Practical considerations

How do you find the most parsimonious tree?

One way:
1. Try every possible tree.
2. For each tree, count min # of steps
3. Choose the tree with the fewest # steps

Too many trees.

Table 3.1: The number of rooted, bifurcating, labeled trees for n species, for various values of n. The numbers for more than 20 species are approximate.

Species	Number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	*34,459,425*
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

Heuristic searches

Don't try every tree.

"Hill-climbing algorithm"

Wander around under certain rules. Hope that what you eventually stop at is the best.

Not guaranteed to find the best tree.

Can be improved by lots of random starting points.

(Not necessary for parsimony, exact algorithmic solution possible with a "branch and bound" search)

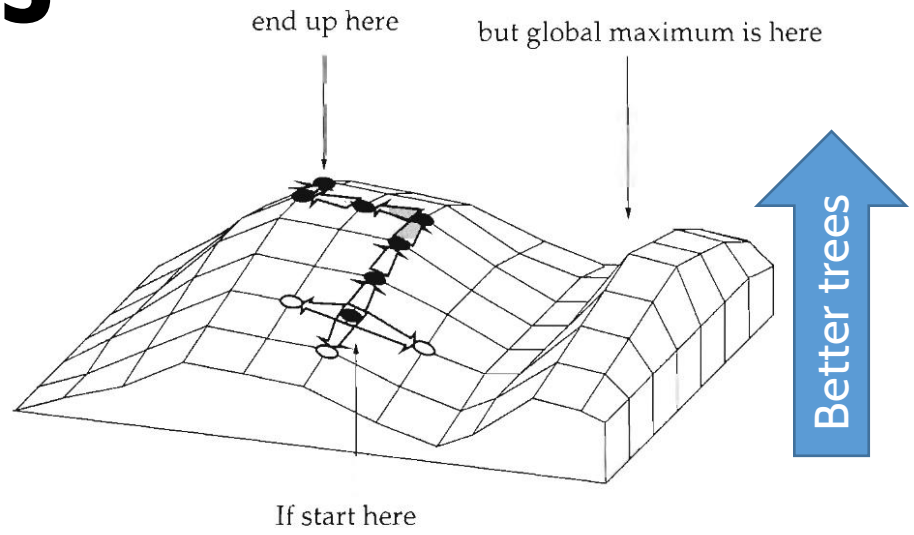


Figure 4.1: A surface rising above a two-dimensional plain (or plane). The process of climbing uphill on the surface is illustrated, as well as the failure to find a higher peak by this "greedy" method.

"Tree space"
Each point on the grid is a possible tree.
Trees next to each other are "1 step away" via algorithmic manipulation
Start at a random tree.
Try all trees that are nearby, choose the one that's best
Keep doing that til you can't make the tree any better.

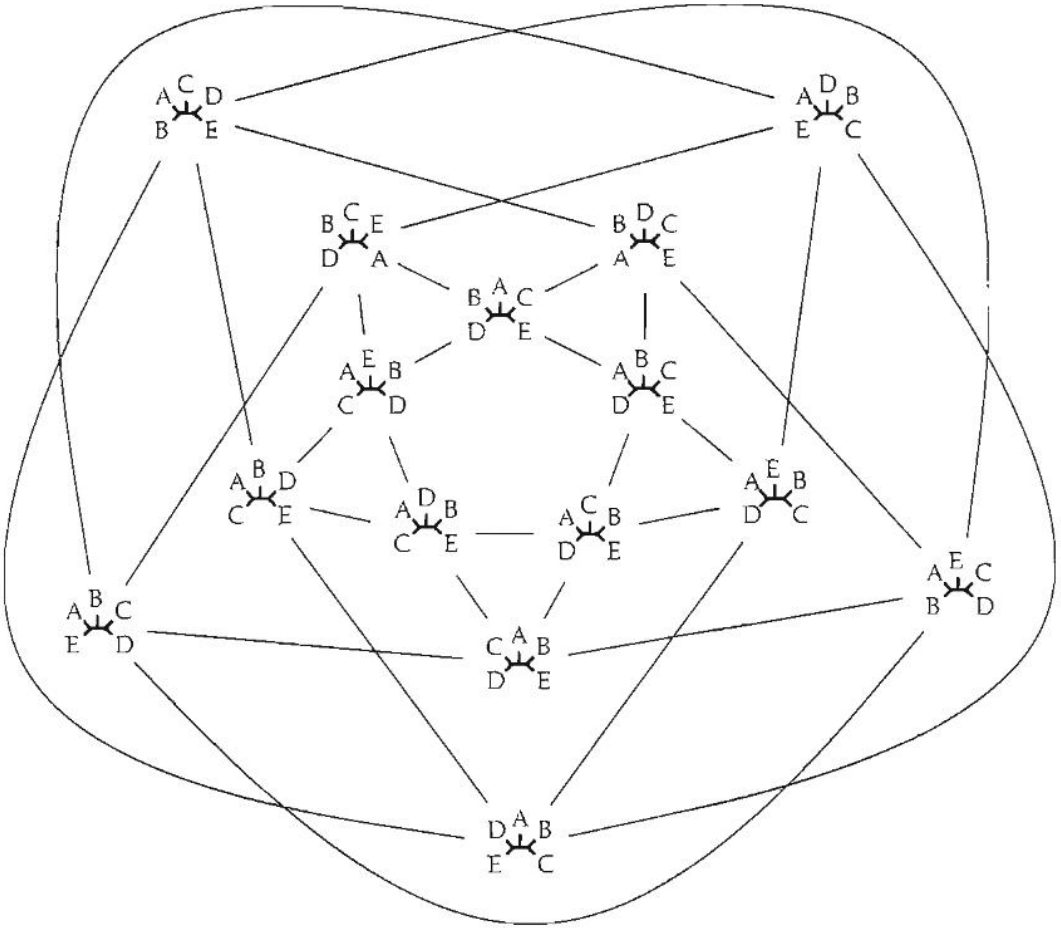


Figure 4.3: The space of all 15 possible unrooted trees with 5 tips. Neighbors are connected by lines when a nearest-neighbor interchange can convert one into the other. The labels A–E correspond to the species names Alpha through Epsilon in that data set. This symmetric arrangement of nodes was discovered by Ben Rudd Schoenberg (personal communication), and we thus denote this graph the Schoenberg graph.

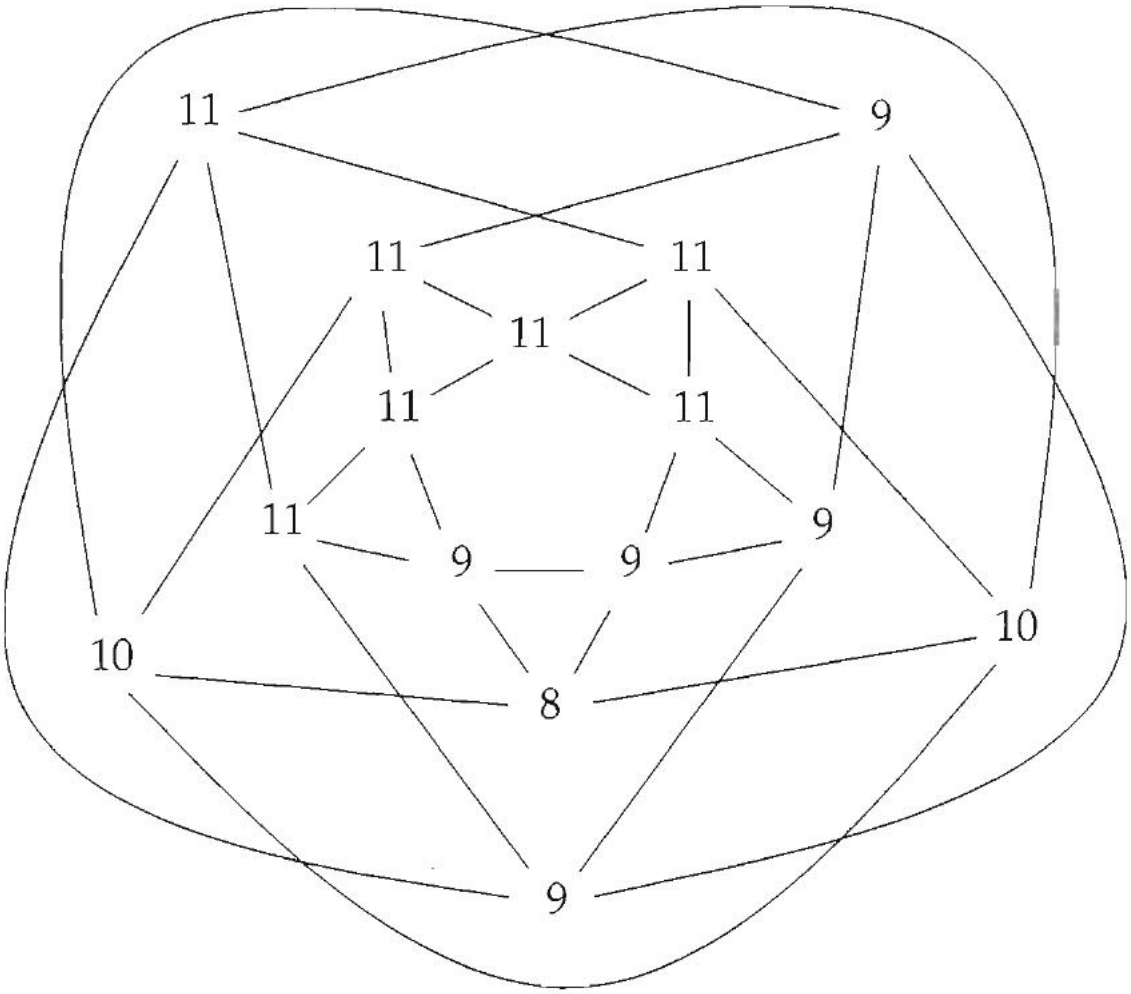
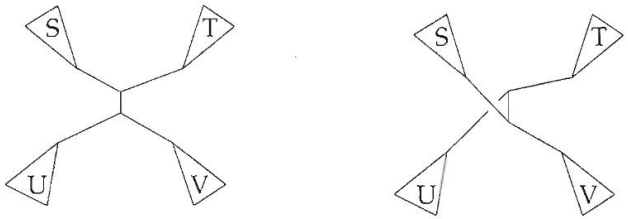
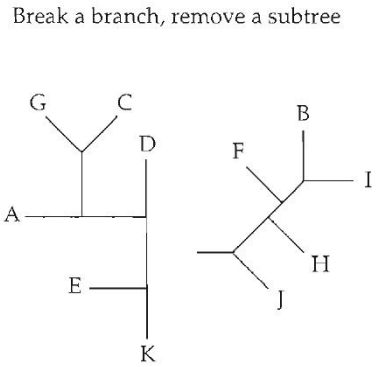


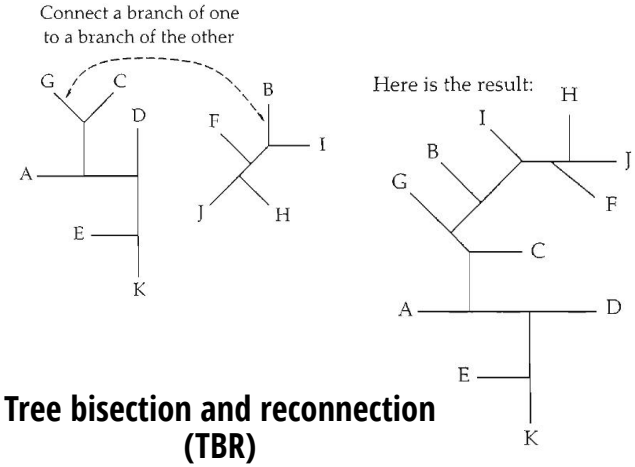
Figure 4.4: The space of all 15 possible trees, as in Figure 4.3, where the number of changes of state on the data set of Table 1.1 is shown. Nearest-neighbor interchanges search for the most parsimonious tree by moving in this graph.

Tree space "moves"

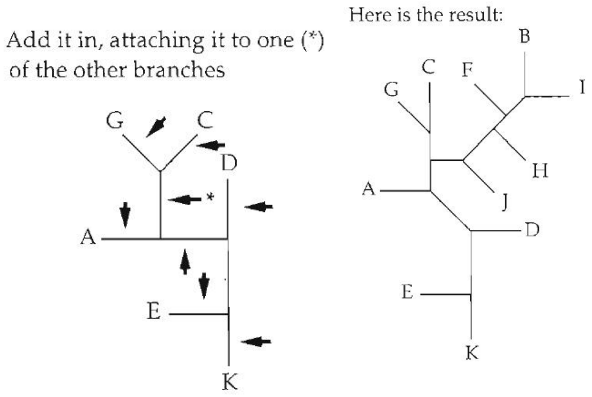


Nearest Neighbor Interchange (NNI)





Tree bisection and reconnection (TBR)



Subtree pruning and regrafting (SPR)

Distance between tree spaces changes depending on which moves you use

NNI fastest, least computationally intensive, but the largest distances between trees

TBR slowest, most computationally intensive, but shortest distances between trees

SPR intermediate

Branch and bound

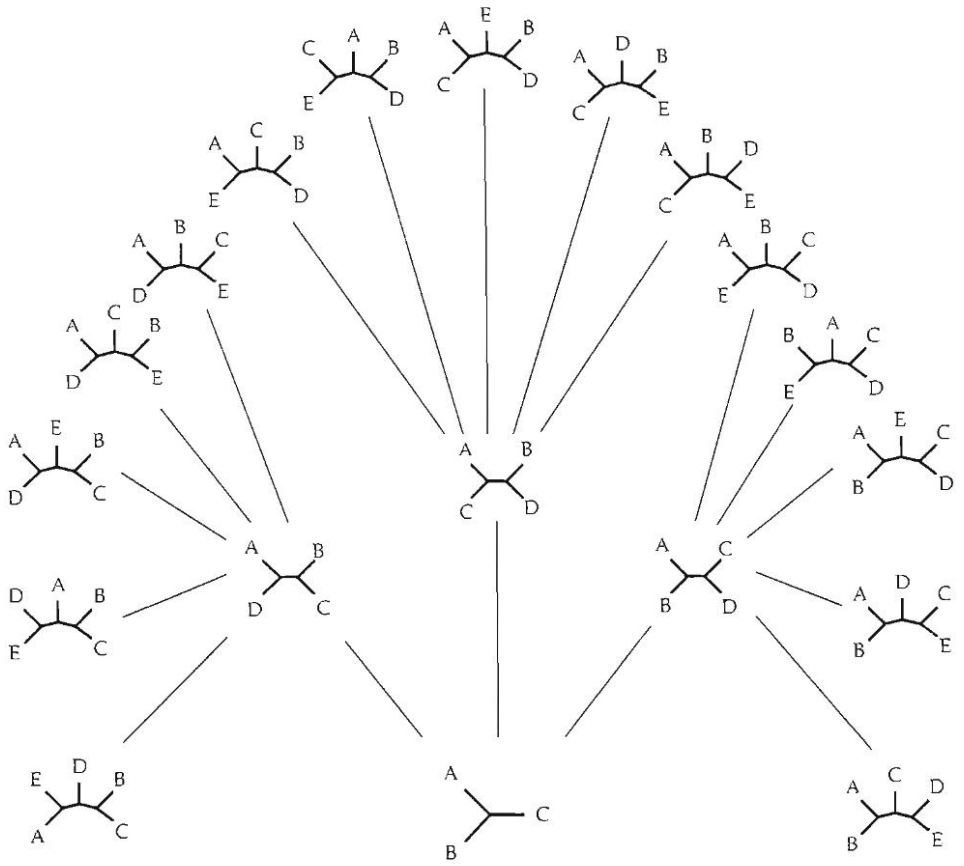


Figure 5.3: Search tree for most parsimonious tree in a five-species case.

Branch and bound

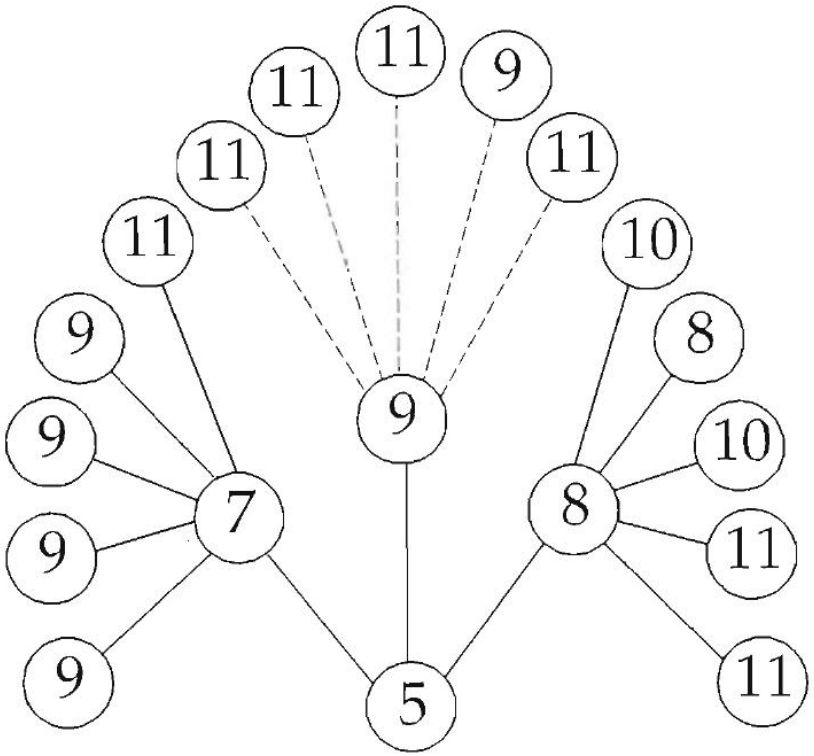


Figure 5.4: Search tree for most parsimonious tree for five species, using the data of Table 1.1. Trees are shown in Figure 5.3. Dashed lines are those not traversed by a branch and bound method. The species names in the data set correspond to labels A through E in Figure 5.3.

Felsenstein & the birth of statistical phylogenetics

How do we think about probabilities & trees?

