Practical considerations

How do you find the most parsimonious tree?

Heuristic searches

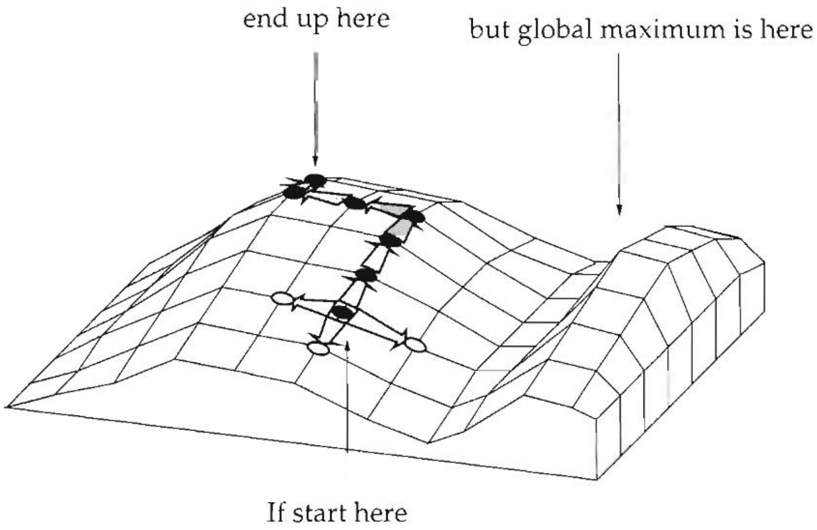


Figure 4.1: A surface rising above a two-dimensional plain (or plane). The process of climbing uphill on the surface is illustrated, as well as the failure to find a higher peak by this "greedy" method.

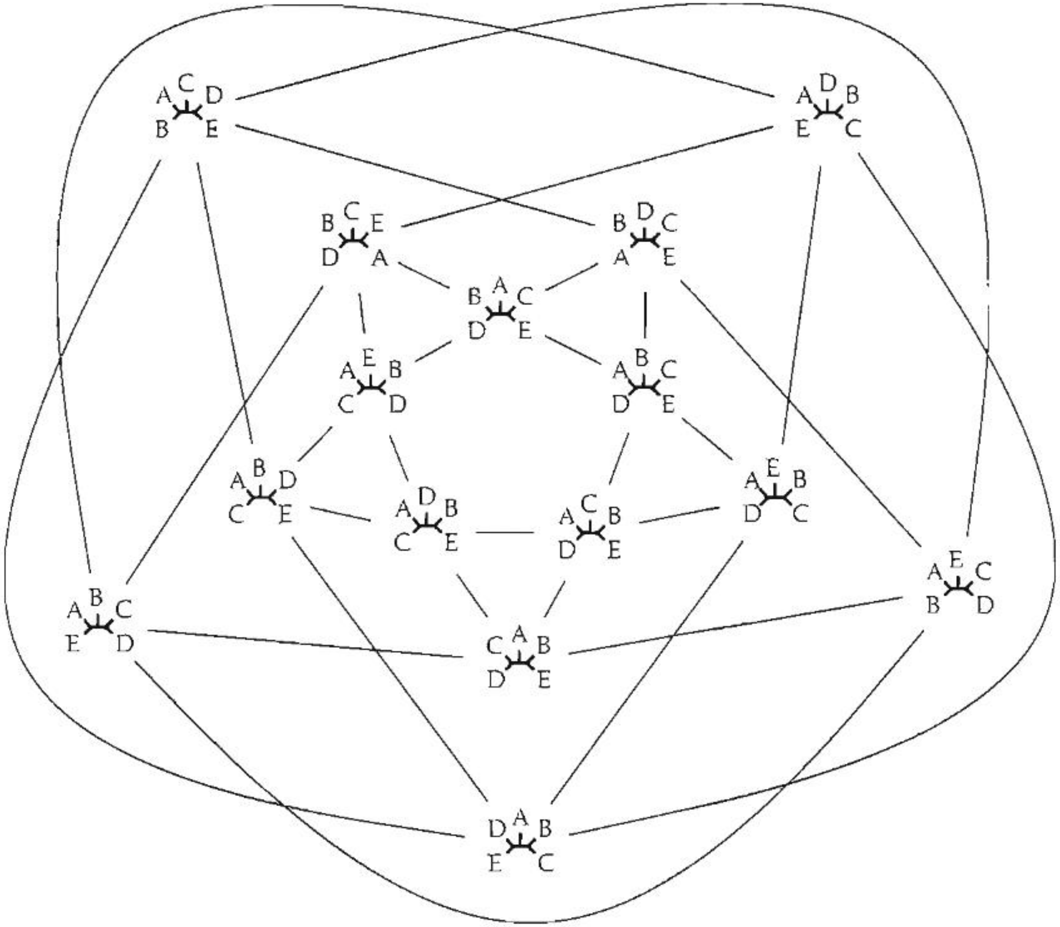


Figure 4.3: The space of all 15 possible unrooted trees with 5 tips. Neighbors are connected by lines when a nearest-neighbor interchange can convert one into the other. The labels A–E correspond to the species names Alpha through Epsilon in that data set. This symmetric arrangement of nodes was discovered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

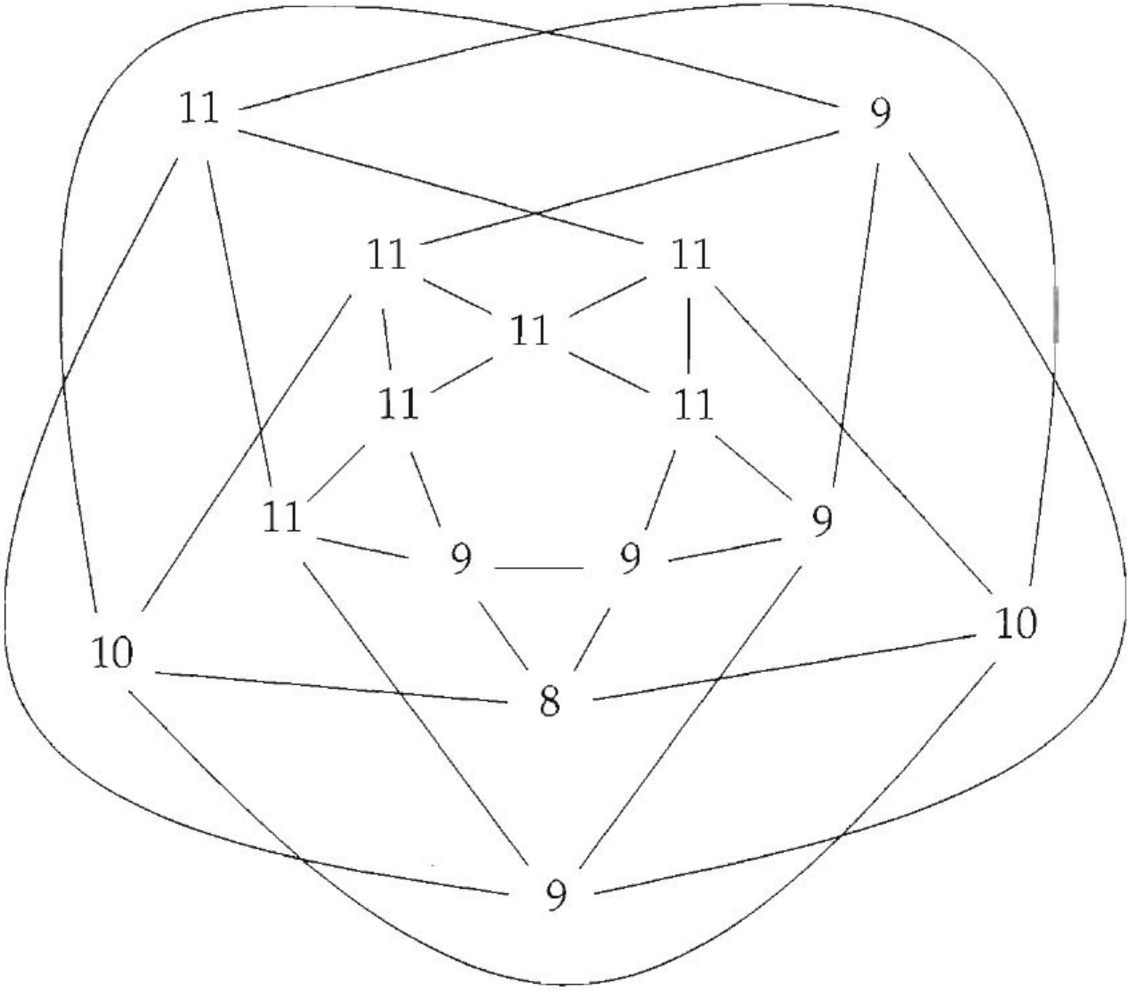
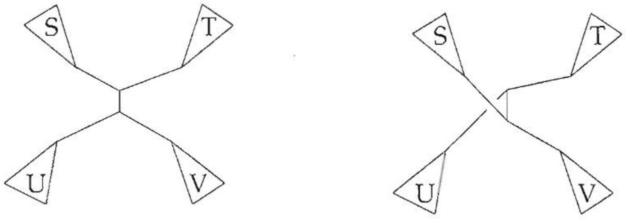
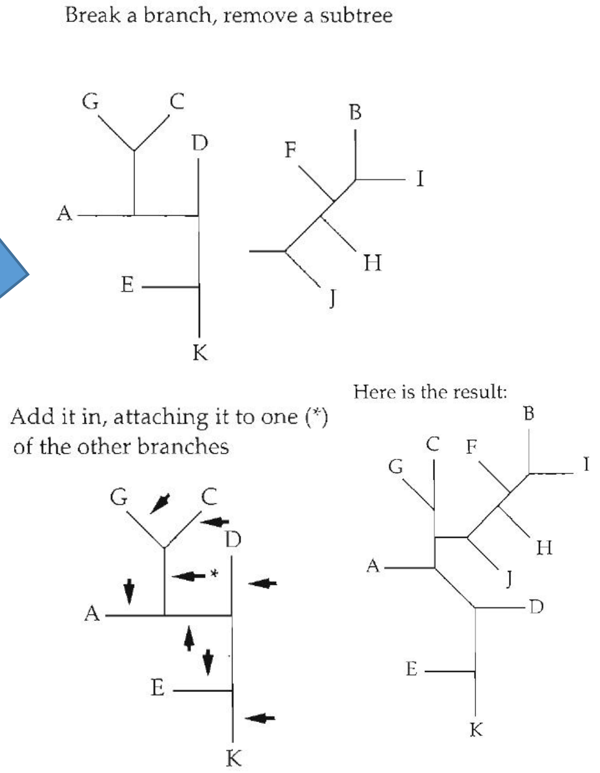


Figure 4.4: The space of all 15 possible trees, as in Figure 4.3, where the number of changes of state on the data set of Table 1.1 is shown. Nearest-neighbor interchanges search for the most parsimonious tree by moving in this graph.

Tree space "moves"



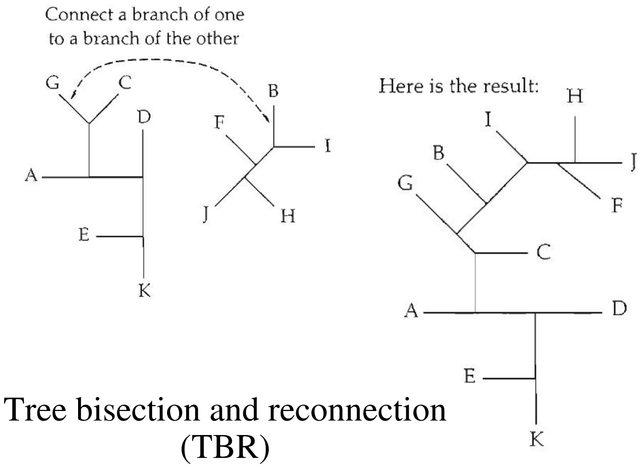
Nearest Neighbor Interchange
(NNI)


Break a branch, remove a subtree

Add it in, attaching it to one (*) of the other branches

Here is the result:

Subtree pruning and regrafting
(SPR)


Connect a branch of one to a branch of the other

Here is the result:

Tree bisection and reconnection
(TBR)

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

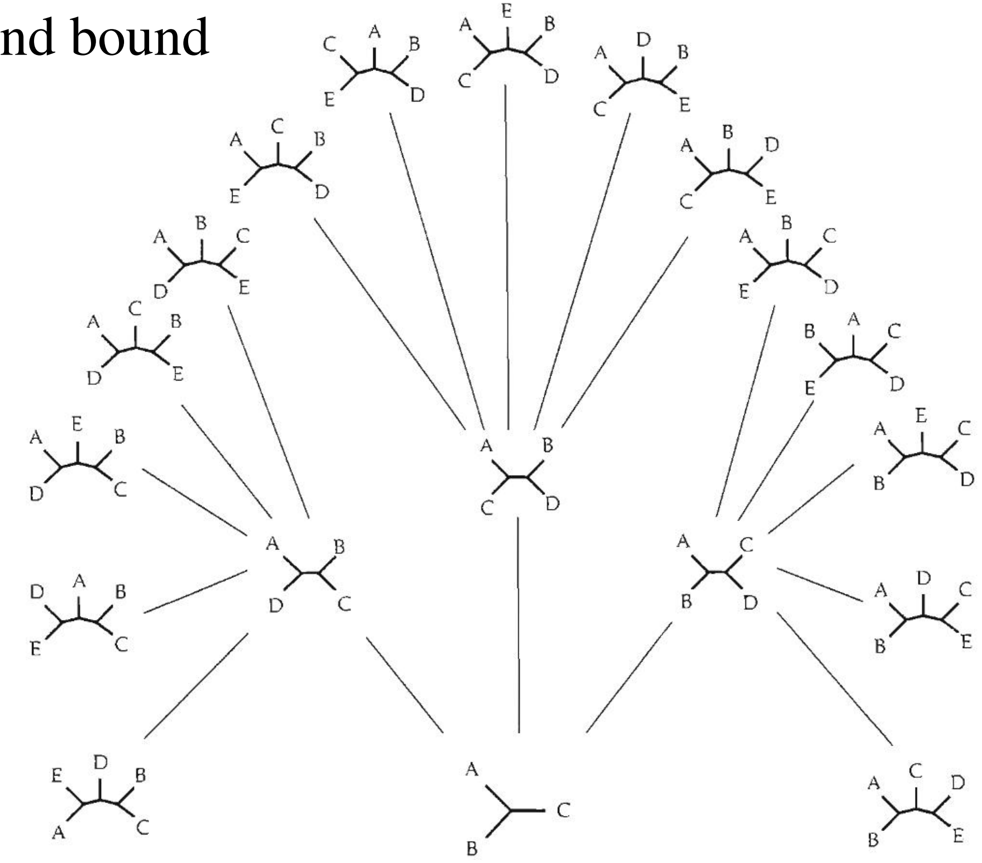


Figure 5.3: Search tree for most parsimonious tree in a five-species case.

Branch and bound

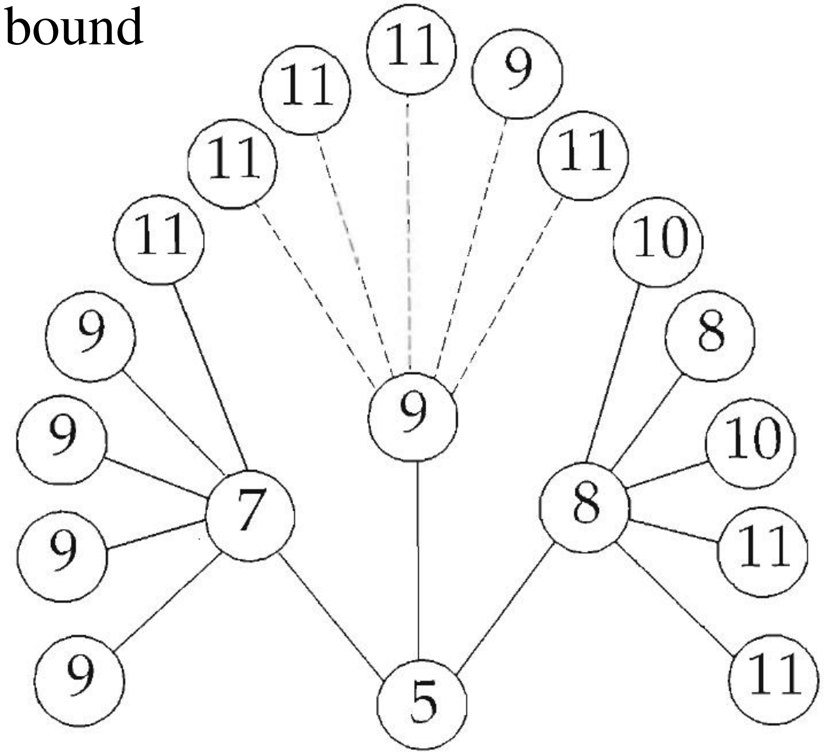
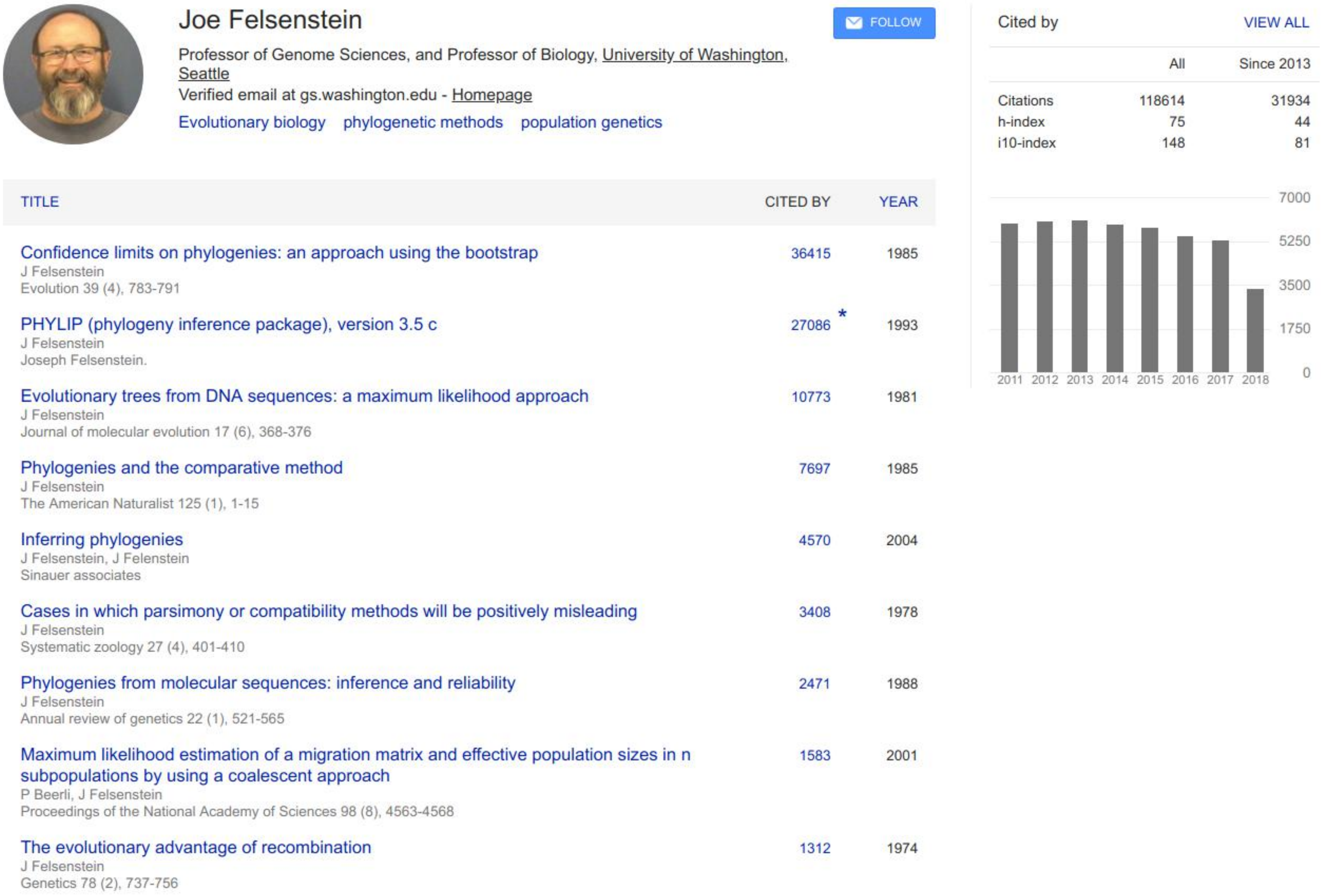


Figure 5.4: Search tree for most parsimonious tree for five species, using the data of Table 1.1. Trees are shown in Figure 5.3. Dashed lines are those not traversed by a branch and bound method. The species names in the data set correspond to labels A through E in Figure 5.3.

