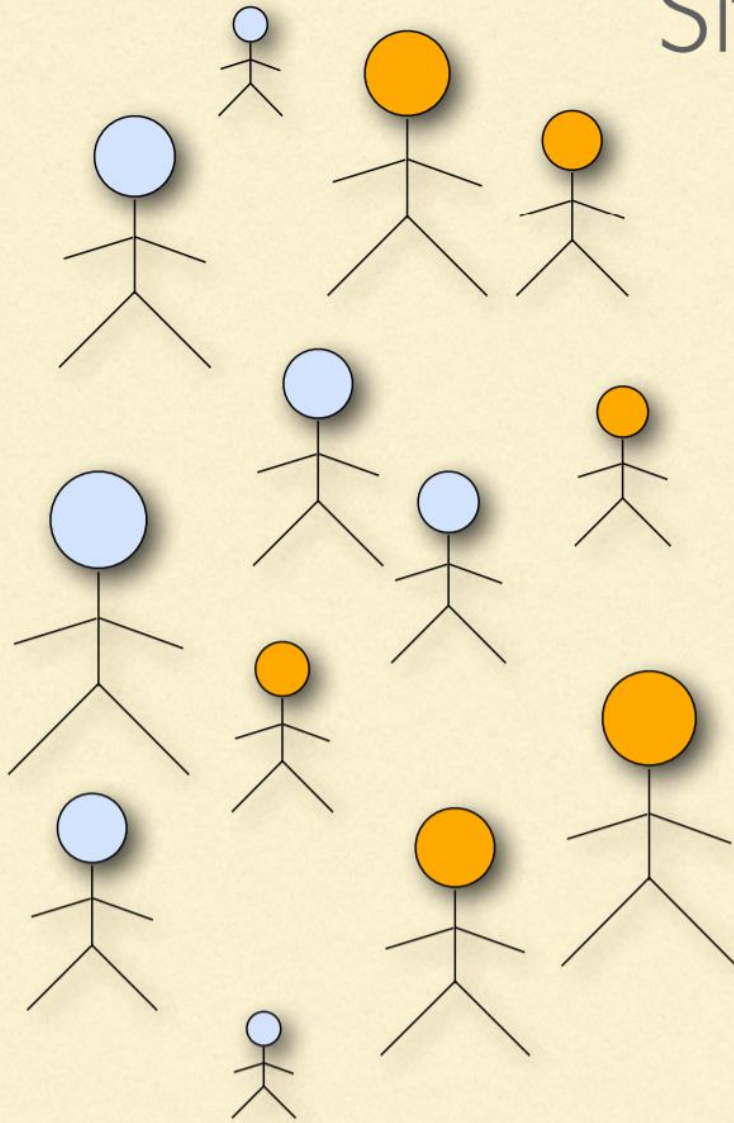
OK, I am going to divide you into 2 groups based on the color of your head, and everyone in each group will get a coat of the average size for their group. Very sorry if this does not work well for some people who are unusually large or small compared to their group.
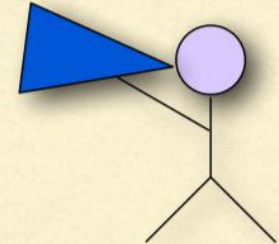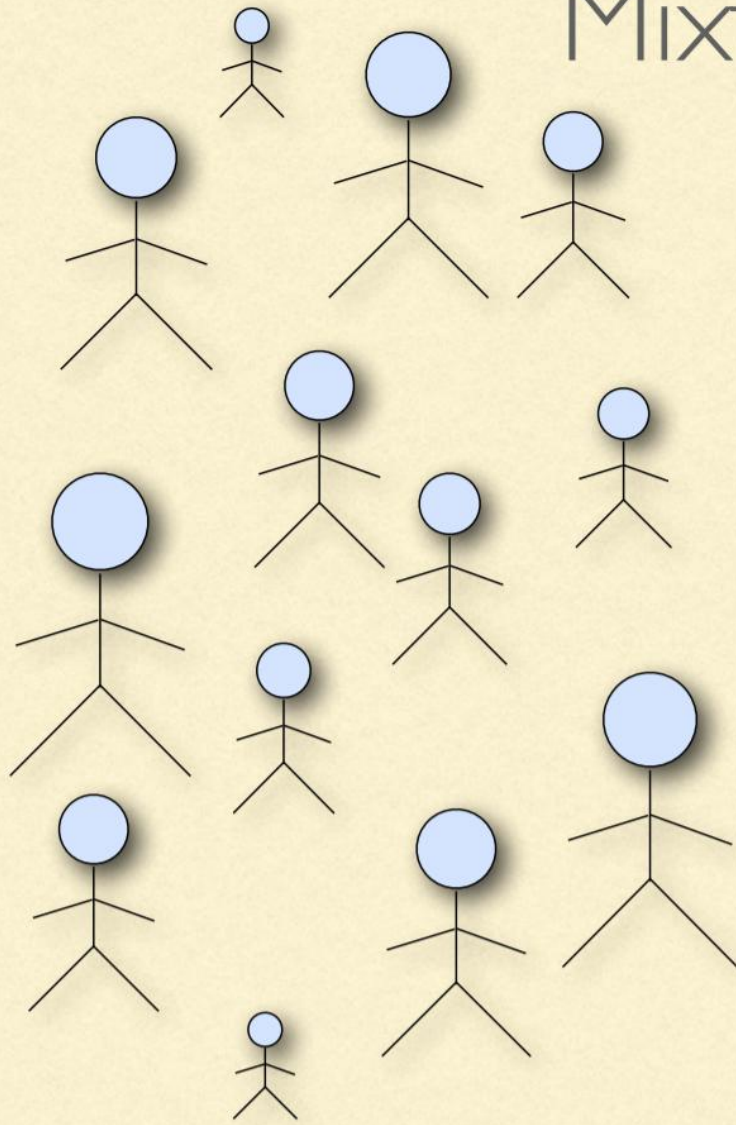
Site-specific approach