Scholar

Joe Felsenstein

FOLLOW

Professor of Genome Sciences, and Professor of Biology, University of Washington, Seattle
Verified email at gs.washington.edu - Homepage

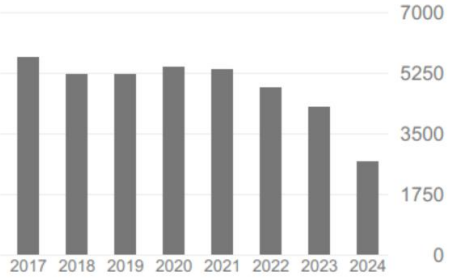
Evolutionary biology phylogenetic methods population genetics

TITLE	CITED BY	YEAR
Confidence limits on phylogenies: an approach using the bootstrap J Felsenstein evolution 39 (4), 783-791	50997	1985
PHYLIP (phylogeny inference package), version 3.5 c J Felsenstein Joseph Felsenstein.	31741 *	1993
Evolutionary trees from DNA sequences: a maximum likelihood approach J Felsenstein Journal of molecular evolution 17, 368-376	16181	1981
Phylogenies and the comparative method J Felsenstein The American Naturalist 125 (1), 1-15	10816	1985
Inferring phylogenies J Felsenstein Inferring phylogenies, 664-664	6115	2004
Cases in which parsimony or compatibility methods will be positively misleading J Felsenstein Systematic zoology 27 (4), 401-410	4166	1978
Phylogenies from molecular sequences: inference and reliability J Felsenstein	2780	1988

ogle.com/citations?view_op=search_authors&hl=en&...

Cited by VIEW ALL

	All	Since 2019
Citations	155027	27831
h-index	79	37
i10-index	146	69



Public access VIEW ALL

0 articles 7 articles

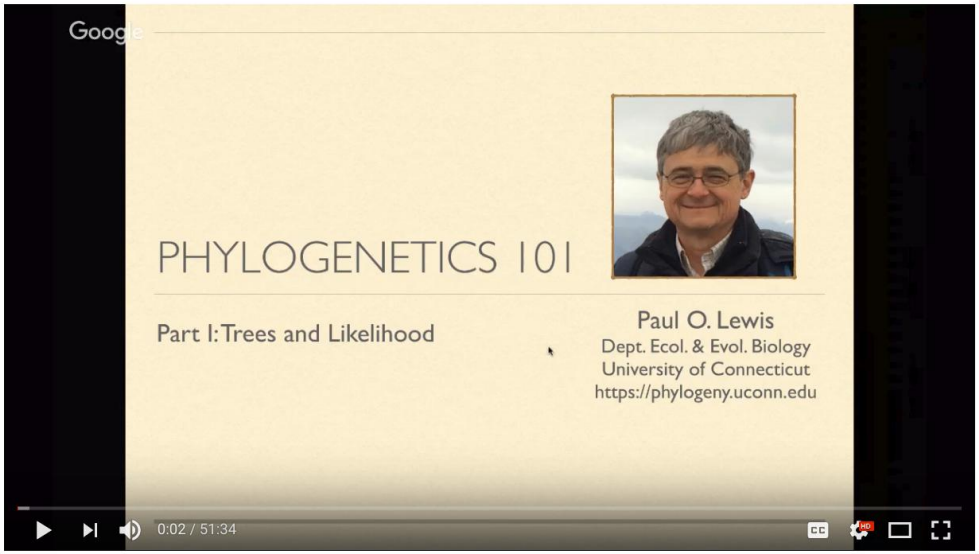
not available available

Based on funding mandates

Co-authors

Mary Kuhner
Department of Genome Science…

Full disclaimer: I'm borrowing A LOT of this material from Paul Lewis (Uconn) (Check out his teaching materials!)



Phyloseminar #76: Paul Lewis (UConn) Primer part 1

701 views LIKE DISLIKE SHARE

phyloseminar.org
Streamed live on Apr 18, 2018

SUBSCRIBED 973

Primer part 1: tree terminology and substitution models

Slides: https://git.io/vpIW9

SHOW MORE

If two DNA sequences are unrelated, what % of bases (aligned sites) do you expect to be identical?

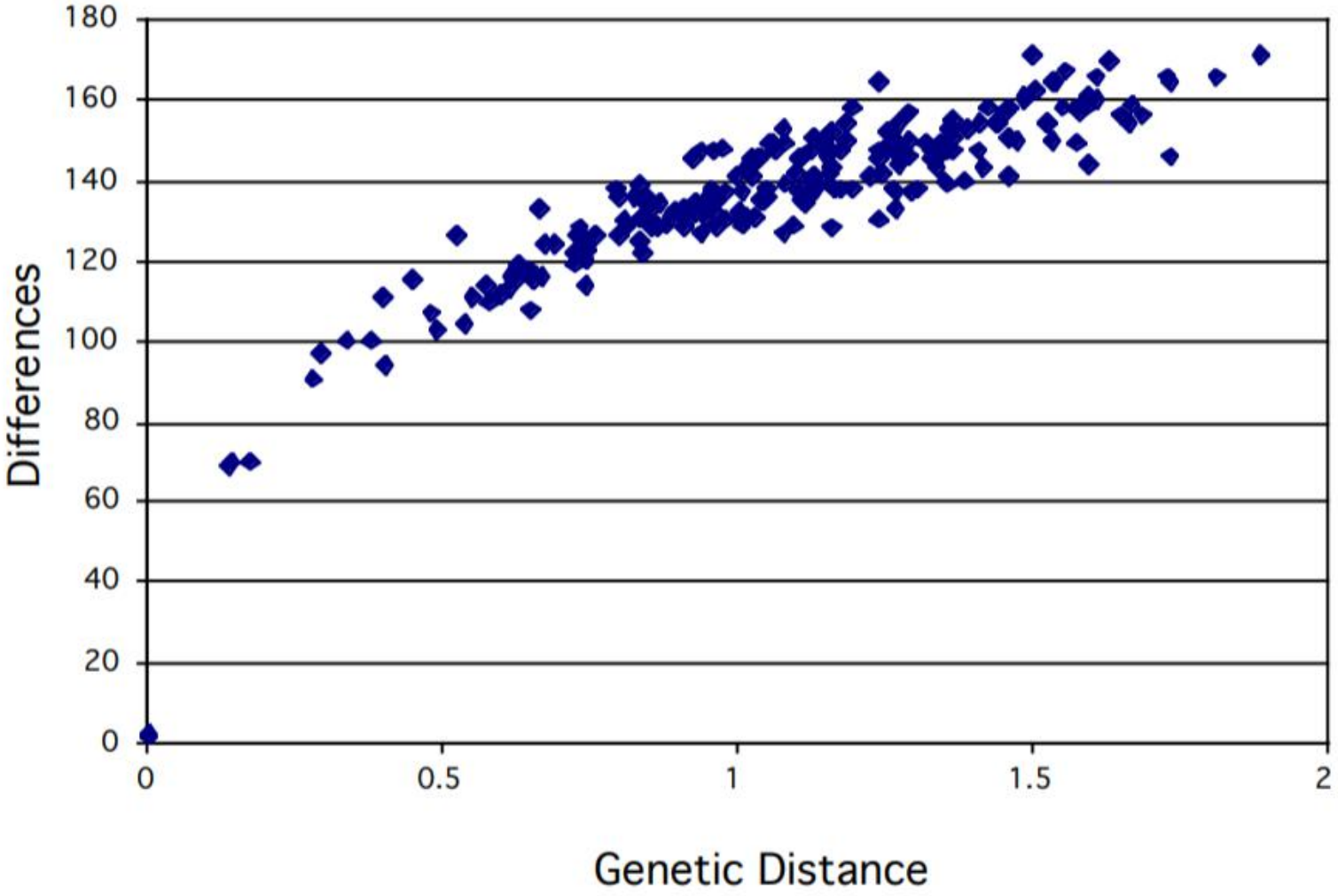
A. 50%

B. 25%

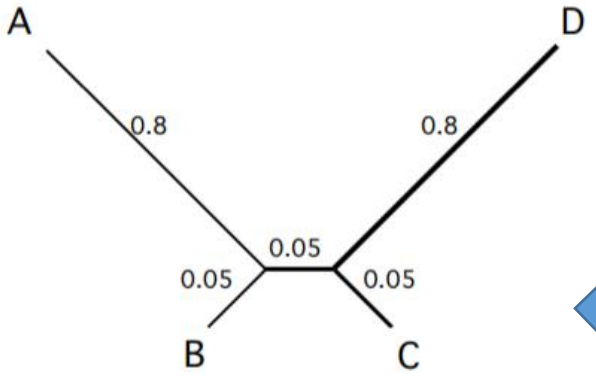
C. 0%

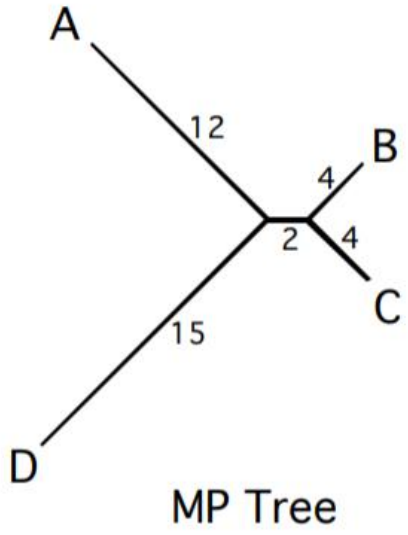
D. I need more information

Why do we need statistics?



"Long-branch attraction"





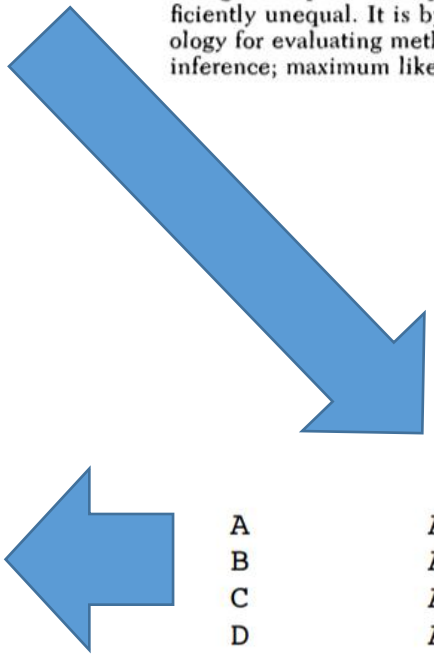
MP Tree

CASES IN WHICH PARSIMONY OR COMPATIBILITY METHODS WILL BE POSITIVELY MISLEADING[1]

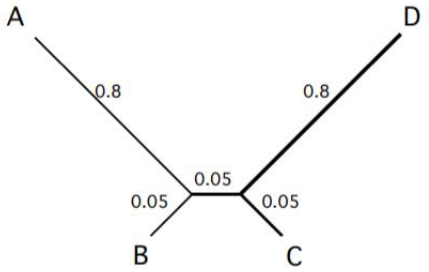
JOSEPH FELSENSTEIN

Abstract

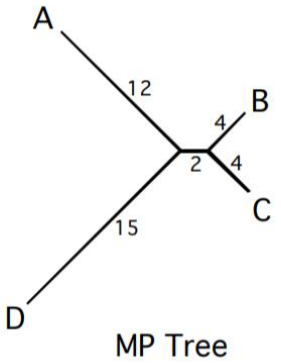
Felsenstein, J. (Department of Genetics, University of Washington, Seattle, WA 98195) 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.—For some simple three- and four-species cases involving a character with two states, it is determined under what conditions several methods of phylogenetic inference will fail to converge to the true phylogeny as more and more data are accumulated. The methods are the Camin-Sokal parsimony method, the compatibility method, and Farris's unrooted Wagner tree parsimony method. In all cases the conditions for this failure (which is the failure to be statistically consistent) are essentially that parallel changes exceed informative, nonparallel changes. It is possible for these methods to be inconsistent even when change is improbable a priori, provided that evolutionary rates in different lineages are sufficiently unequal. It is by extension of this approach that we may provide a sound methodology for evaluating methods of phylogenetic inference. [Numerical cladistics; phylogenetic inference; maximum likelihood estimation; parsimony; compatibility.]




```
A   ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
B   ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
C   ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
D   AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC
```

P(ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC | 

)

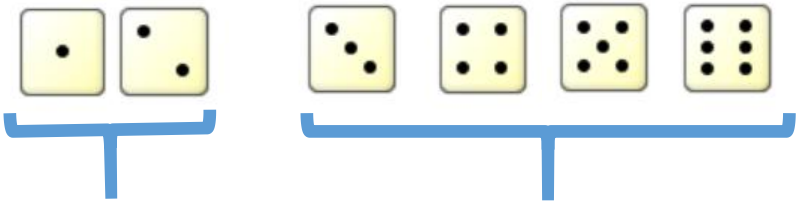
P(ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC | 

MP Tree
)

read as: The probability of this alignment (the data) GIVEN or CONDITIONAL UPON this being the true tree.

How do we treat trees probabilistically?

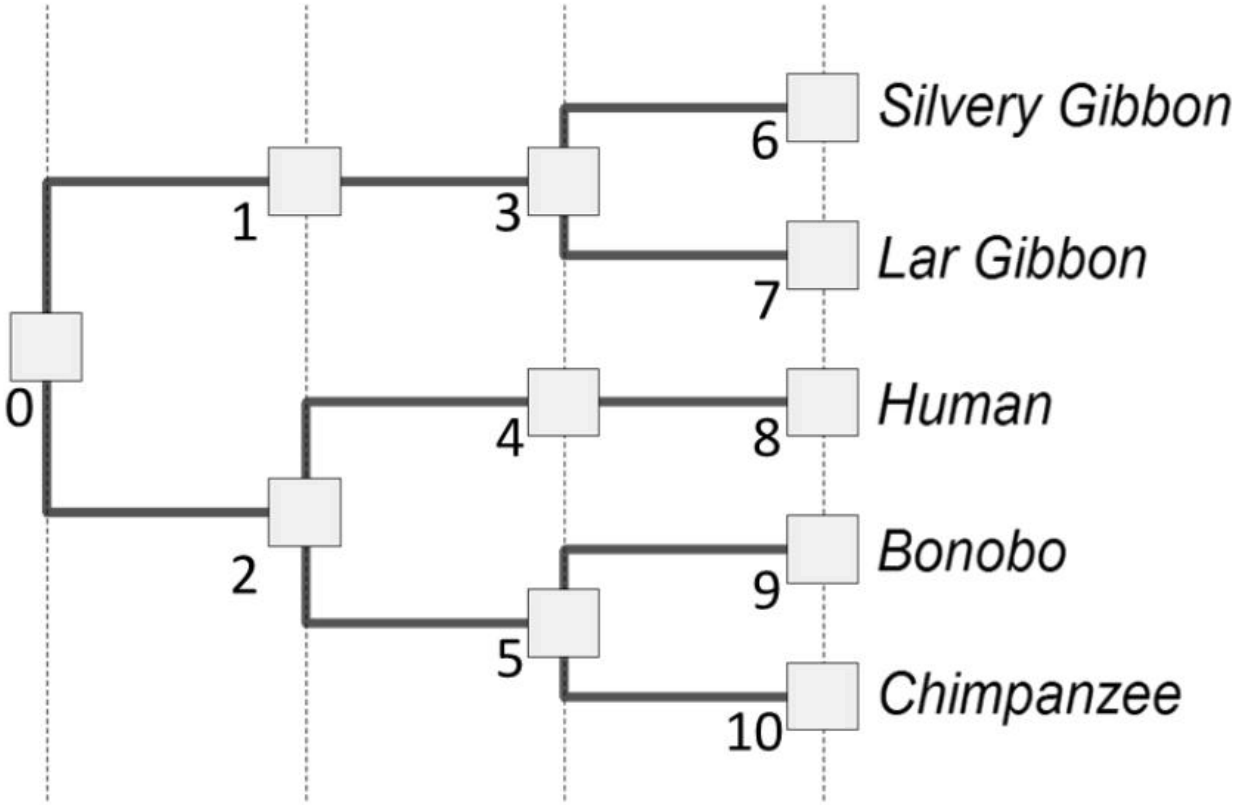
Let's simplify to a binary (0/1) trait (rather than 4 DNA bases)