Software: PAUP*, TNT, Mesquite, others...

Felsenstein & the birth of statistical phylogenetics



Joe Felsenstein

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

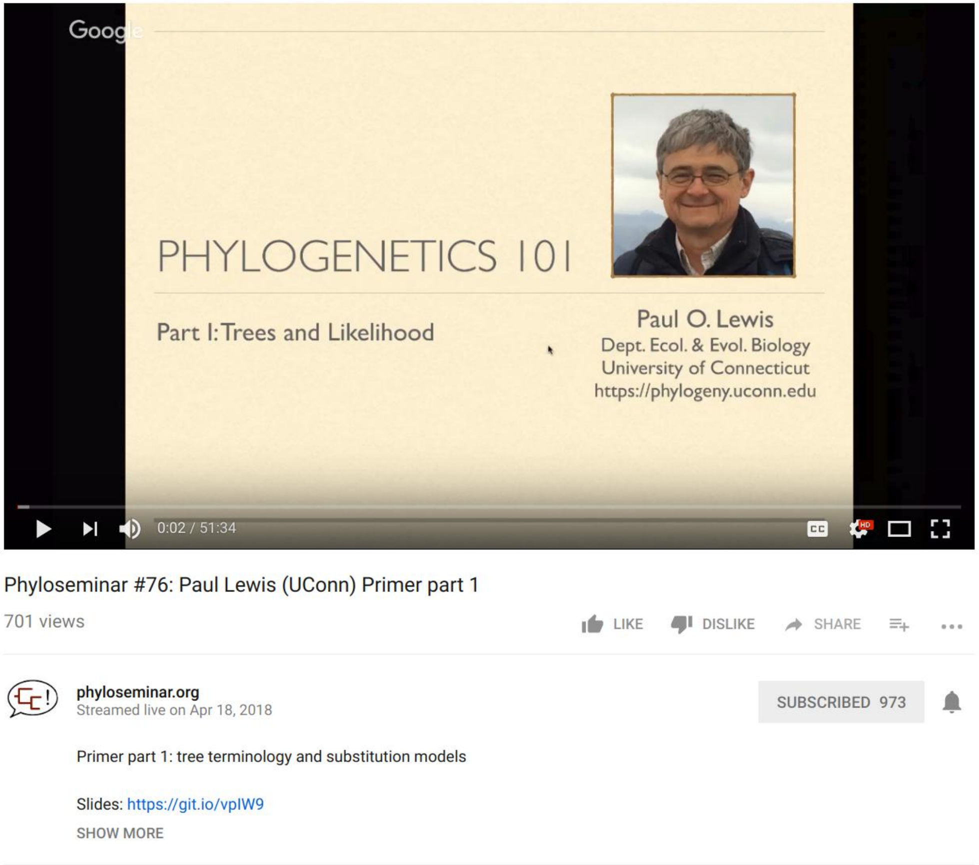
Evolutionary biology phylogenetic methods population genetics

Cited by VIEW ALL

	All	Since 2013
Citations	118614	31934
h-index	75	44
i10-index	148	81

TITLE	CITED BY	YEAR
Confidence limits on phylogenies: an approach using the bootstrap J Felsenstein Evolution 39 (4), 783-791	36415	1985
PHYLIP (phylogeny inference package), version 3.5 c J Felsenstein Joseph Felsenstein.	27086 *	1993
Evolutionary trees from DNA sequences: a maximum likelihood approach J Felsenstein Journal of molecular evolution 17 (6), 368-376	10773	1981
Phylogenies and the comparative method J Felsenstein The American Naturalist 125 (1), 1-15	7697	1985
Inferring phylogenies J Felsenstein, J Felenstein Sinauer associates	4570	2004
Cases in which parsimony or compatibility methods will be positively misleading J Felsenstein Systematic zoology 27 (4), 401-410	3408	1978
Phylogenies from molecular sequences: inference and reliability J Felsenstein Annual review of genetics 22 (1), 521-565	2471	1988
Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach P Beerli, J Felsenstein Proceedings of the National Academy of Sciences 98 (8), 4563-4568	1583	2001
The evolutionary advantage of recombination J Felsenstein Genetics 78 (2), 737-756	1312	1974

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



PHYLOGENETICS 101

Part I: Trees and Likelihood

Paul O. Lewis
Dept. Ecol. & Evol. Biology
University of Connecticut
https://phylogeny.uconn.edu

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

701 views

phyloseminar.org
Streamed live on Apr 18, 2018

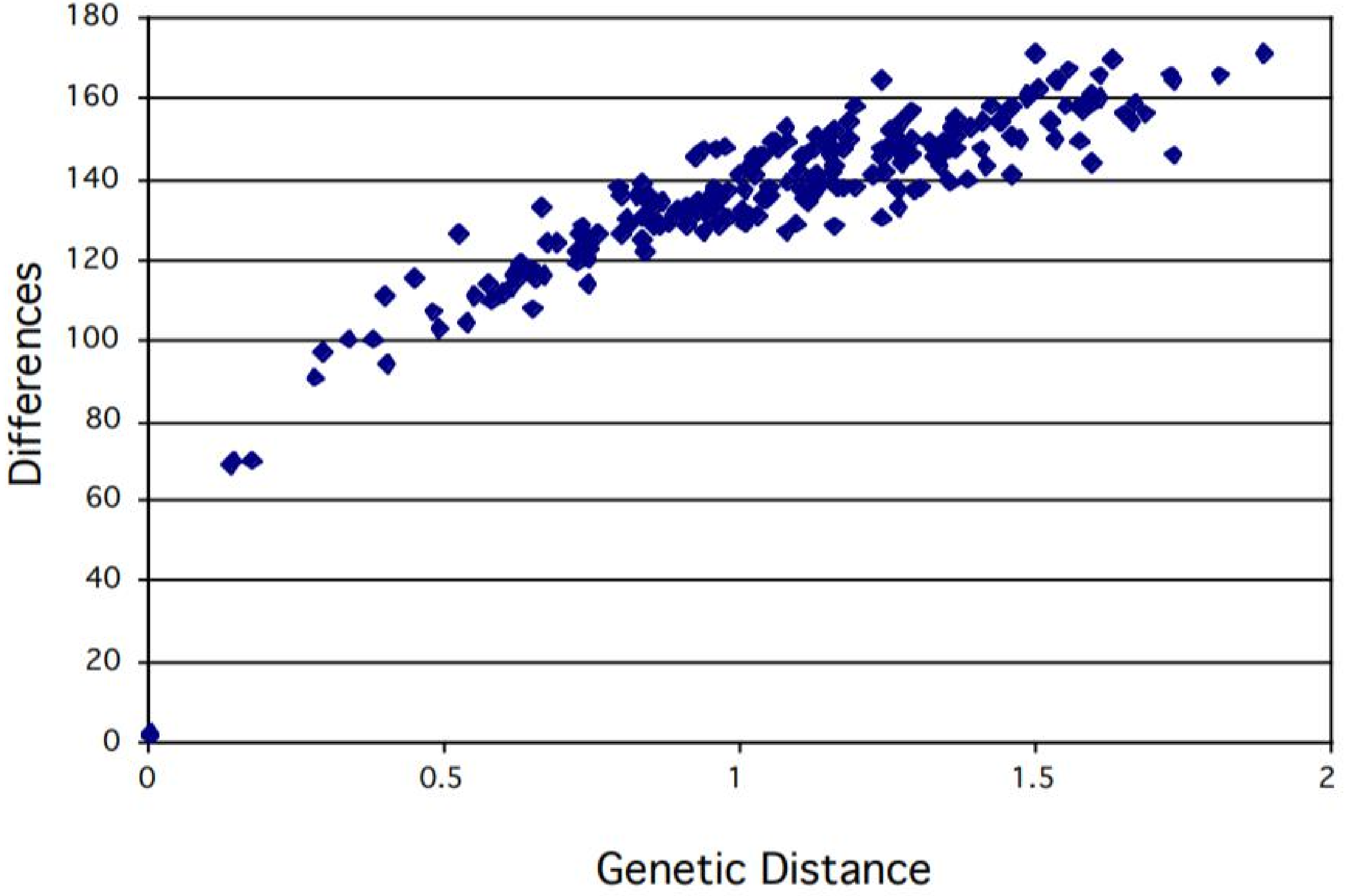
Primer part 1: tree terminology and substitution models

Slides: https://git.io/vpIW9

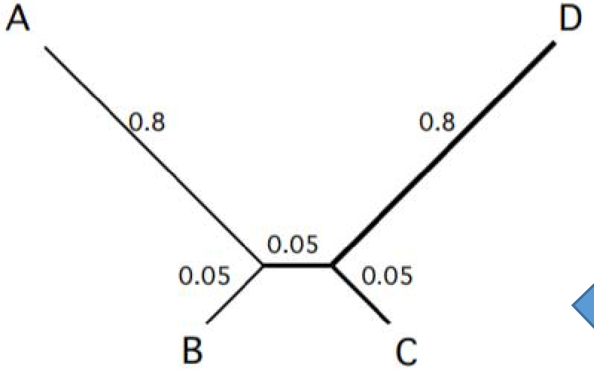
If two sequences are unrelated, what % of bases (aligned sites) do you expect to be identical?

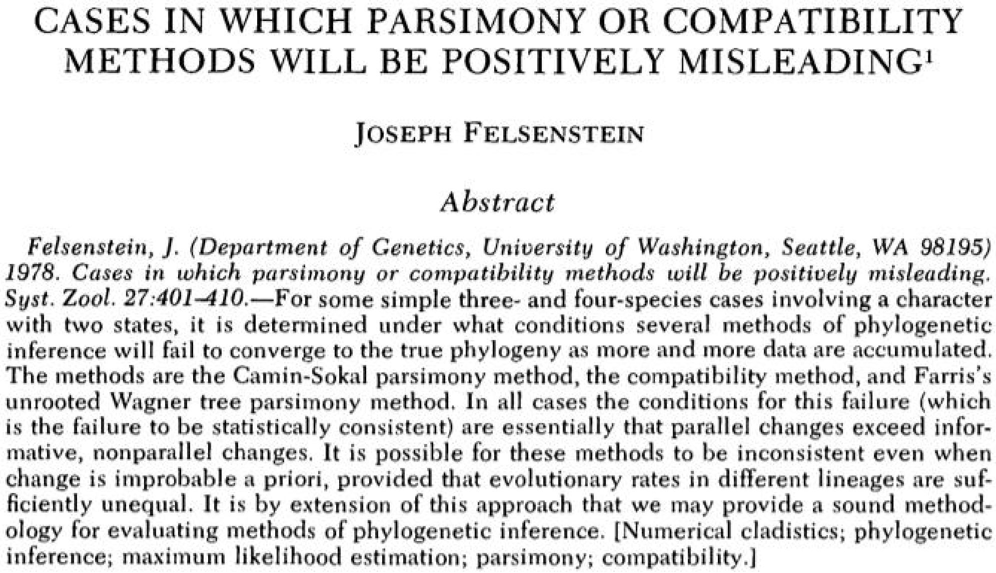
A. 50%
B. 25%
C. 0%
D. I need more information

Why do we need statistics?



"Long-branch attraction"



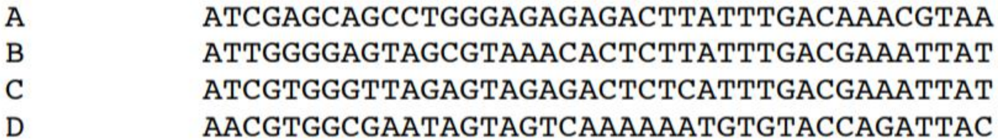


CASES IN WHICH PARSIMONY OR COMPATIBILITY METHODS WILL BE POSITIVELY MISLEADING[1]