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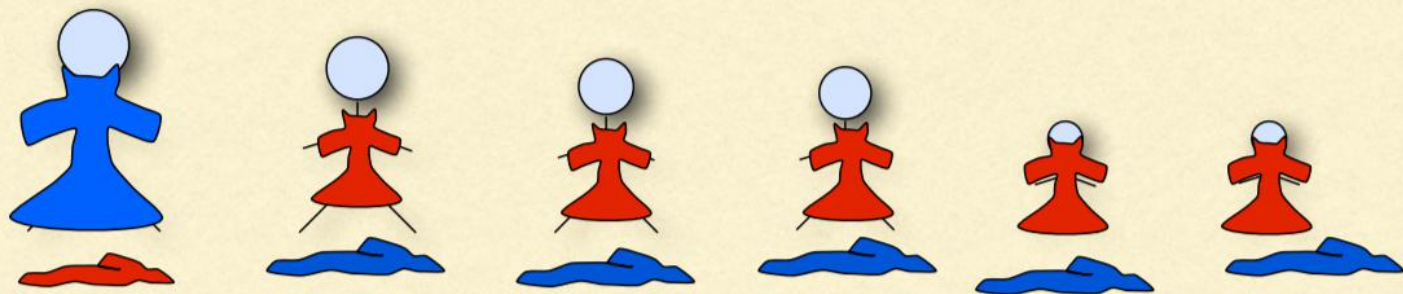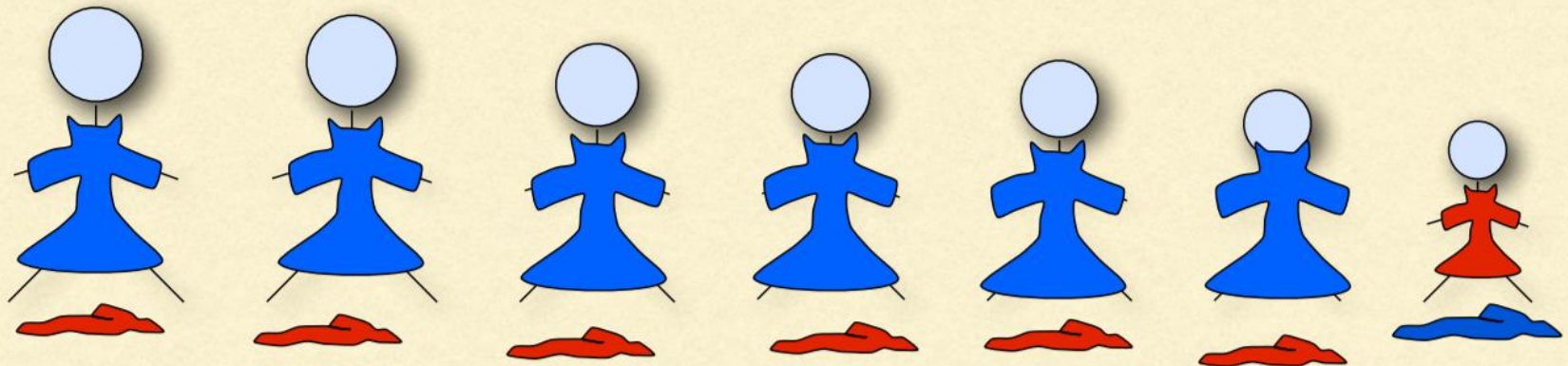Bad: every person in a group has to wear the same size coat