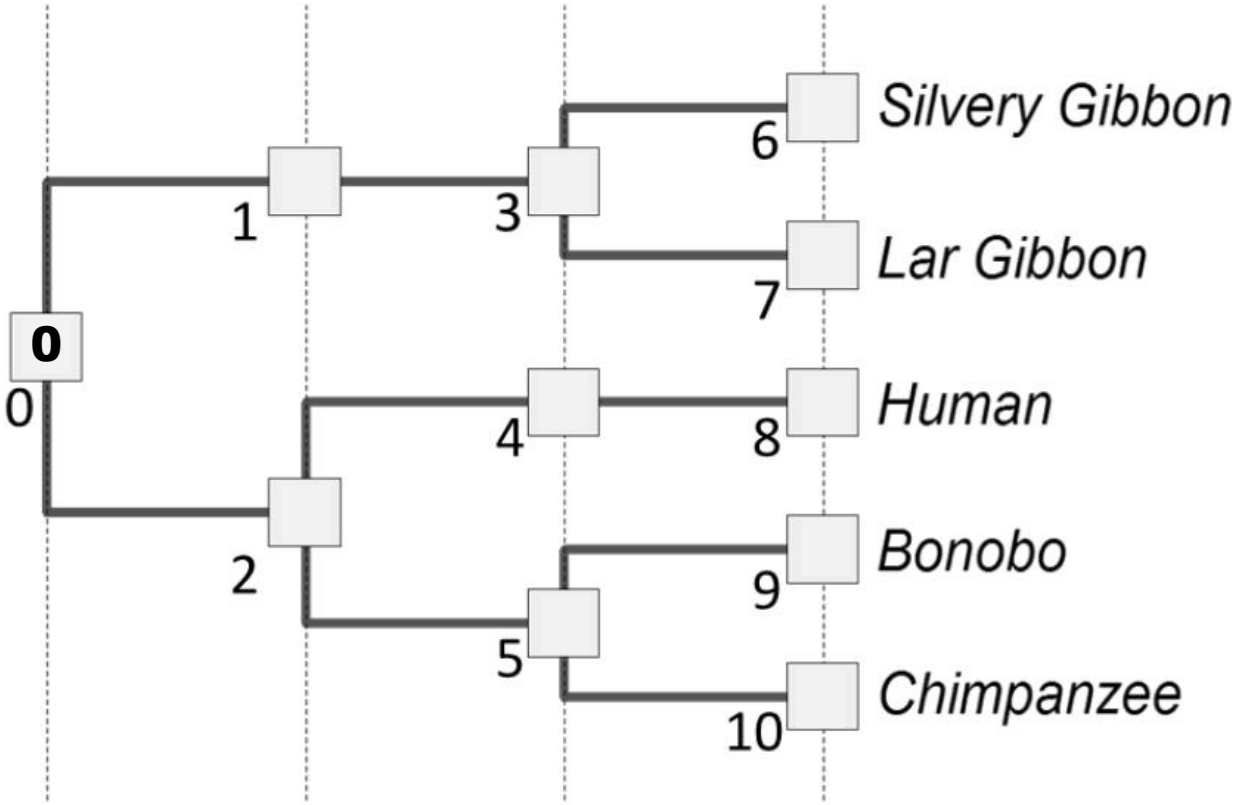


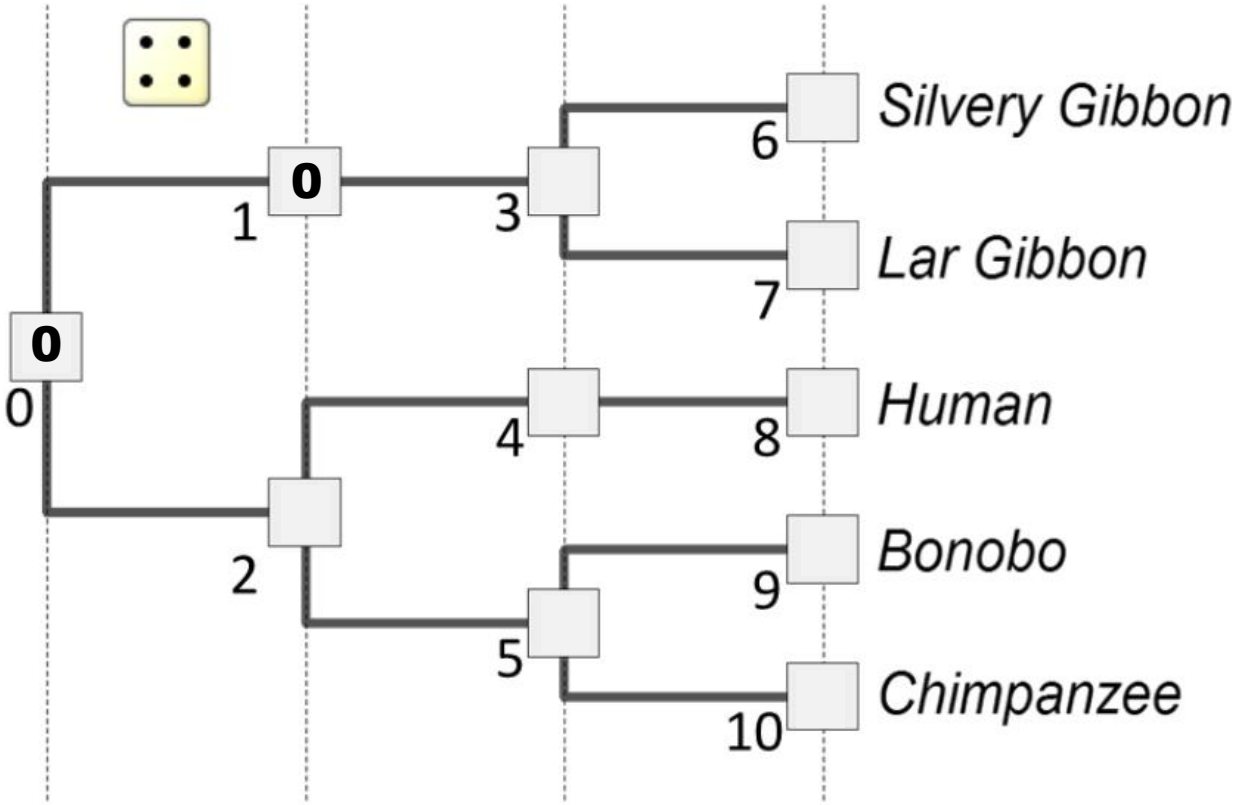
If these come up, a substitution occurs

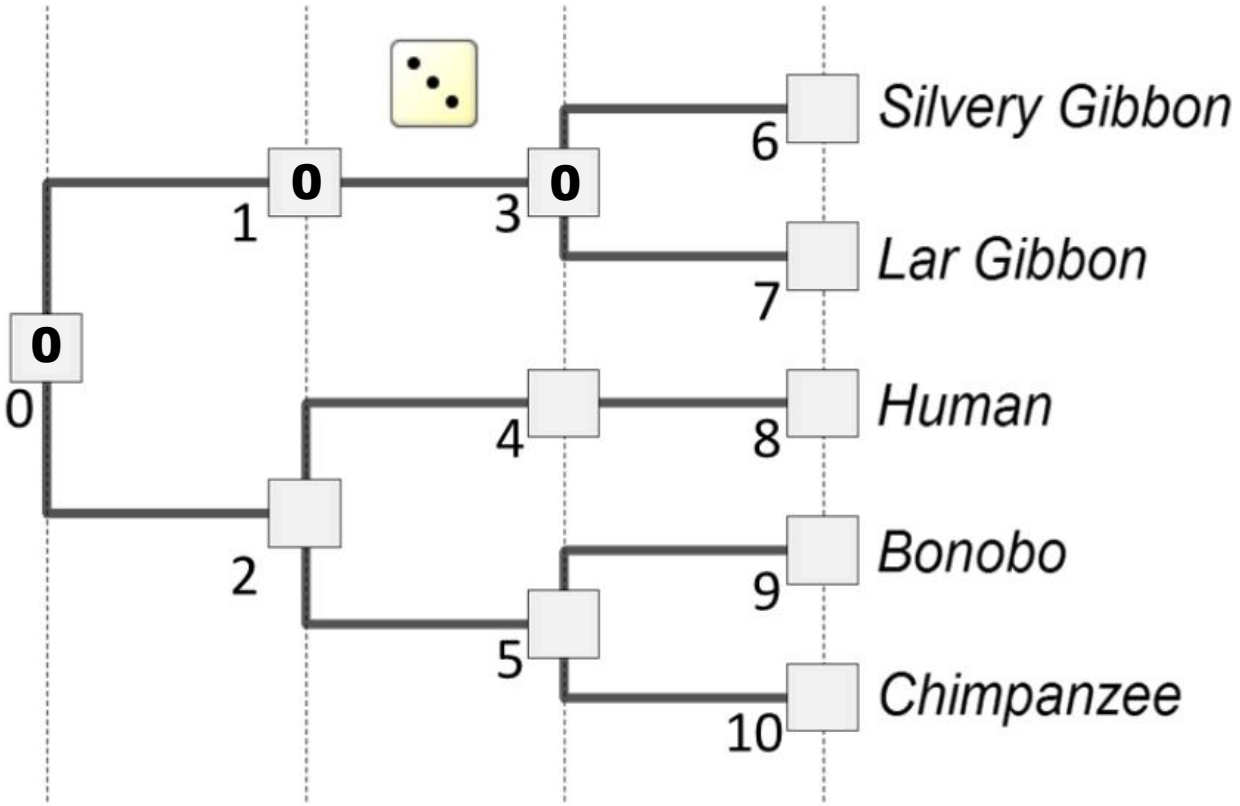
If these come up, no substitution occured

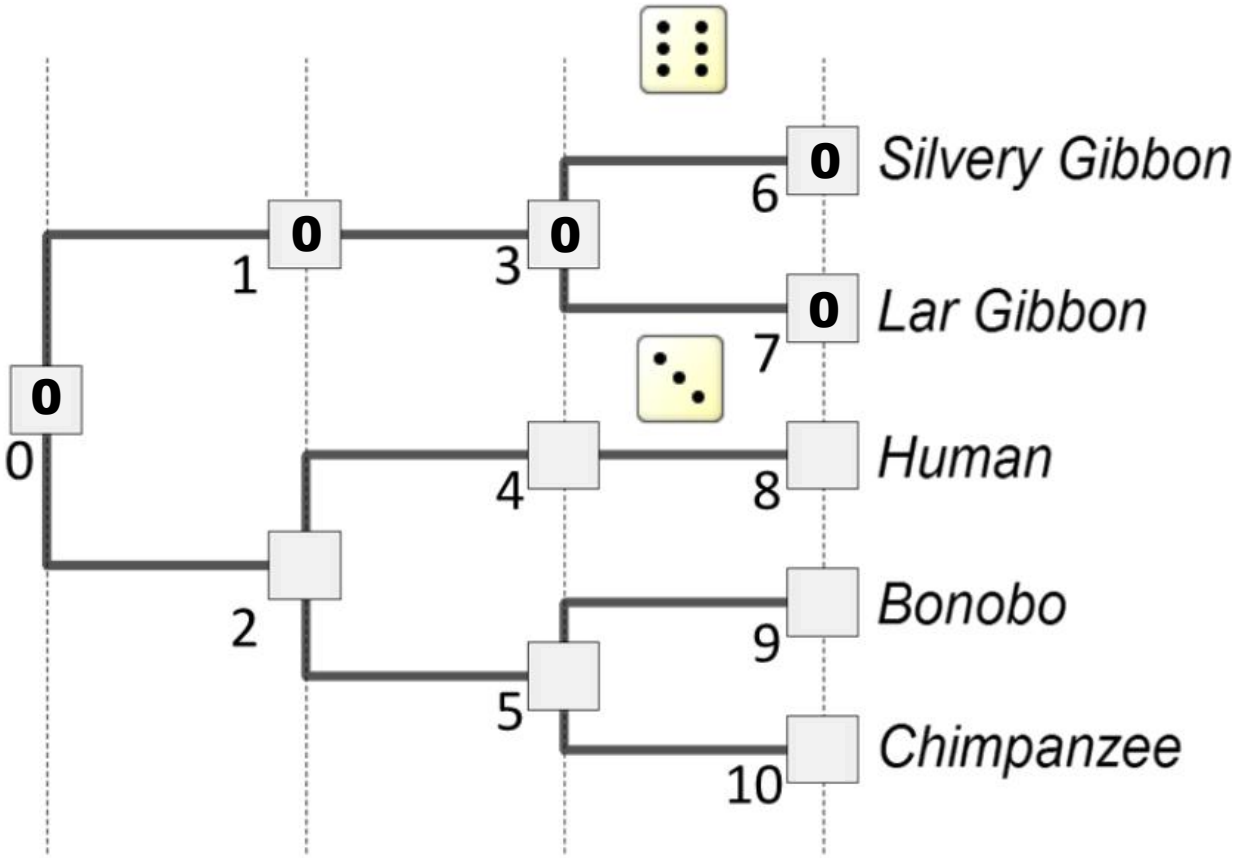
If a substitution occurs, choose *which* substitution occurs by a coin flip.