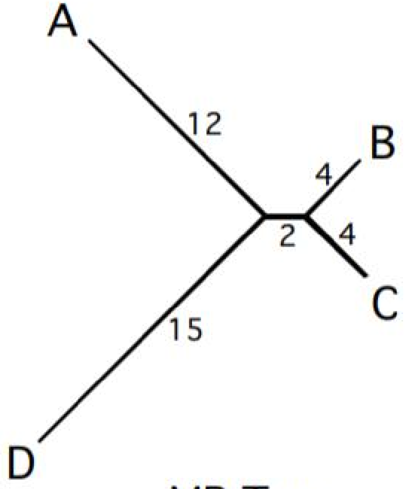
JOSEPH FELSENSTEIN

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
```

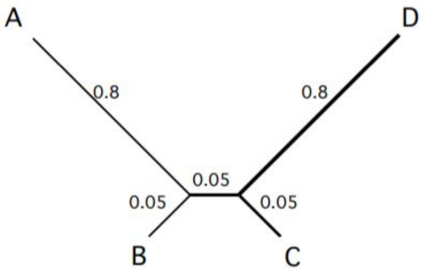


MP Tree

P(

ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC

|

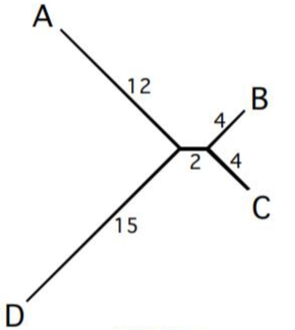


)

P(

ATCGAGCAGCCTGGGAGAGAGACTTATTTGACAAACGTAA
ATTGGGGAGTAGCGTAAACACTCTTATTTGACGAAATTAT
ATCGTGGGTTAGAGTAGAGACTCTCATTTGACGAAATTAT
AACGTGGCGAATAGTAGTCAAAAAATGTGTACCAGATTAC

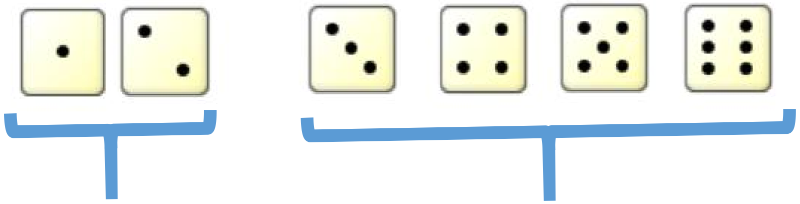
|



MP Tree

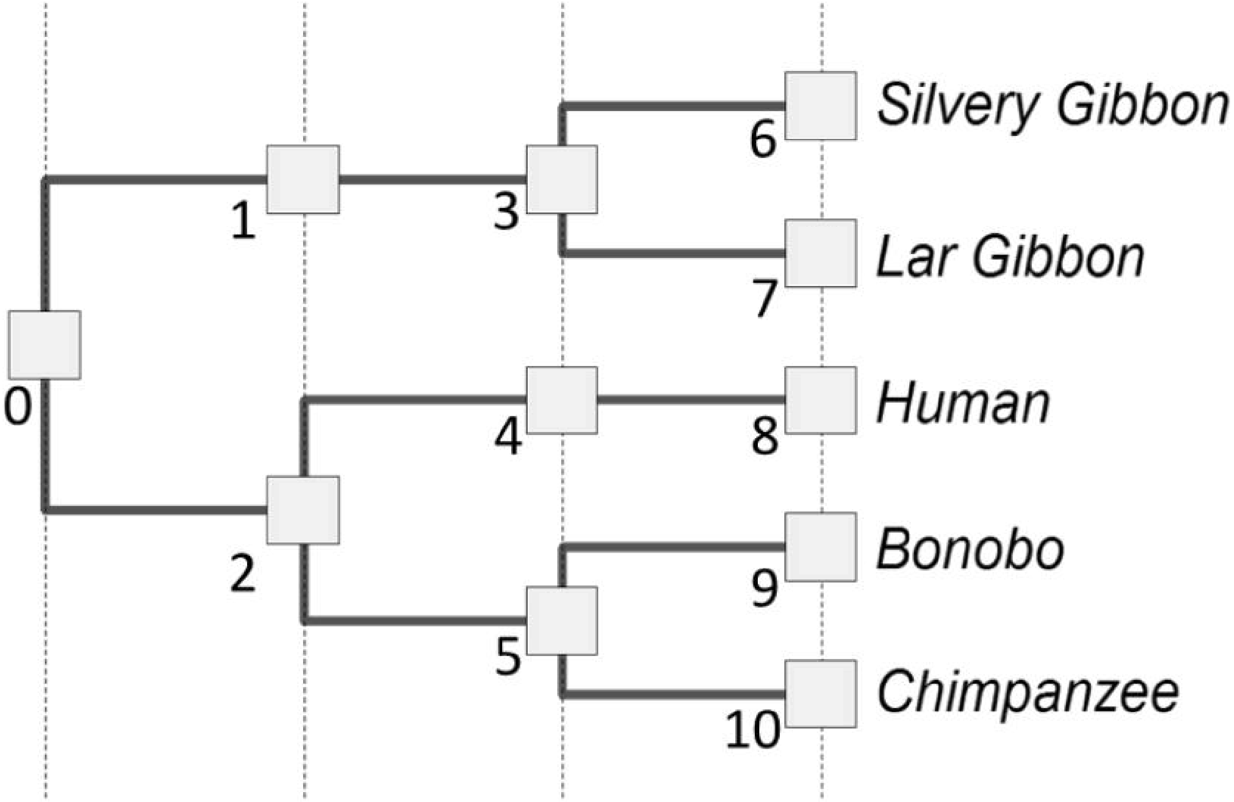
)

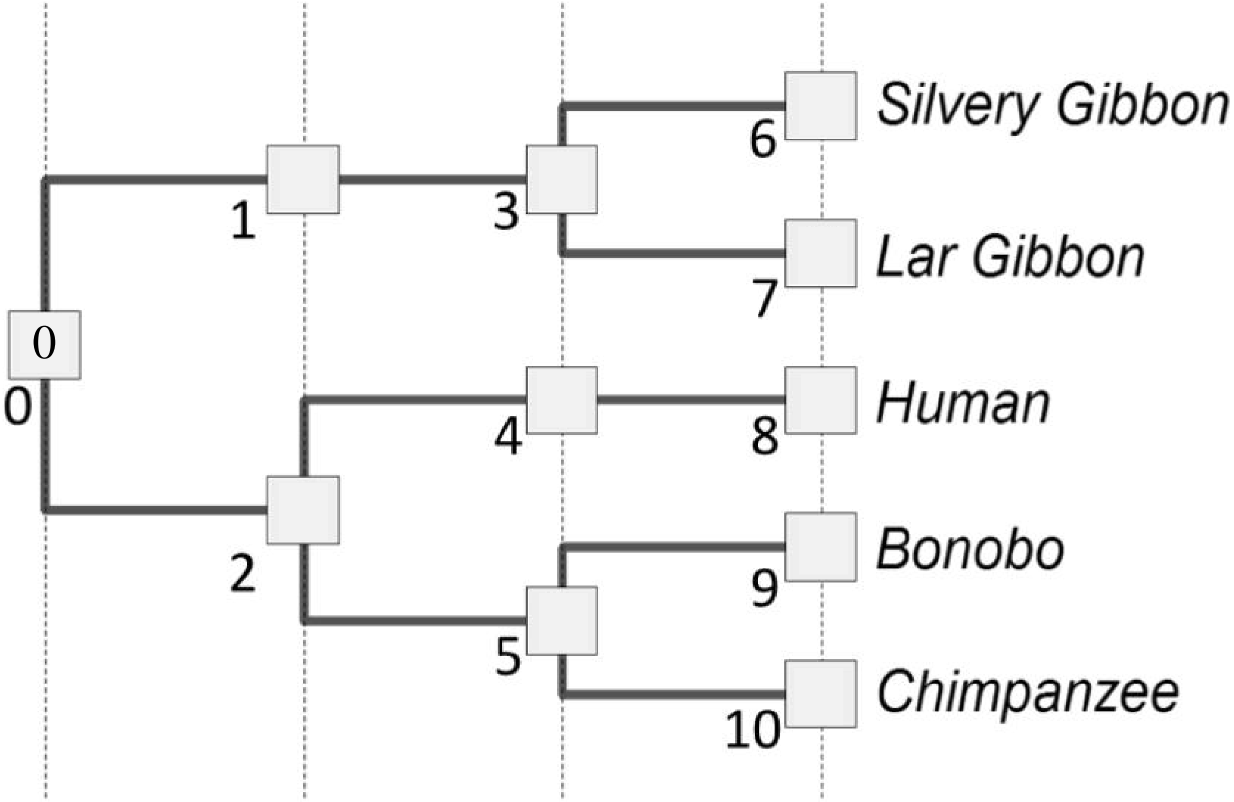
How do we treat trees probabilistically?

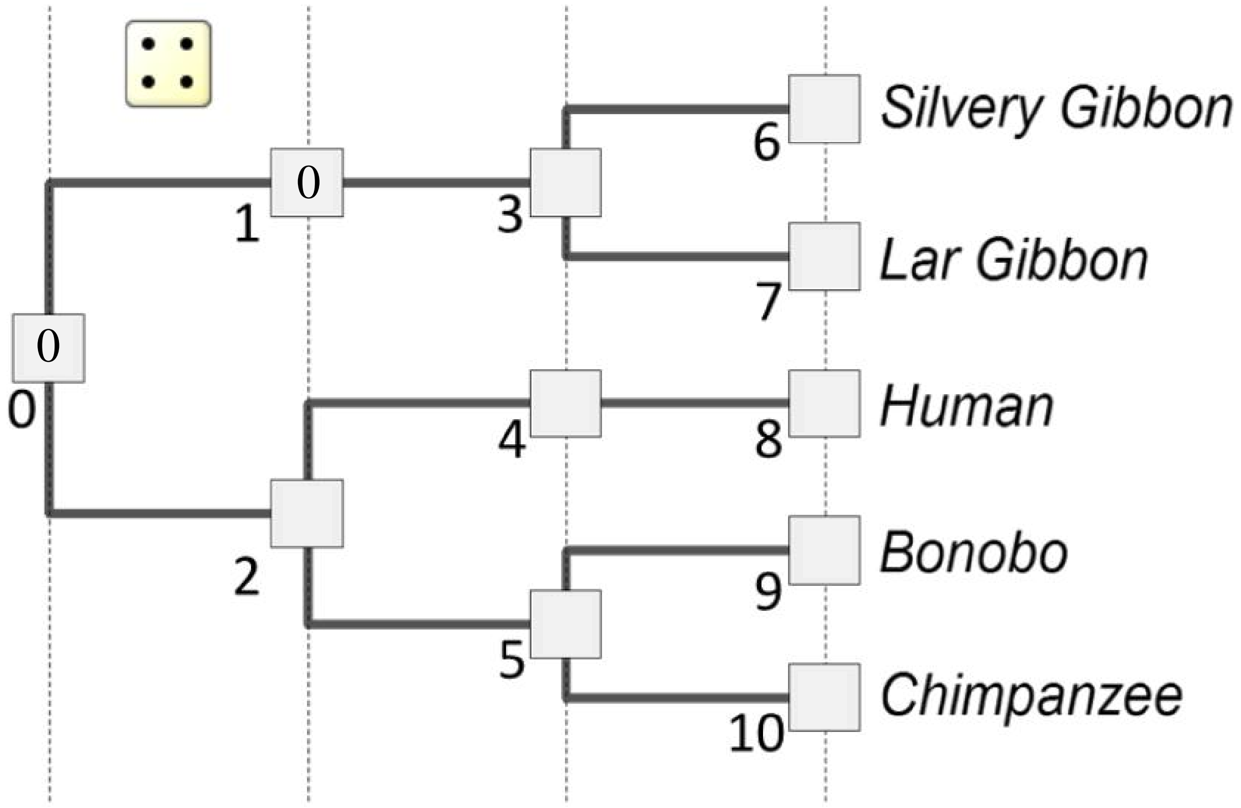


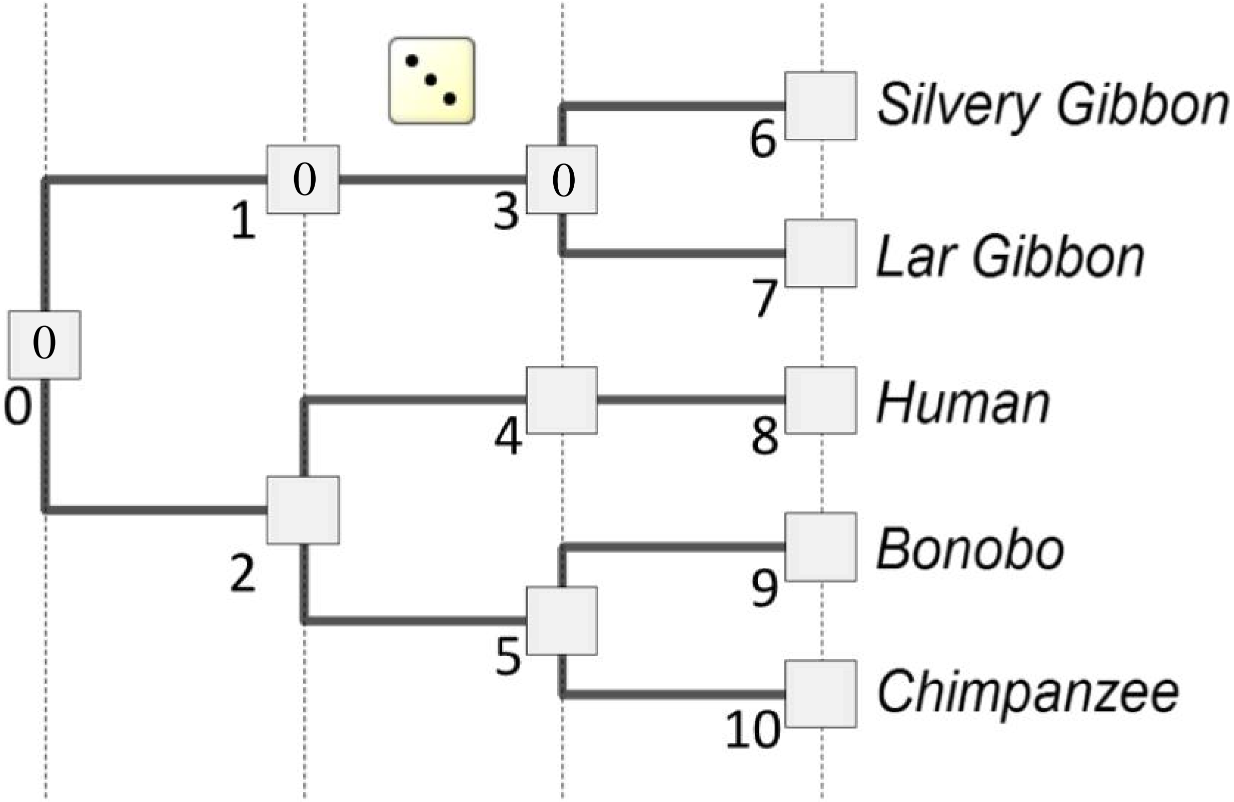
If these come up, change state

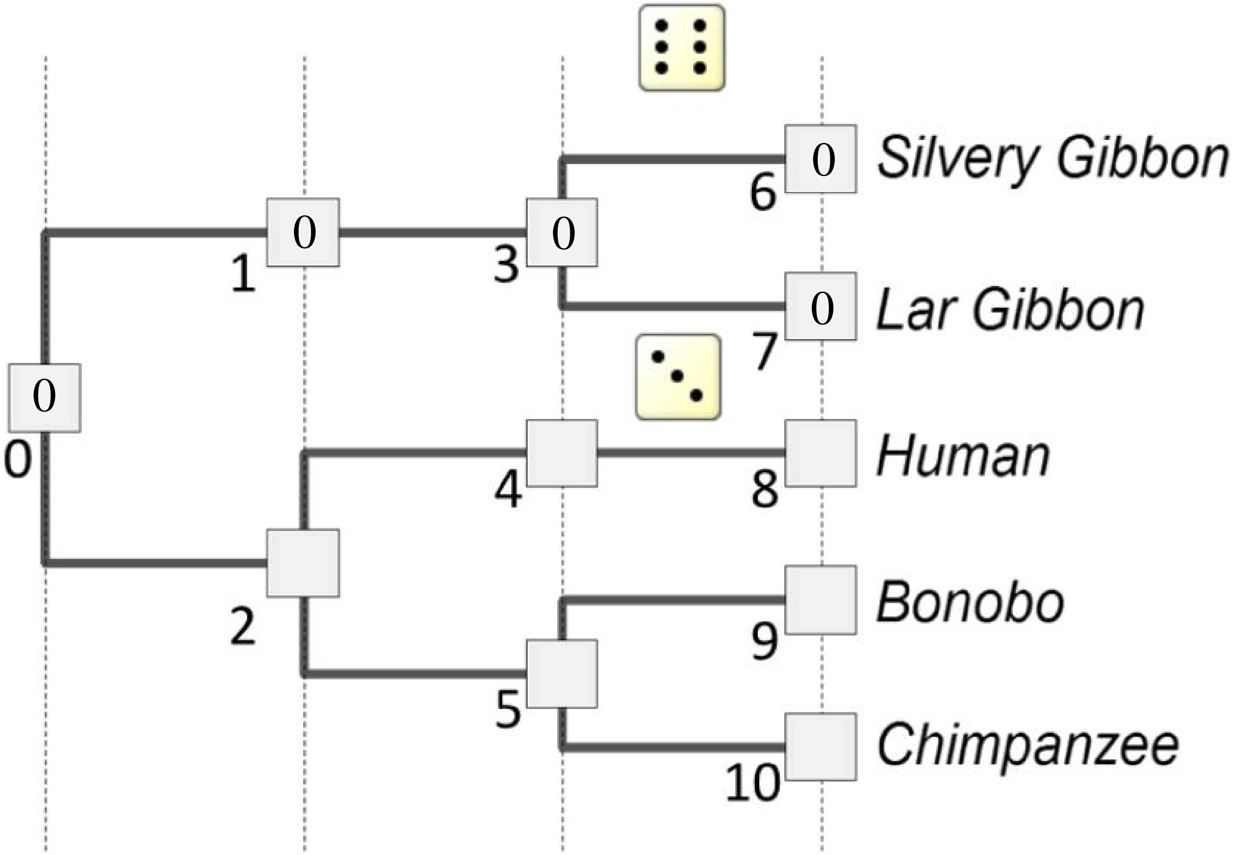
If these come up, stay the same

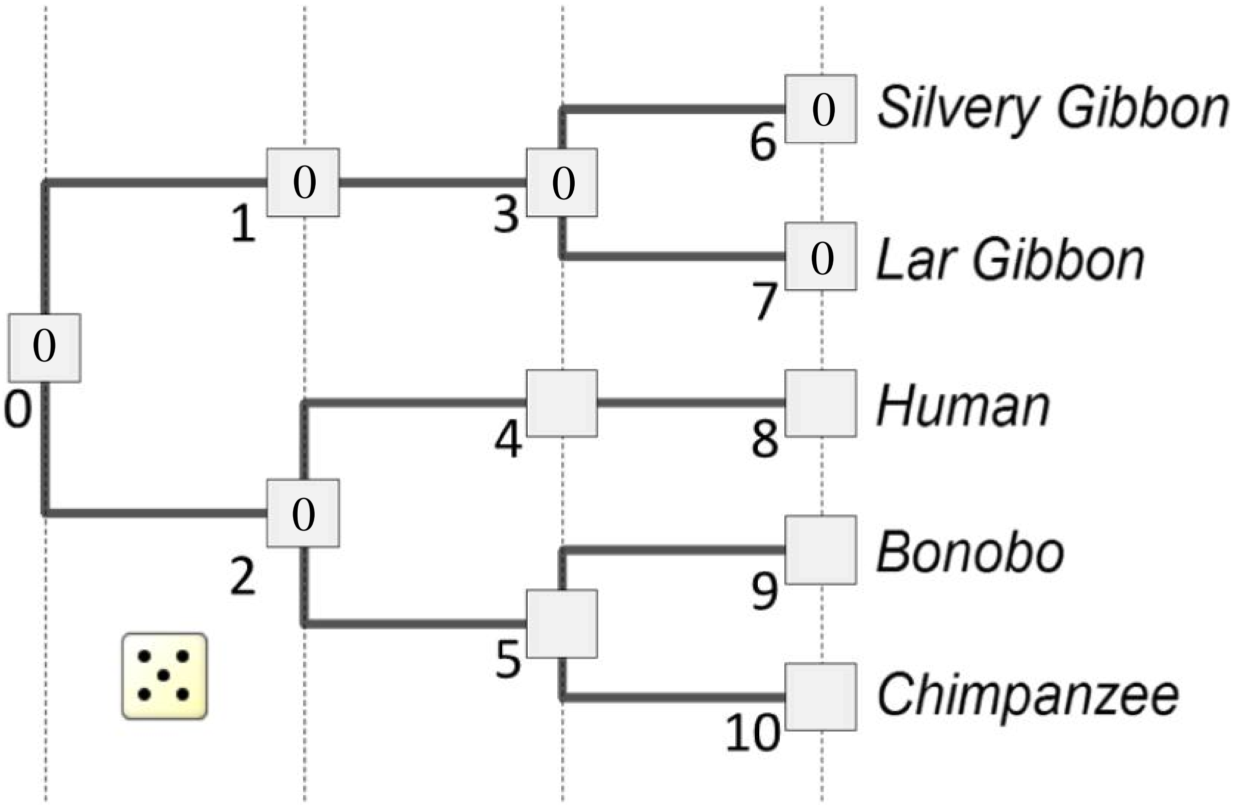

Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee
0
1
2
3
4
5
6
7
8
9
10

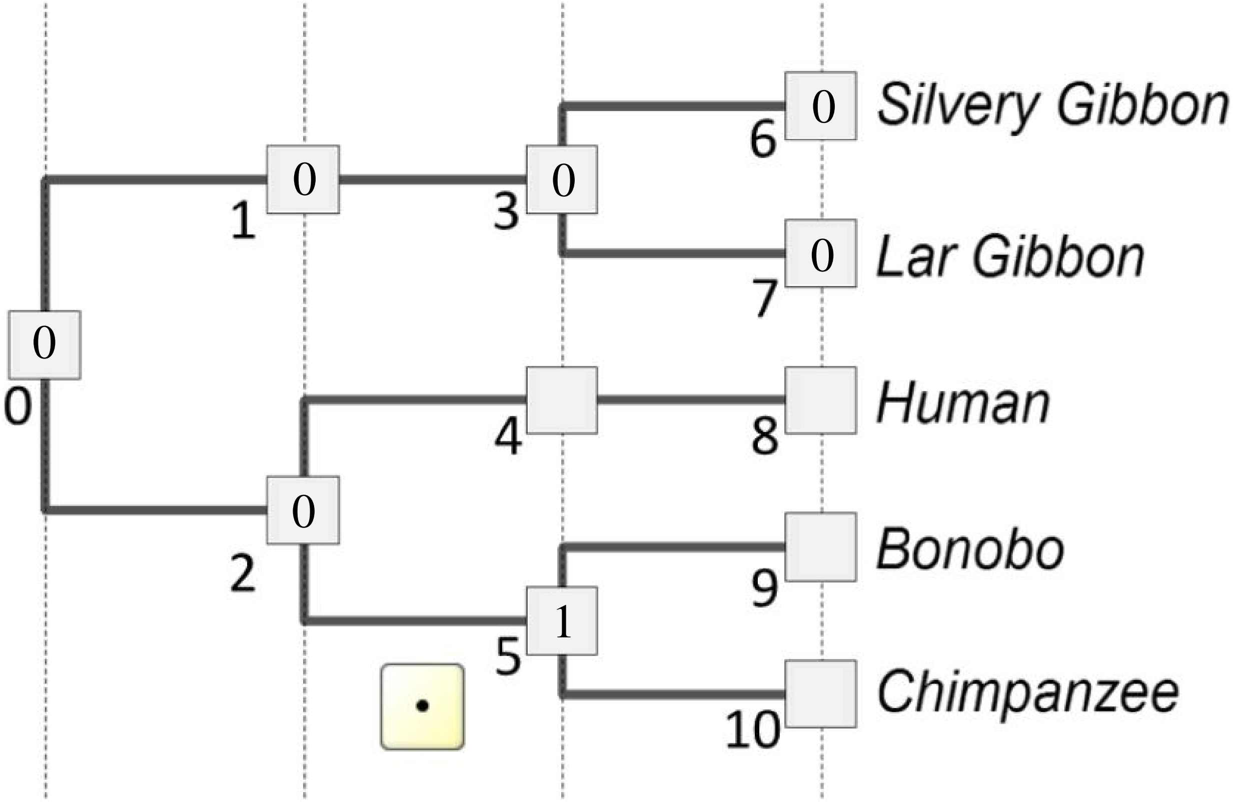

Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee
0
0
1
2
3
4
5
6
7
8
9
10

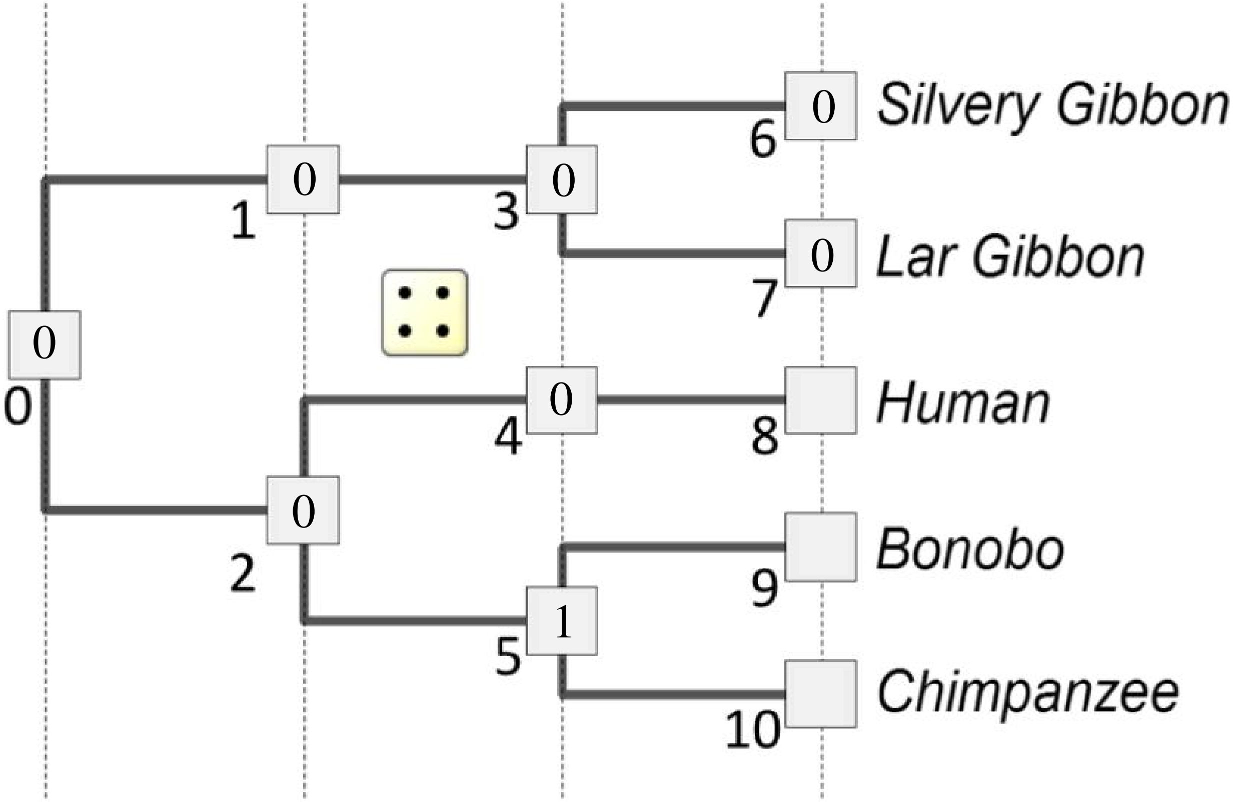

0
1
0
3
6
Silvery Gibbon
7
Lar Gibbon
0
0
4
8
Human
2
9
Bonobo
5
10
Chimpanzee

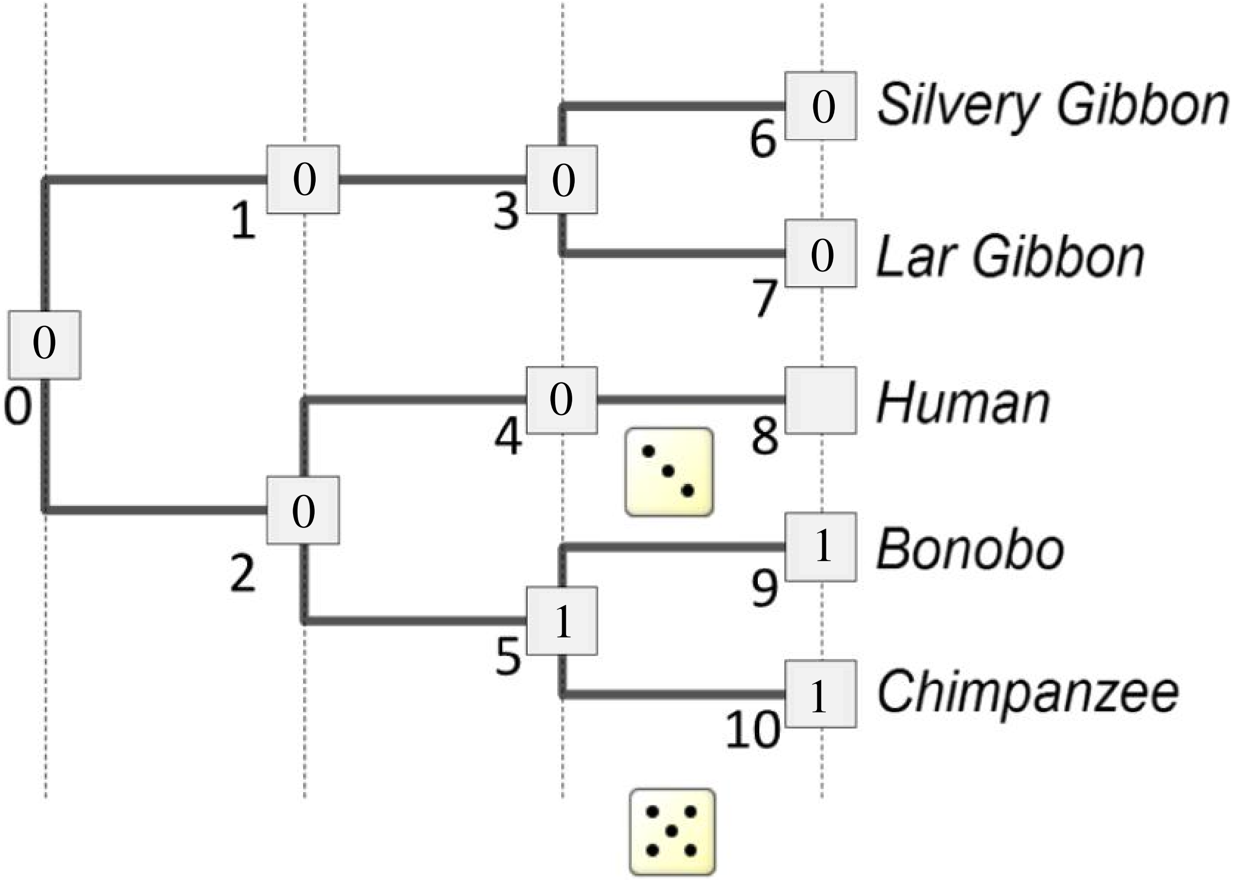

0
0
0
0
1
3
6
7
8
9
10
2
4
5
Silvery Gibbon
Lar Gibbon
Human
Bonobo
Chimpanzee

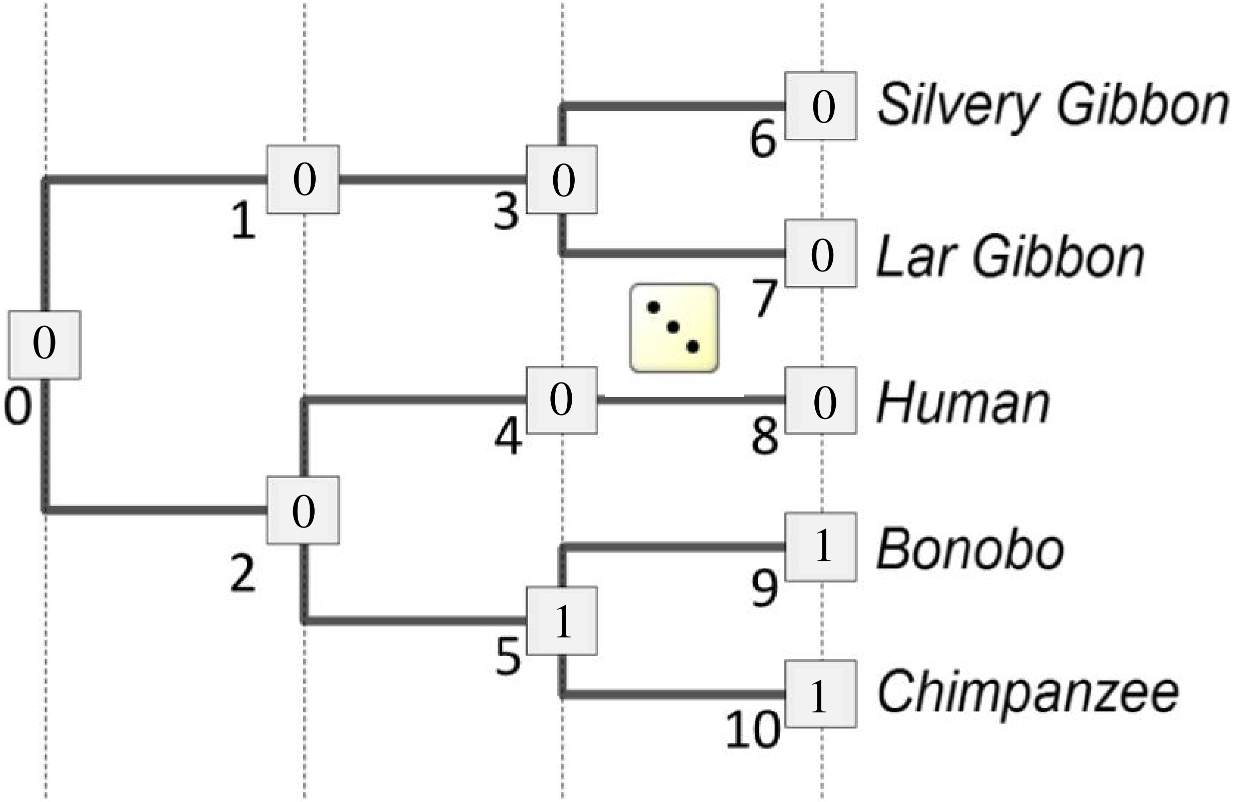

0
0
1
0
3
0
6
Silvery Gibbon
0
7
Lar Gibbon
0
2
4
8
Human
5
9
Bonobo
10
Chimpanzee


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


