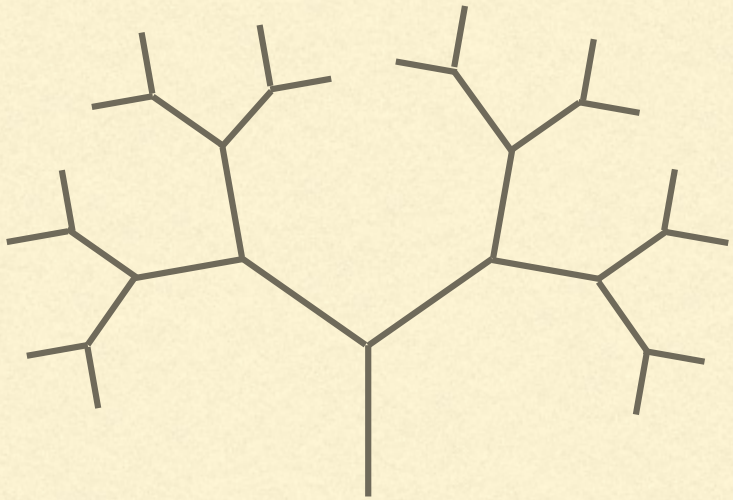
PHYLOGENETICS 101

Part I: Trees and Likelihood

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

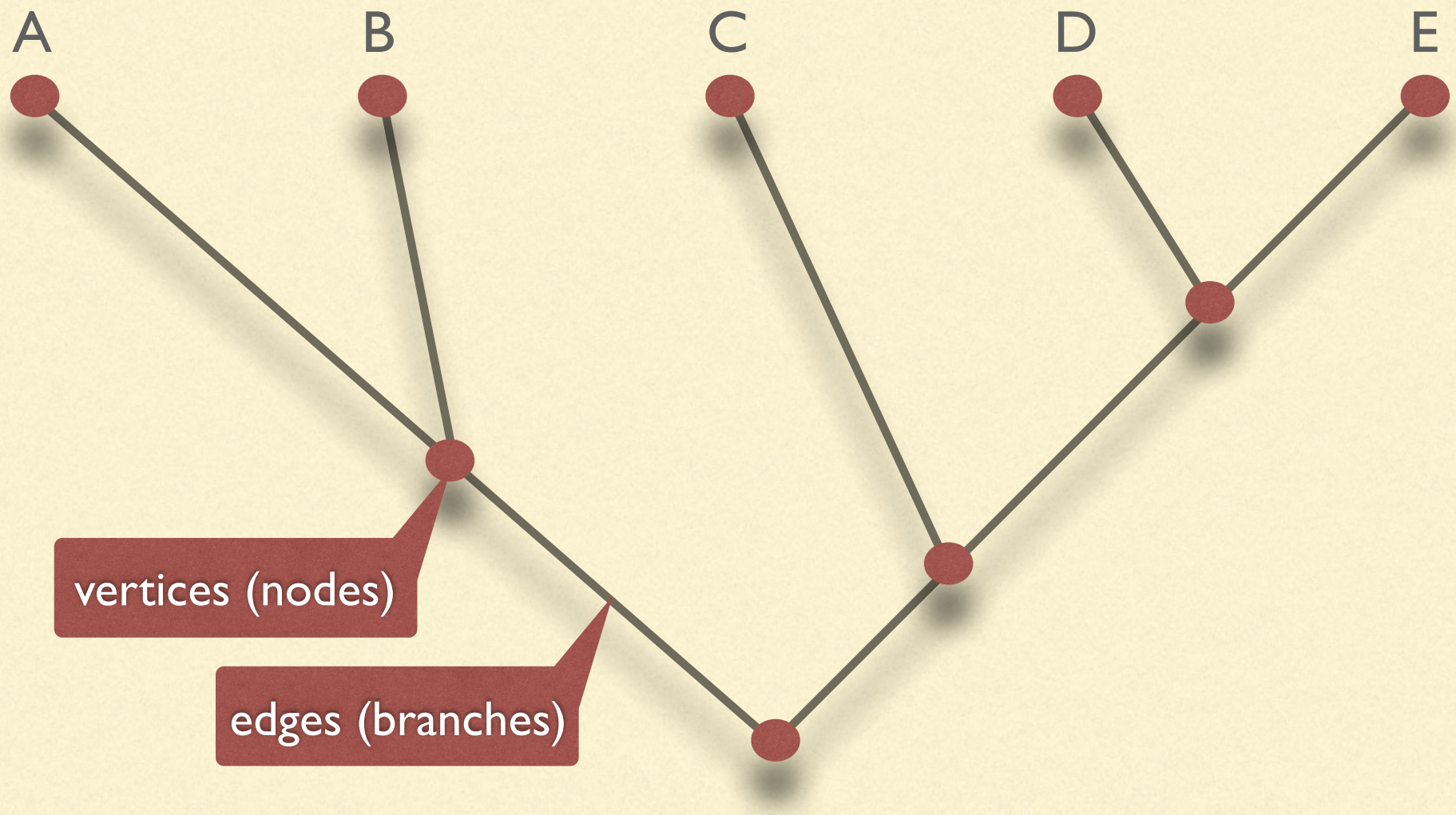


Trees

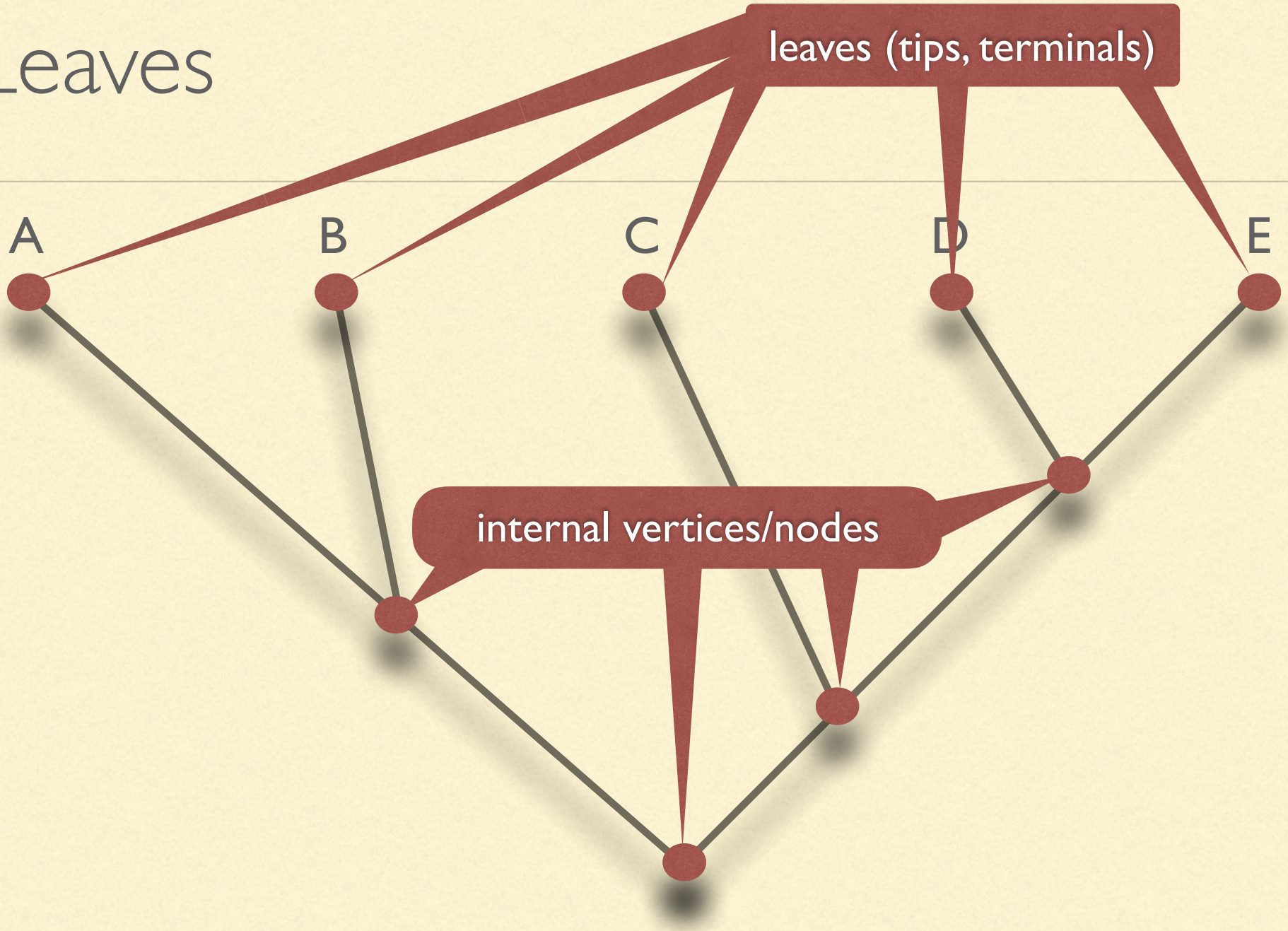
Graphs that represent phylogenies

Phylogeny: the history of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms (https://www.britannica.com/science/phylogeny)