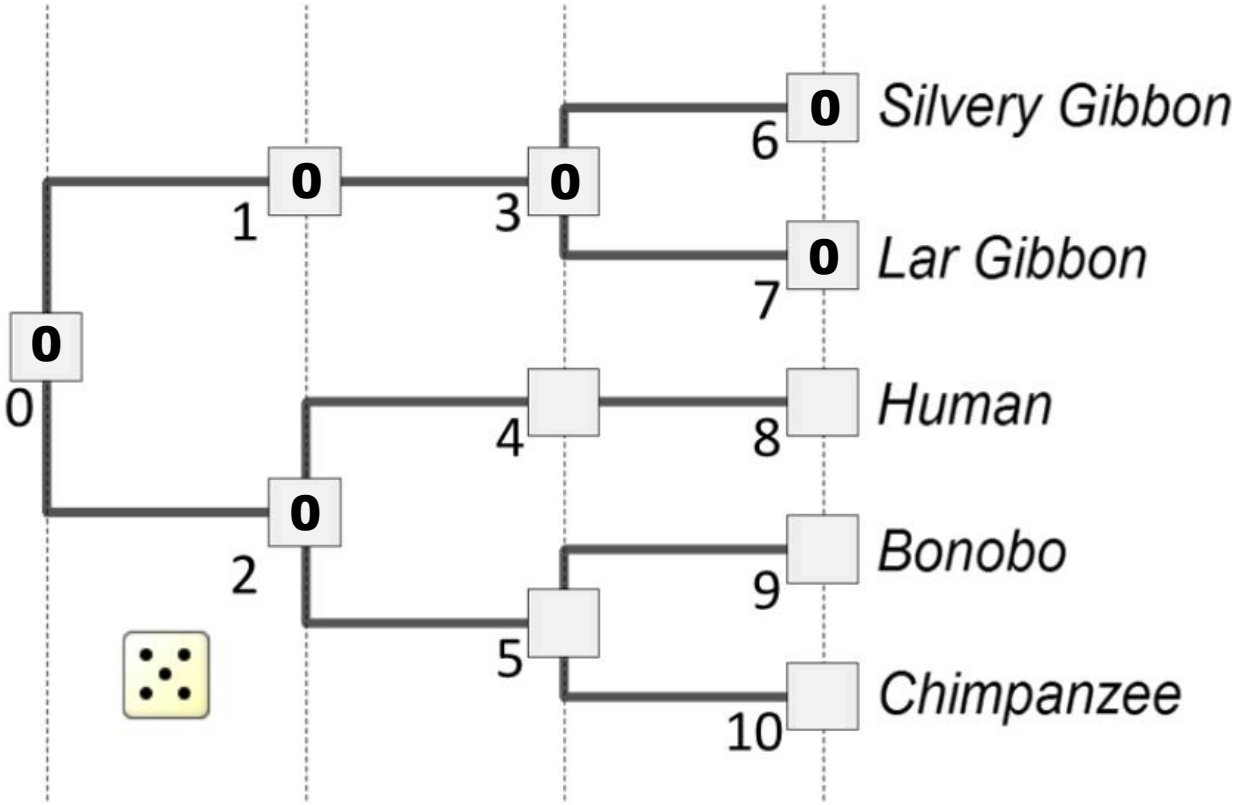

Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee
0
1
2
3
4
5
6
7
8
9
10

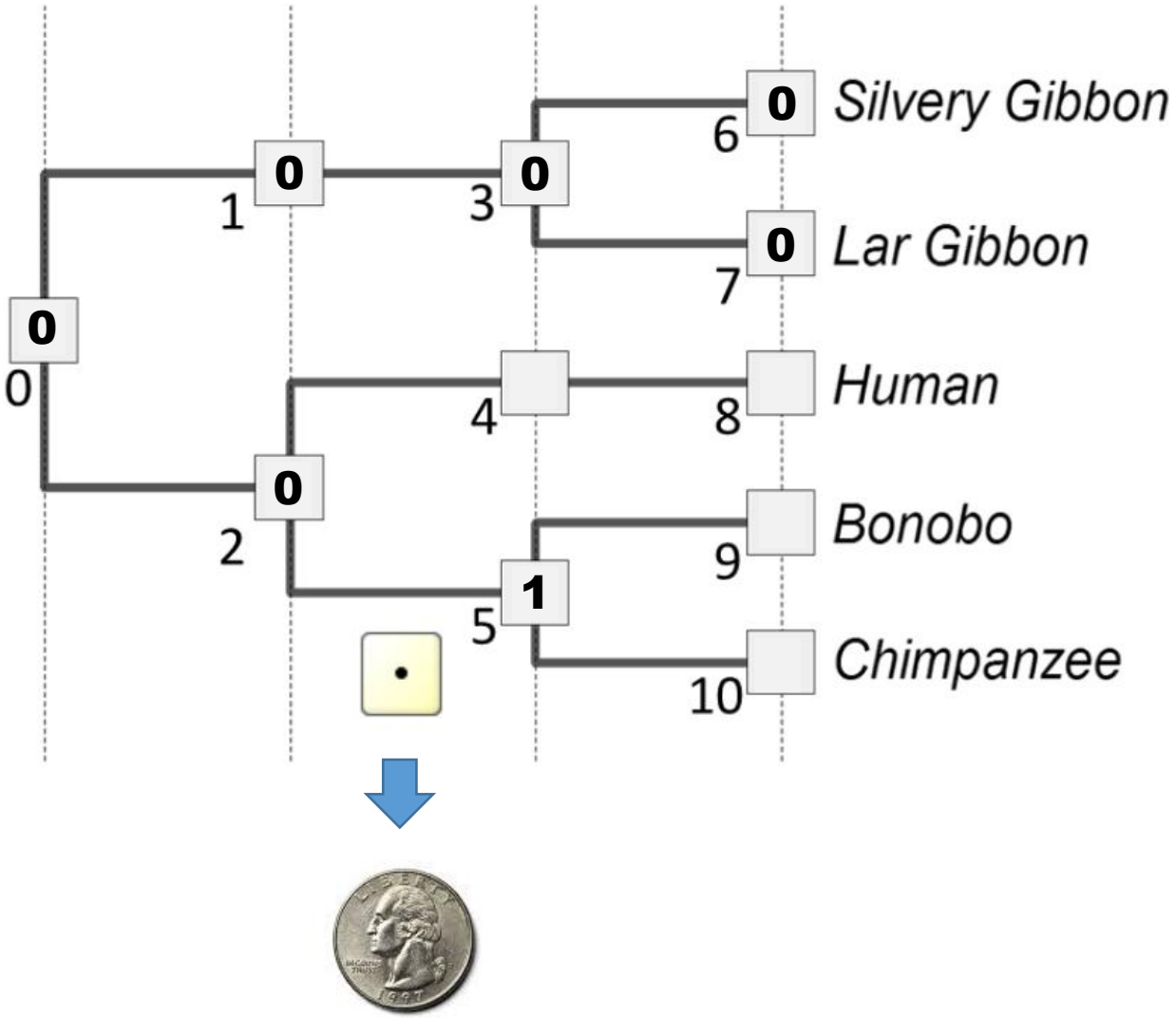

0
1
2
3
4
5
6
7
8
9
10
Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee

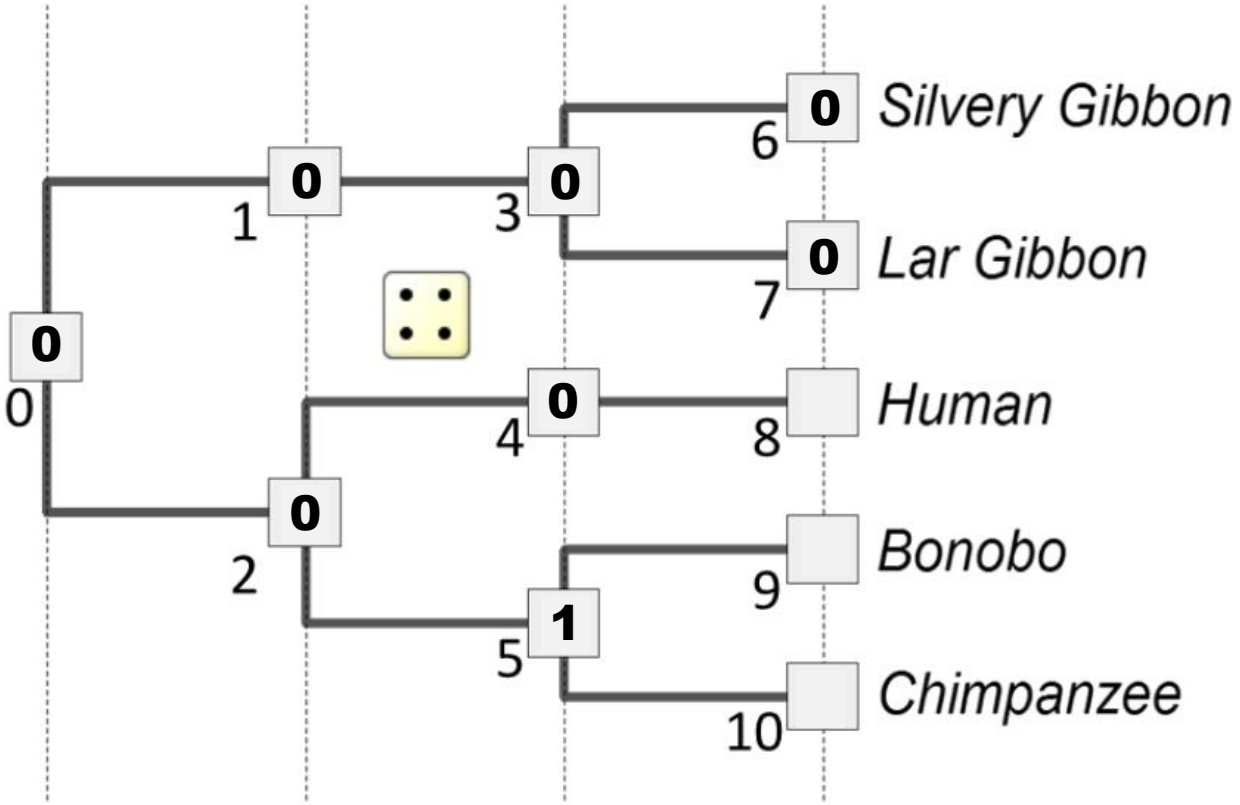

0
0
1
2
3
4
5
6
7
8
9
10
Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee

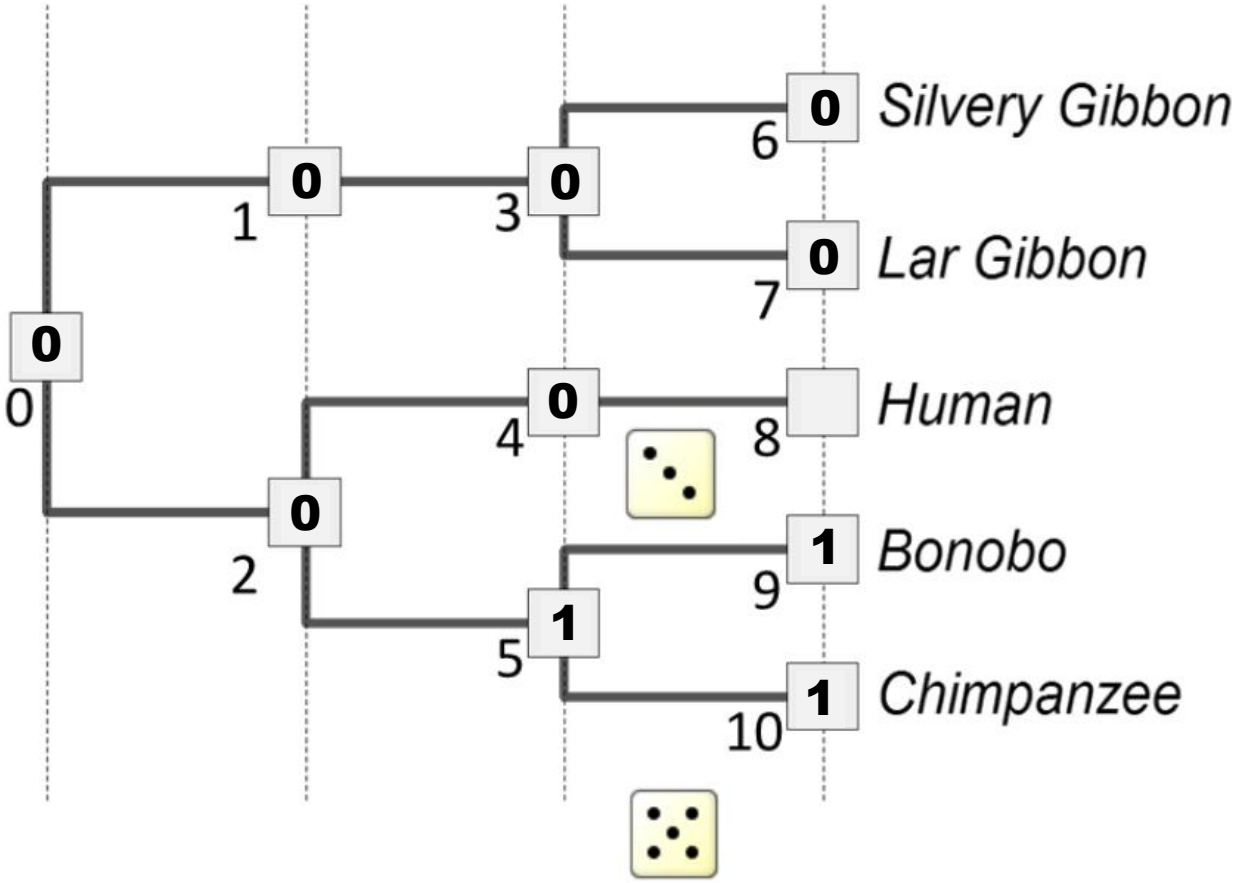

0
1
2
3
4
5
6
7
8
9
10
Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee

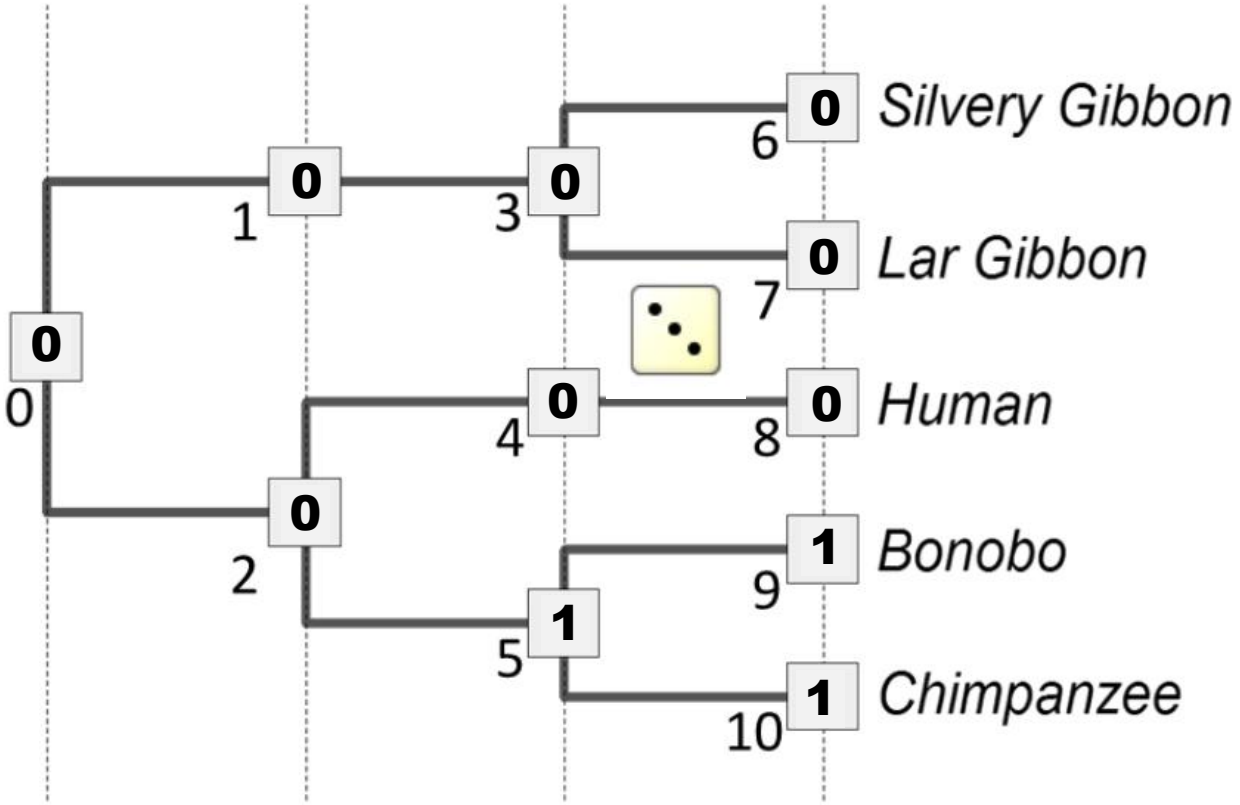

0
0
1
0
3
0
6
0
Silvery Gibbon
7
0
Lar Gibbon
0
4
8
Human
2
9
Bonobo
5
10
Chimpanzee


0
0
1
0
2
0
3
0
4
5
6
0
Silvery Gibbon
7
0
Lar Gibbon
8
Human
9
Bonobo
10
Chimpanzee


0
0
1
0
2
0
3
0
4
5
1
6
0
Silvery Gibbon
7
0
Lar Gibbon
8
Human
9
Bonobo
10
Chimpanzee


0
0
1
0
2
0
3
0
4
0
5
1
6
0
7
0
8
9
10
Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee


0
0
1
0
2
0
3
0
4
0
5
1
6
0
7
0
8
9
1
10
1
Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee


0
0
1
0
2
0
3
0
4
0
5
1
6
0
Silvery Gibbon
7
0
Lar Gibbon
8
0
Human
9
1
Bonobo
10
1
Chimpanzee

Combining probabilities: The AND rule

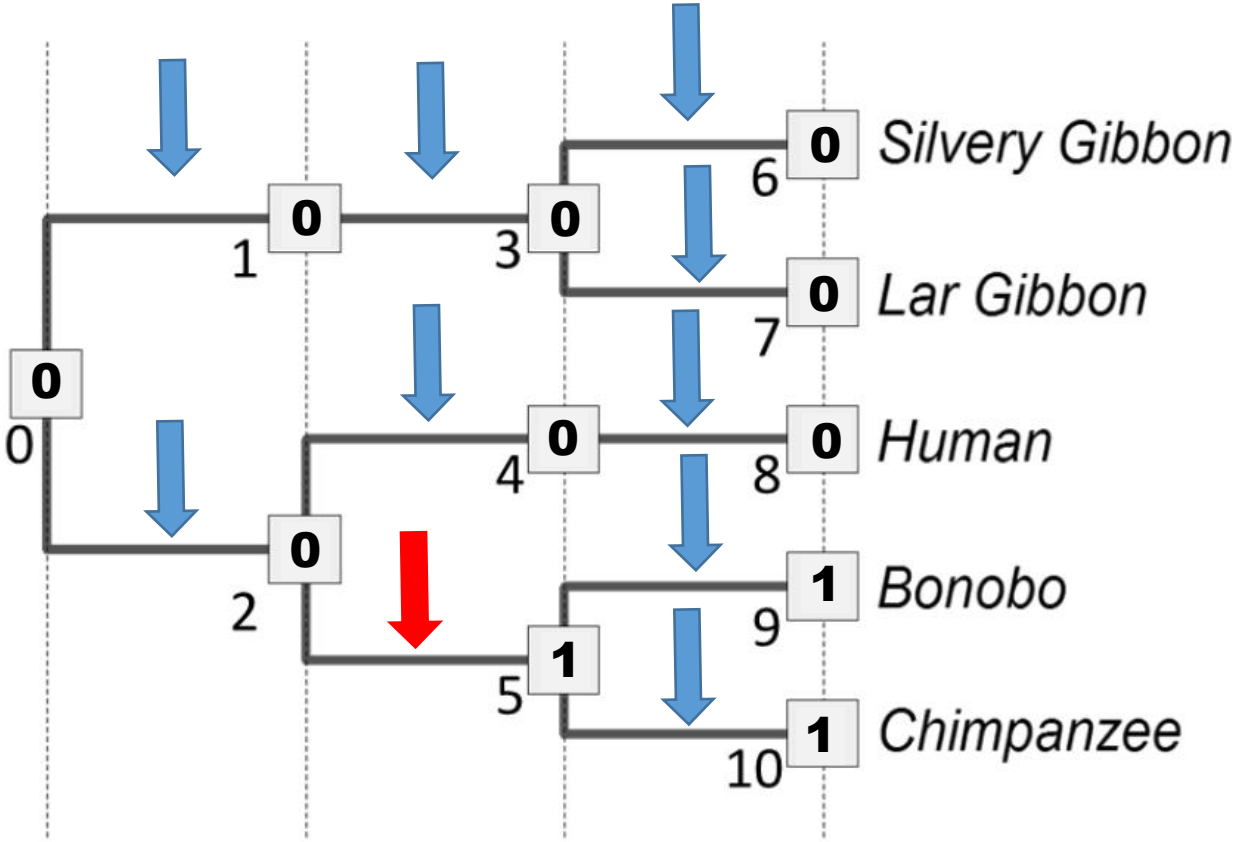
If two independent events occur, MULTIPLY their individual probabilities to get the full probability of an event. Each roll of the dice (and each flip of the coin) is *independent.*

Using 2 dice, what is the probability of

 AND ?

$(1/6) \times (1/6) = 1/36$

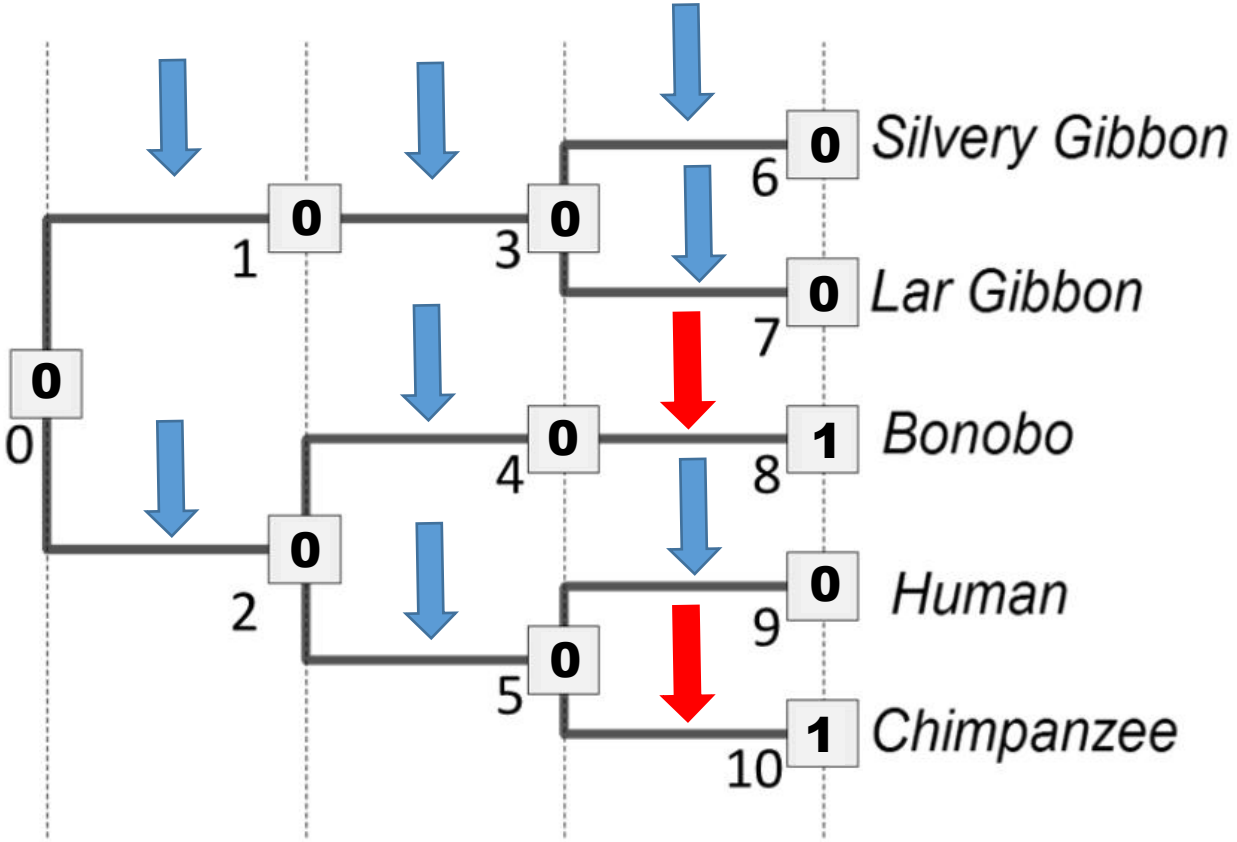
Combining probabilities: The AND rule in phylogenetics



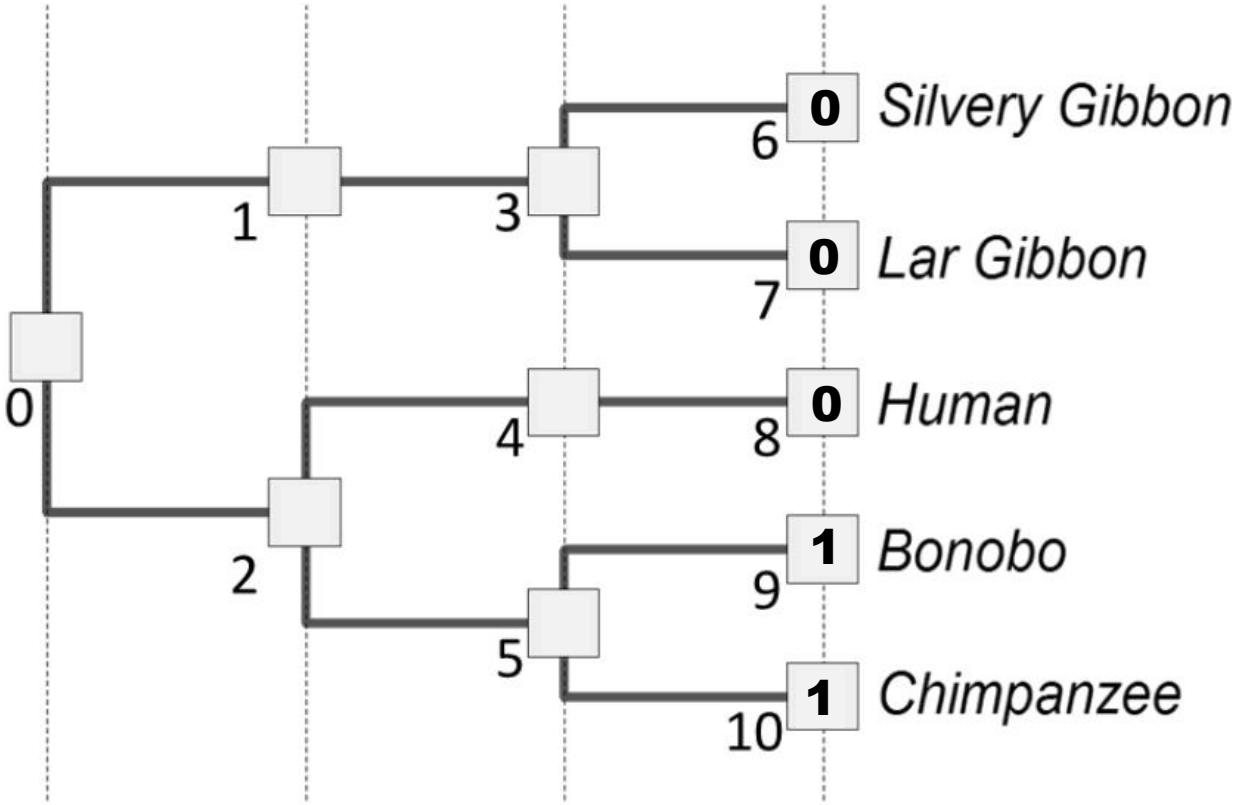
$$L(T_1 \mid D, N_i) = P(D, N_i \mid T_1)^* = (4/6)^9 \times [(2/6)(1/2)]^1 = 0.0043$$