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

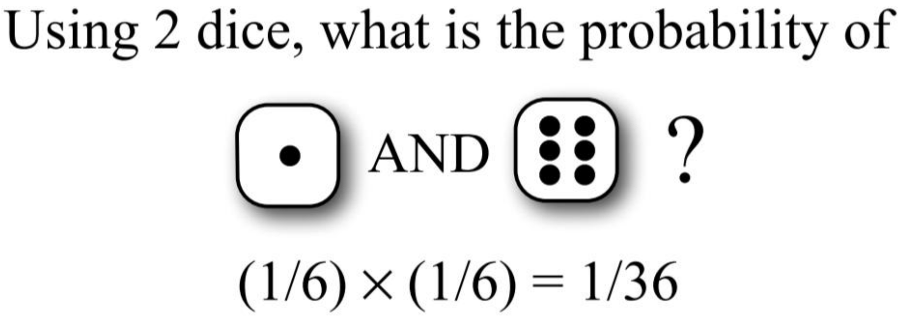

0
Silvery Gibbon
6
0
0
1
3
0
Lar Gibbon
7
0
0
0
Human
8
4
0
2
Bonobo
9
1
5
Chimpanzee
10


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


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

Combining probabilities:
The AND rule

If two independent events occur, multiply their individual probabilities to get the full probability of an event

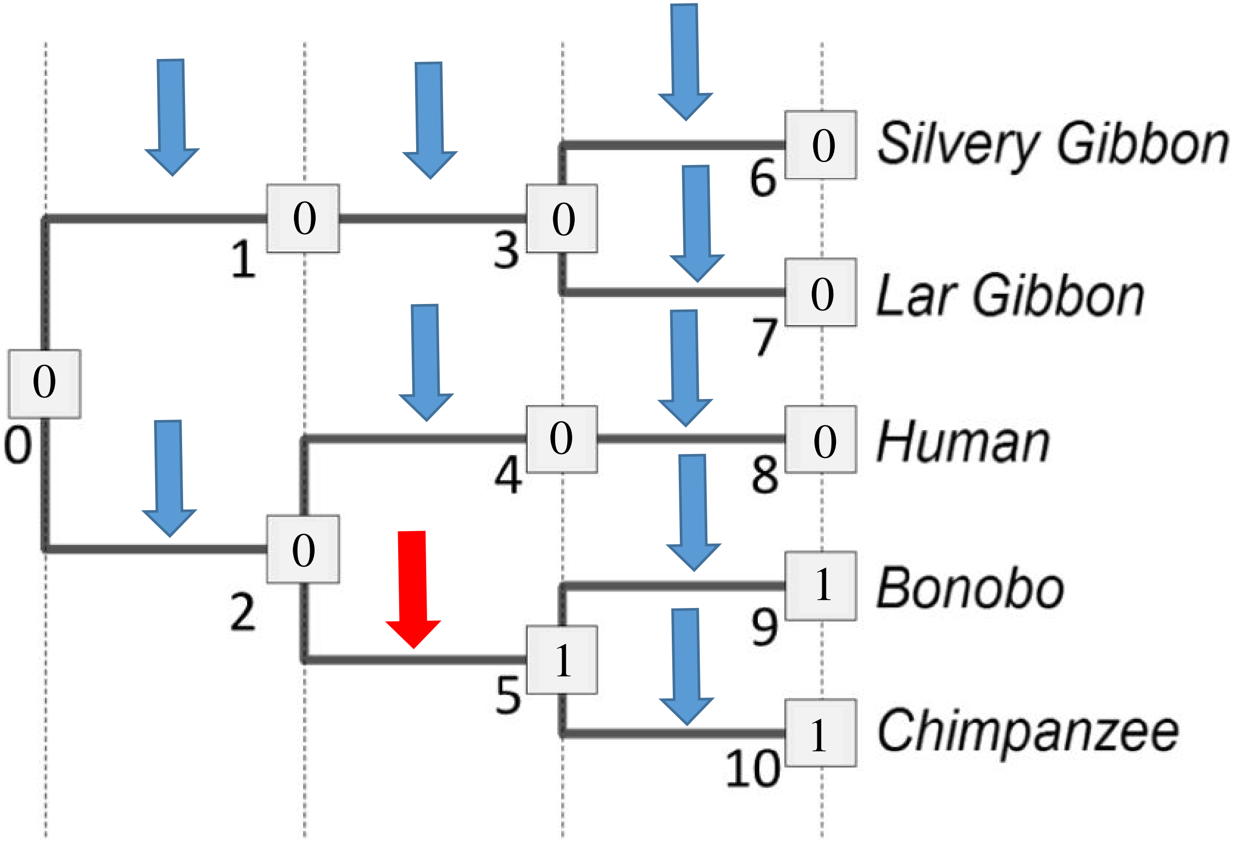


Using 2 dice, what is the probability of

⚀ AND ⚅ ?

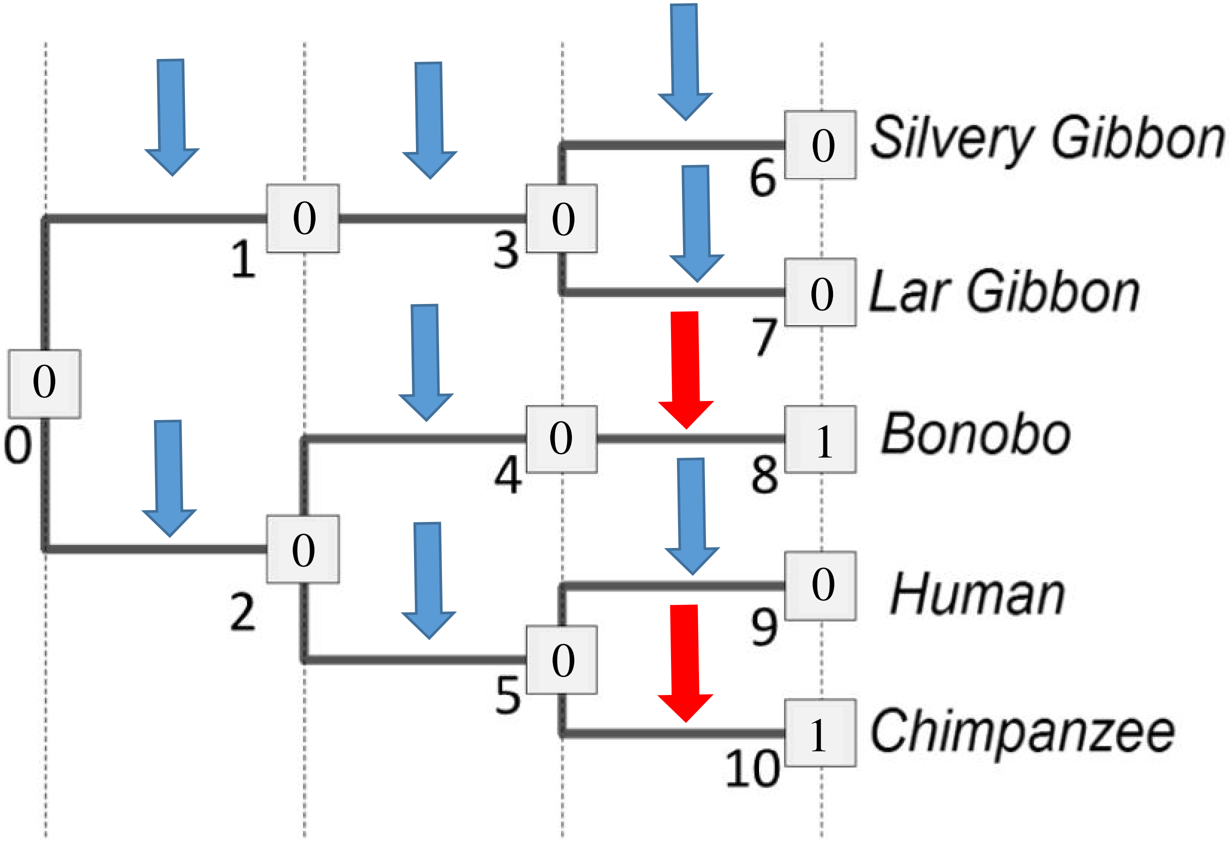
(1/6) × (1/6) = 1/36

Combining probabilities:
The AND rule in phylogenetics

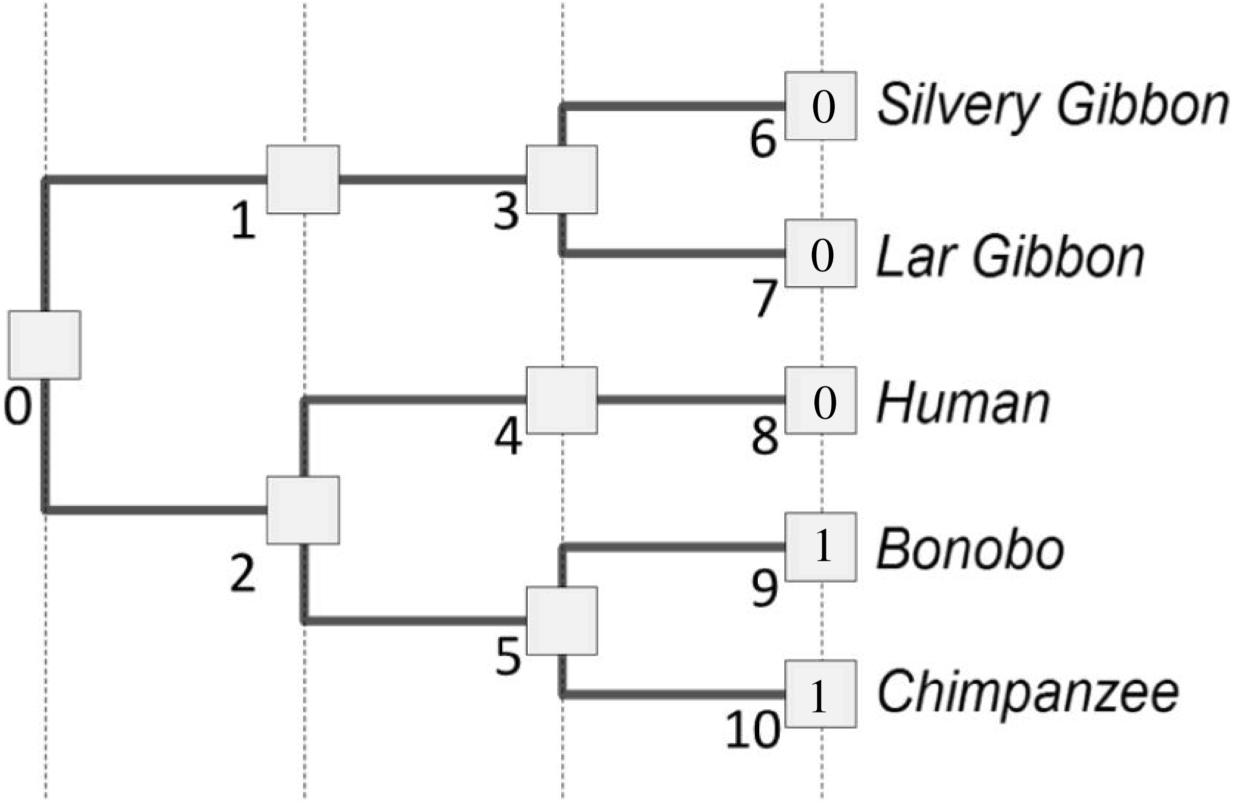


Likelihood(D, $N_i|T_1$)* = $(4/6)^9$ x $(2/6)^1$=0.0087

Combining probabilities: The AND rule in phylogenetics

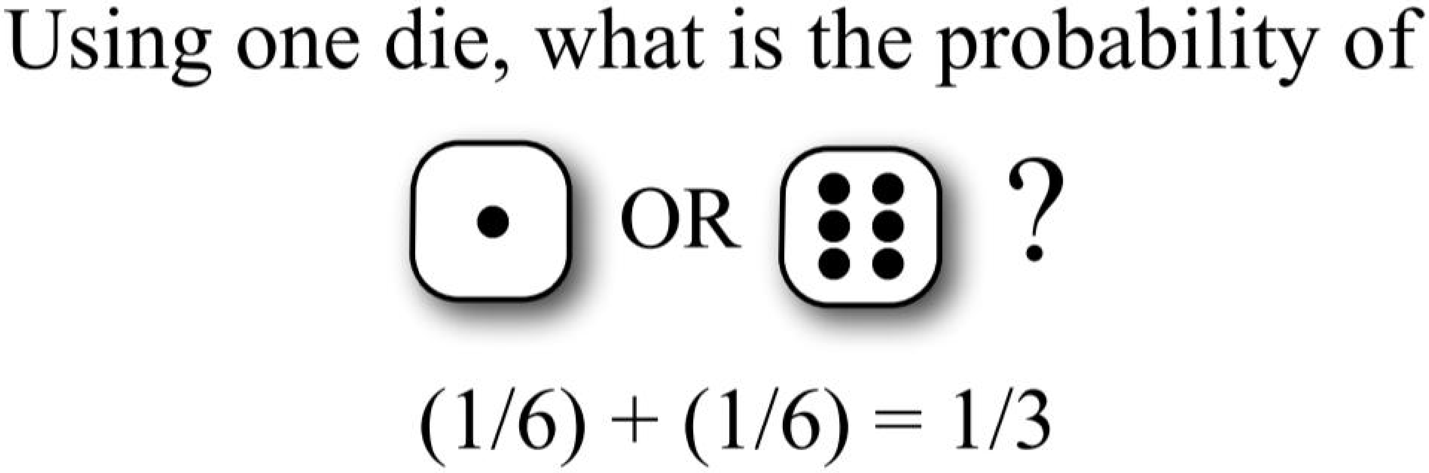


$$\text{Likelihood}(D, N_j|T_2)^* = (4/6)^8 \times (2/6)^2 = 0.0043$$


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

Combining probabilities: The OR rule

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

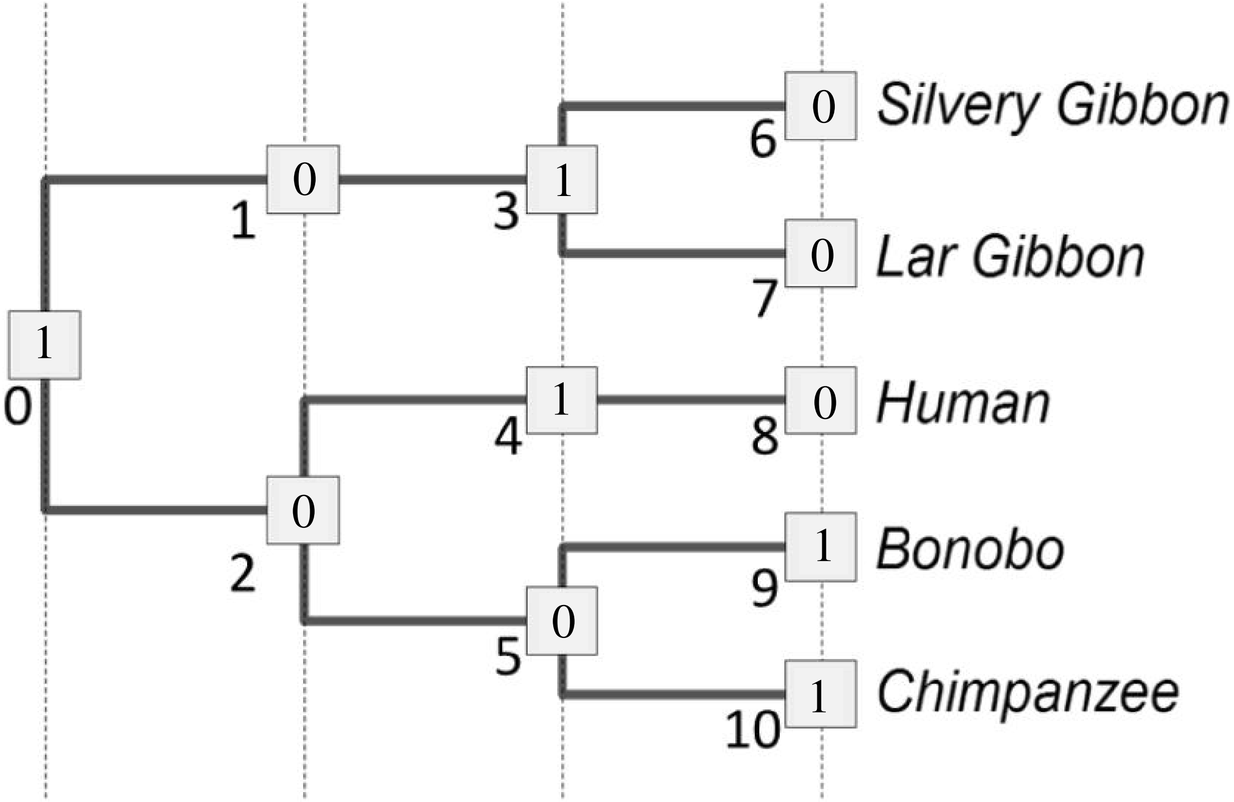


Using one die, what is the probability of

⚀ OR ⚅ ?

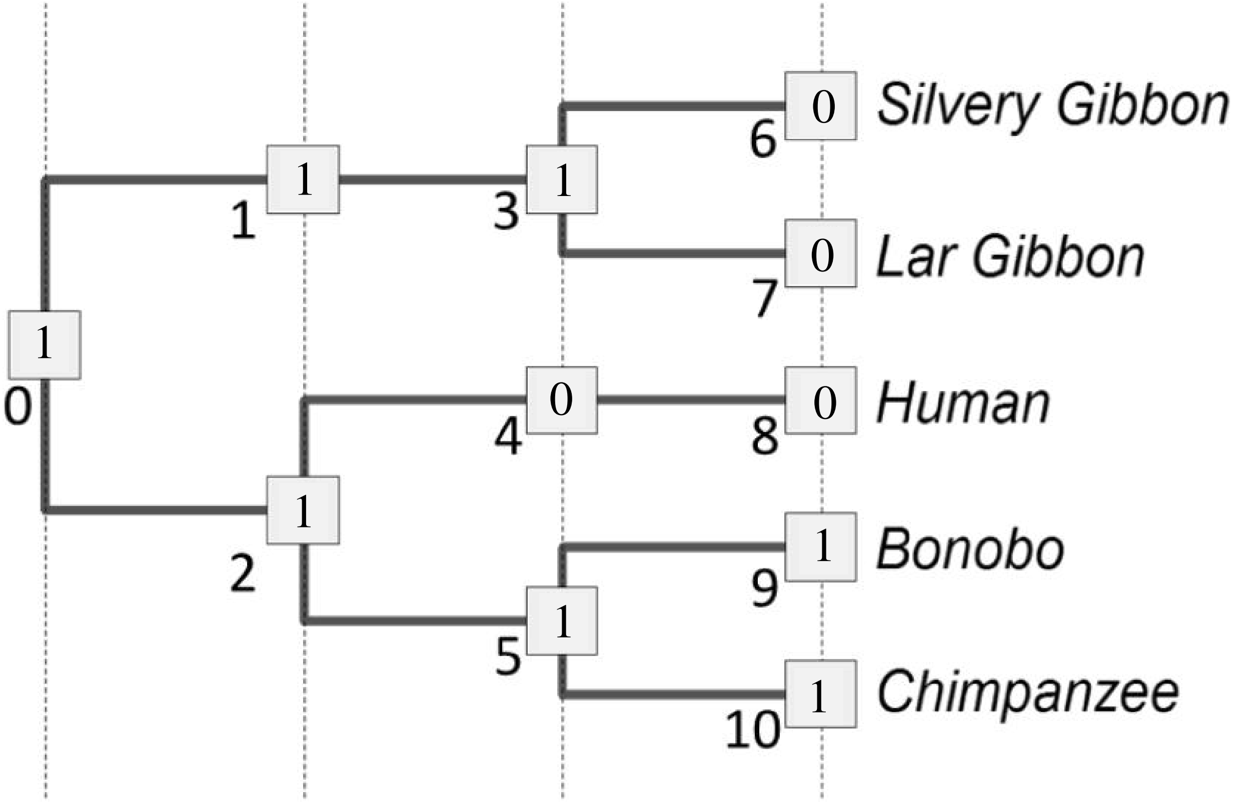
$(1/6) + (1/6) = 1/3$

Combining probabilities: The OR rule in phylogenetics

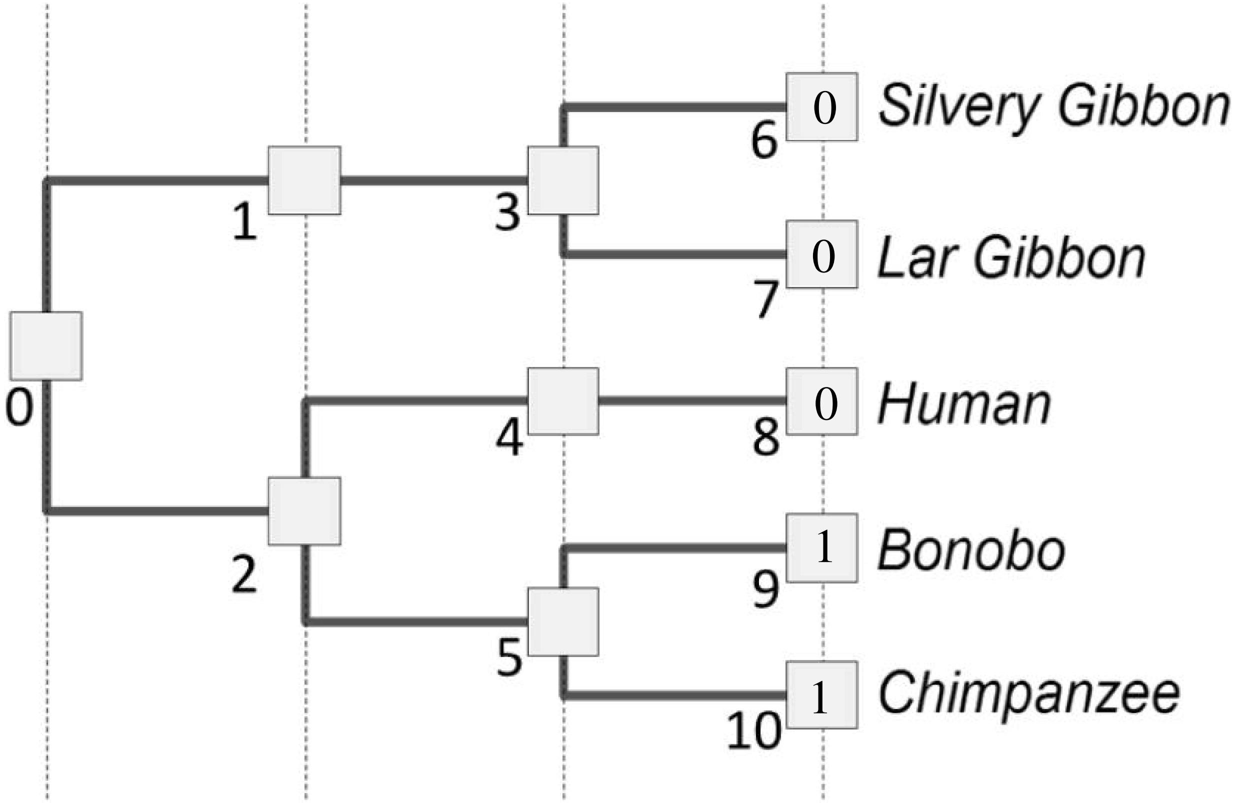


Likelihood (D, $N_k|T_1$) = $(4/6)^1$ x $(2/6)^9$=0.000033

Combining probabilities: The OR rule in phylogenetics