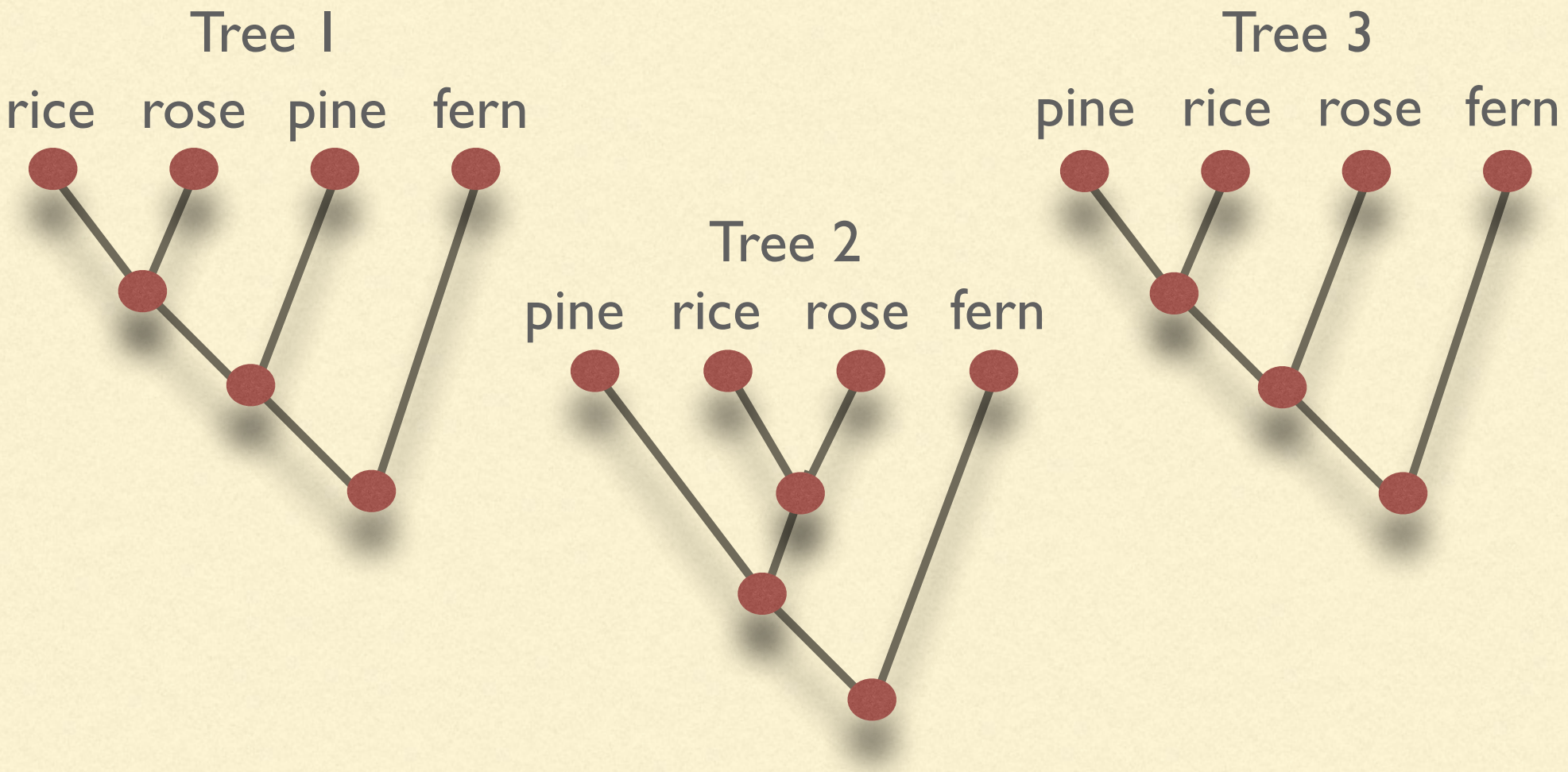
Vertices and edges



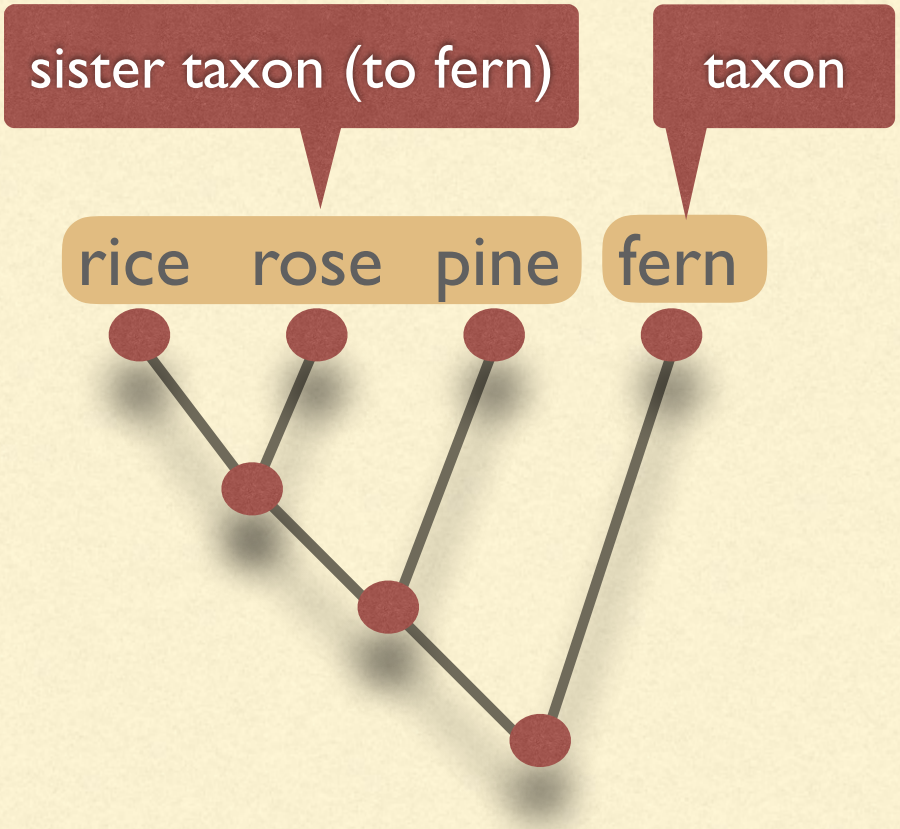
Leaves



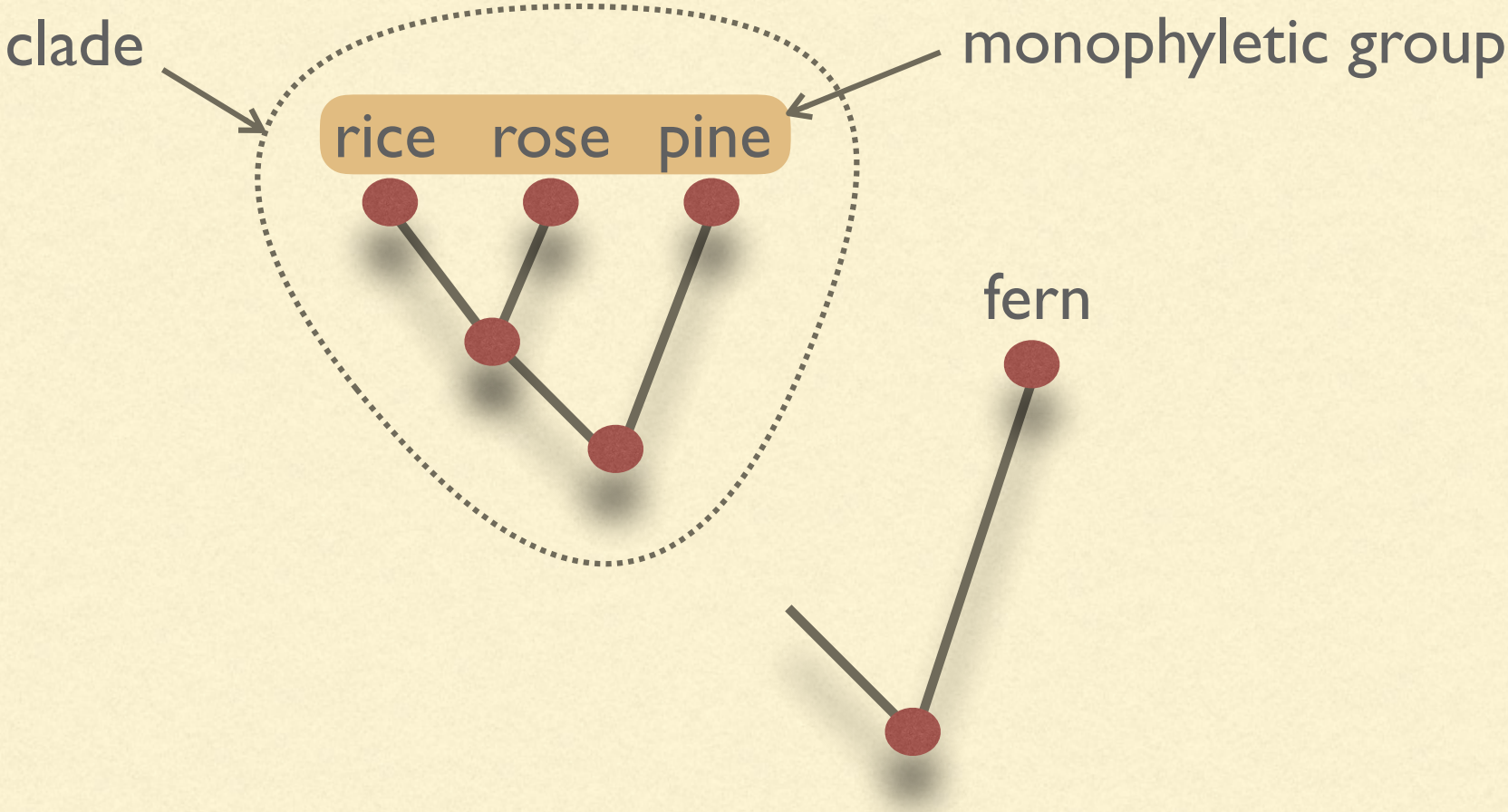
Topology



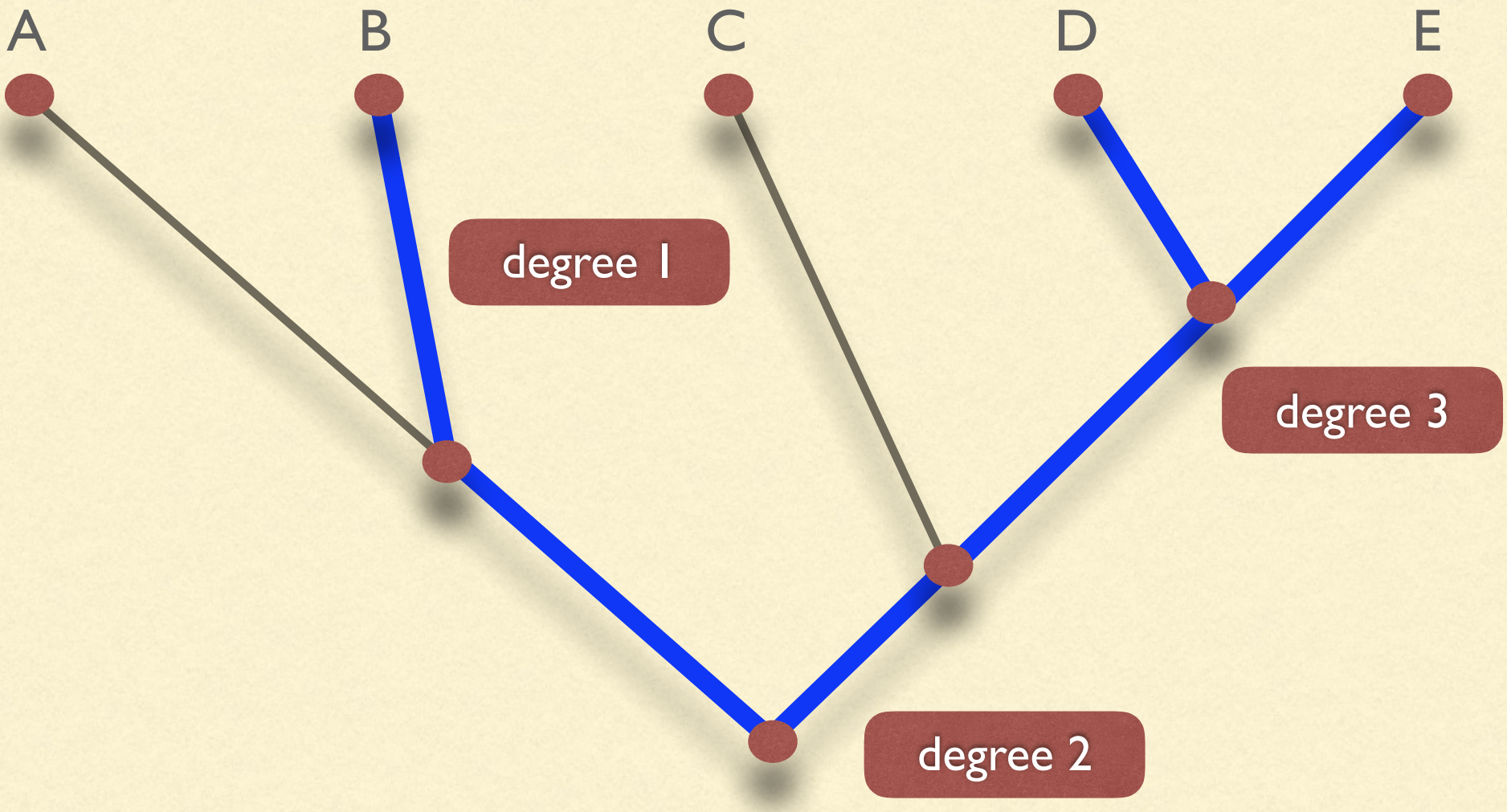
Sister taxa