Read as: The likelihood of Tree 1 given the data at the tips (D) and nodes (N_i) is equal to the probability of the data and node states given Tree 1 is true. (We can't ask "is Tree 1 true" so simply, so we use the word Likelihood instead)

Combining probabilities: The AND rule in phylogenetics



Likelihood(D, $N_j|T_2$)* = $(4/6)^8$ x $[(2/6)(1/2)]^2$=0.0011


0
1
2
3
4
5
6
7
8
9
10
0
Silvery Gibbon
0
Lar Gibbon
0
Human
1
Bonobo
1
Chimpanzee

Combining probabilities: The OR rule

Two mutually exclusive probabilities should be ADDED together to get the total probability of the two events

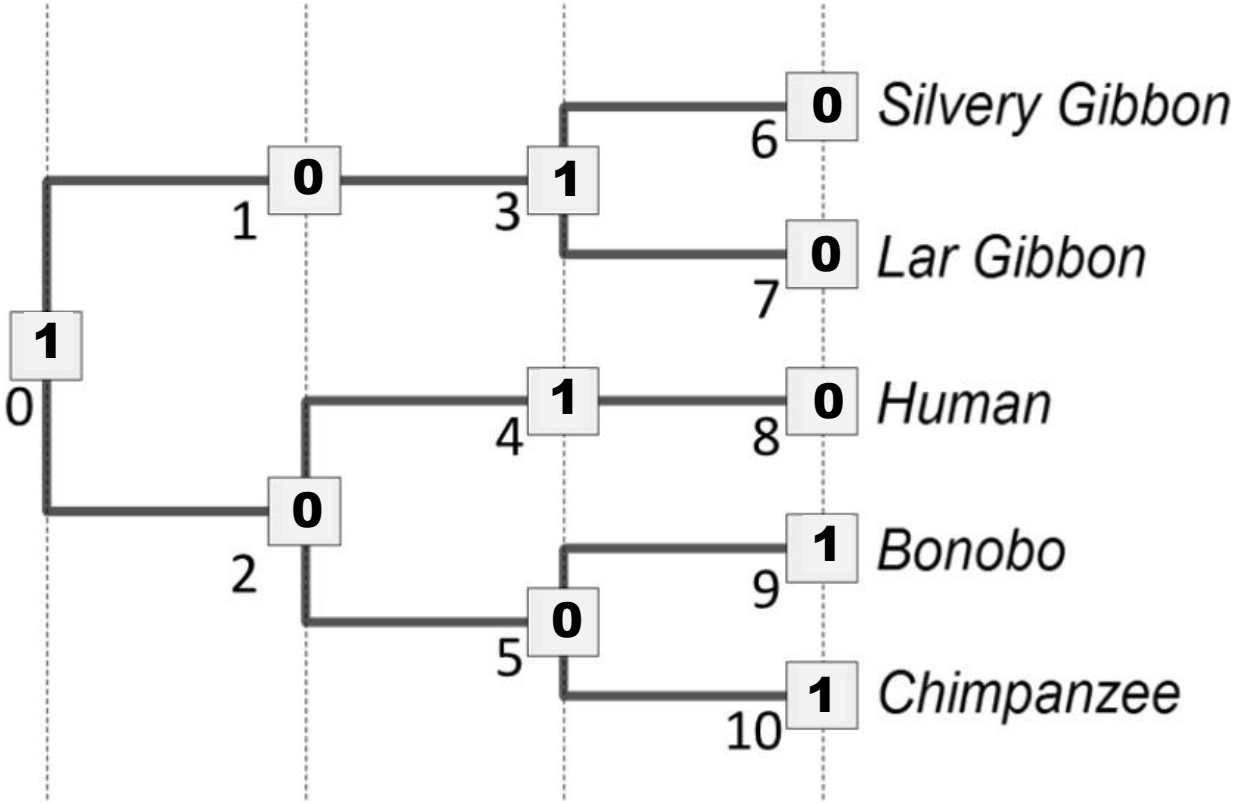
Using one die, what is the probability of

 OR ?

$(1/6) + (1/6) = 1/3$

Combining probabilities: The OR rule in phylogenetics

What's the likelihood of the data under "Tree$_1$"

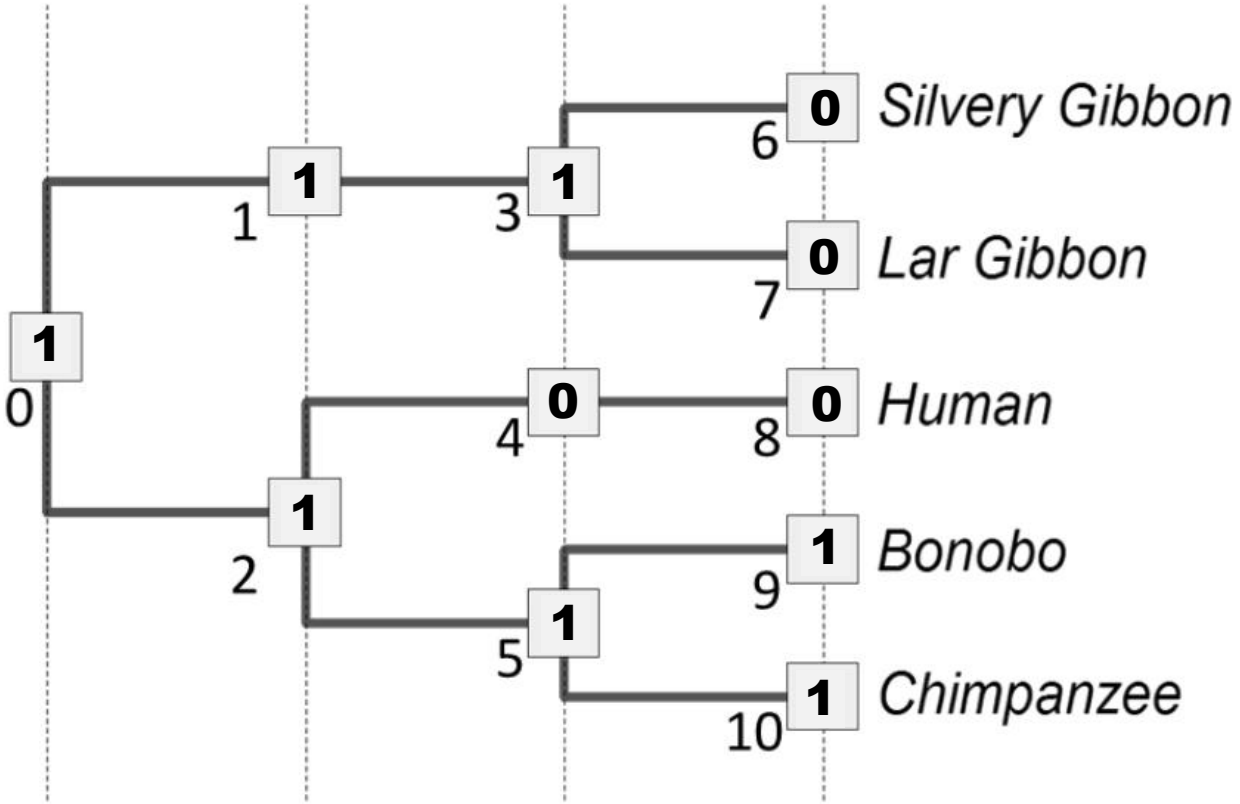


Likelihood (D, $N_k|T_1$)=$(4/6)^1$ x $[(2/6)(1/2)]^9$=0.000000066

This is a possible outcome under Tree$_1$, but not especially likely!

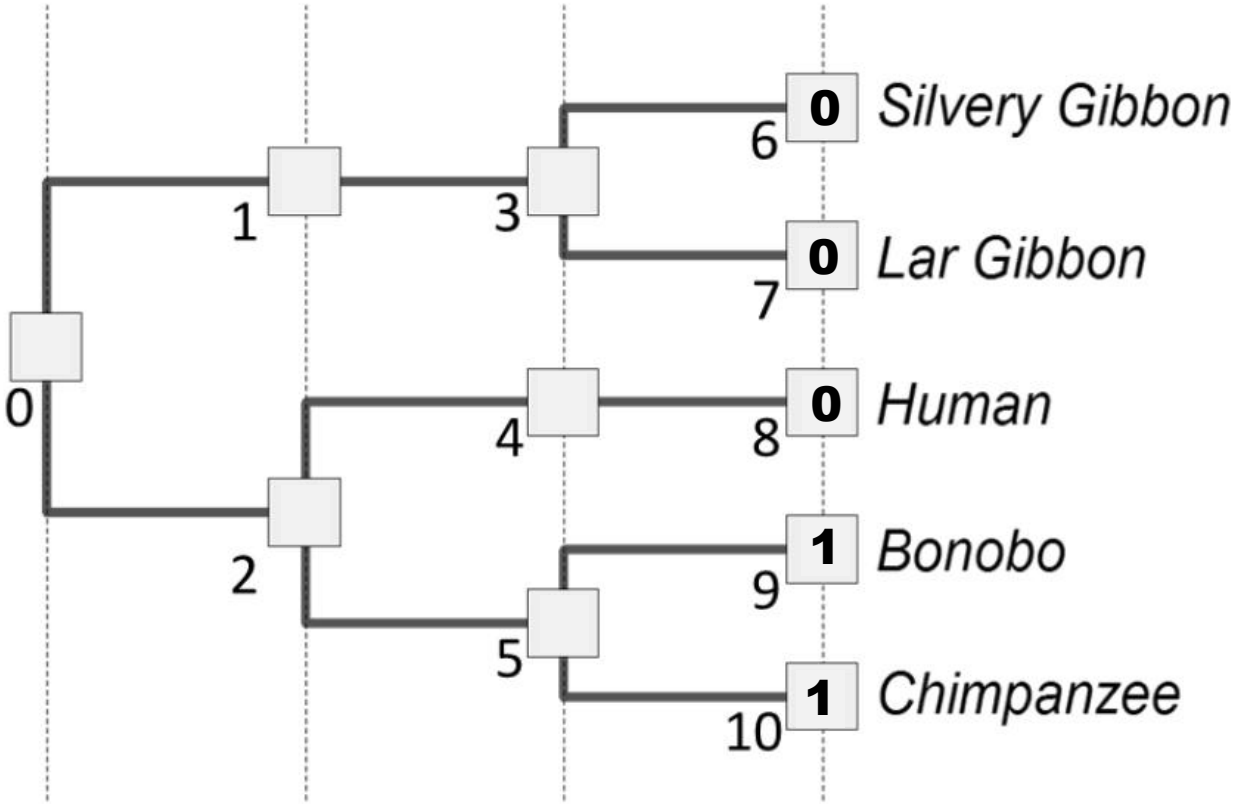
Combining probabilities: The OR rule in phylogenetics

What's the likelihood of the data under "Tree$_1$"



Likelihood $(D, N_k|T_1) = (4/6)^7 \times [(2/6)(1/2)]^3 = 0.00027$

This is also a possible outcome, a little bit more likely!



There are $2^6 = 64$ possible node assignments for $Tree_1$. We could calculate the likelihood of each one, then _____ them together, to get the total Likelihood($Tree_1$ | Data).

Now, like parsimony, you can do either a heuristic or exhaustive search for the tree with the maximum likelihood)

We used discrete time units. Biology will want *continuous time*

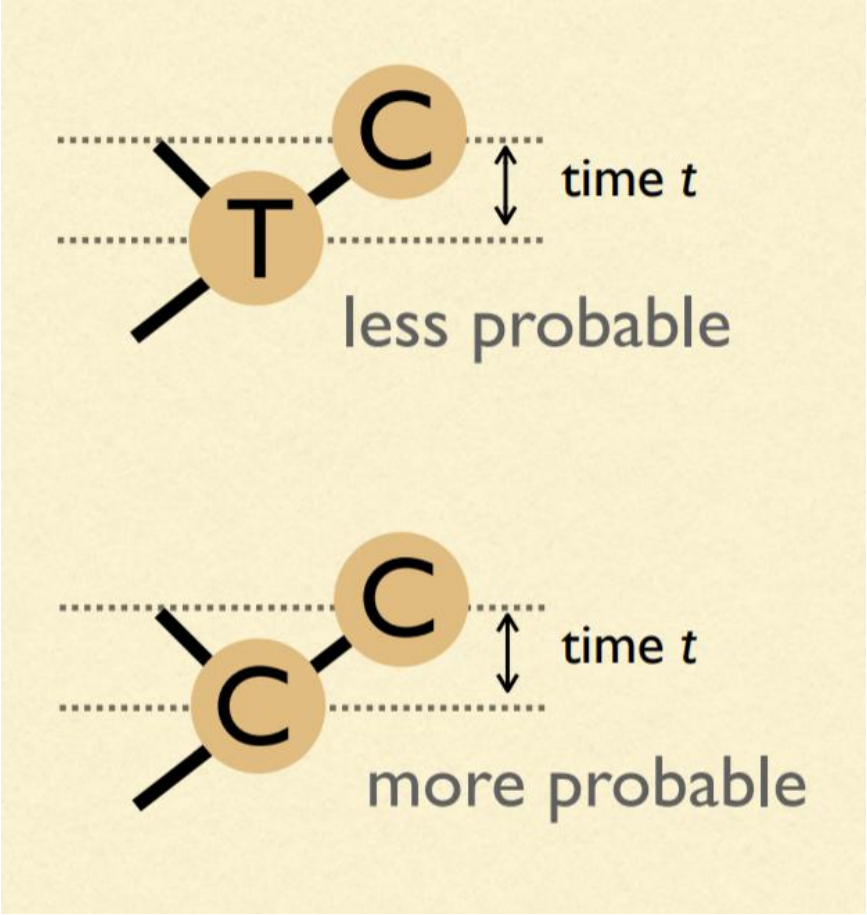
Continuous-Time Markov Models

Markov assumption **= probability of change depends only on current state, not how long it has been in that state**

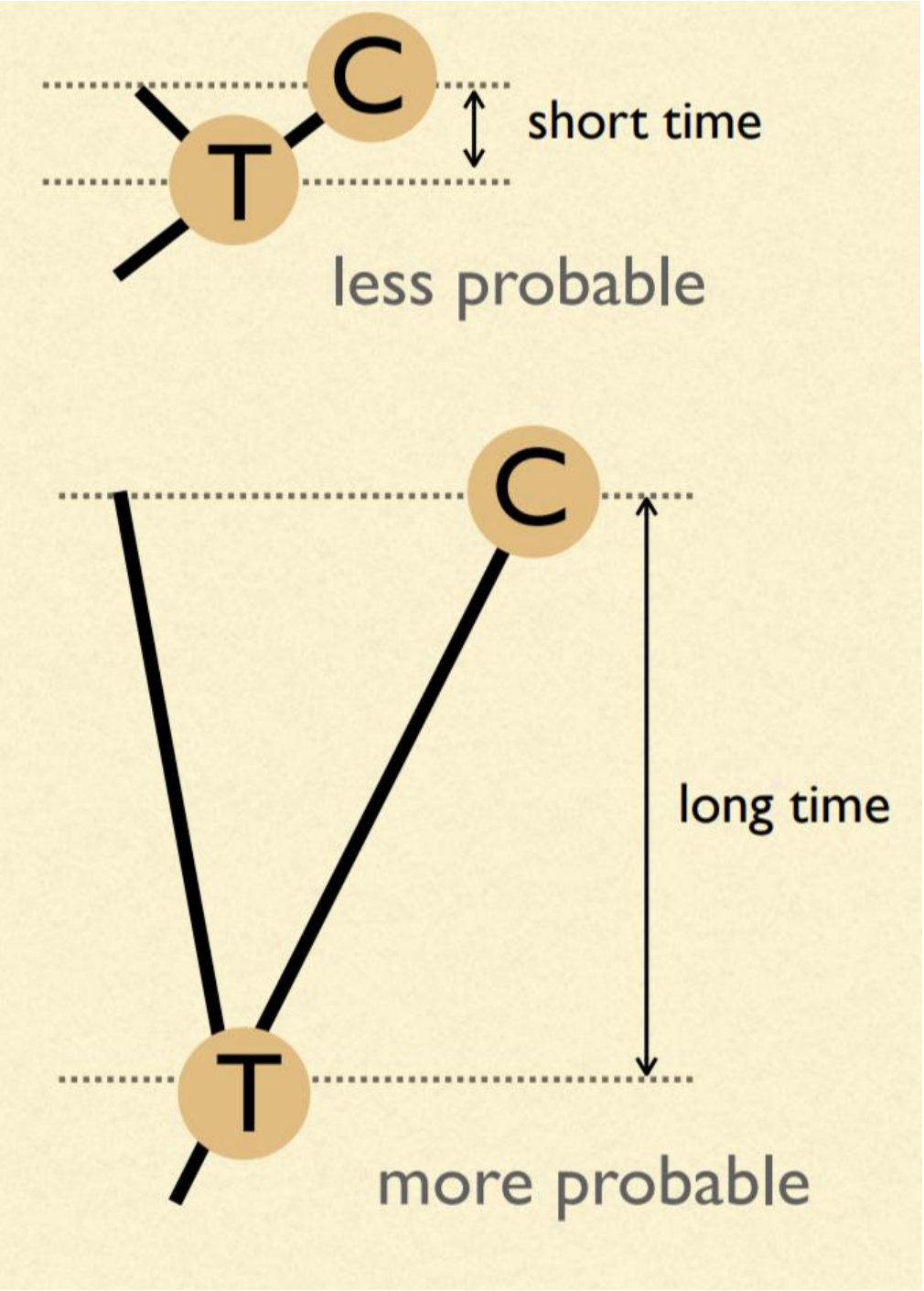
Our model of change depends on time: We must estimate branch lengths

Units of branch length will be expected number of substitutions per site

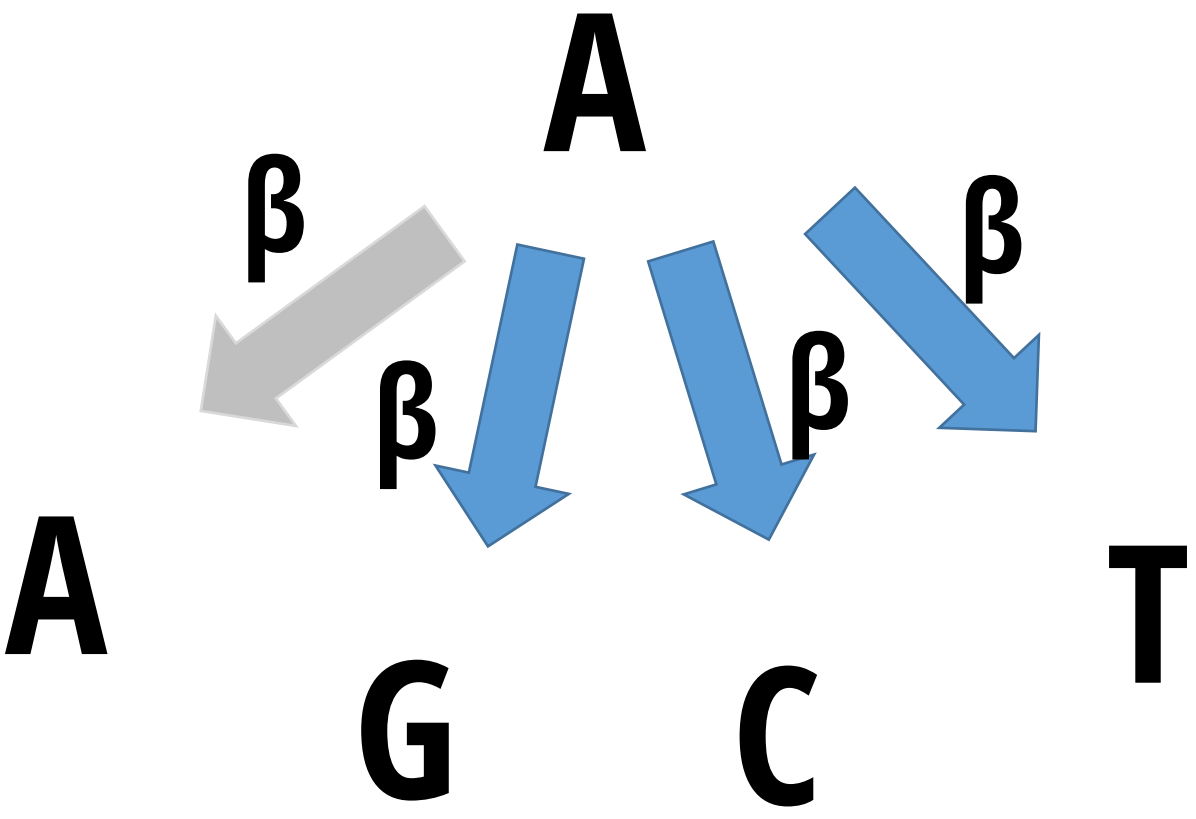
(= rate of substitions x time)



Probabilities are dependent on time



$\mu = 4\beta$



A (very) simple phylogeny…

P_{AA} =

Probability nothing happened +

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

A (very) simple phylogeny...