Likelihood (D, $N_k|T_1$) = $(4/6)^7$ x $(2/6)^3$=0.00217



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

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

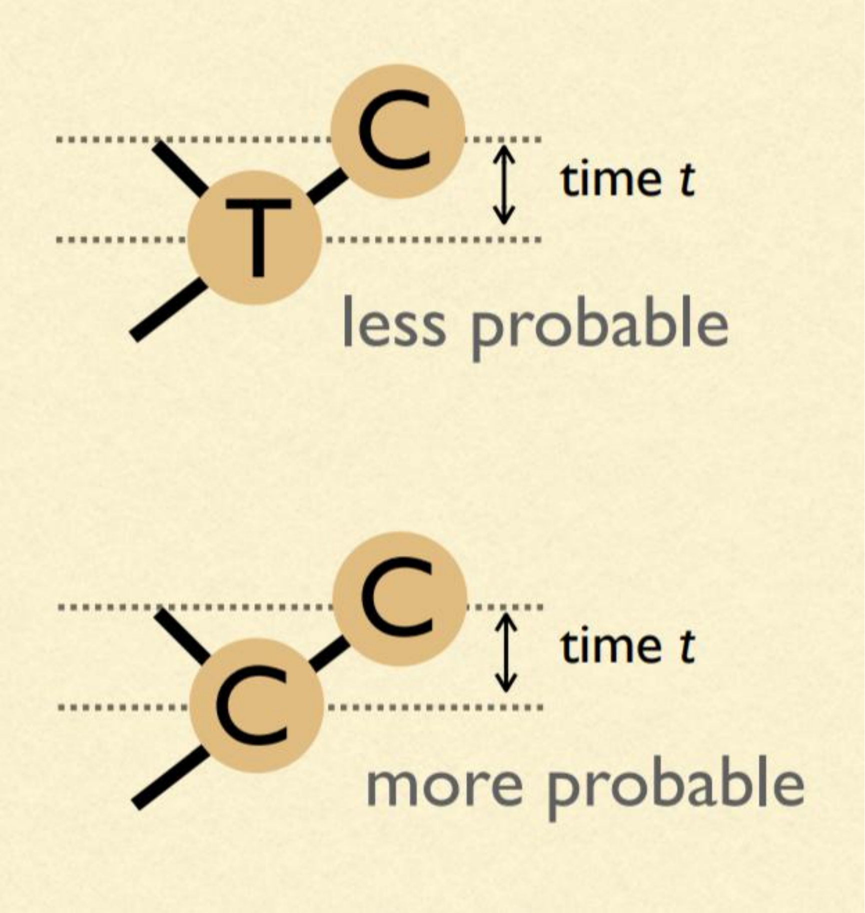
Continuous-Time Markov Models

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

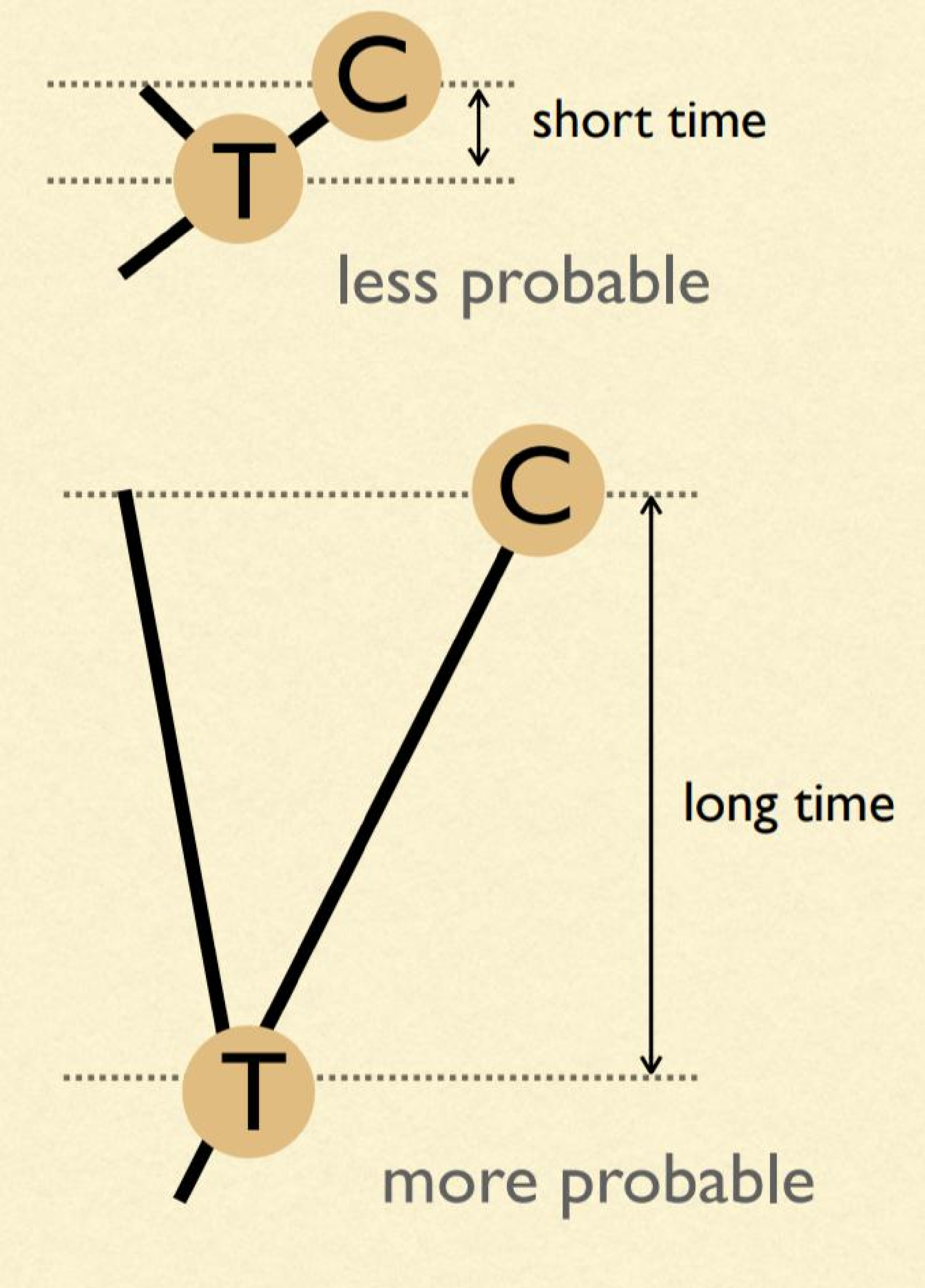
Our model of change depends on time: We must estimate branch lengths

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

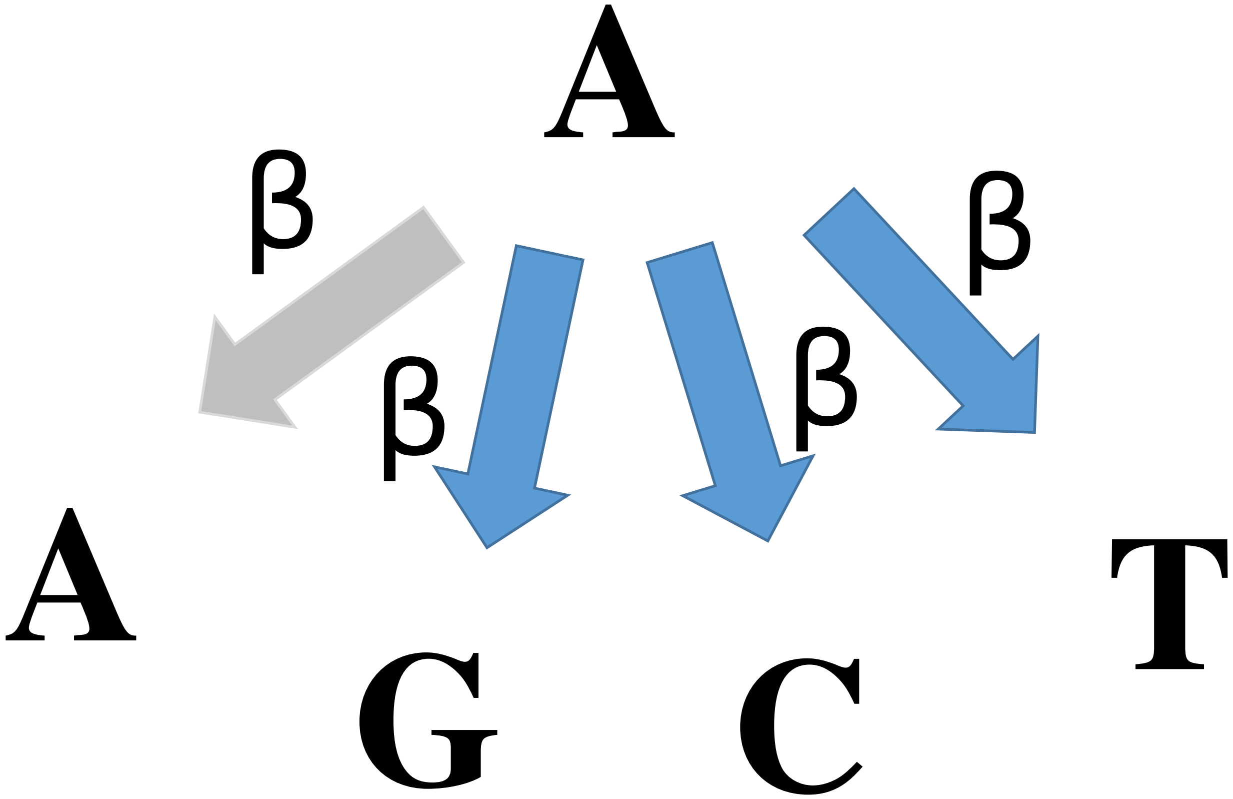
(= rate of substitions x time)



Probabilities are dependent on time



$\mu = 4\beta$



A

β

β

β

β

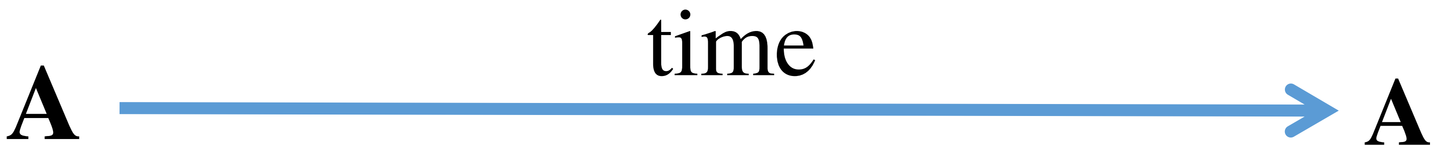
A

G

C

T

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...

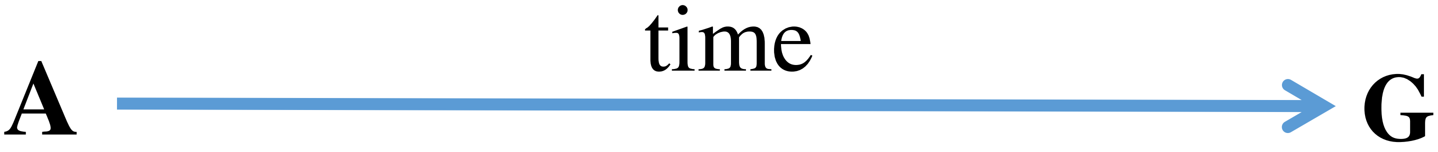
$$P_{AA} = (e^{-\mu t}) + (1 - e^{-\mu t})(1/4)$$

Probability something doesn't happen

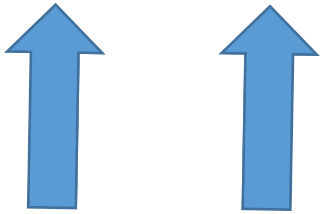
Probability at least one thing happens

Probability that the last thing that happened ends in an A

A (very) simple phylogeny...



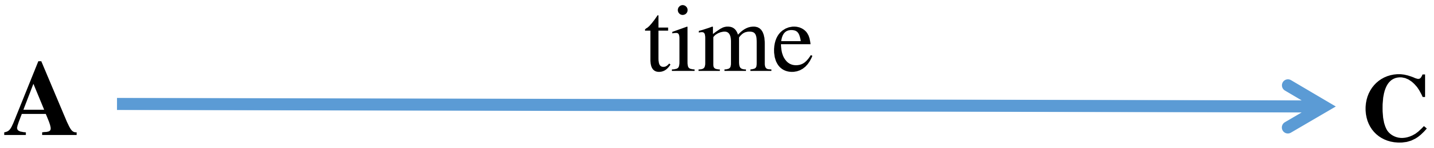
$$P_{AG} = \qquad (1 - e^{-\mu t})(1/4)$$



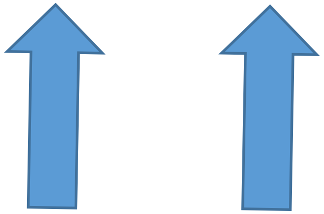
Probability at least one thing happens

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...

A

time

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"

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

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