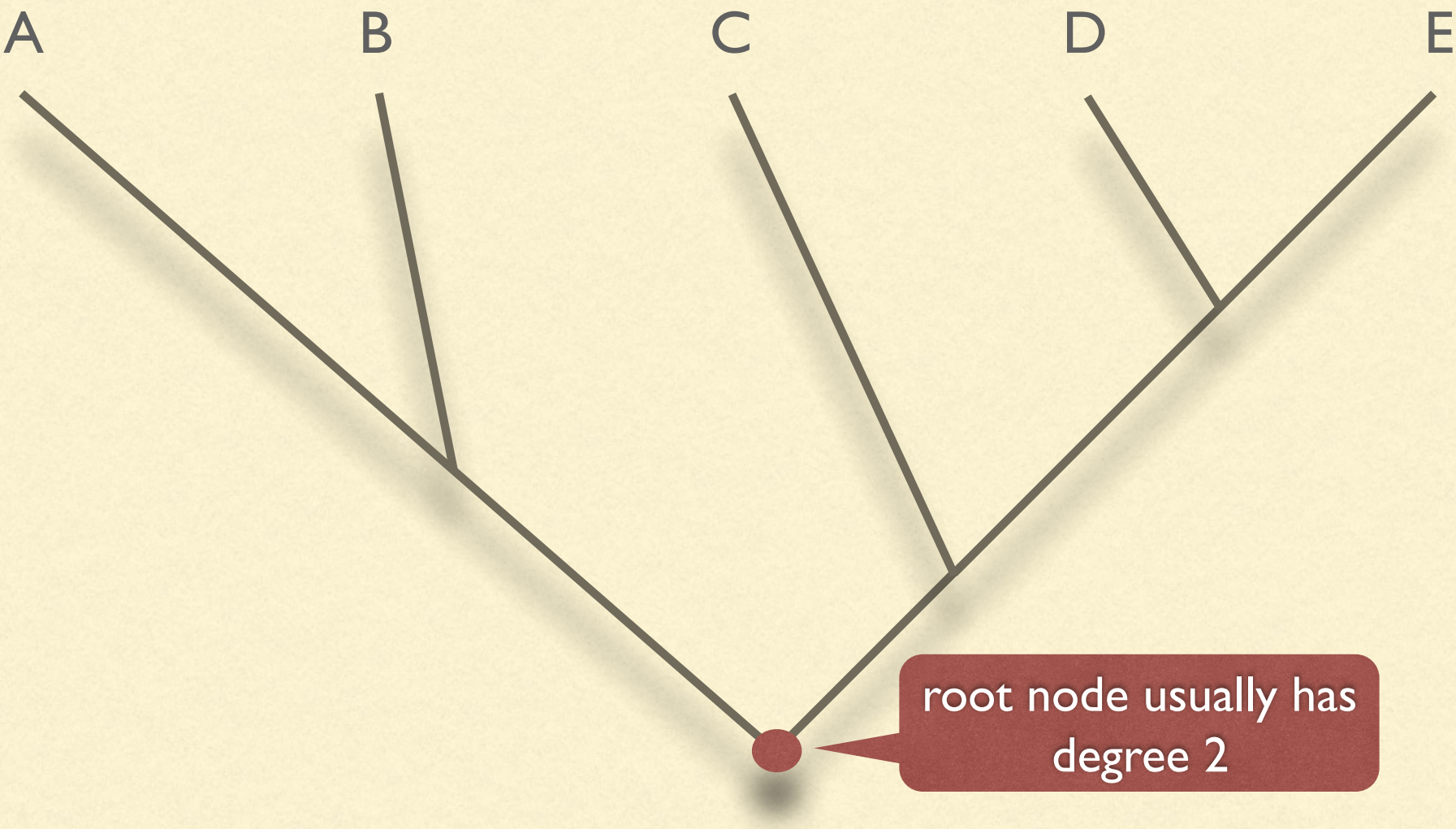
Clades



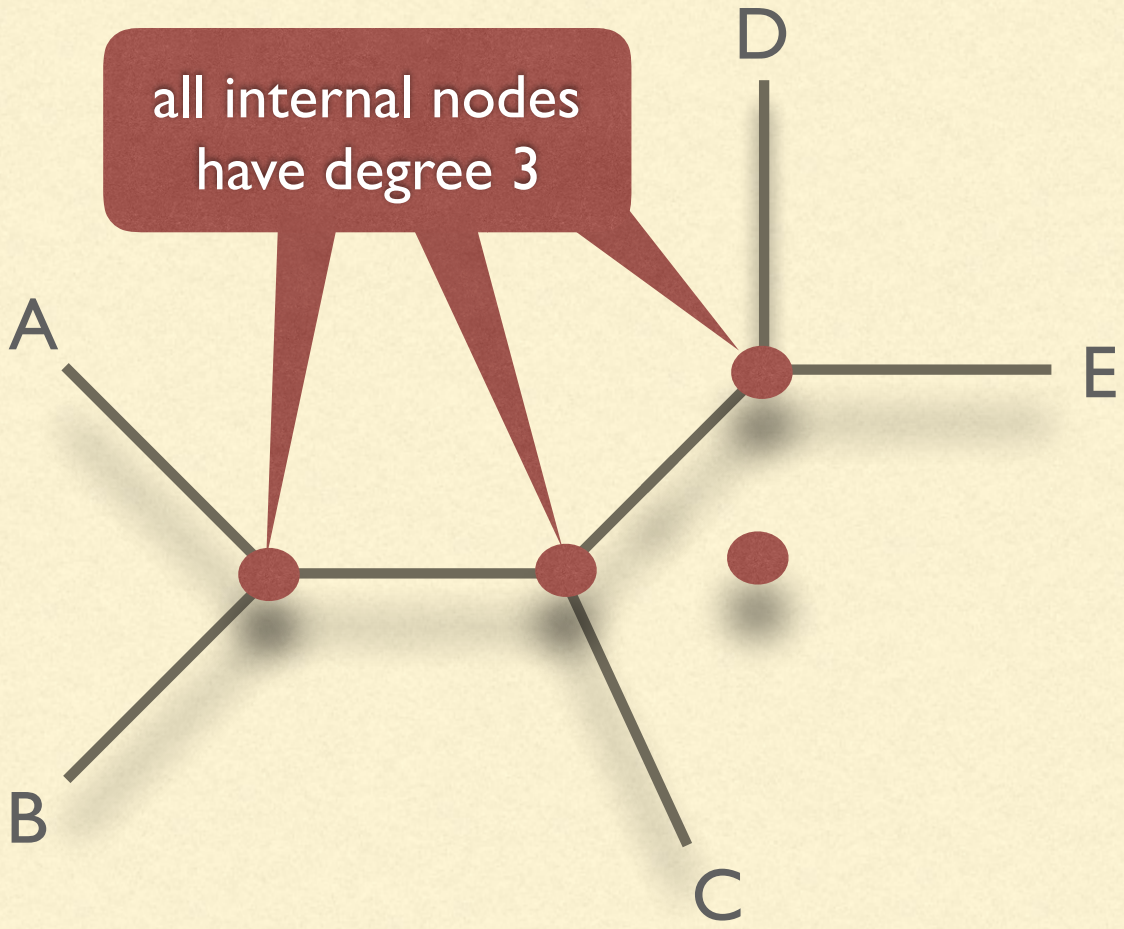
Degree of a vertex



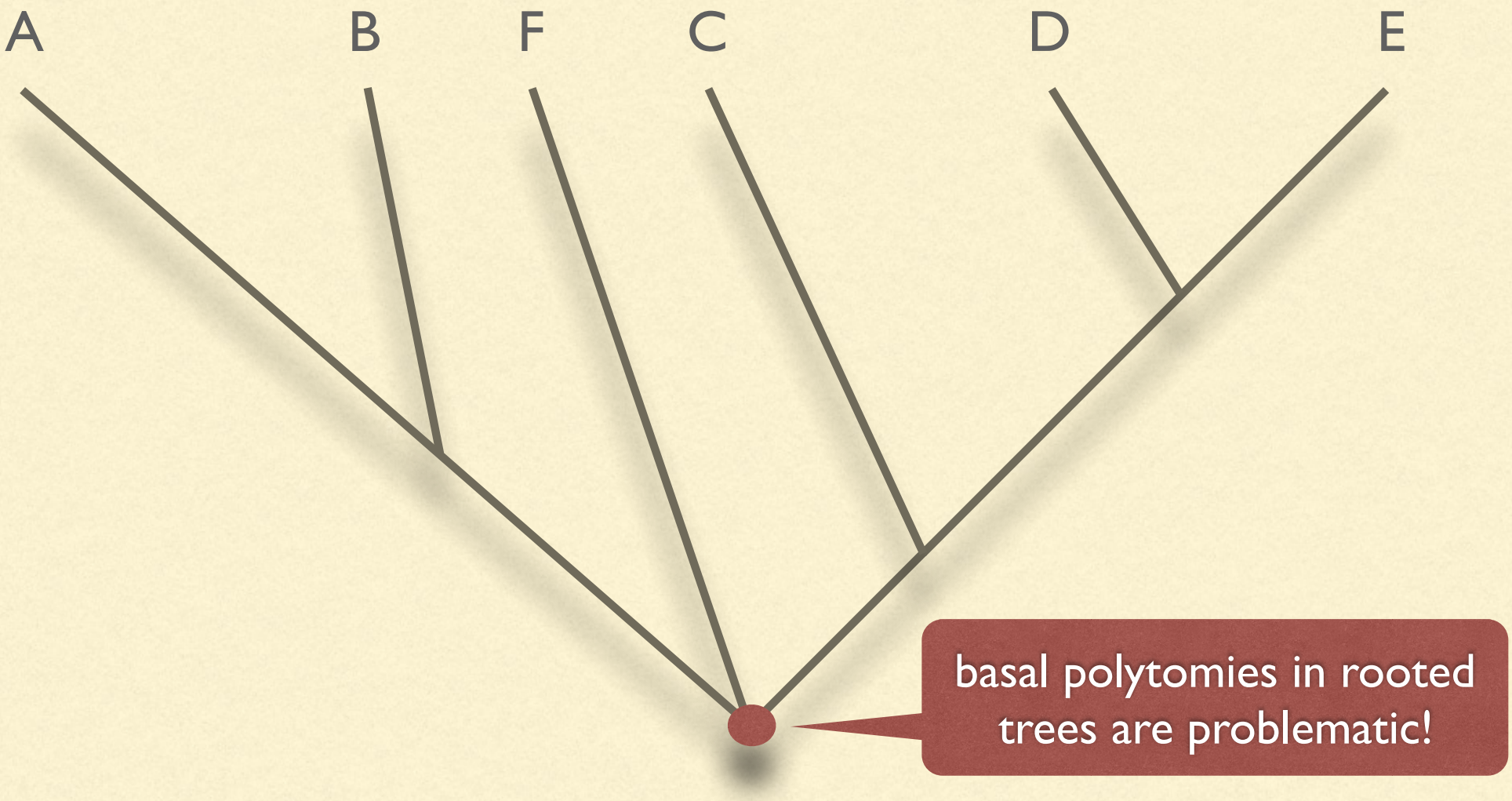
Rooted trees



Unrooted trees

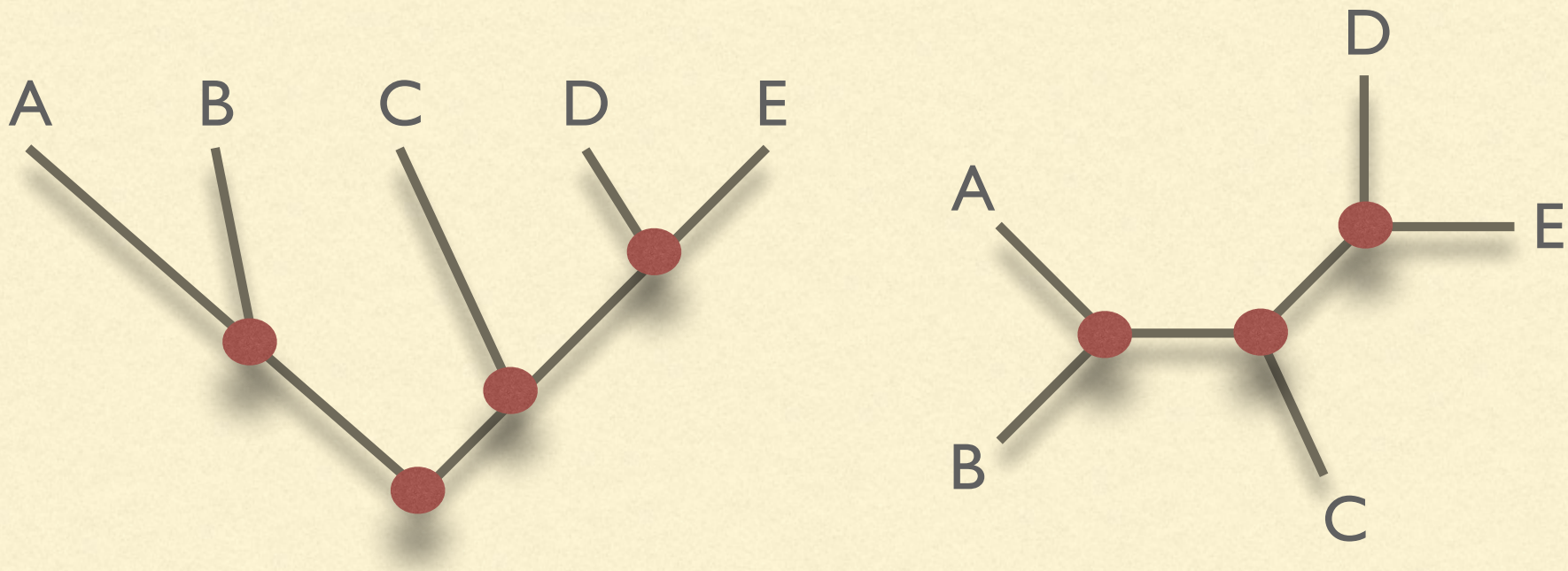


Rooted or unrooted?