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.
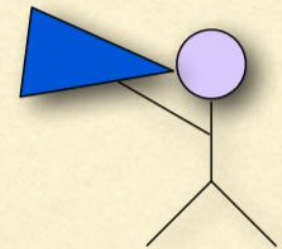
Mixture model approach

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)
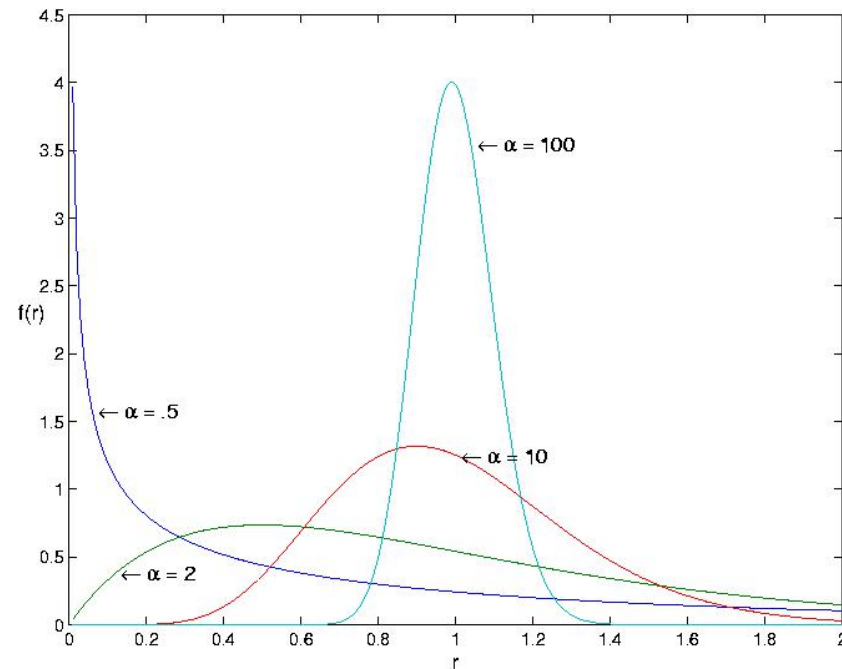
Gamma-distributed rates