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

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

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

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

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

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

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

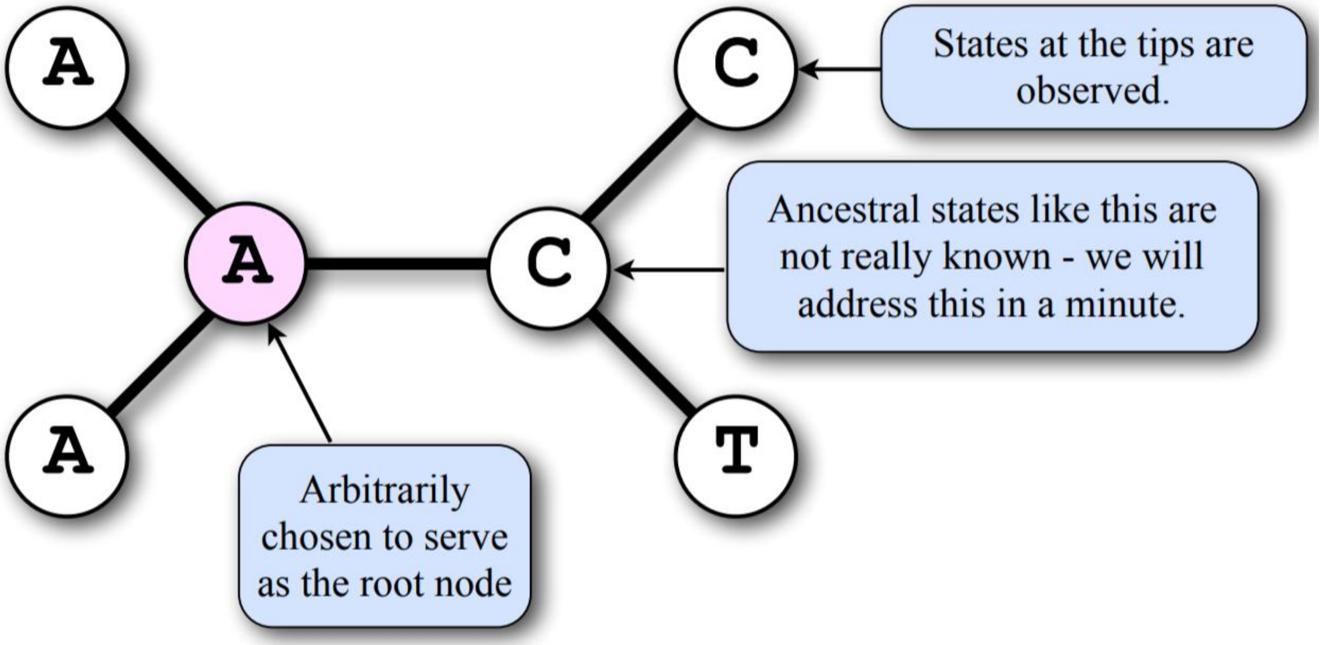
$$P_{AC} = (1/4) - (1/4)(e^{-4\nu/3})$$

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

Sanity check:
Do they all add to 1?

Likelihood of an unrooted tree

(data shown for only one site)



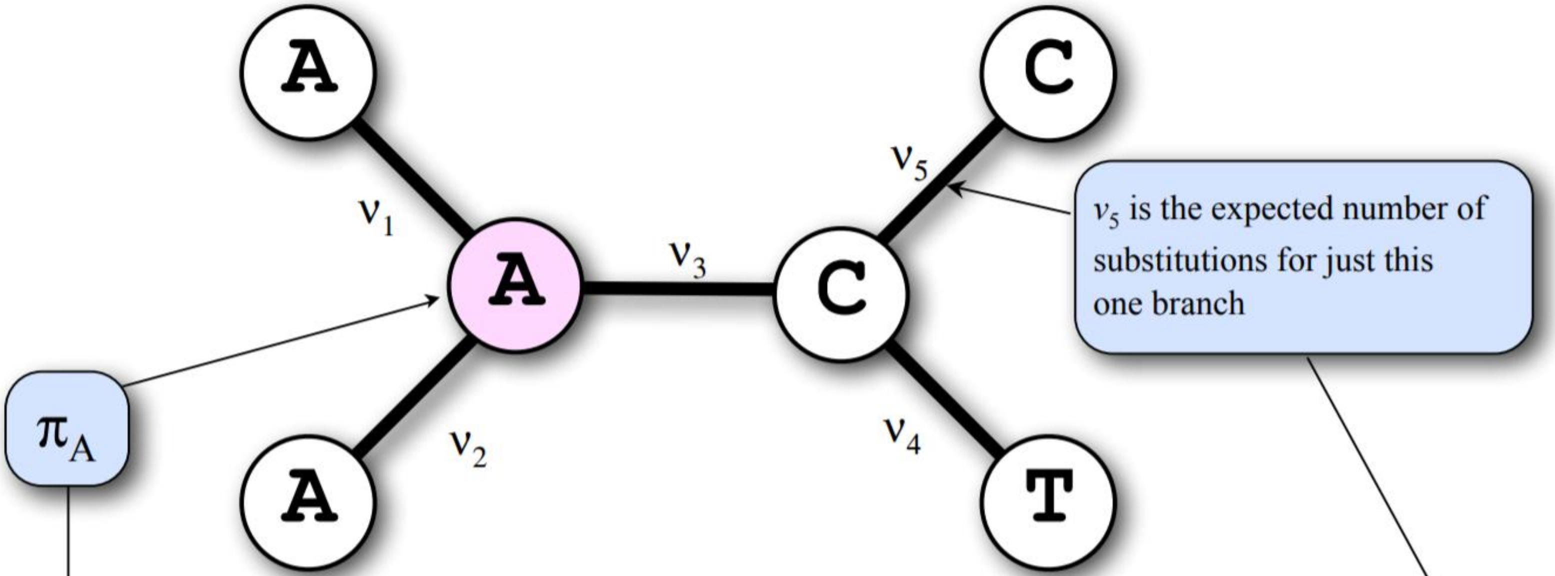
States at the tips are observed.

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

Arbitrarily chosen to serve as the root node

Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

Likelihood for site k



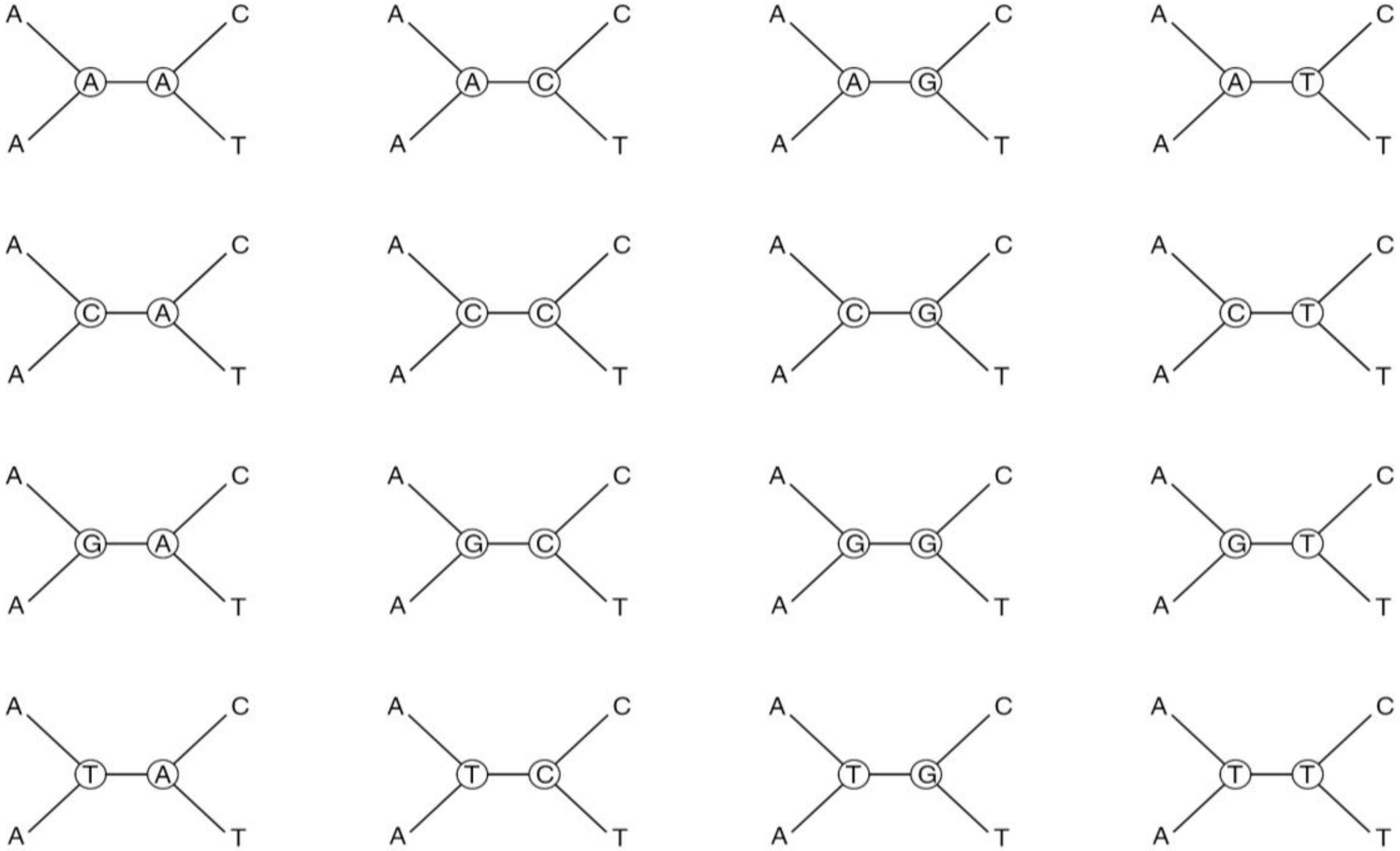
ν_5 is the expected number of substitutions for just this one branch

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

$P_{AA}(\nu_1)$ $P_{AA}(\nu_2)$ $P_{AC}(\nu_3)$ $P_{CT}(\nu_4)$ $P_{CC}(\nu_5)$

Note use of the AND probability rule

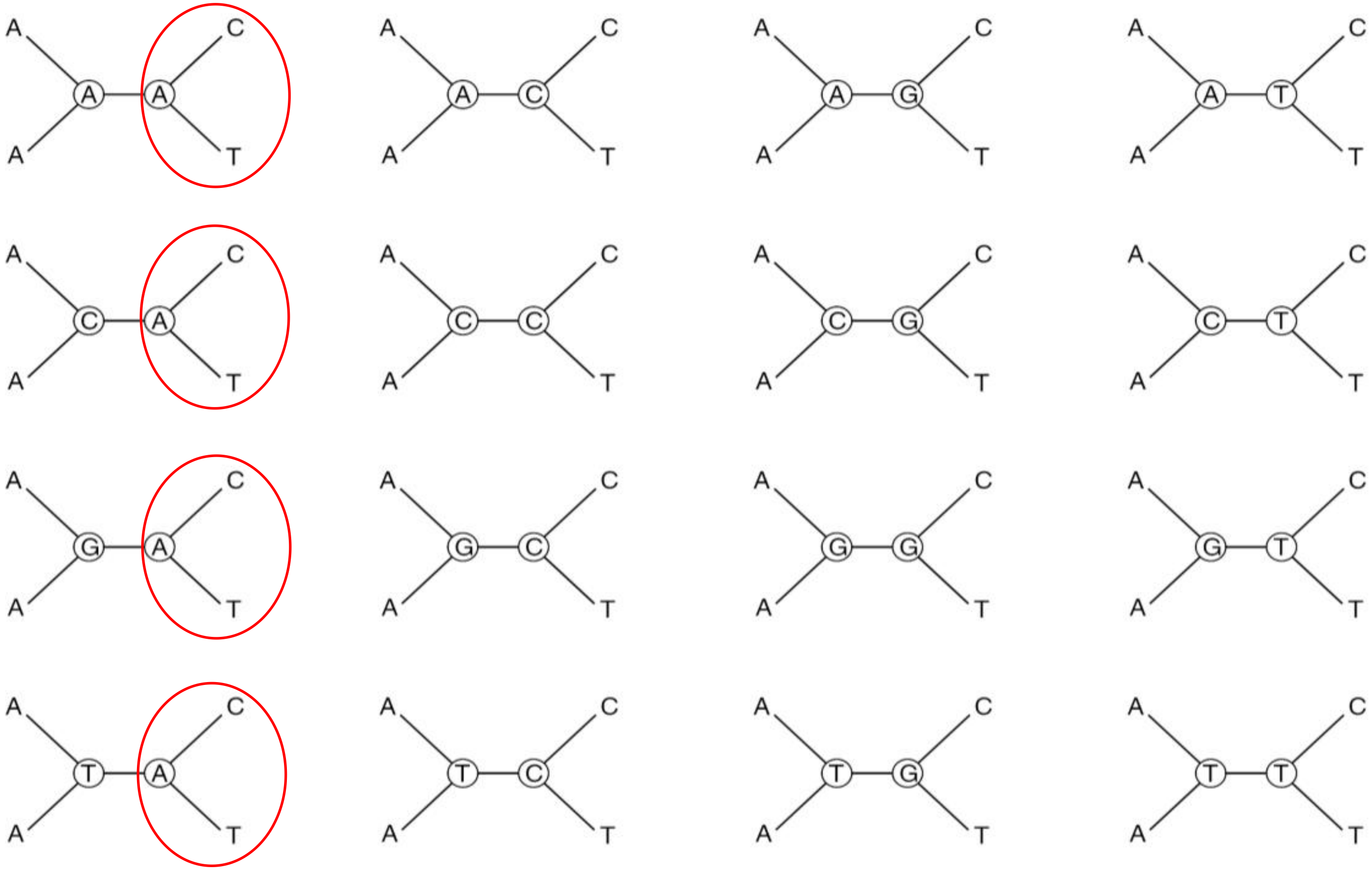
Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them



Note use of the OR probability rule

Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

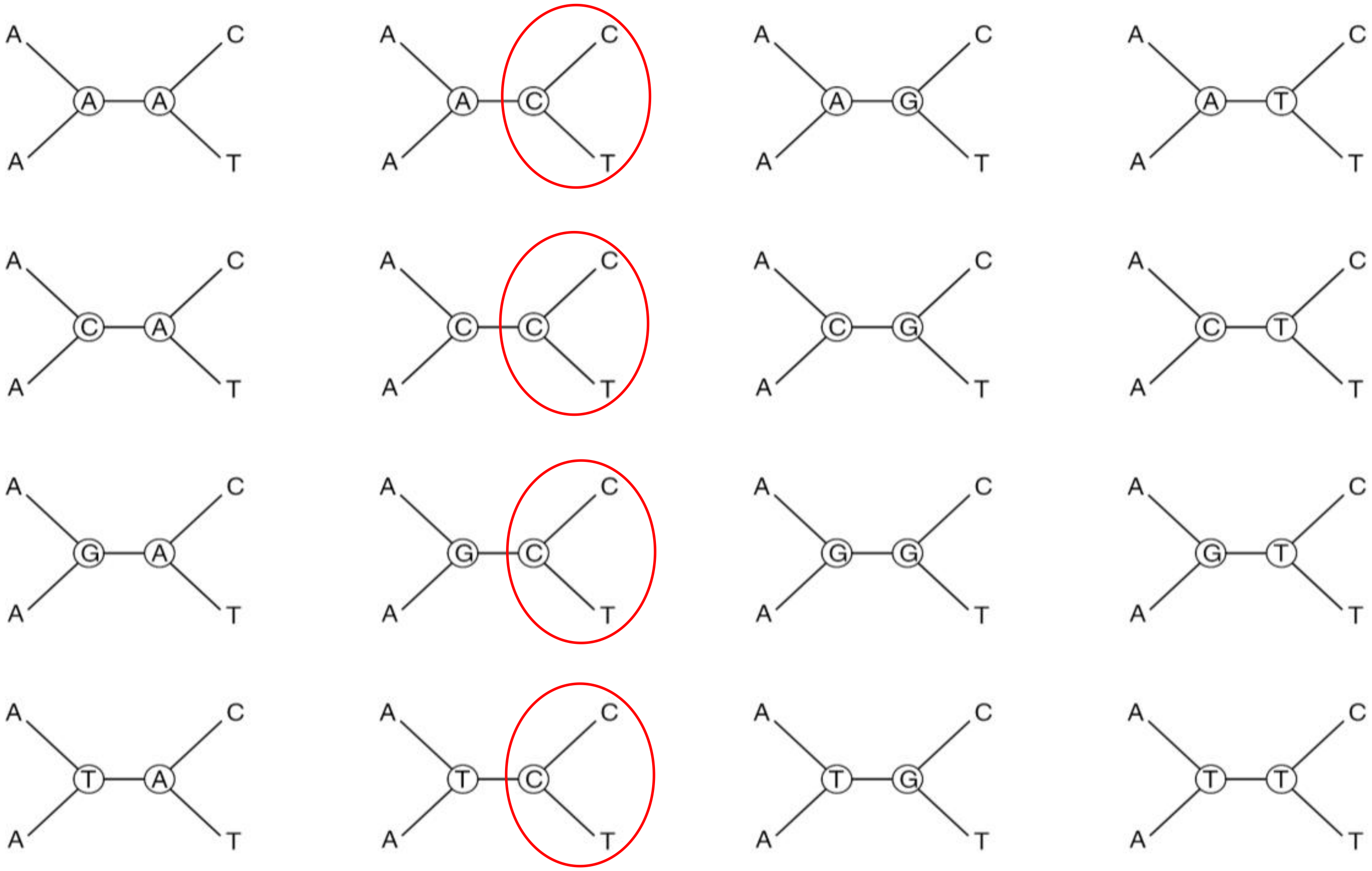
Pruning algorithm



Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