Binary trees

In binary trees, all internal nodes have degree 2 or 3

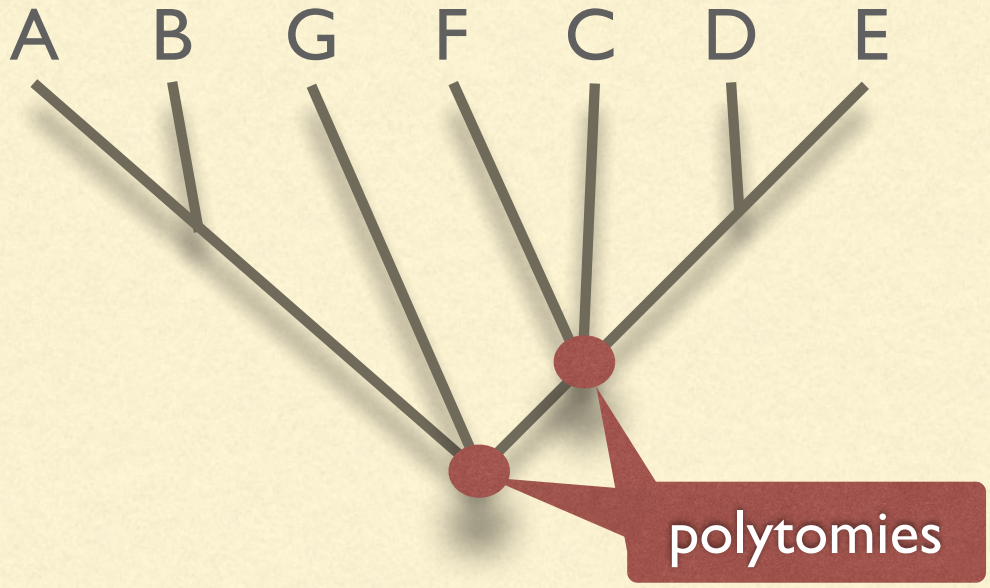


rooted binary

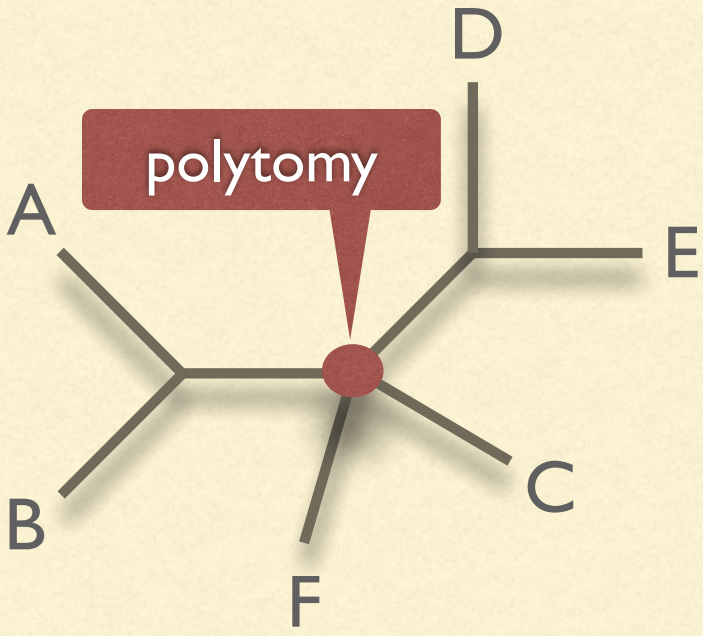
unrooted binary

Multifurcating trees

In multifurcating trees, at least one internal node has degree 4+

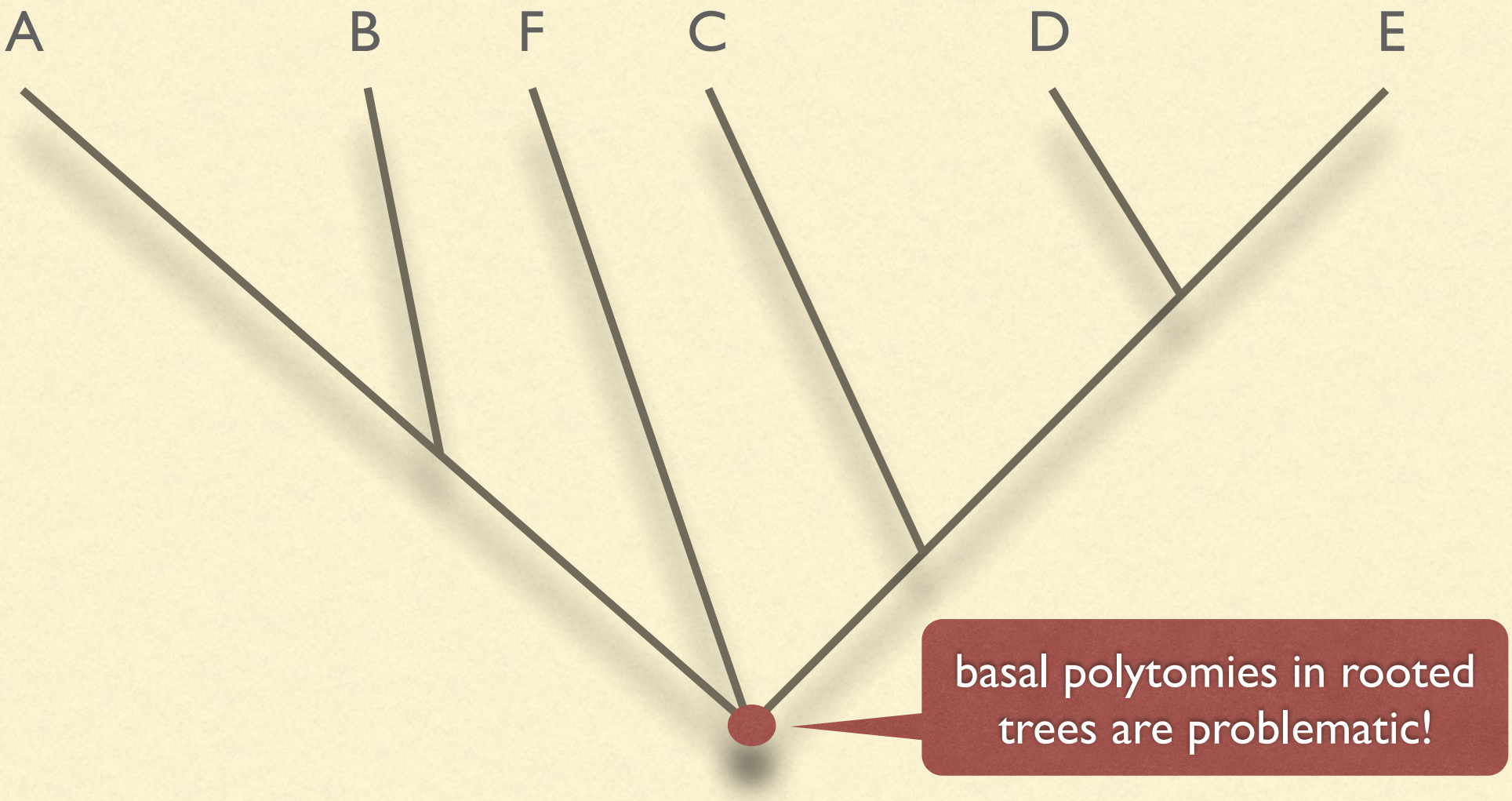


rooted multifurcating

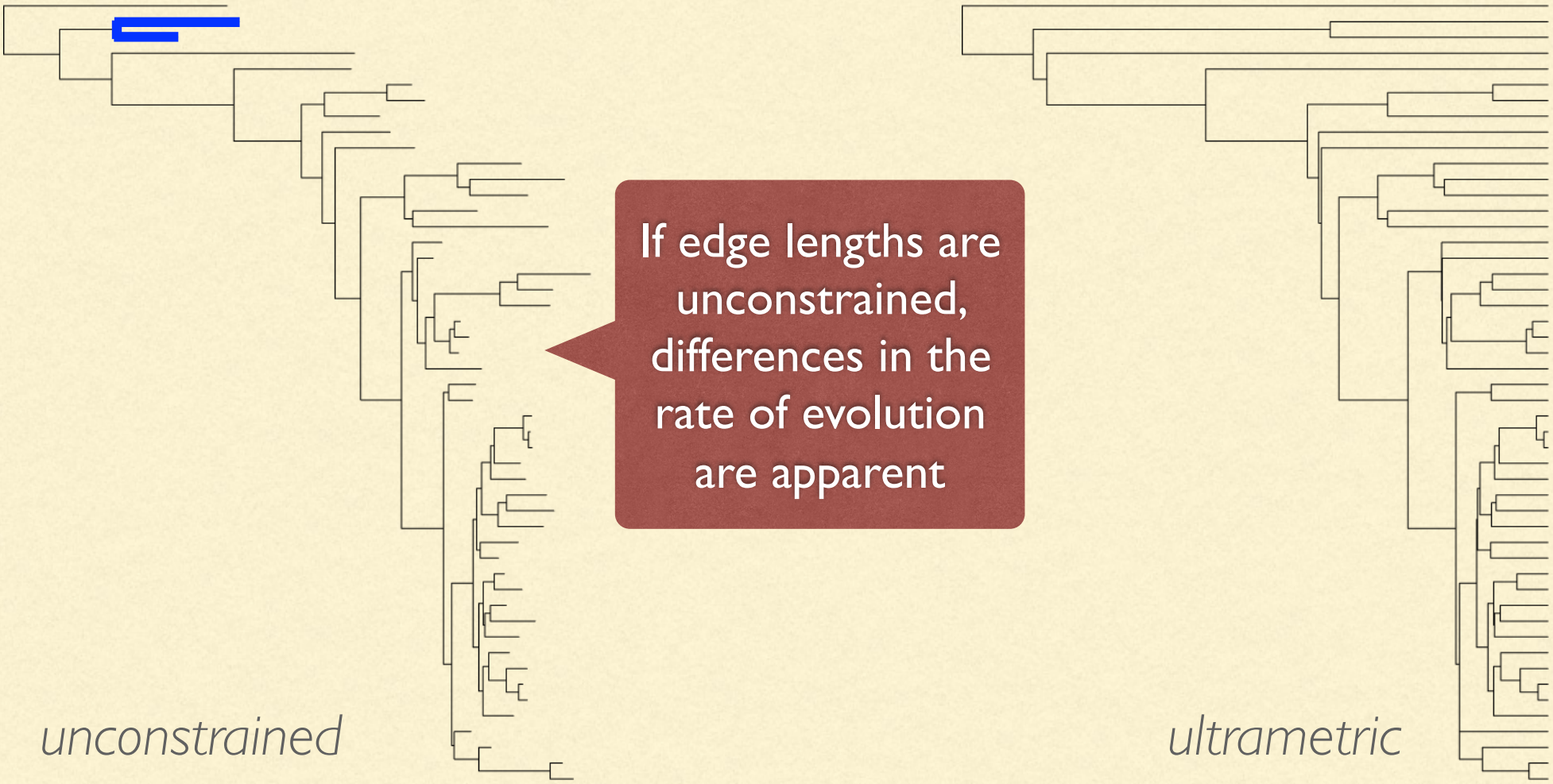


unrooted multifurcating

Rooted or unrooted?



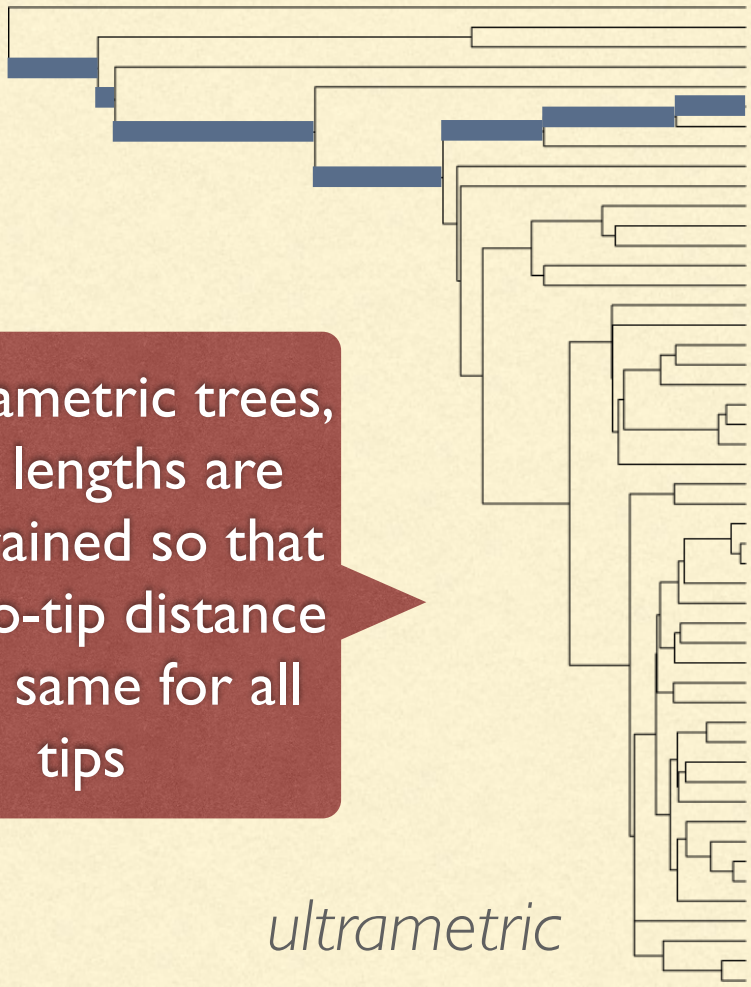
Unconstrained trees



Ultrametric trees

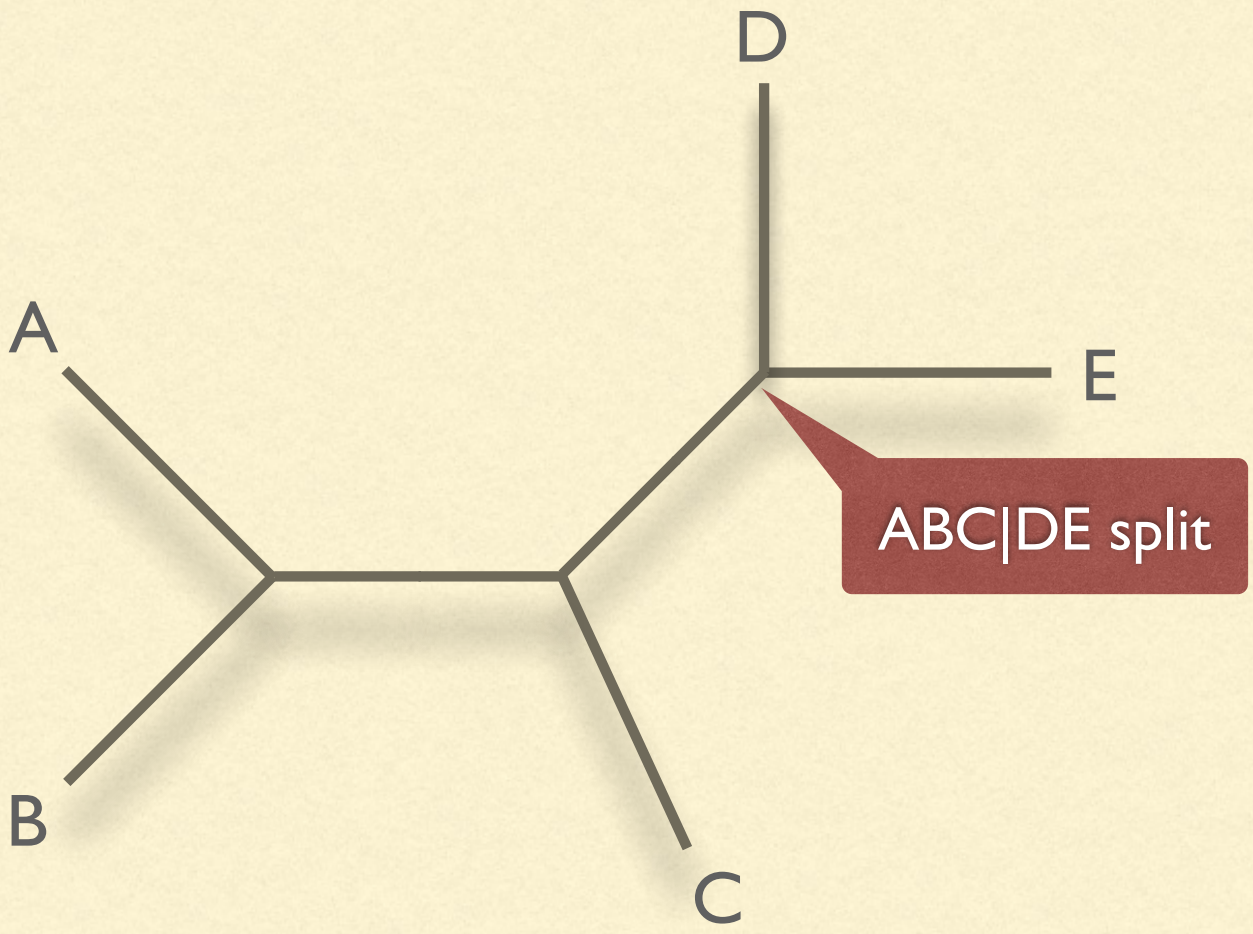
unconstrained

In ultrametric trees, edge lengths are constrained so that root-to-tip distance is the same for all tips

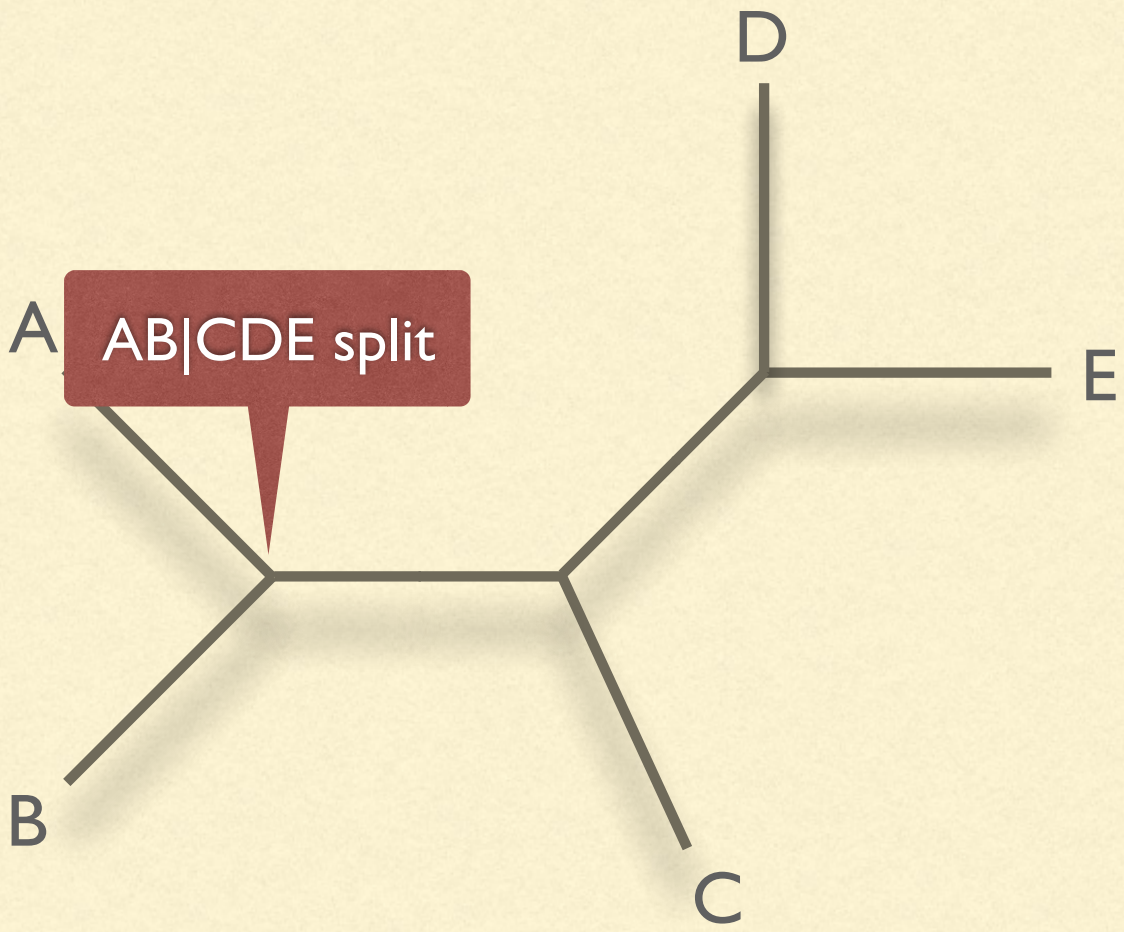


ultrametric

Splits



Splits



Splits

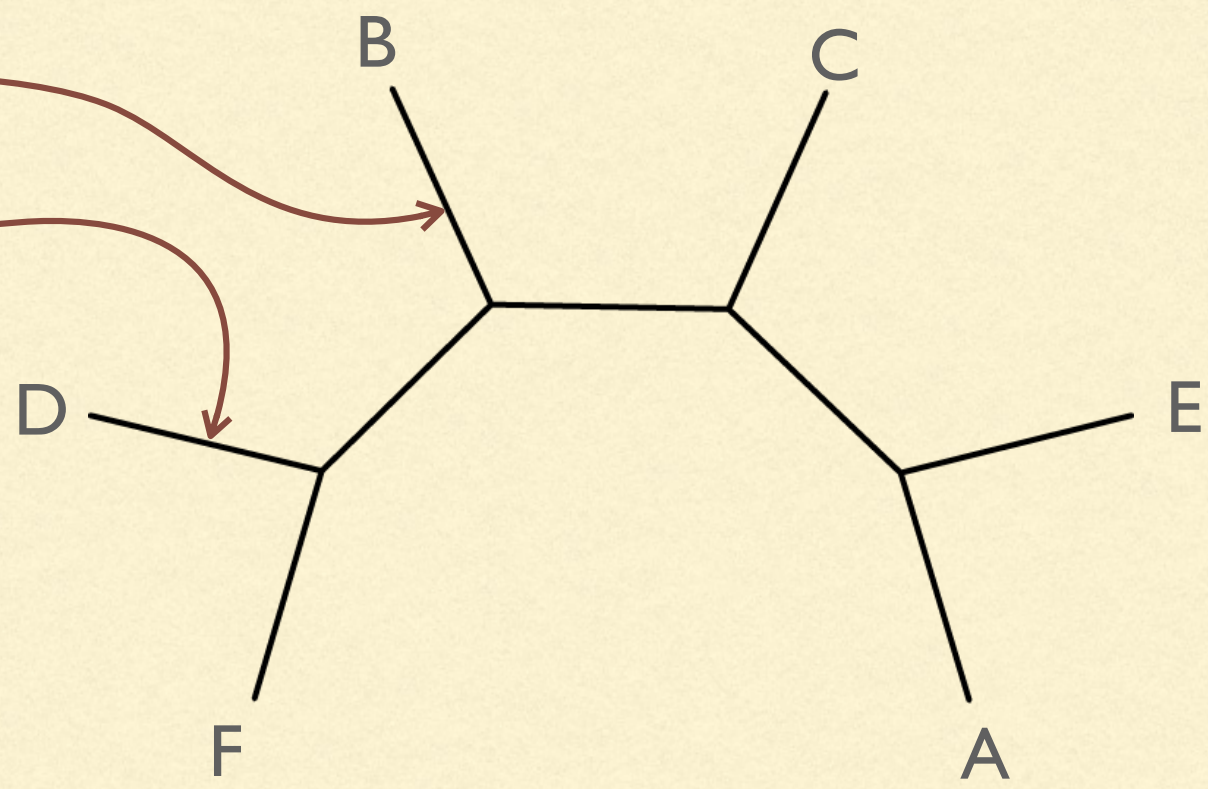
Split (split), split representation (pattern), frequency (freq.), posterior probability (prob.), mean edge length (weight), first sojourn start (s0), last sojourn end (sk), and number of sojourns (k):