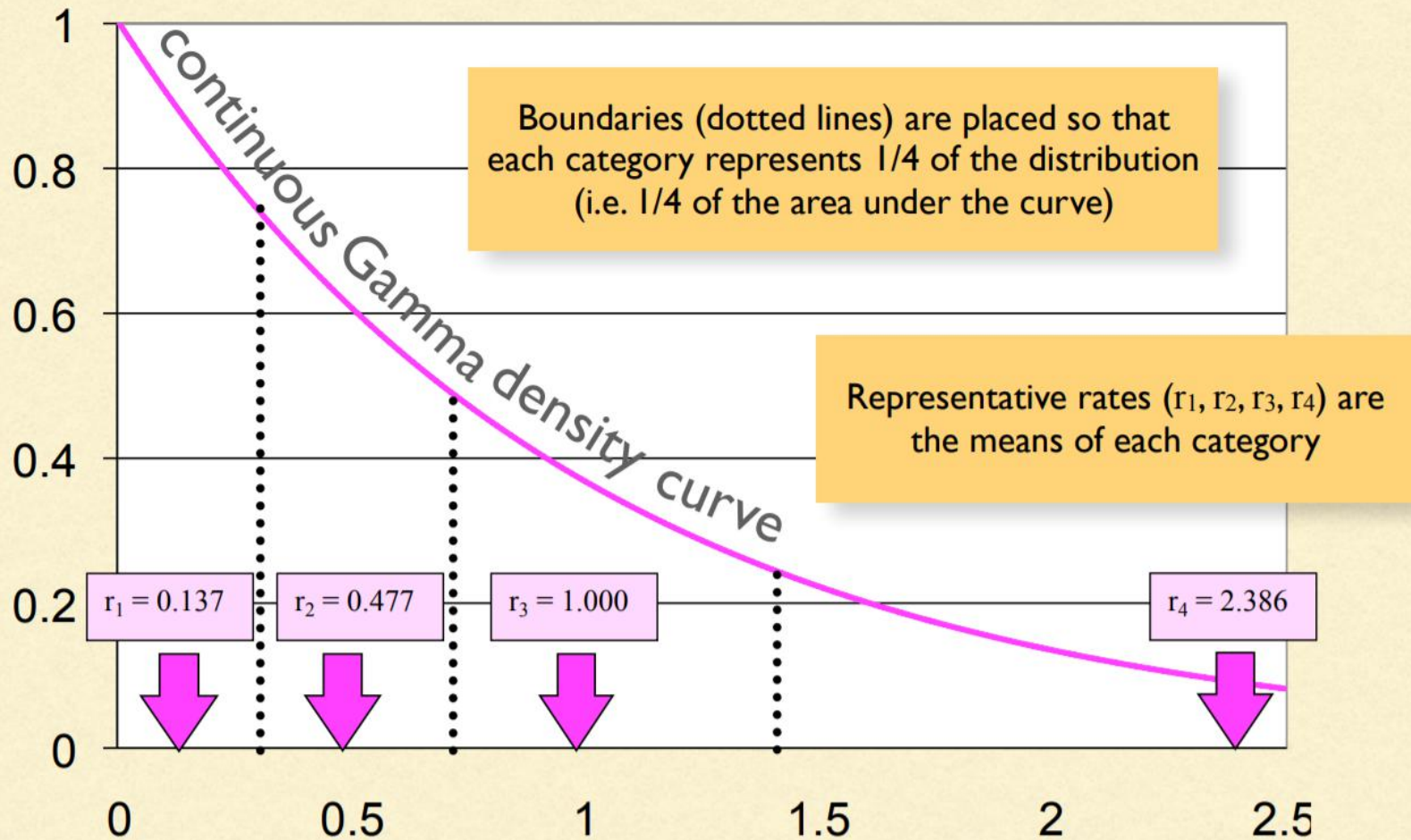Rate categories set by a discrete gamma distribution (e.g. 4 rate categories)

Estimate α

Site likelihoods calculated for each rate category and summed

"+G"

(e.g. HKY85 + I + G)

Discretizing the Gamma distribution

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

continuous Gamma density curve

$r_1 = 0.137$ $r_2 = 0.477$ $r_3 = 1.000$ $r_4 = 2.386$

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