Note use of the OR probability rule

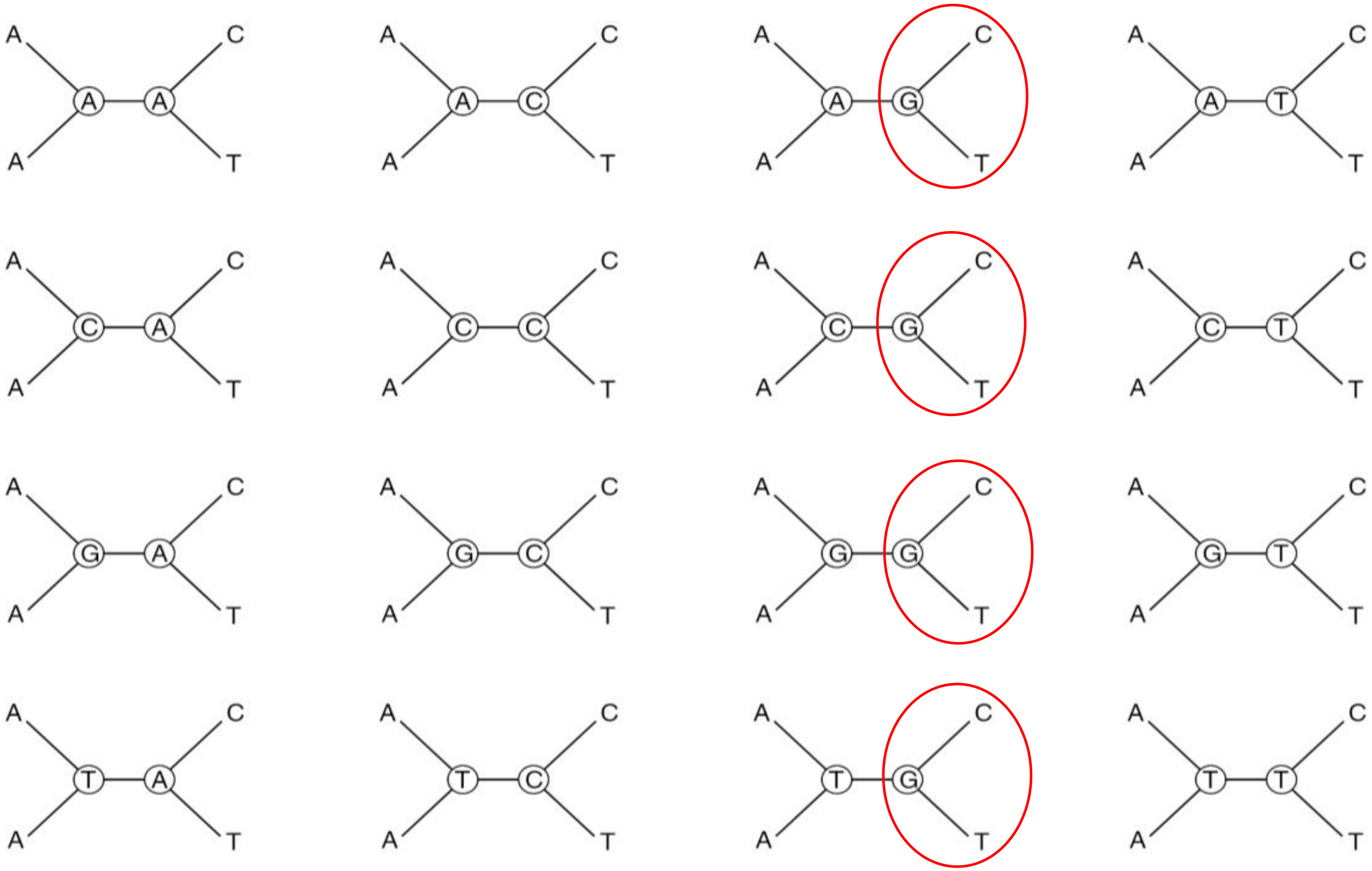
Pruning algorithm



Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

Note use of the OR probability rule

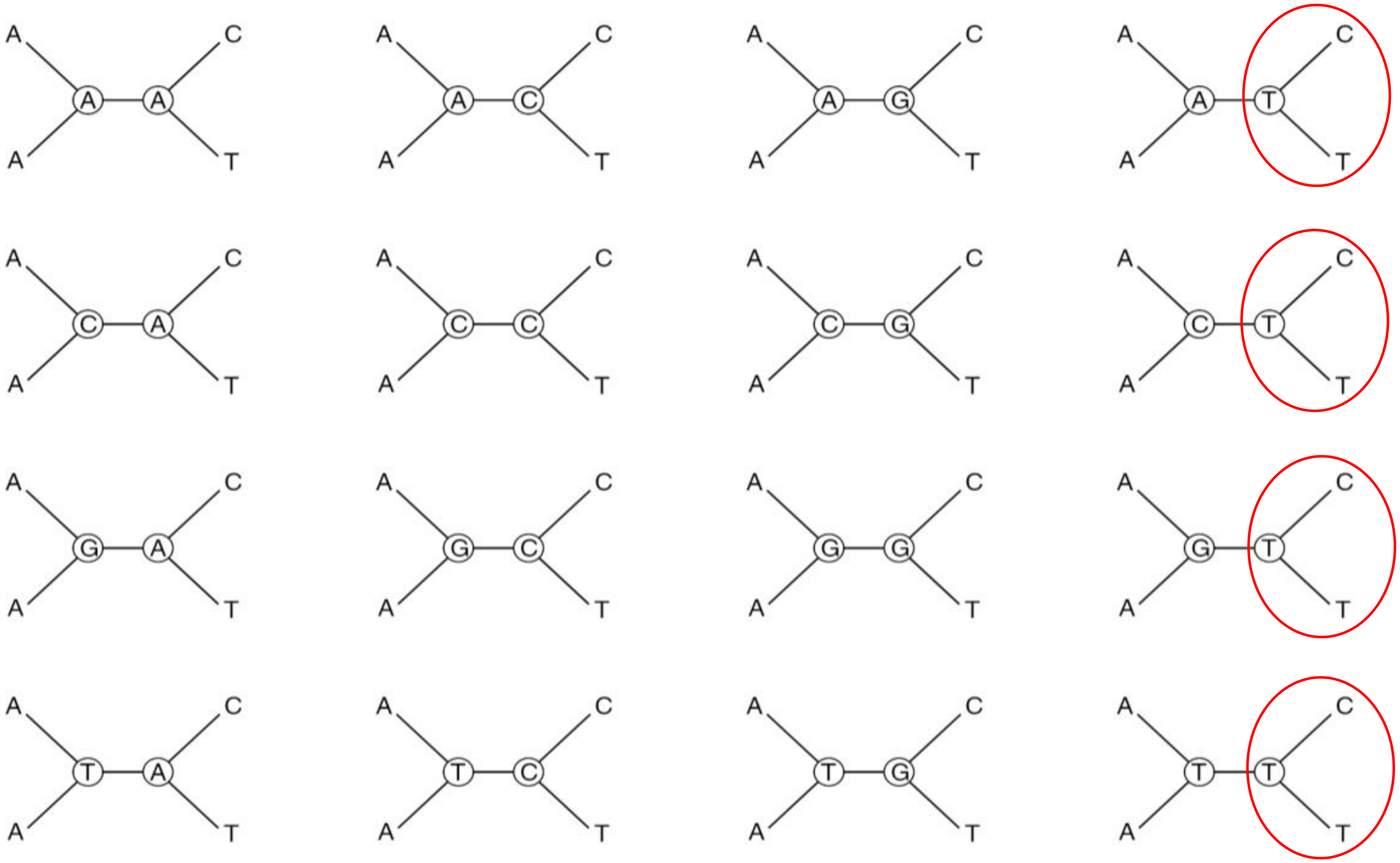
Pruning algorithm



Note use of the OR probability rule

Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

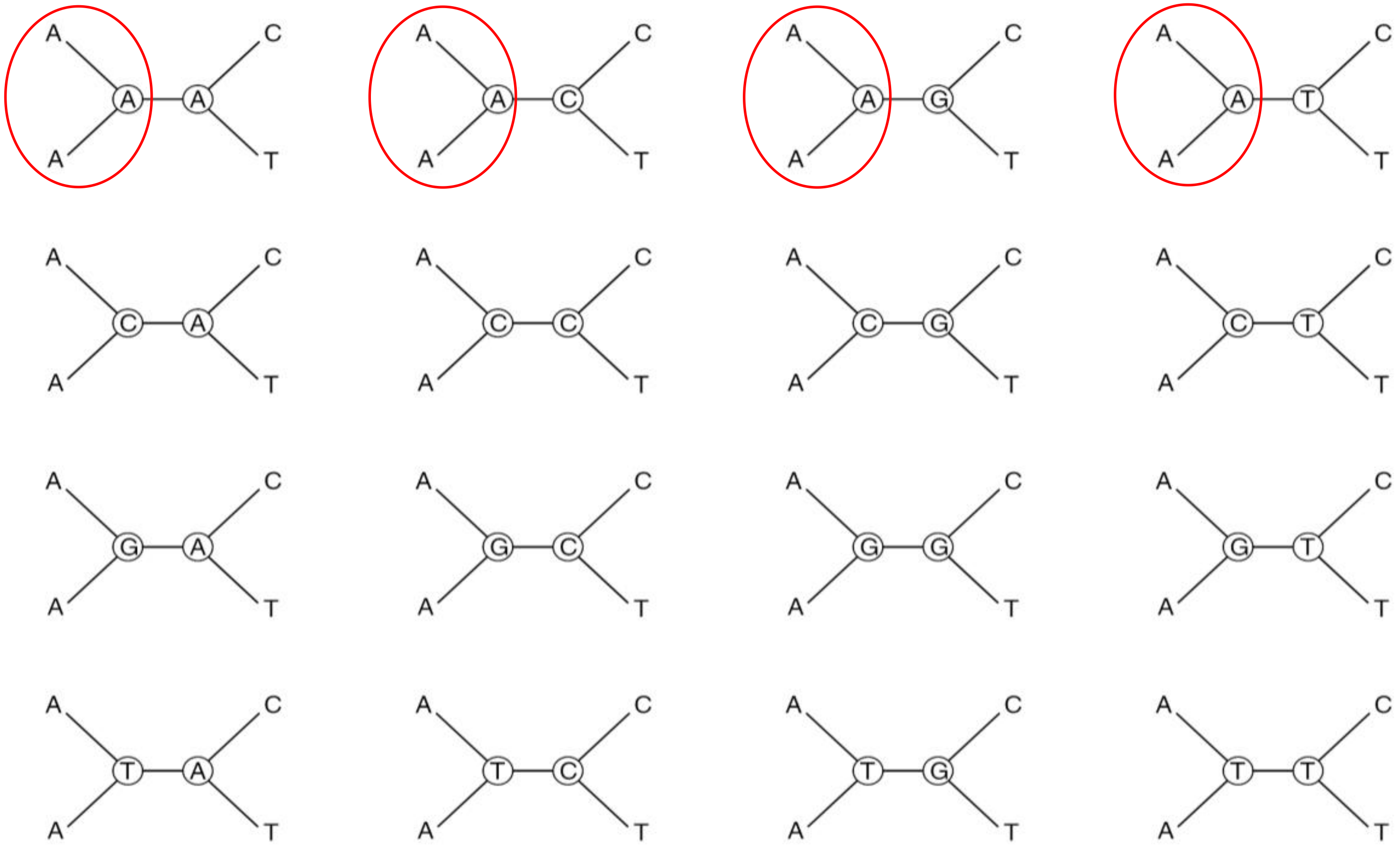
Pruning algorithm



Note use of the OR probability rule

Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

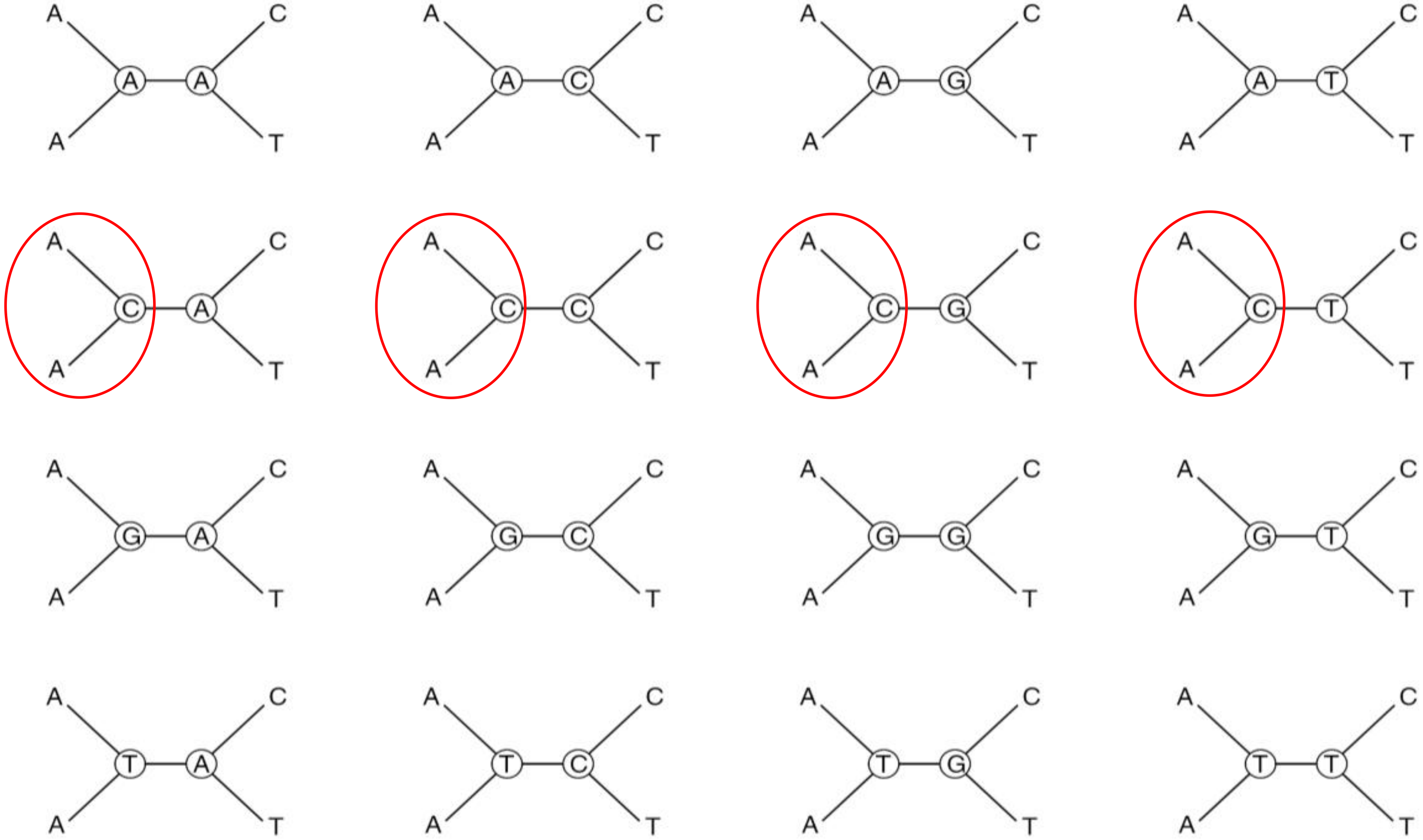
Pruning algorithm



Note use of the OR probability rule

Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

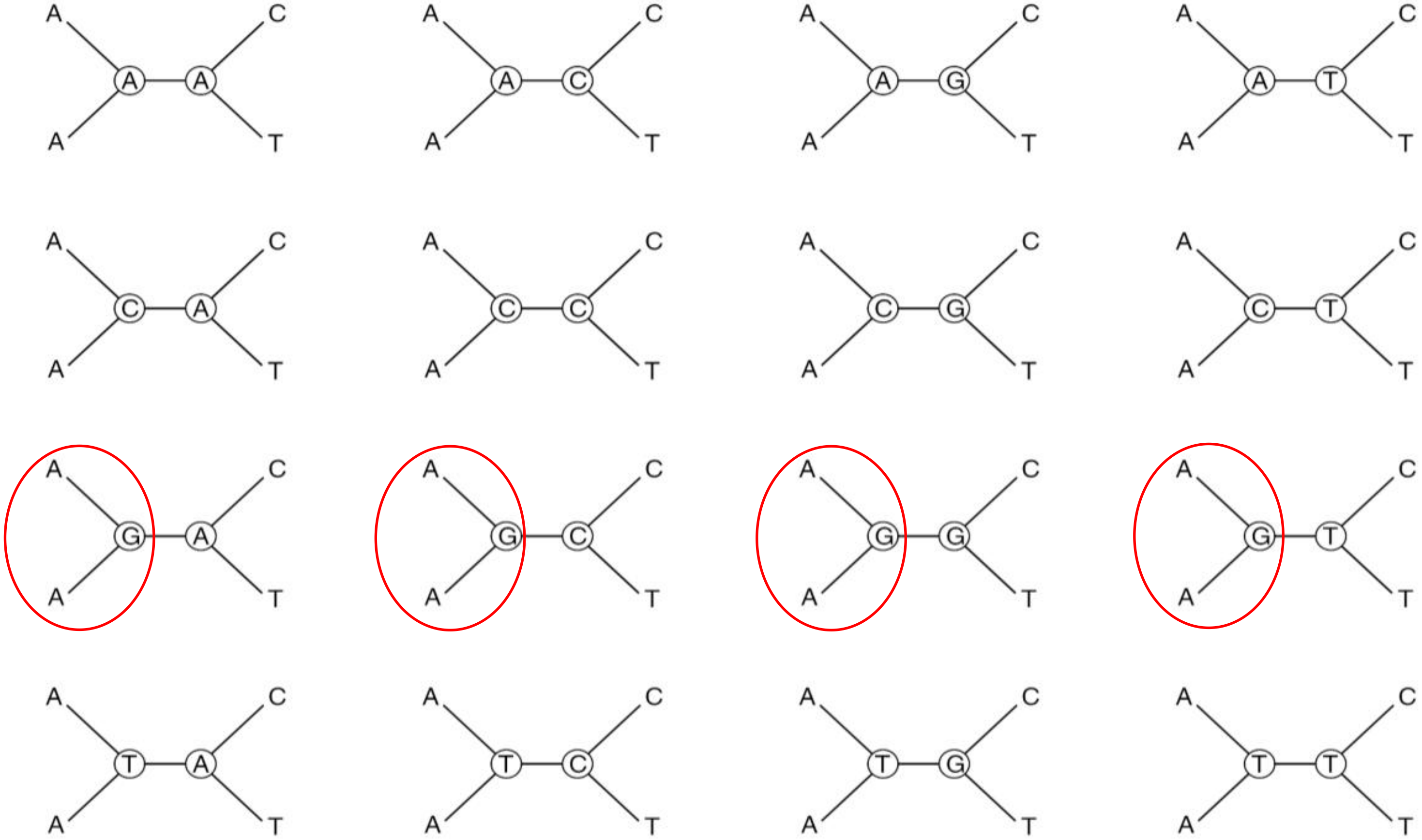
Pruning algorithm



Note use of the OR probability rule

Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

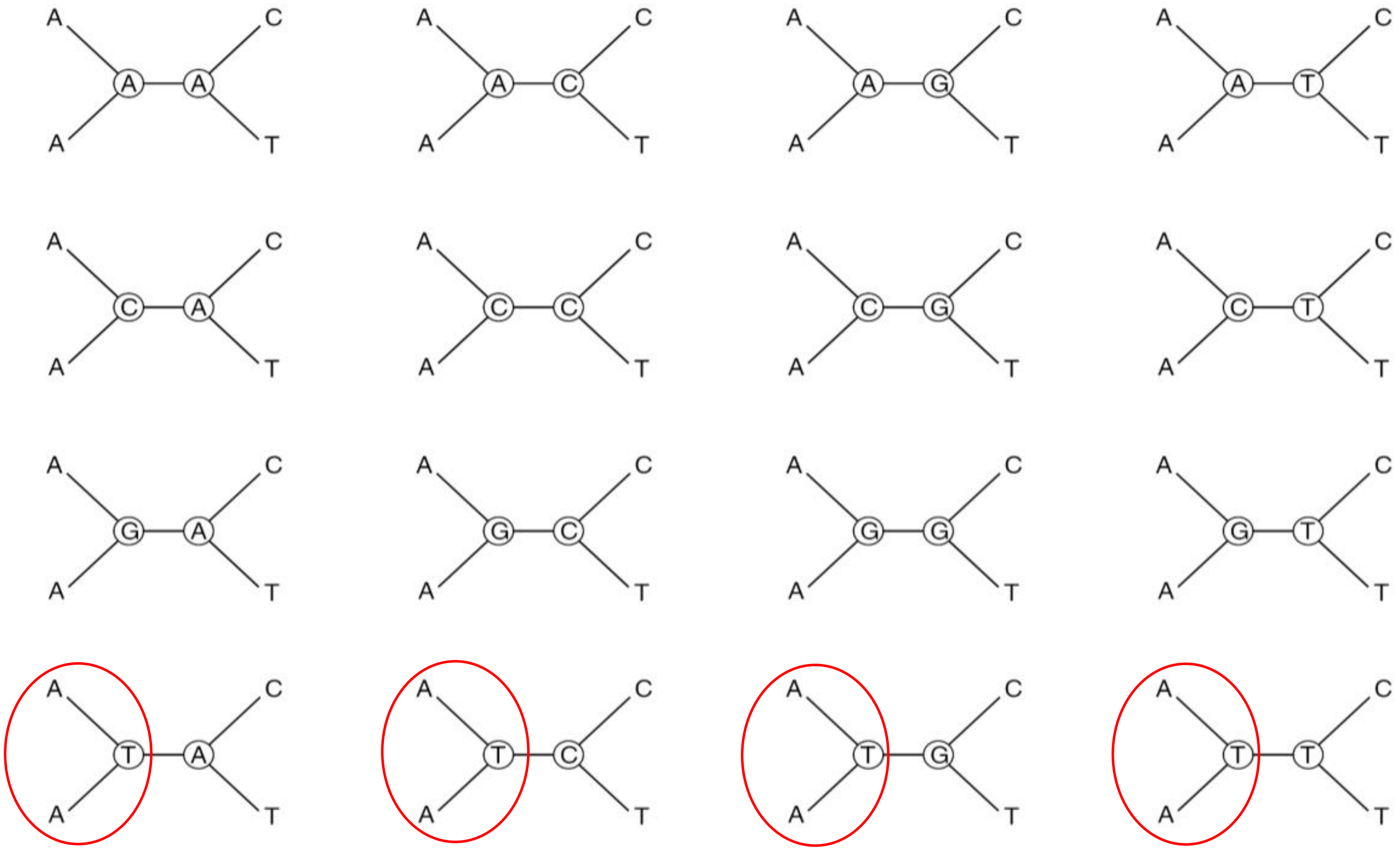
Pruning algorithm



Note use of the OR probability rule

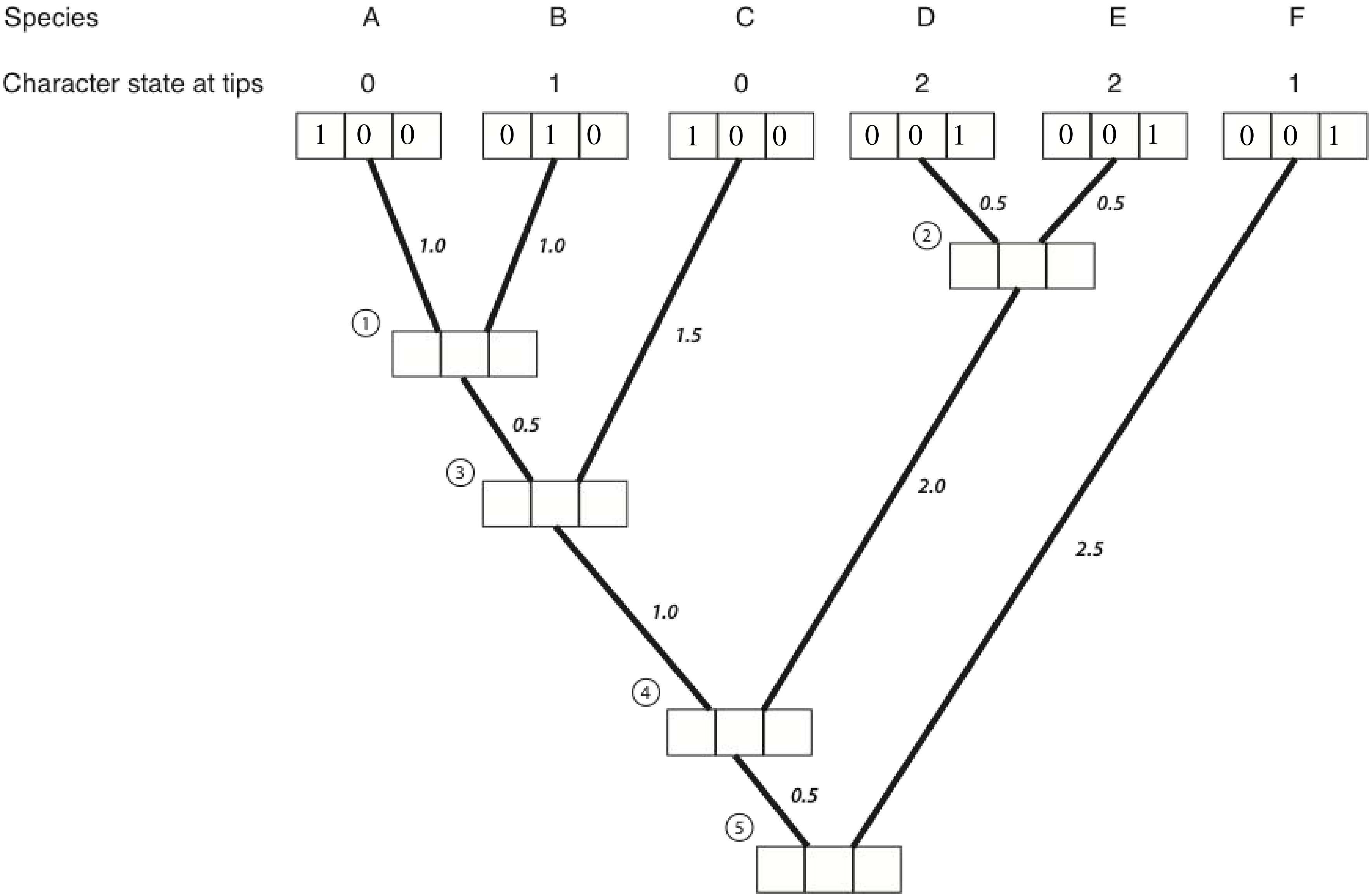
Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

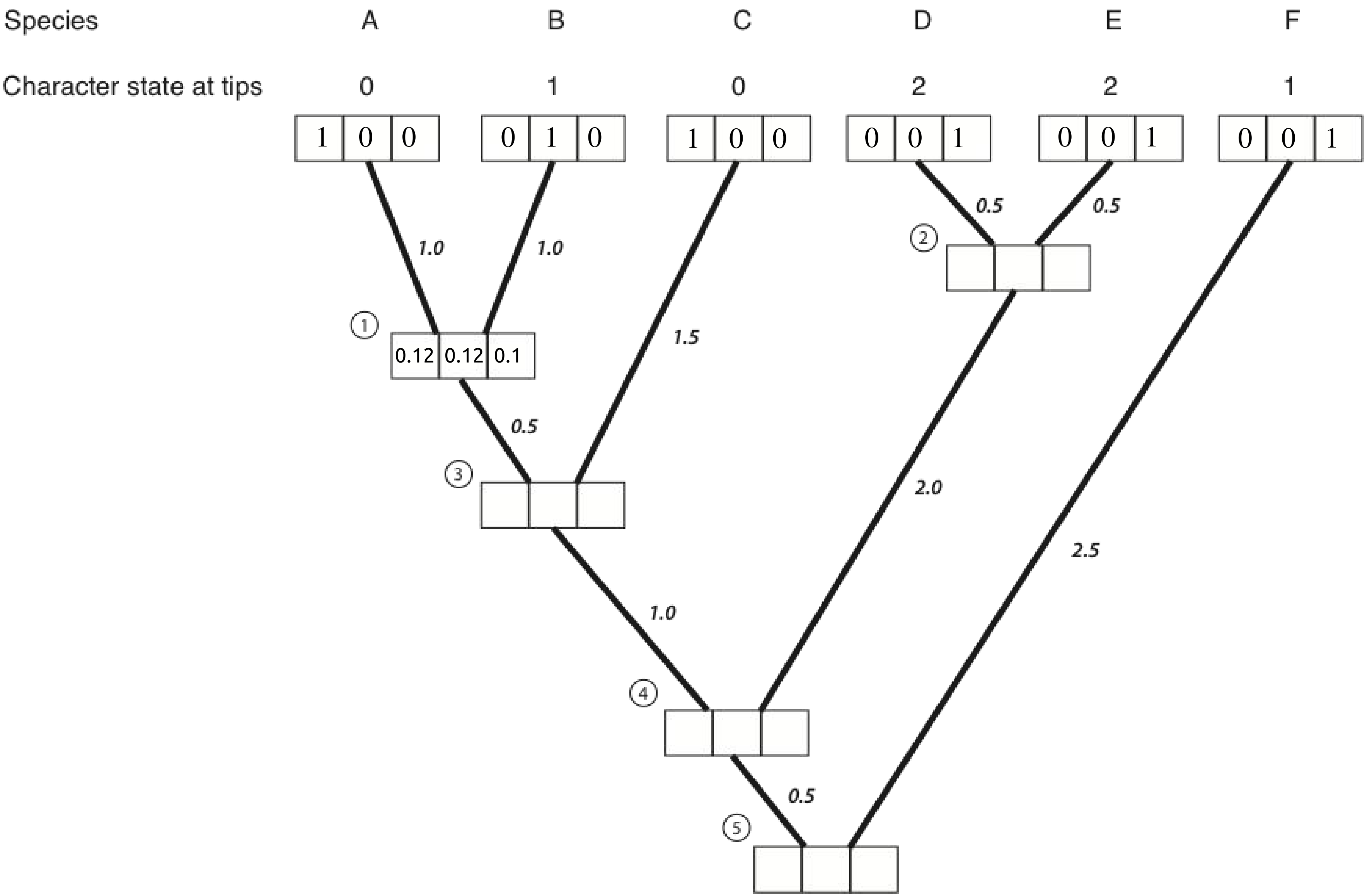
Pruning algorithm

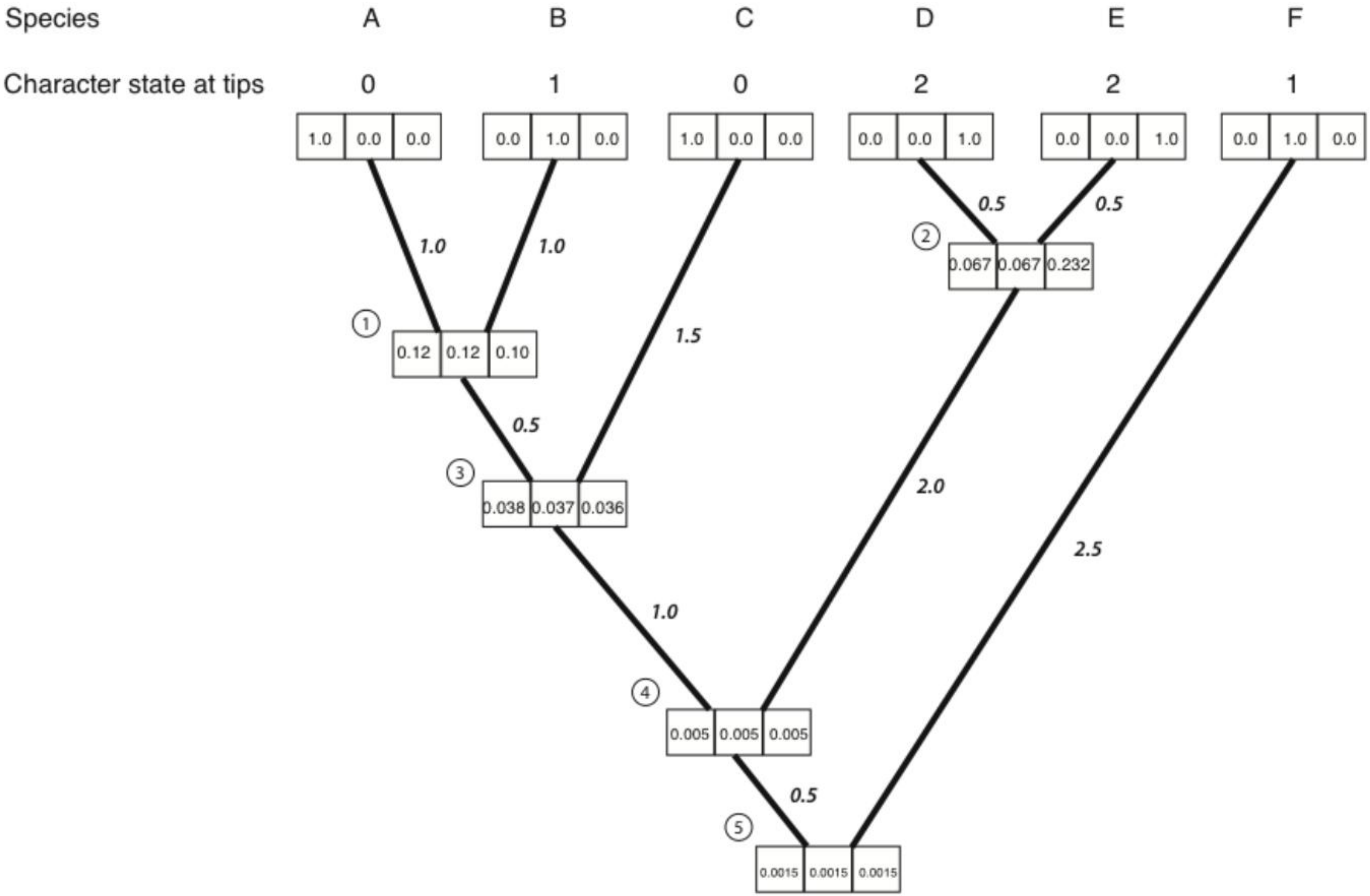


Paul O. Lewis (2014 Woods Hole Workshop in Molecular Evolution)

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
1.5
0.5
0.5
2.5
2.0
0.5
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
1.5
0.5
0.5
①
0.12 0.12 0.1
②
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
0.5
0.5
1.0
1.0
2
0.067 0.067 0.232
1.5
1
0.12 0.12 0.10
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