split	pattern	freq	prob.	weight	s0	sk	k
1	-*****	1000	1.00000	0.09664	1	1000	1
2	----*-	1000	1.00000	0.00298	1	1000	1
3	-*----	1000	1.00000	0.01030	1	1000	1
4	---*-*	1000	1.00000	0.03016	1	1000	1
5	--*---	1000	1.00000	0.00727	1	1000	1
6	---*--	1000	1.00000	0.00203	1	1000	1
7	-----*	1000	1.00000	0.00149	1	1000	1
8	-*-*-*	969	0.96900	0.00646	2	1000	29
9	-***-*	745	0.74500	0.00347	4	1000	184
10	-*-***	162	0.16200	0.00242	2	993	140
11	--*-*-	84	0.08400	0.00229	1	967	75
12	-**-*-	14	0.01400	0.00896	1	686	14
13	--****	9	0.00900	0.00485	181	939	9
14	--**-*	8	0.00800	0.00358	152	864	8
15	---***	5	0.00500	0.00261	181	939	5
16	-**---	3	0.00300	0.00276	3	680	3
17	-*--*-	1	0.00100	0.00172	456	456	1

Splits

trivial splits are those that cut off a single tip
(not that interesting)

```
split pattern freq
    1 -*****  1000
    2 ----*-  1000
    3 -*----  1000
    4 ---*-*  1000
    5 --*---  1000
    6 ---*--  1000
    7 -----*  1000
    8 -*-*-*   969
    9 -***-*   745
   10 -*-***   162
   11 --*-*-    84
   12 -**-*-    14
   13 --****     9
   14 --**-*     8
   15 ---***     5
   16 -**---     3
   17 -*--*-     1
      ABCDEF
```

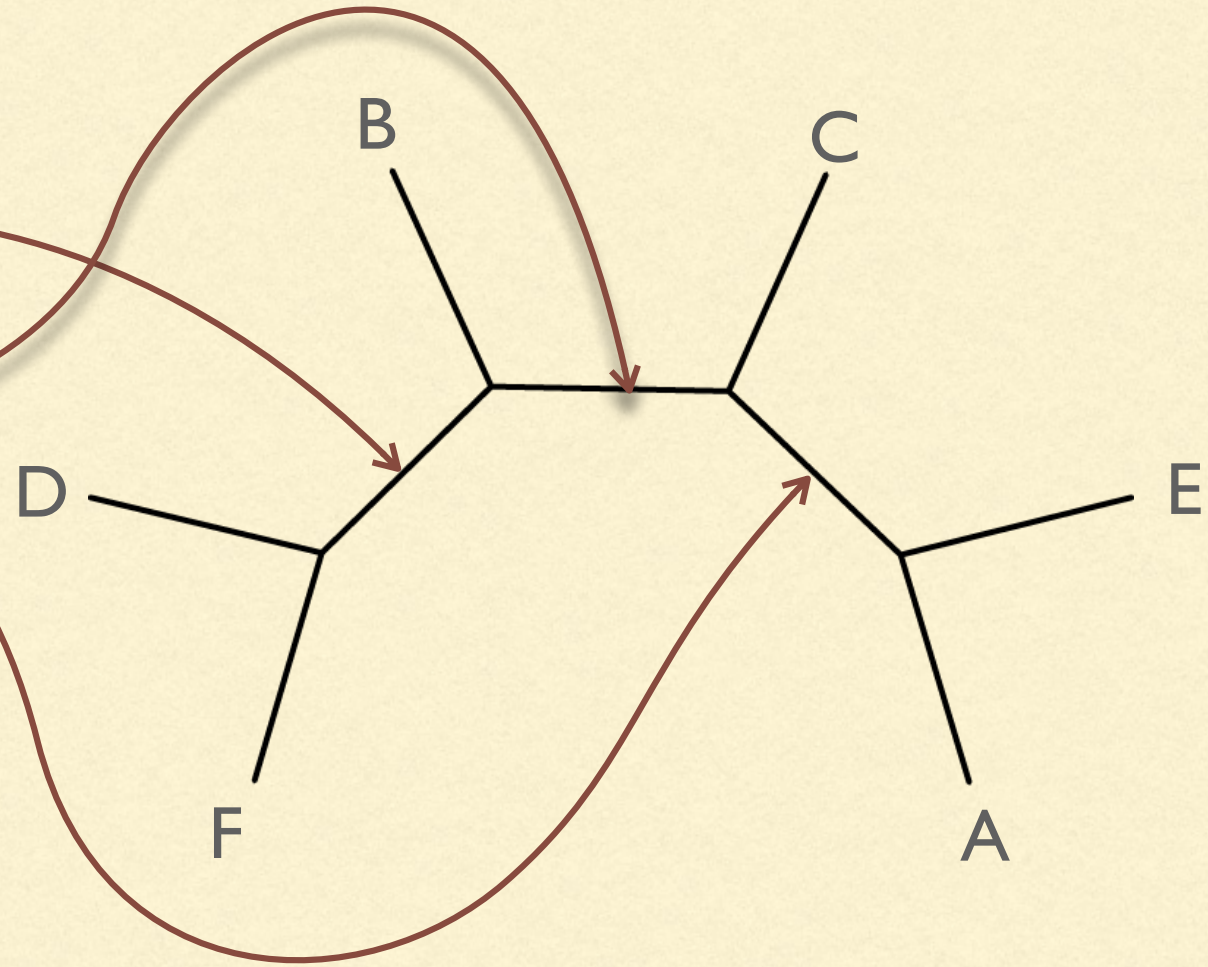


frequencies in which each split appears in trees sampled from a Bayesian MCMC analysis

Splits

majority-rule consensus tree includes all splits with frequency > 50%

split	pattern	freq
1	-*****	1000
2	----*-	1000
3	-*----	1000
4	---*-*	1000
5	--*---	1000
6	---*--	1000
7	-----*	1000
8	-*-*-*	969
9	-***-*	745
10	-*-***	162
11	--*-*-	84
12	-**-*-	14
13	--****	9
14	--**-*	8
15	---***	5
16	-**---	3
17	-*--*-	1
	ABCDEF	