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

Probability something doesn't happen

Probability at least one thing happens

Probability that the last thing that happened ends in an A

A (very) simple phylogeny…

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…

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…

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

Probability at least one thing happens

Probability that the last thing that happened ends in an T

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

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

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

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

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

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

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

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

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

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

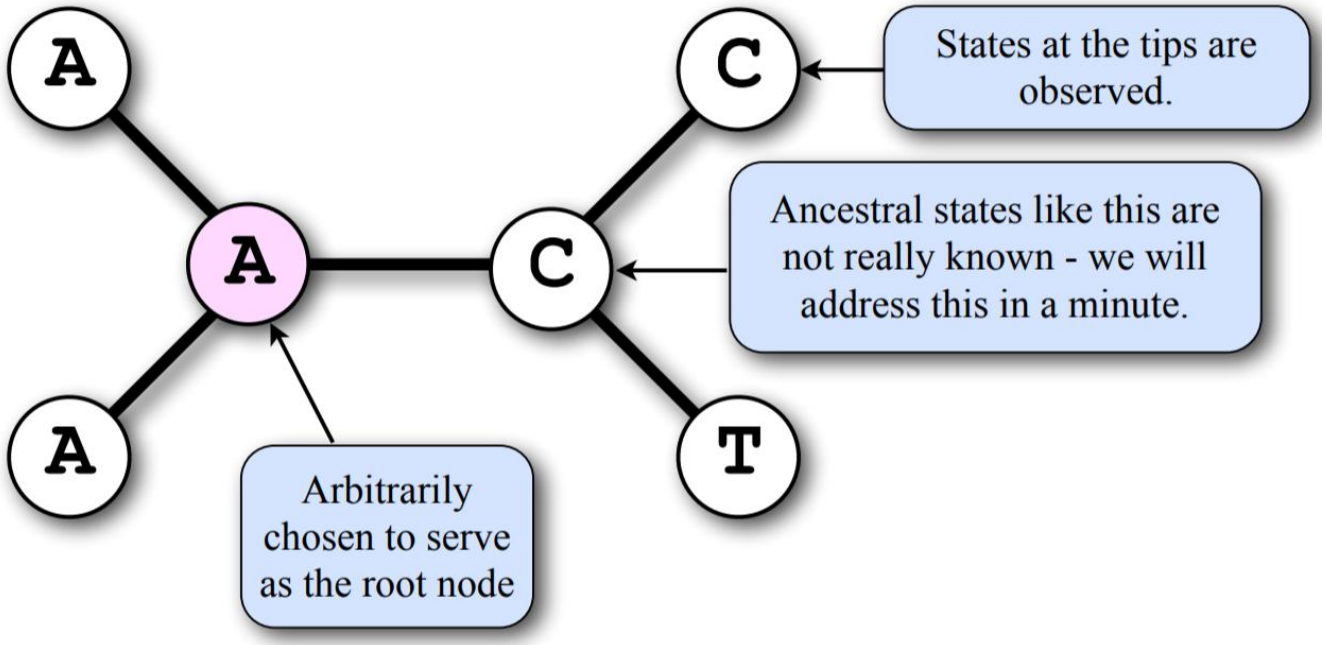
$$P_{AC} = (1/4) - (1/4)(e^{-4v/3})$$

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

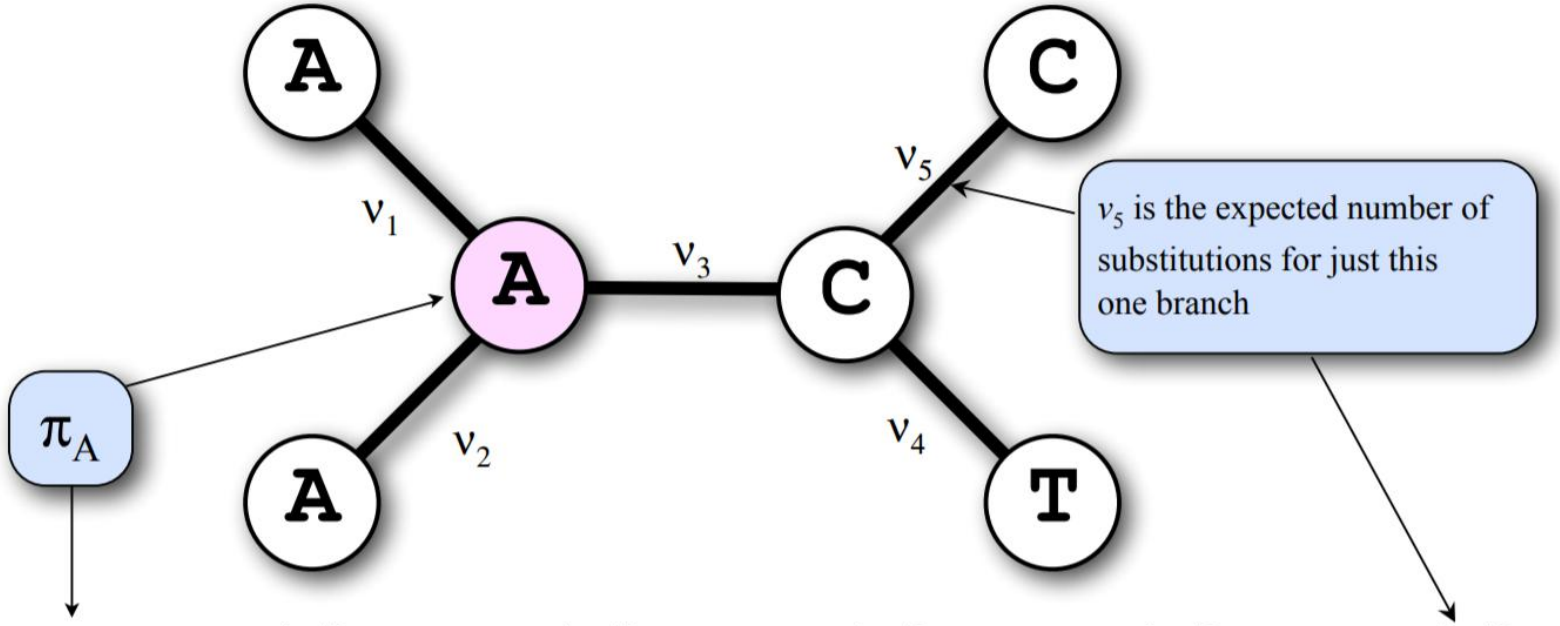
Sanity check:
Do they all add to 1?

Likelihood of an unrooted tree

(data shown for only one site)