Jukes-Cantor Model (JC69)

Q matrix *(instantaneous rates)*

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

Jukes-Cantor Model (JC69)

Transition probabilities:

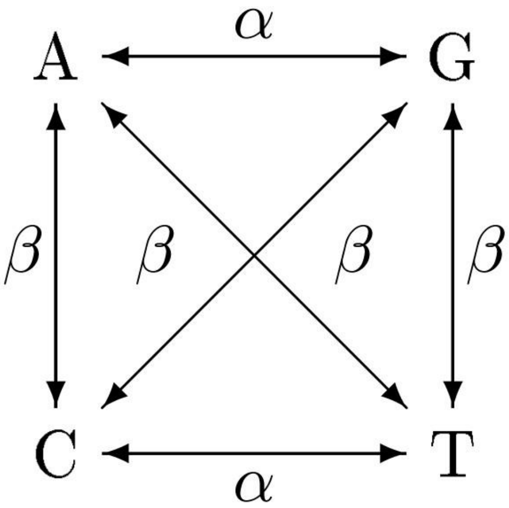
$$P = e^{Qt}$$

Matrix exponentiation

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

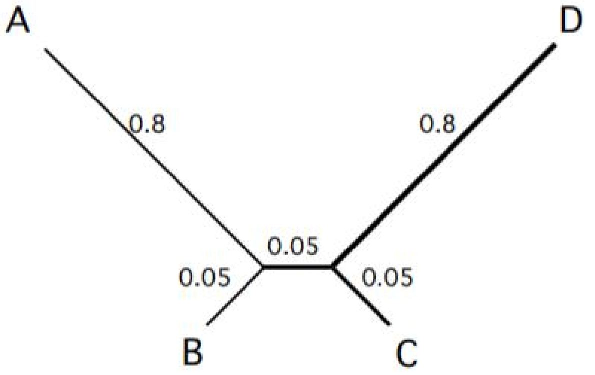
$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & \beta & \beta & \beta \\ C & \beta & - & \beta & \beta \\ G & \beta & \beta & - & \beta \\ T & \beta & \beta & \beta & - \end{array}$$

Kimura 2 Parameter model: K2P

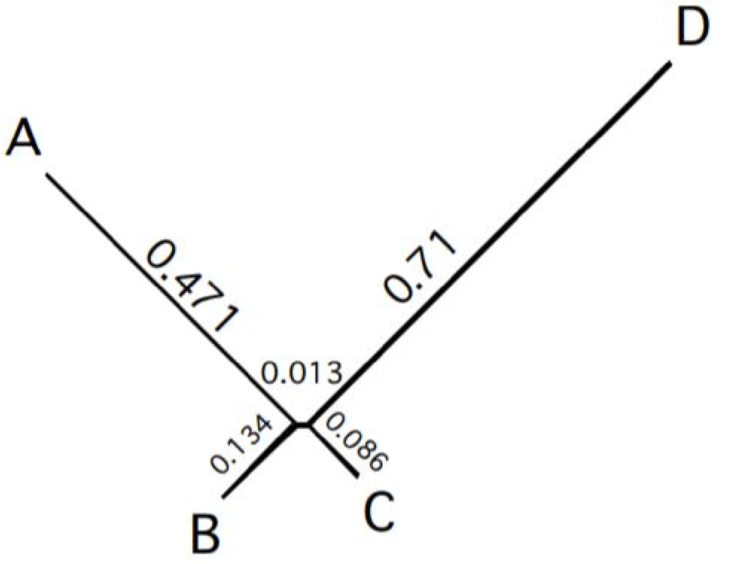


$$Q = \begin{array}{c|cccc|} & A & C & G & T \\ A & - & \beta & \alpha & \beta \\ C & \beta & - & \beta & \alpha \\ G & \alpha & \beta & - & \beta \\ T & \beta & \alpha & \beta & - \end{array}$$

Returning to our original problem:



True Tree



ML Tree

Differences between statistical phylogenetics and parsimony

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

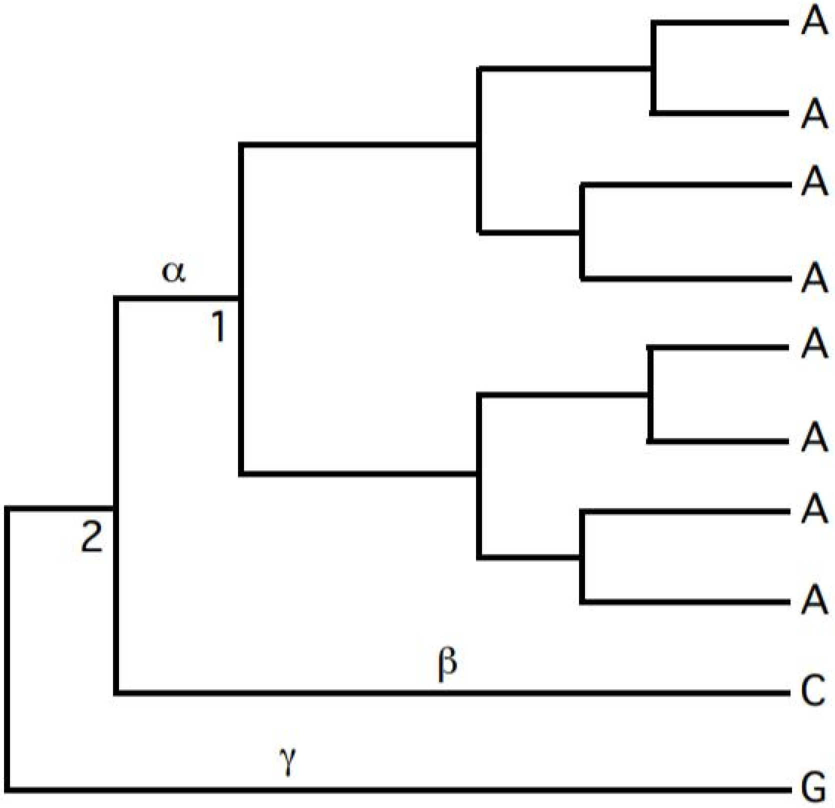
We expect and probabilistically incorporate all possible paths to the data, not just the shortest path

We have the flexbility to modify and compare models

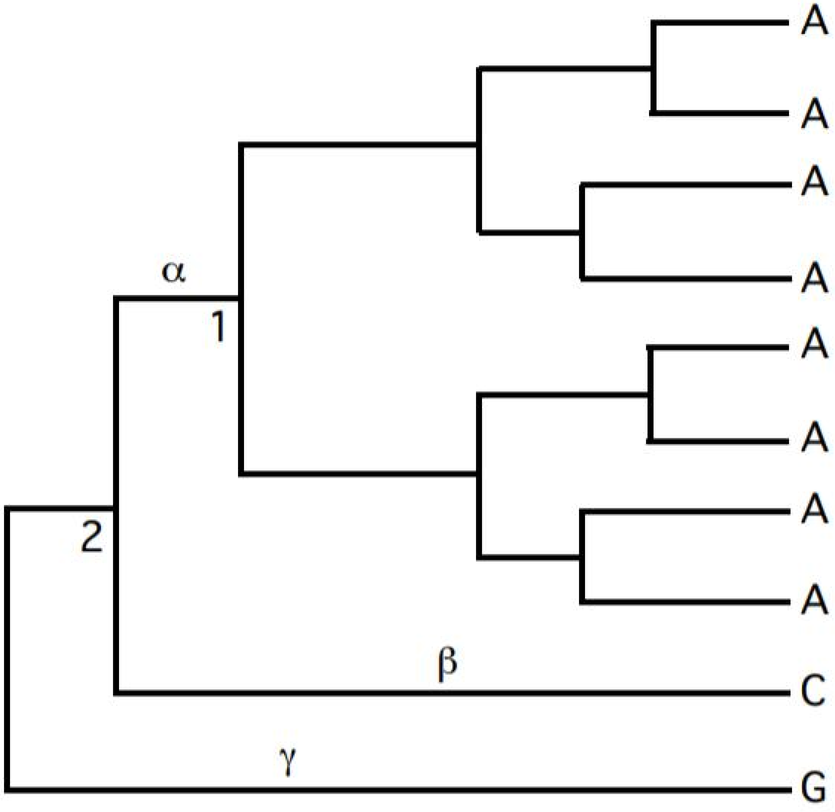
We have a straight-forward way to convert branch lengths to time (with fossils or other constraints)

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: Why might we argue node 2 is most likely NOT A?


A
A
A
A
α
1
A
A
A
2
A
β
C
γ
G