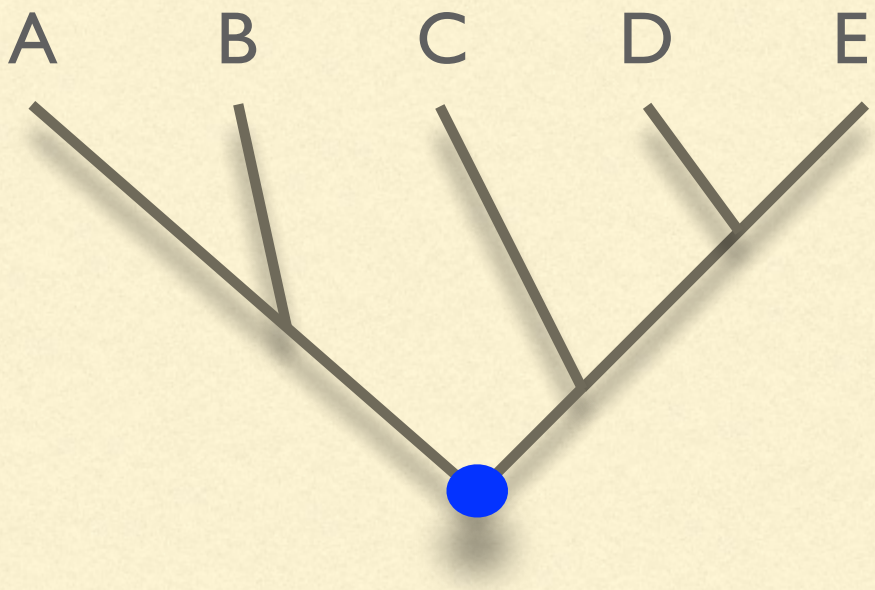
Newick trees


```
#NEXUS

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

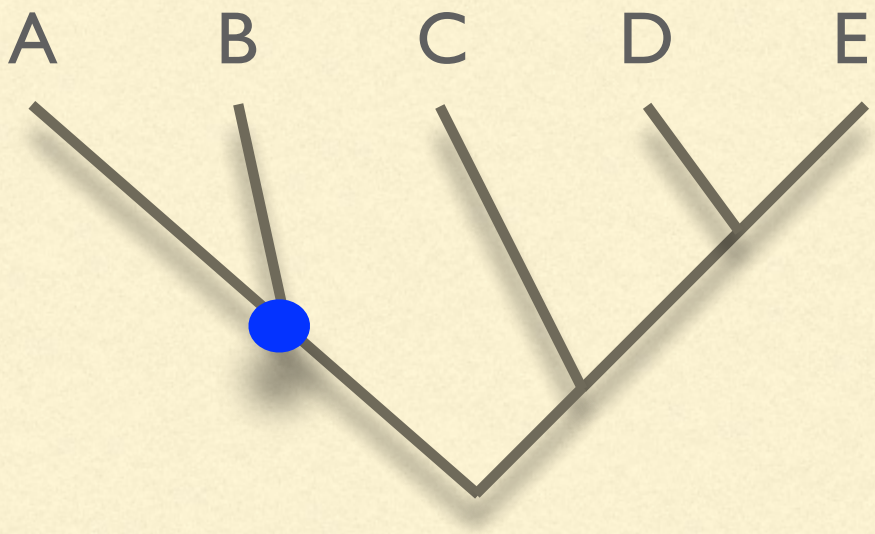
https://en.wikipedia.org/wiki/Newick_format

Newick tree descriptions



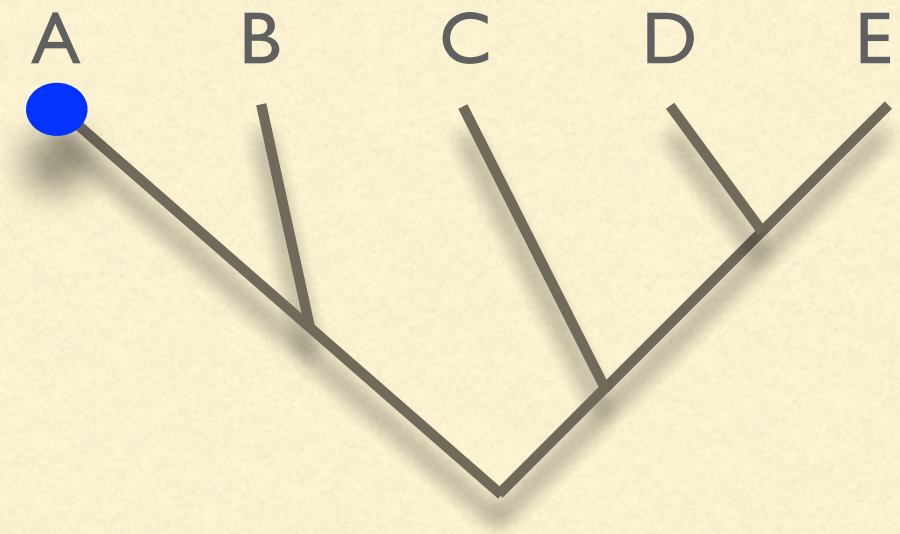
((A,B),(C,(D,E)))

Newick tree descriptions



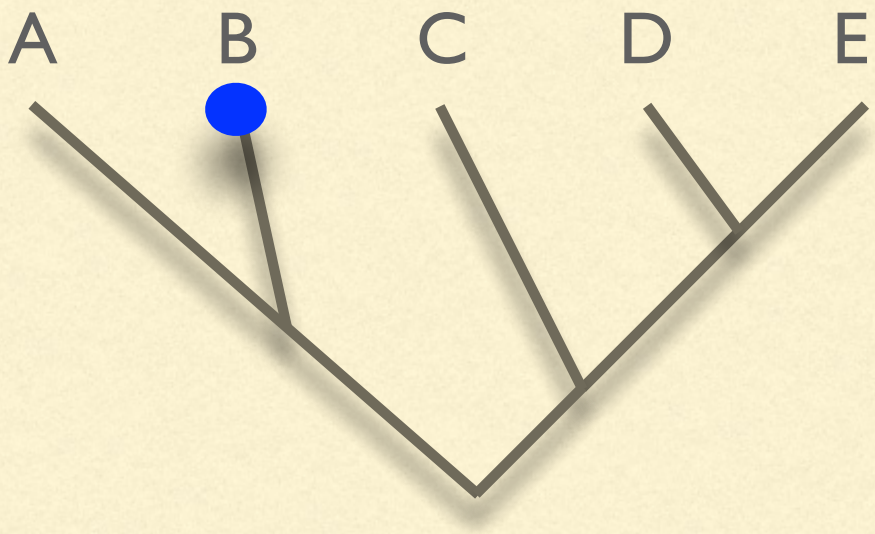
((A,B),(C,(D,E)))

Newick tree descriptions



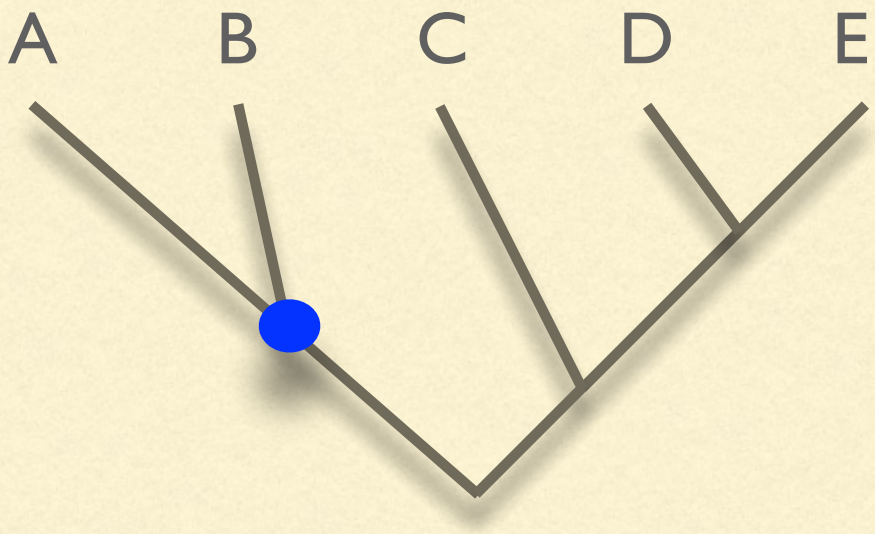
((A,B),(C,(D,E)))

Newick tree descriptions



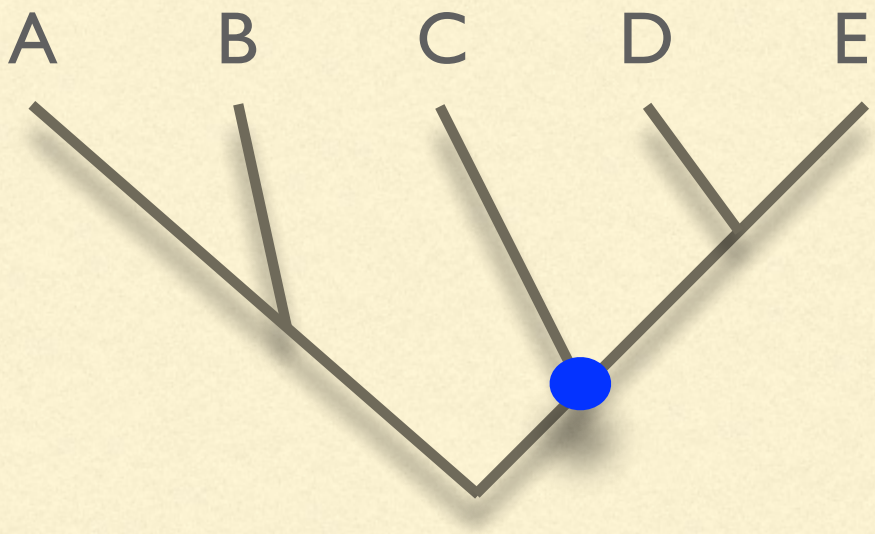
((A,B),(C,(D,E)))

Newick tree descriptions



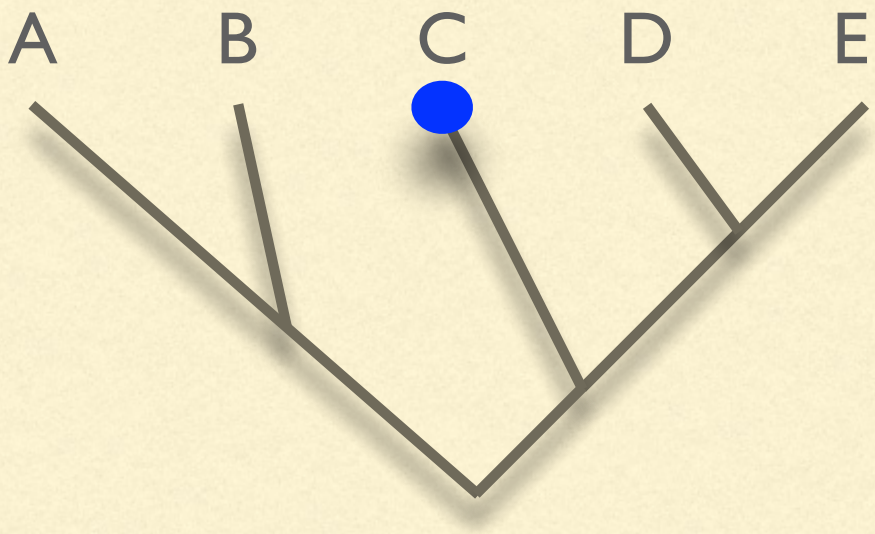
((A,B),(C,(D,E)))

Newick tree descriptions



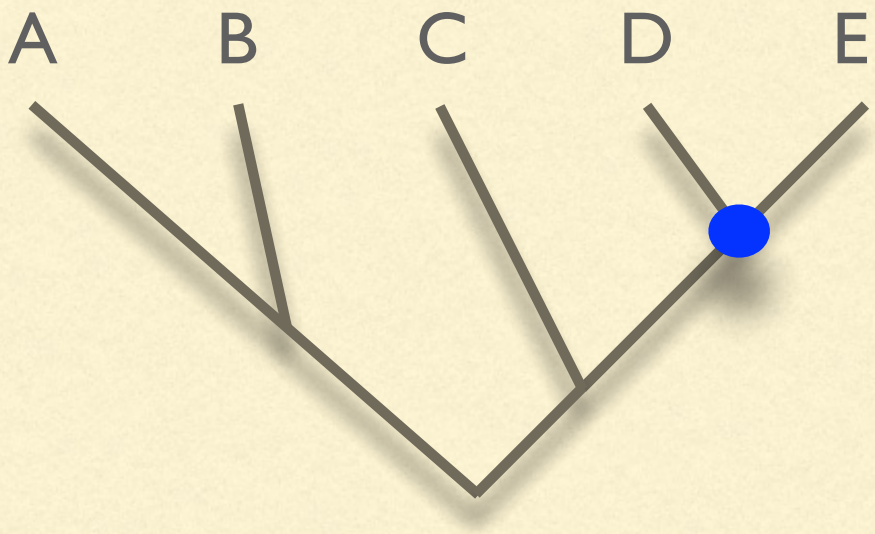
((A,B),(C,(D,E)))

Newick tree descriptions



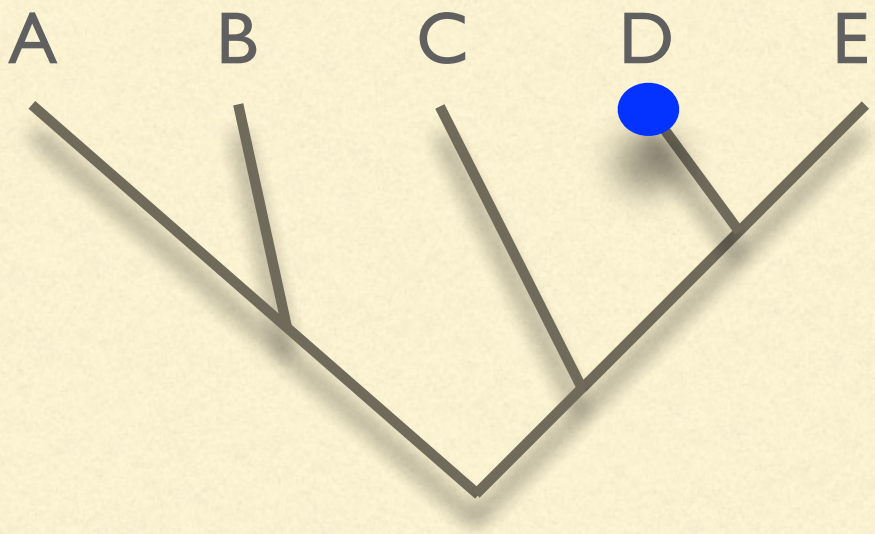
((A,B),(C,(D,E)))

Newick tree descriptions



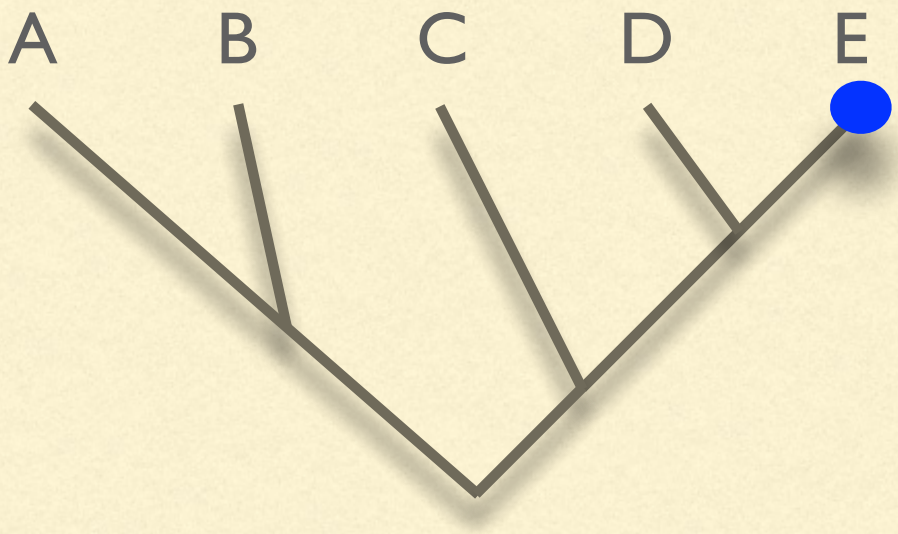
((A,B),(C,(D,E)))

Newick tree descriptions



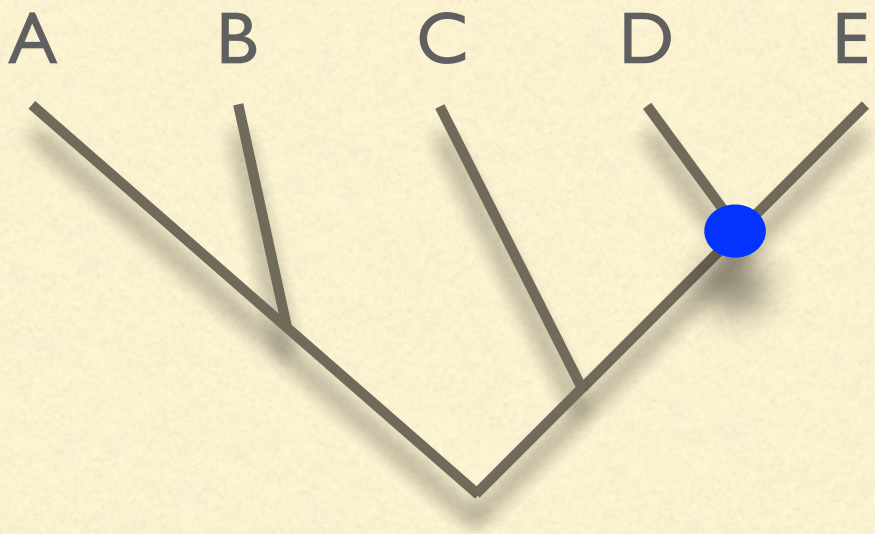
((A,B),(C,(D,E)))

Newick tree descriptions



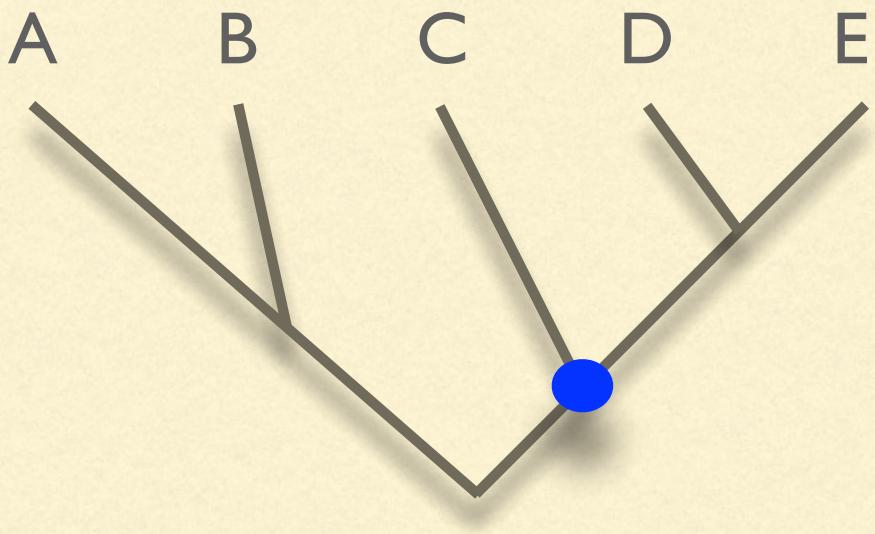
((A,B),(C,(D,E)))

Newick tree descriptions



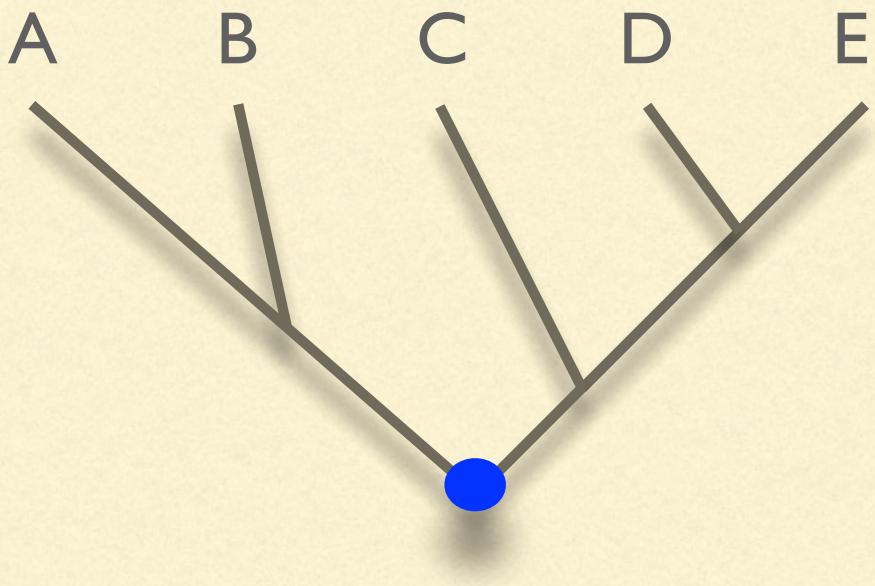
((A,B),(C,(D,E)))

Newick tree descriptions



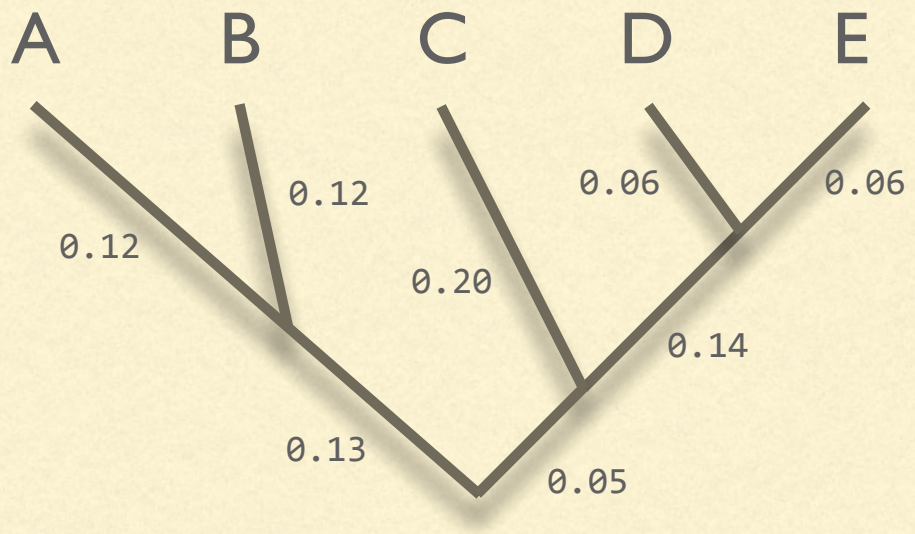
((A,B),(C,(D,E)))

Newick tree descriptions



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

Newick tree descriptions




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

edge lengths follow colon after node name (if present)

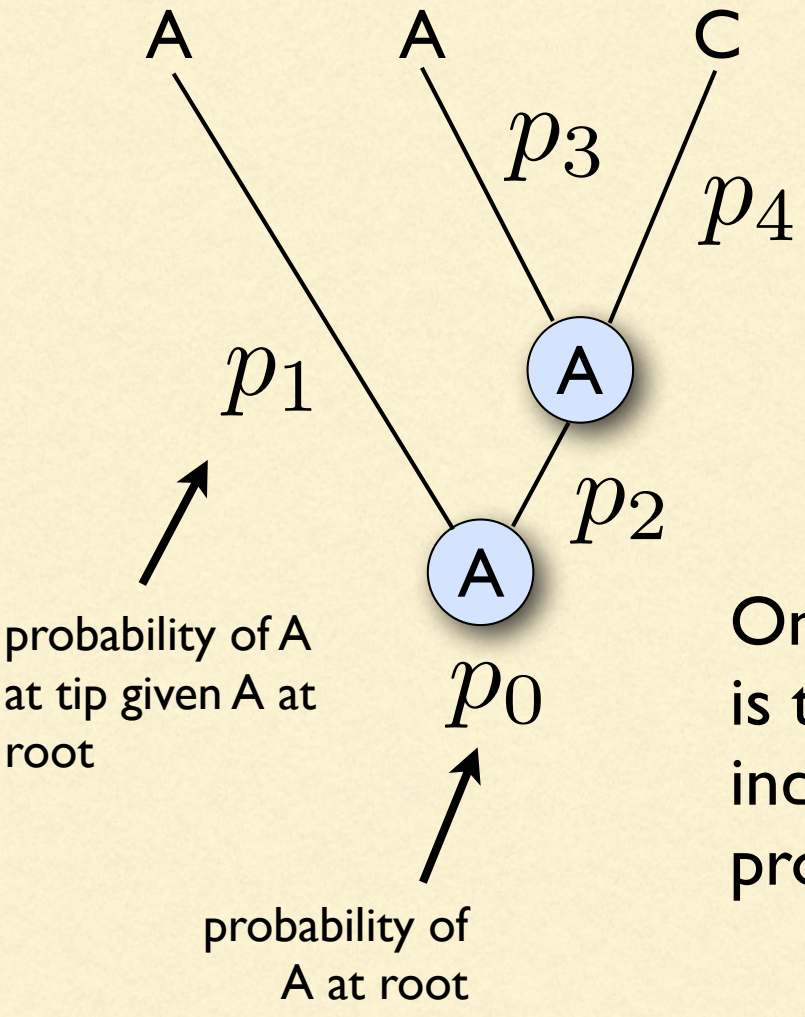
Probability

Probabilities: the AND rule

Rolling 2 dice, what is the probability of seeing (simultaneously) a 1 on the first die and a 6 on the second die?

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