Likelihood for site k

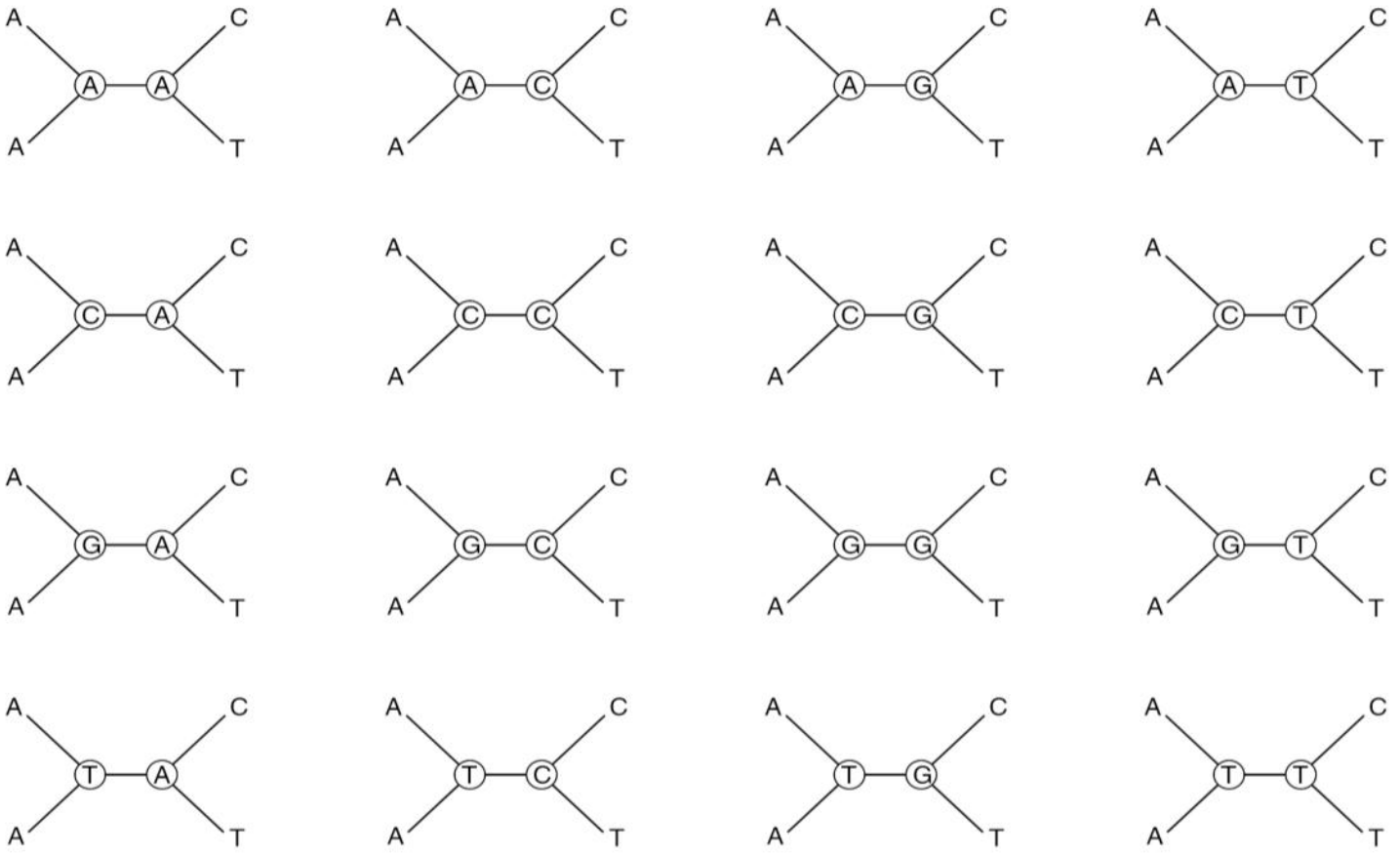


$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_5/3}\right]$$

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

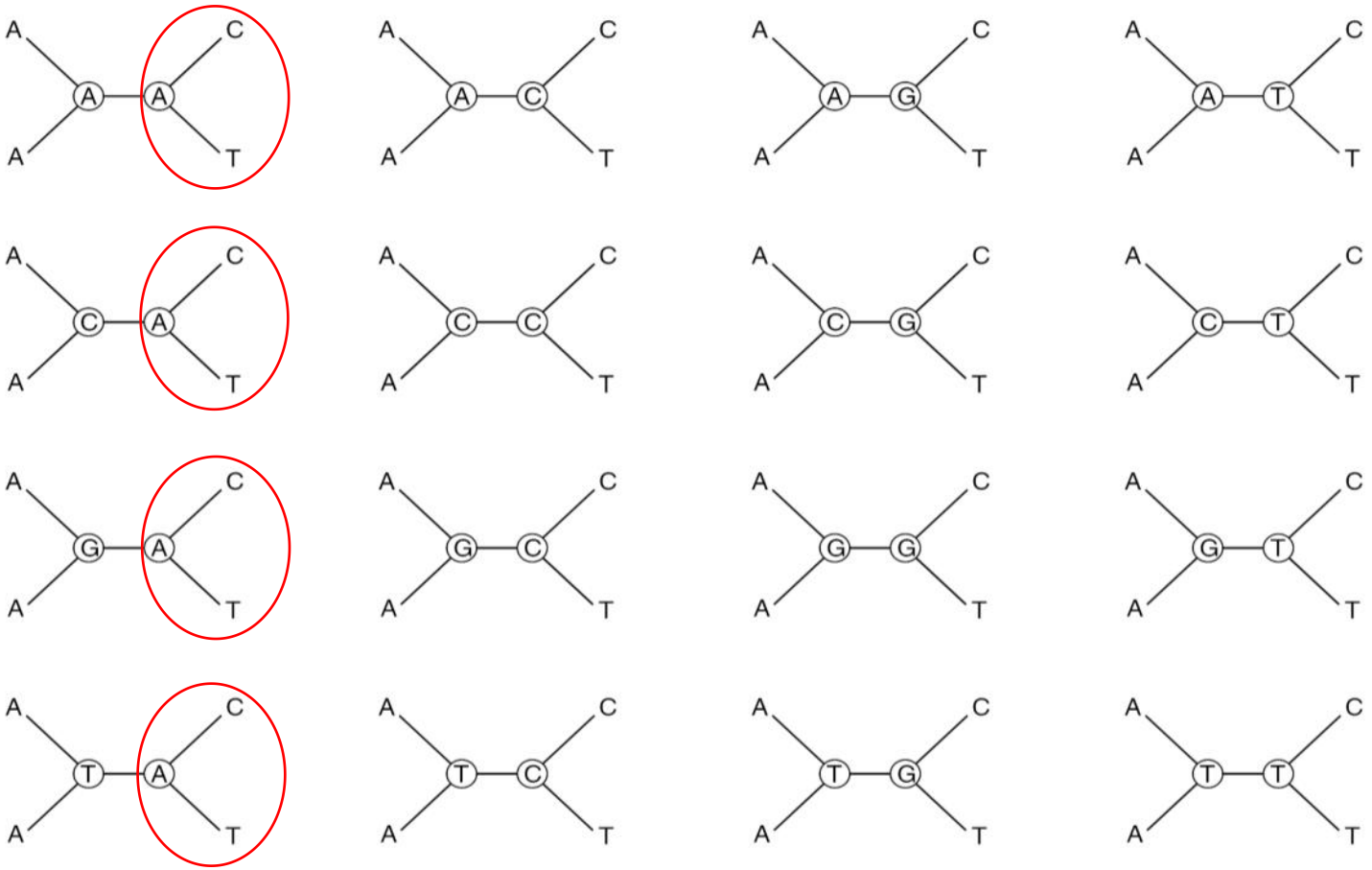
Note use of the AND probability rule

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them



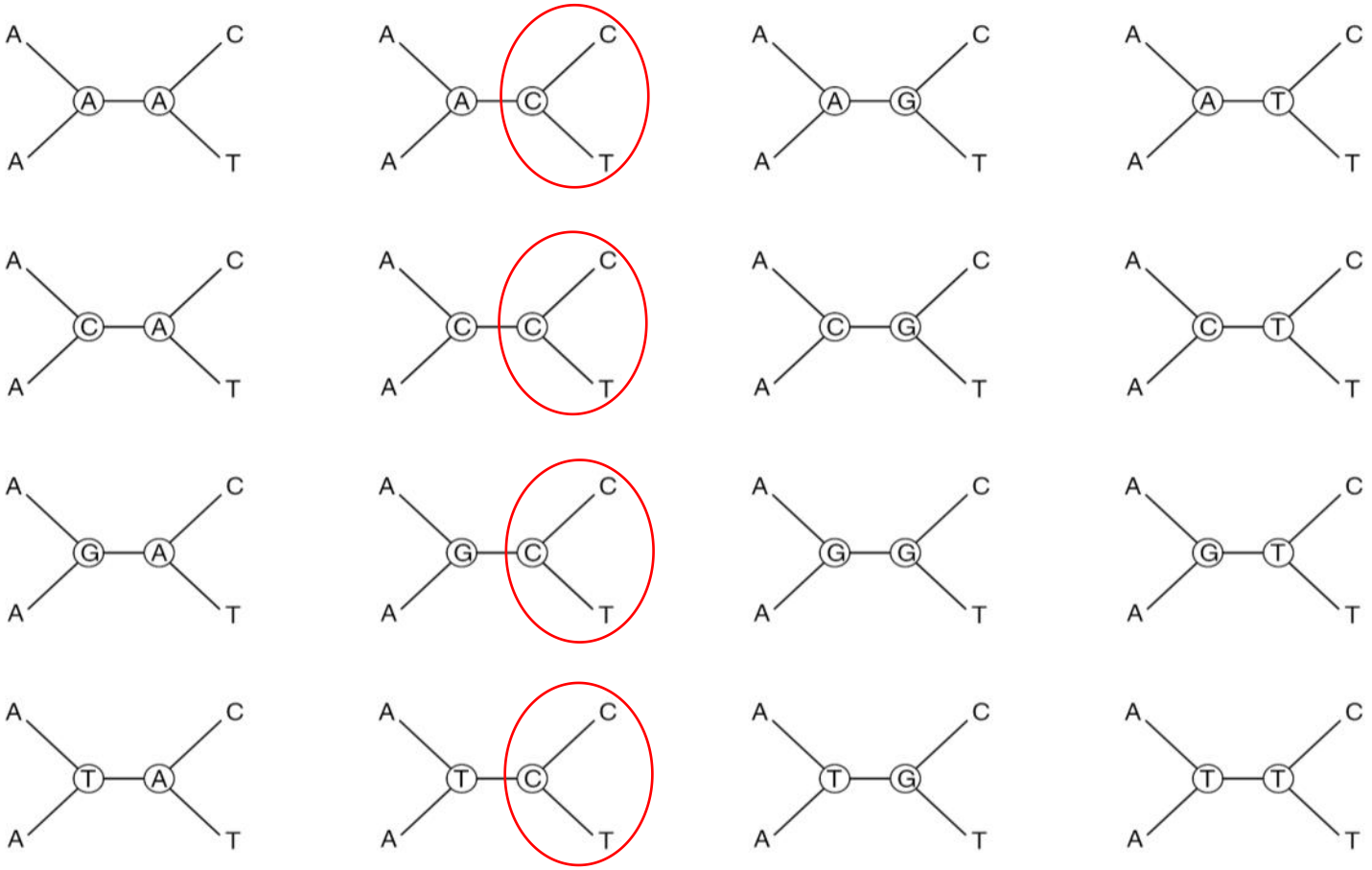
Note use of the OR probability rule

Pruning algorithm



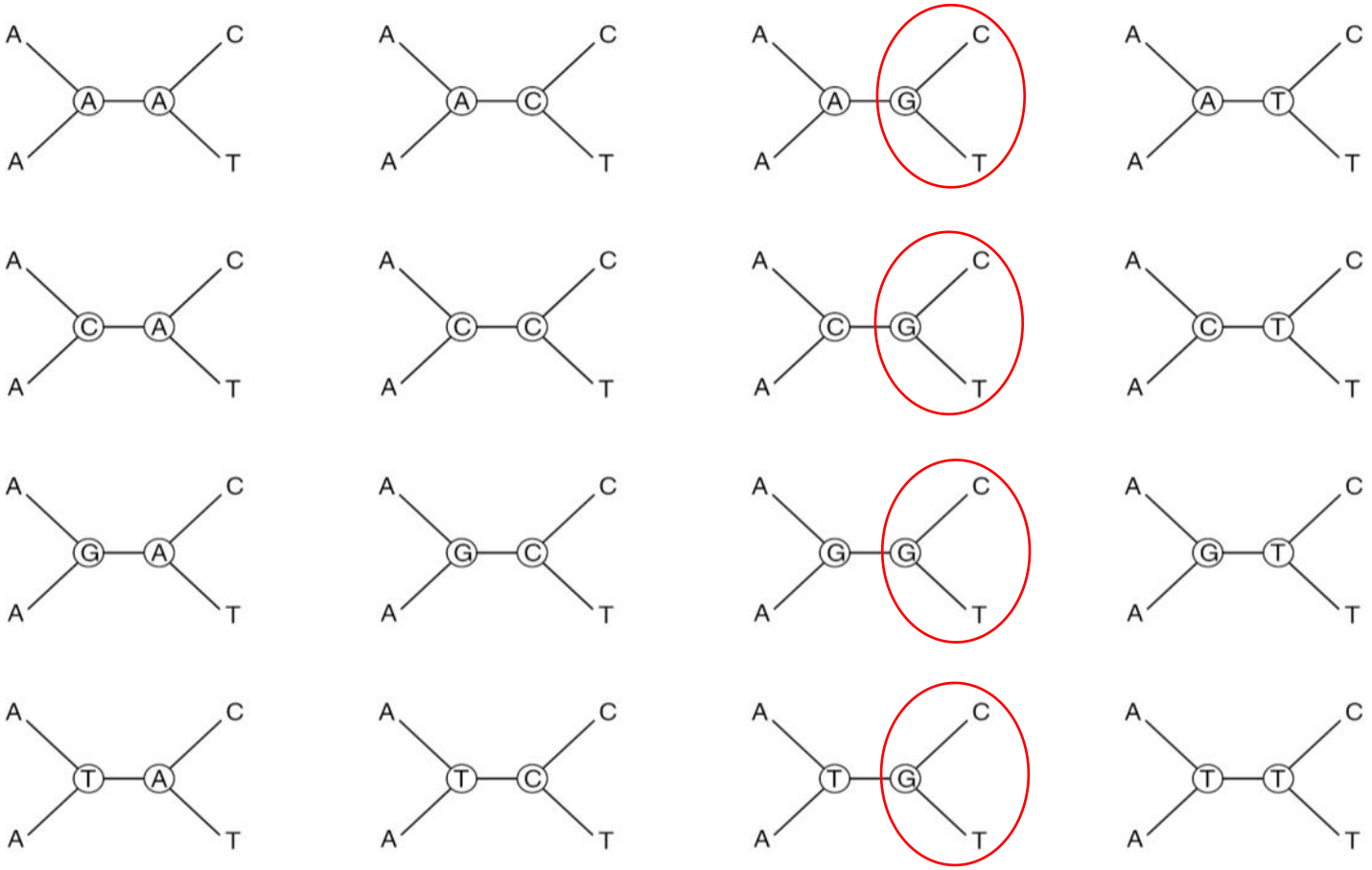
Note use of the OR probability rule

Pruning algorithm



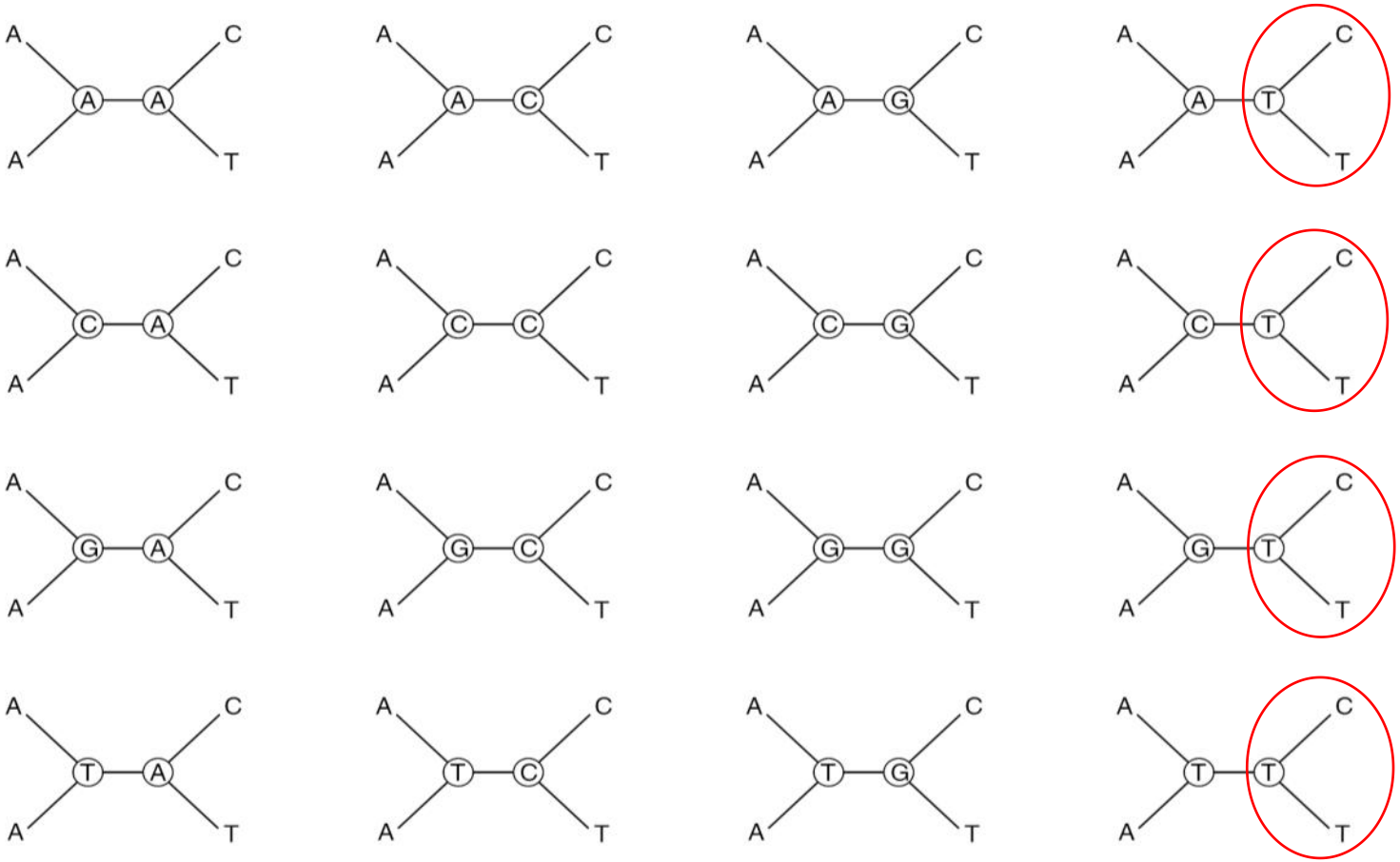
Note use of the OR probability rule

Pruning algorithm



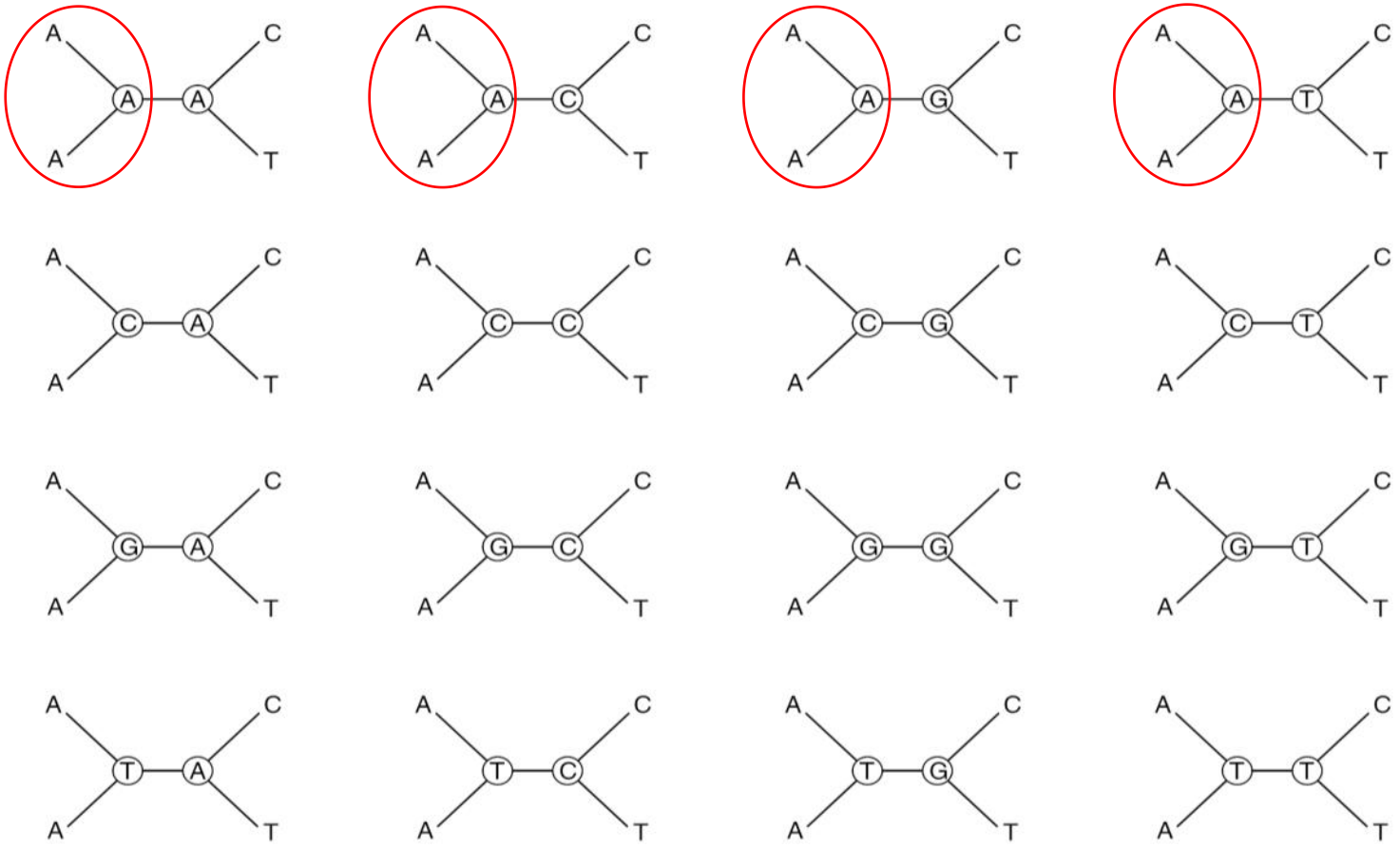
Note use of the OR probability rule

Pruning algorithm



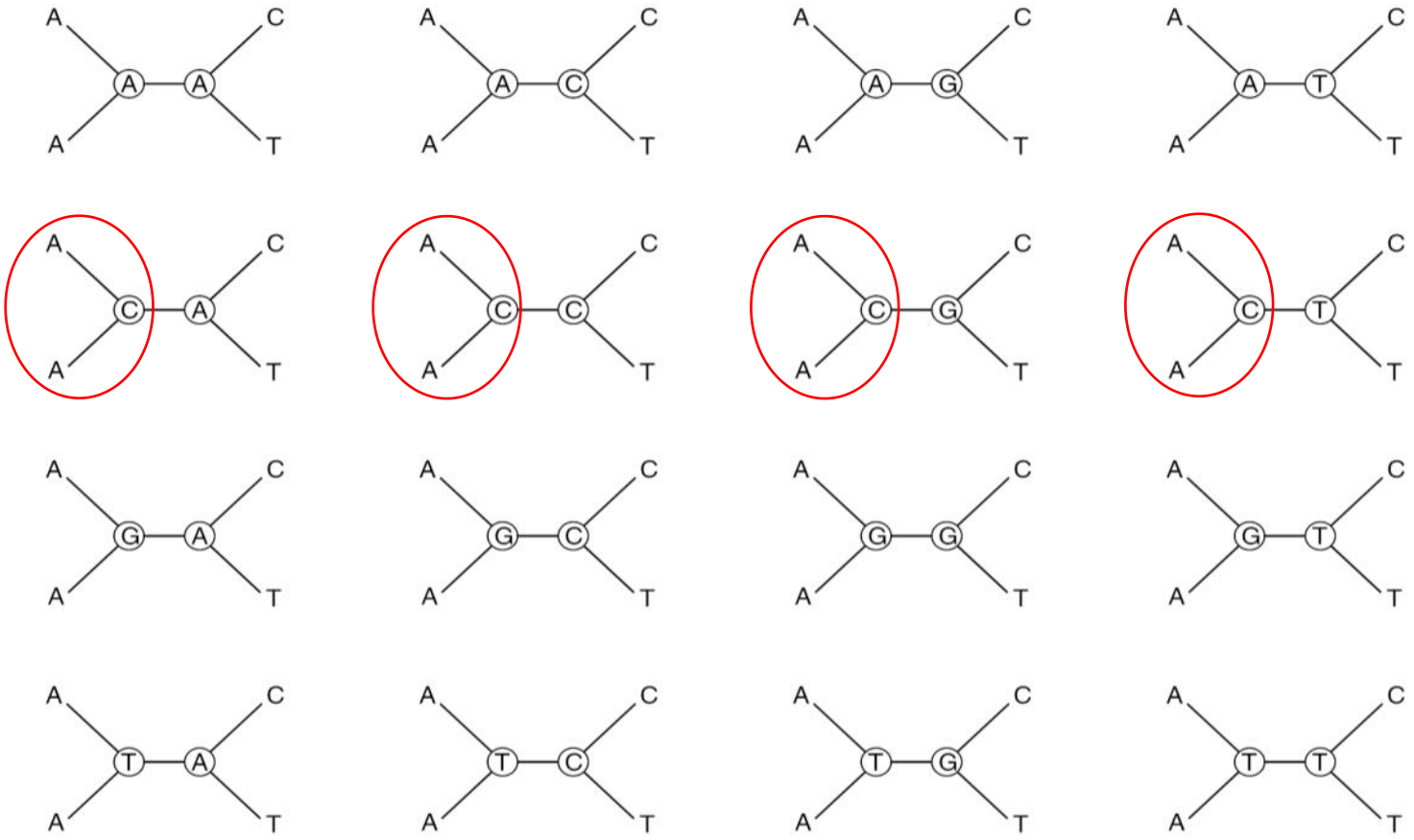
Note use of the OR probability rule

Pruning algorithm



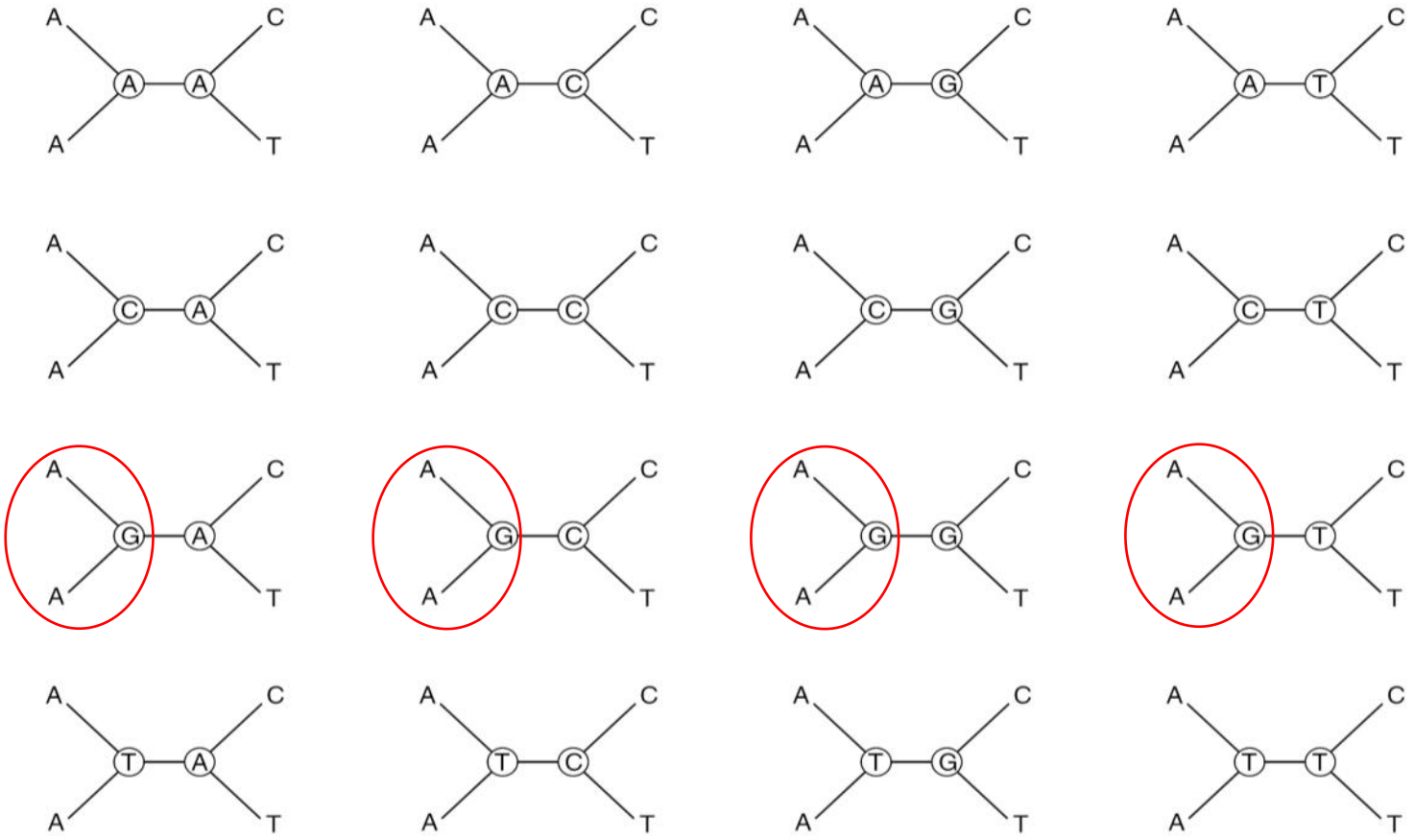
Note use of the OR probability rule

Pruning algorithm



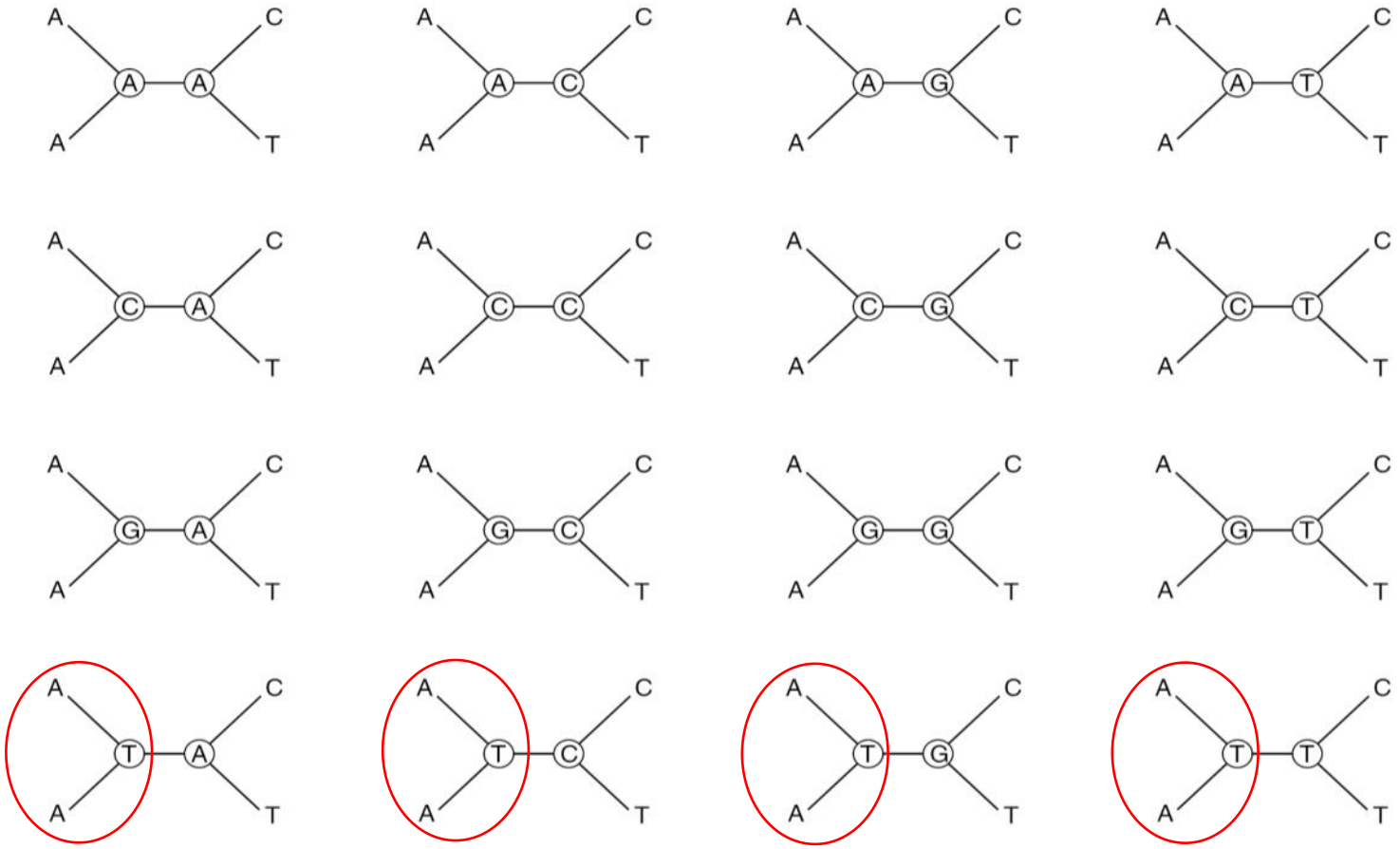
Note use of the OR probability rule

Pruning algorithm

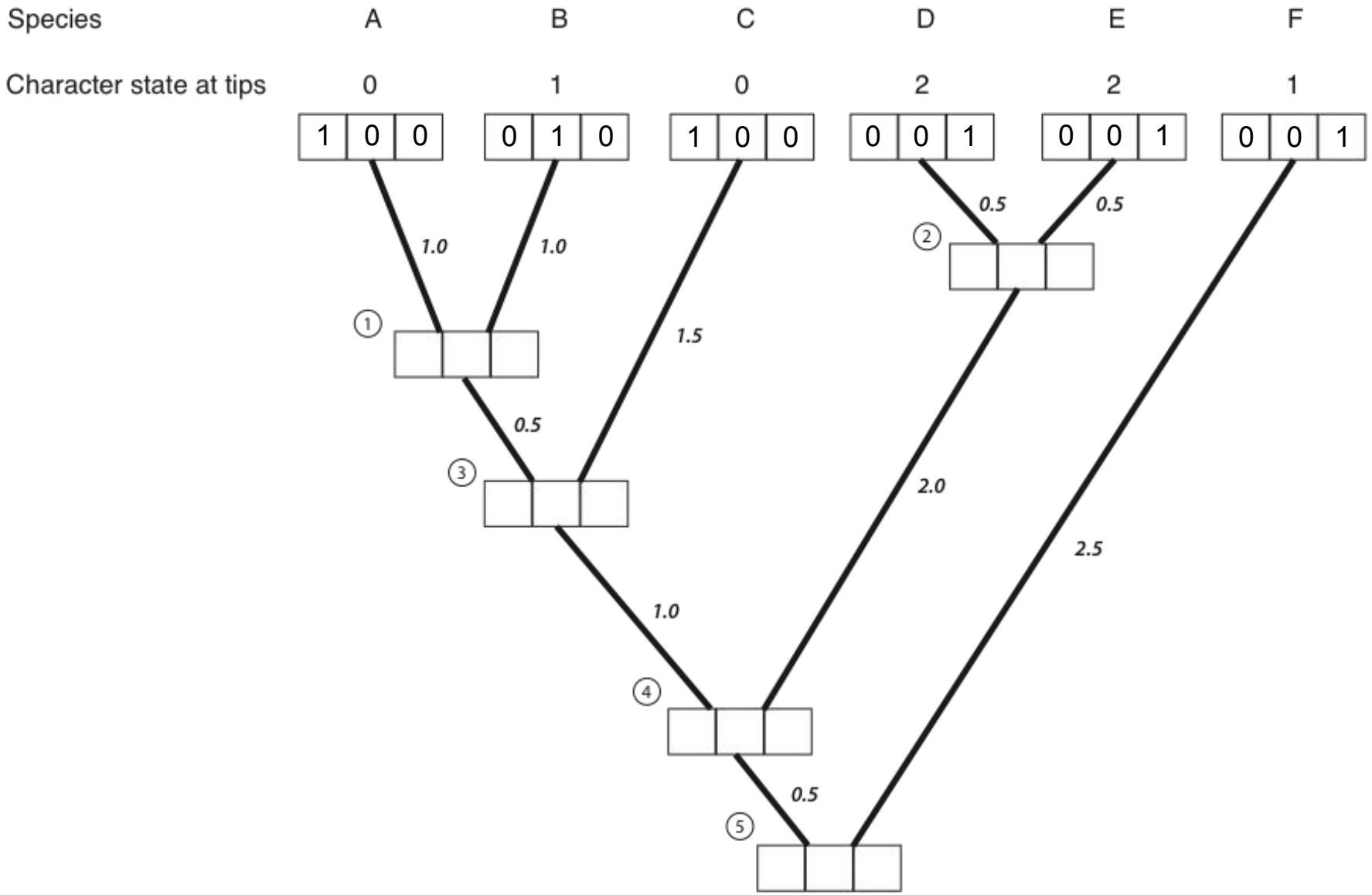


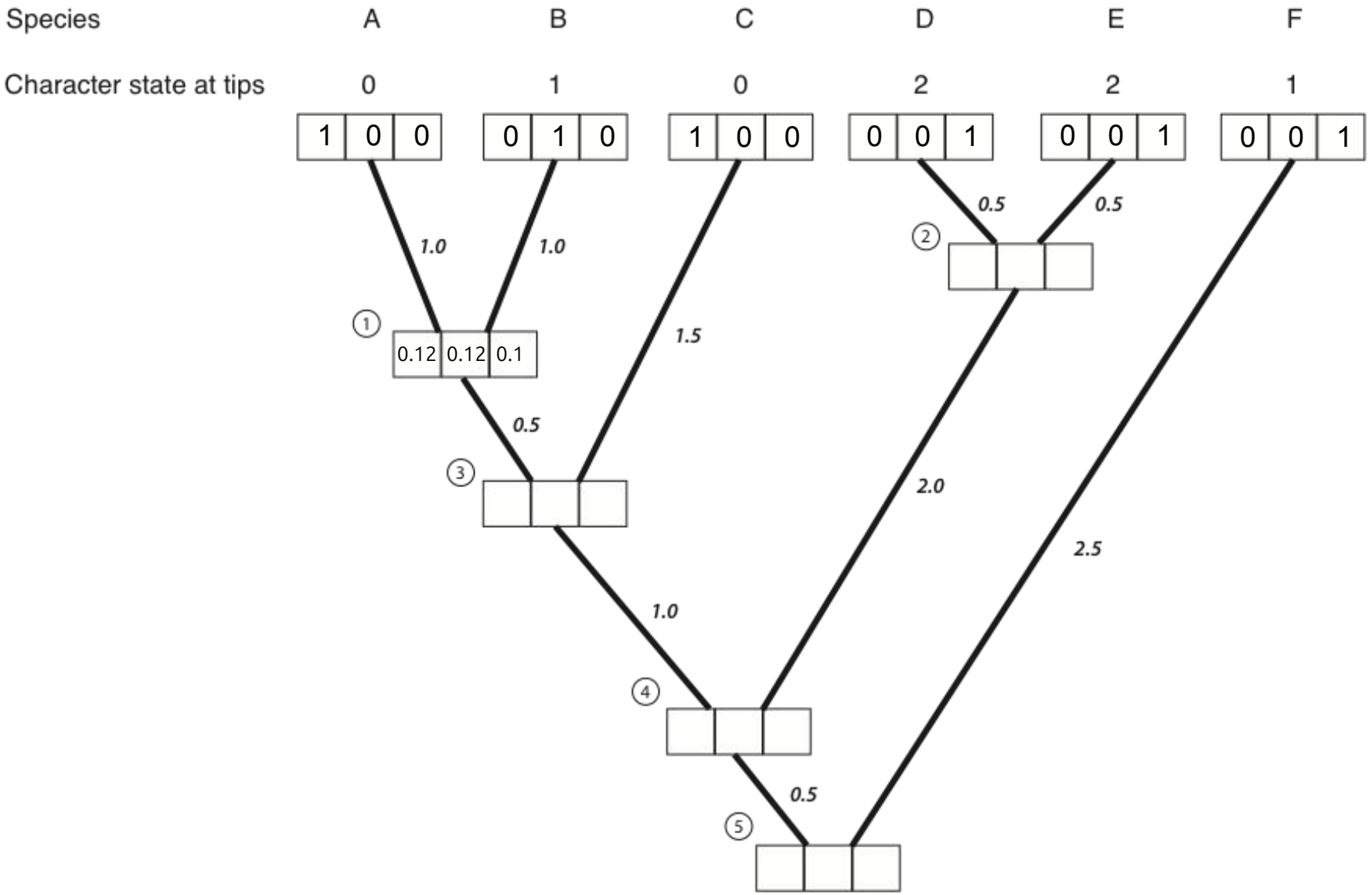
Note use of the OR probability rule

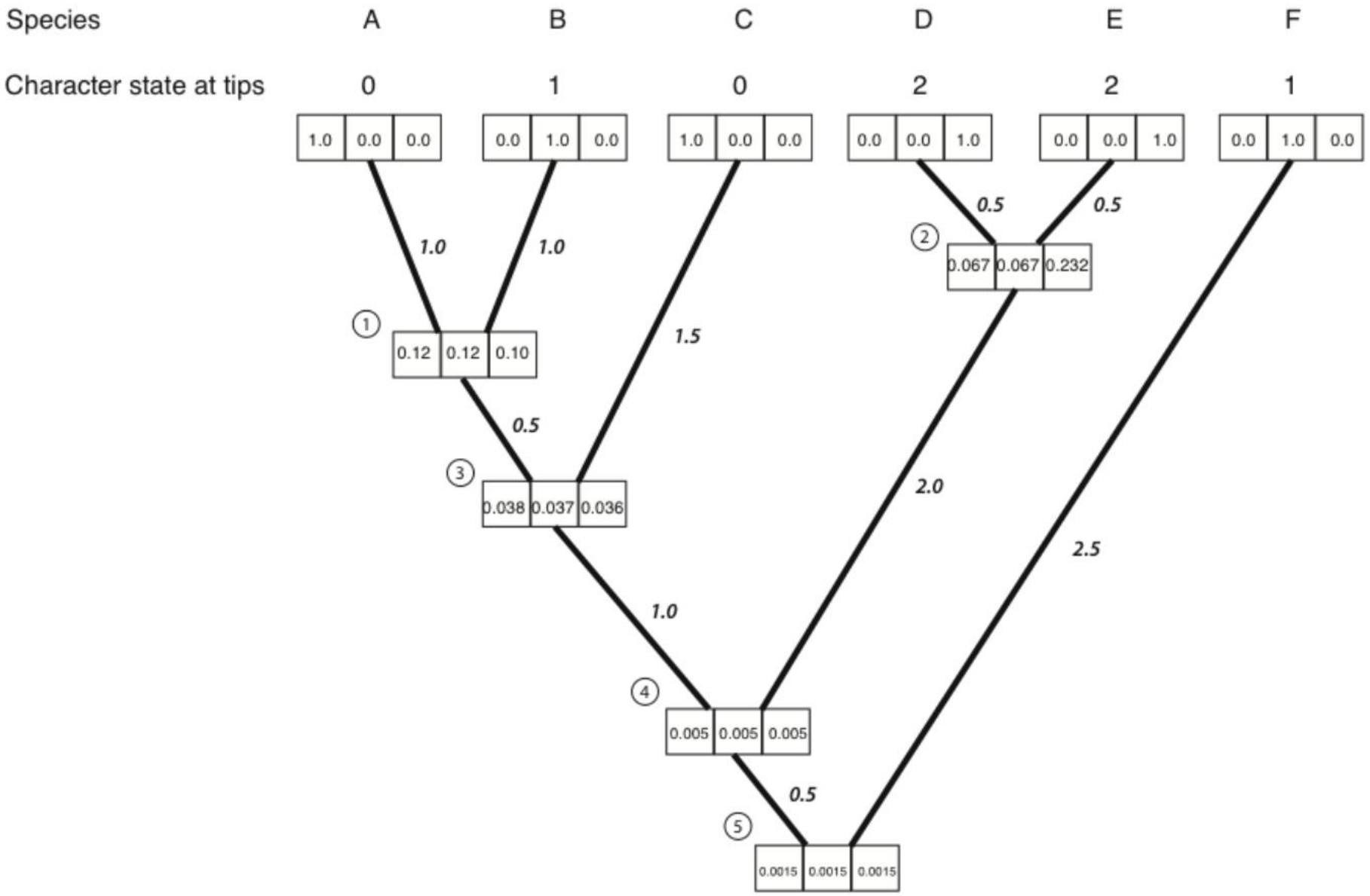
Pruning algorithm



Note use of the OR probability rule


Species
A
B
C
D
E
F
Character state at tips
0
1
0
2
2
1
1 0 0
0 1 0
1 0 0
0 0 1
0 0 1
0 0 1
1.0
1.0
0.5
0.5
1.5
0.5
2.0
2.5
1.0
0.5
1
2
3
4
5


Species
A
B
C
D
E
F
Character state at tips
0
1
0
2
2
1
1 0 0
0 1 0
1 0 0
0 0 1
0 0 1
0 0 1
1.0
1.0
0.5
0.5
1
2
0.12 0.12 0.1
1.5
0.5
3
2.0
2.5
1.0
4
0.5
5


Species
A
B
C
D
E
F
Character state at tips
0
1
0
2
2
1
1.0 0.0 0.0
0.0 1.0 0.0
1.0 0.0 0.0
0.0 0.0 1.0
0.0 0.0 1.0
0.0 1.0 0.0
1.0
1.0
0.5
0.5
1
0.12 0.12 0.10
2
0.067 0.067 0.232
1.5
0.5
3
0.038 0.037 0.036
2.0
2.5
1.0
4
0.005 0.005 0.005
0.5
5
0.0015 0.0015 0.0015