AND rule in phylogenetics



we have observed these states

pretend (for now) we also know these ancestral states

$$\Pr(A, A, C, A, A) = p_0\ p_1\ p_2\ p_3\ p_4$$

One use of the AND rule in phylogenetics is to combine probabilities associated with individual branches to produce the overall probability of the data for one site.

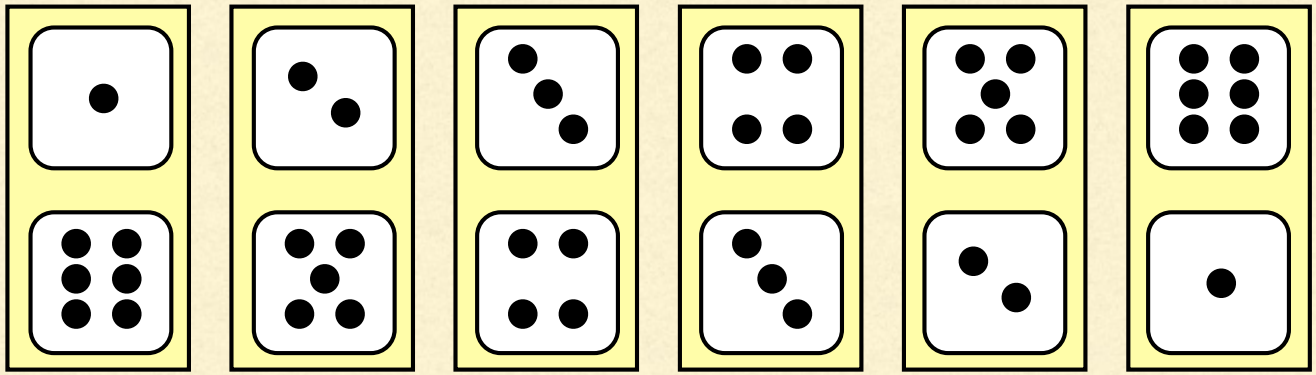
Probabilities: the OR rule

Rolling 1 die, what is the probability of seeing either a 1 or a 6?

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

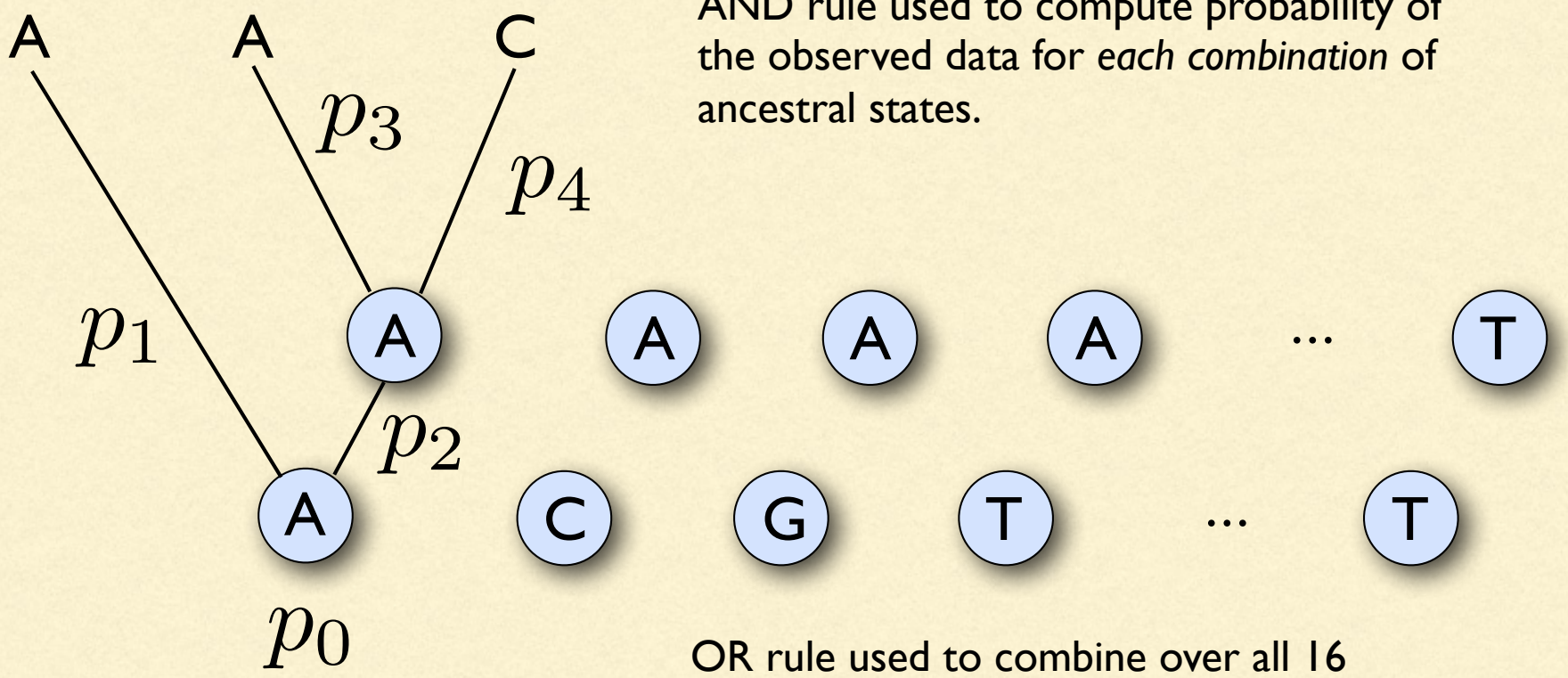
Combining AND and OR

What is the probability that the sum of two dice is 7?



(1/36) + (1/36) + (1/36) + (1/36) + (1/36) + (1/36) = 1/6

Using both AND and OR in phylogenetics



AND rule used to compute probability of the observed data for *each combination* of ancestral states.

OR rule used to combine over all 16 combinations of ancestral states.

Pr(**A**,**A**,**C**) = Pr(**A**,**A**,**C**,A,A)+ Pr(**A**,**A**,**C**,A,C)+...+ Pr(**A**,**A**,**C**,T,T)

Independence

$$\Pr(A, B) = \Pr(A) \Pr(B)$$

Probability of flipping a coin twice and getting heads both times:

Pr(H,H) = Pr(H) Pr(H)

Non-independence

$$\Pr(A, B) = \Pr(A)\Pr(B|A)$$

joint probability of A and B

conditional probability of B given A

Pr(walk to work|sunny) = 0.99
Pr(walk to work|raining) = 0.50

Independence

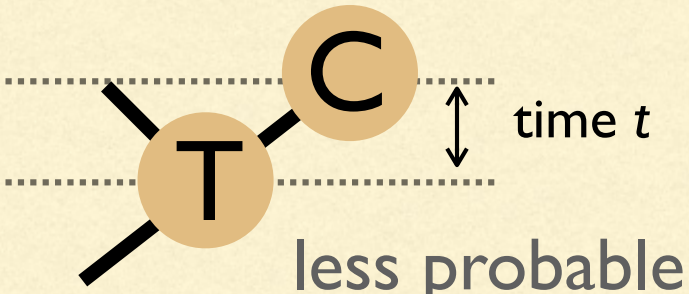
$$\Pr(A, B) = \Pr(A)\Pr(B|A)$$

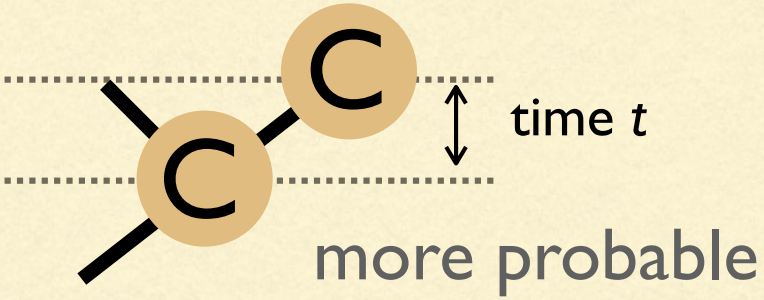
A and B are independent events if this conditional probability has the same value for every state of A

Non-independence in phylogenies

Normally, for a given rate of substitution and time, the probability of the end state is *dependent* on the starting state

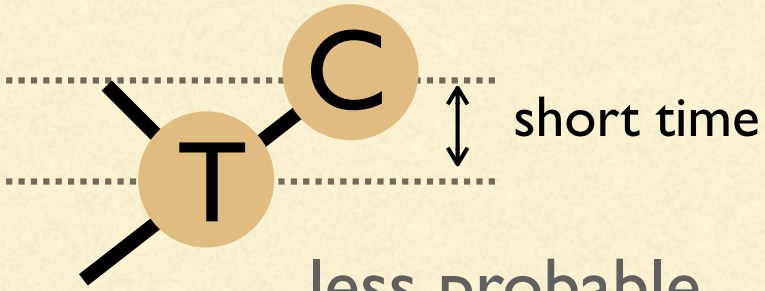
$$Pr(C|C) > Pr(C|T)$$



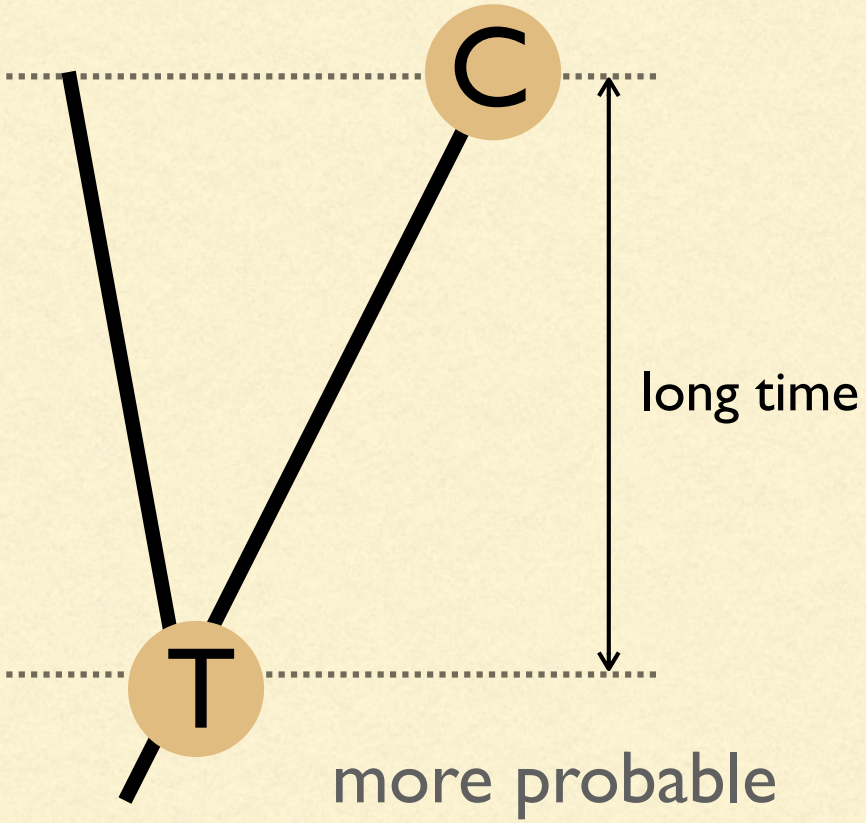


Non-independence in phylogenies

For a given rate of substitution and starting state, the probability of the end state is *dependent* on time



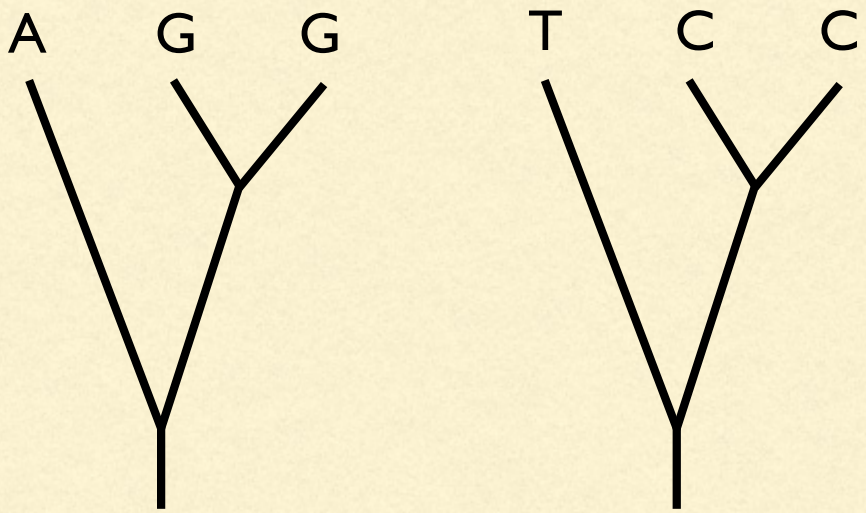
less probable



more probable

Conditional Independence

$$\Pr(A,B|C) = \Pr(A|C)\,\Pr(B|C)$$

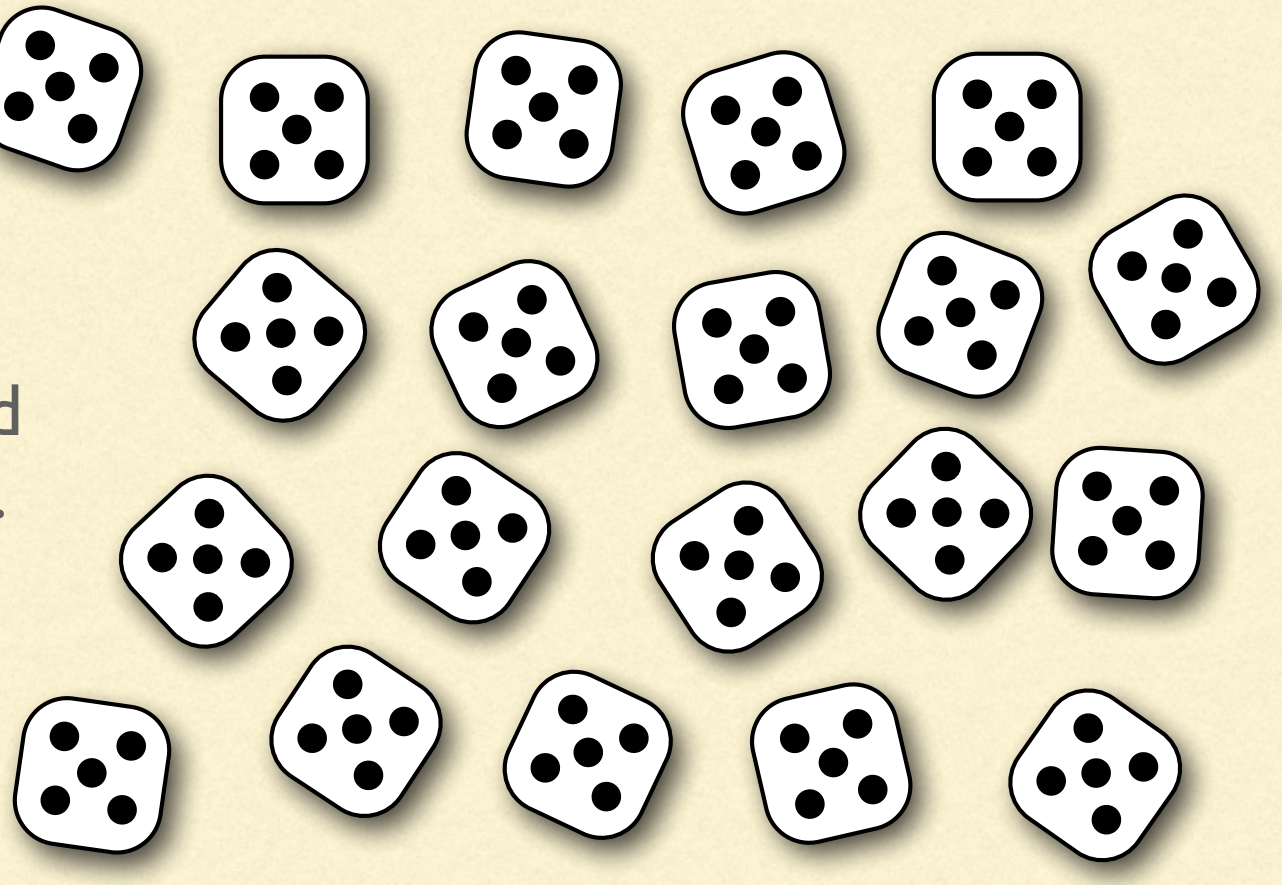


$$\Pr(AGG,TCC|\text{tree}) = \Pr(AGG|\text{tree})\,\Pr(TCC|\text{tree})$$

Likelihood

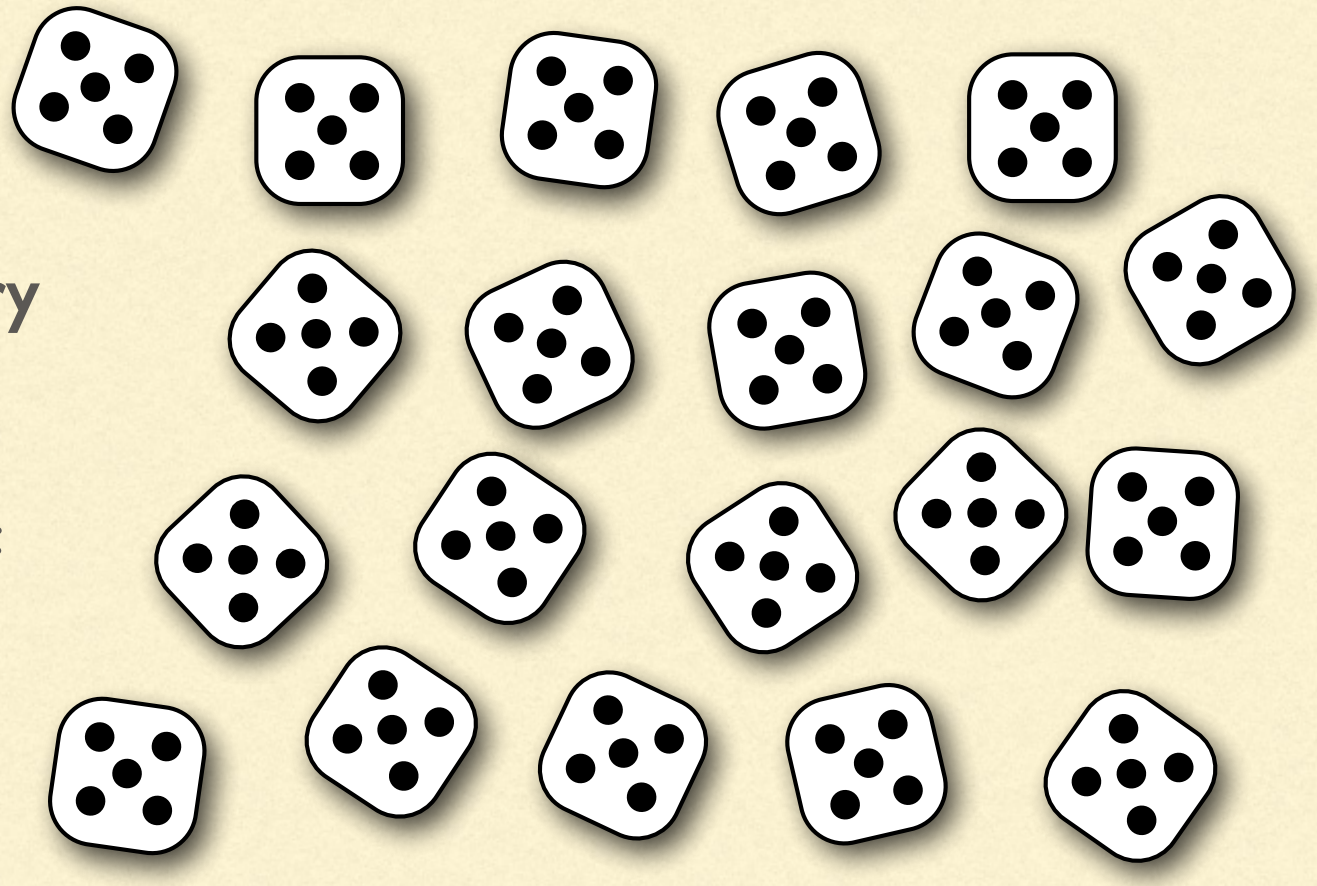
The Likelihood Criterion

Suppose I threw 20 dice on the table and this was the result...



The Fair Dice model

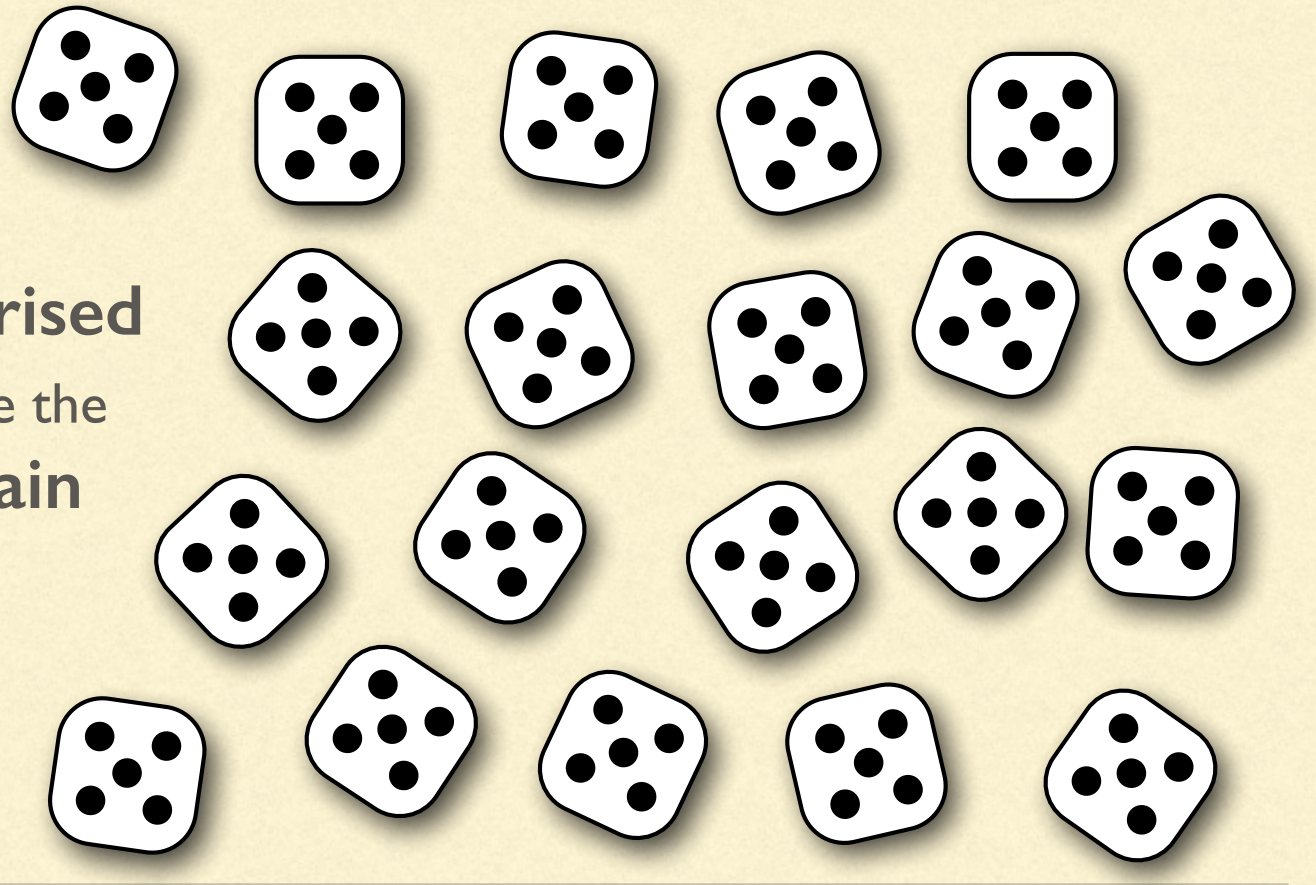
You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!



The Trick Dice model

(assumes dice each have 5 on every side)

You should **not be surprised at all** at this result because the observed outcome is **certain** under this model.



What's changed? (the model)

The **winning model** makes us **least surprised** at the data we've observed

Model	Likelihood	Surprise level
Fair Dice	$\frac{1}{3{,}656{,}158{,}440{,}062{,}976}$	Very, *very*, ***very*** surprised
Trick Dice	1	Not surprised at all

Why do we need the term ***likelihood***?

Outcome	Fair coin model	Two-heads model
H	0.5	1
T	0.5	0
	1	1

Probabilities of data outcomes given one particular model sum to 1.0

Likelihoods of models given one particular data set are not expected to sum to 1.0

Probability of the data
given the model

Likelihood of the model
given the data

Likelihood of a single vertex

First 32 nucleotides of the ψη-globin gene of gorilla:

- **GAAGTCCTTGAGAAATAAACTGCACACACTGG**

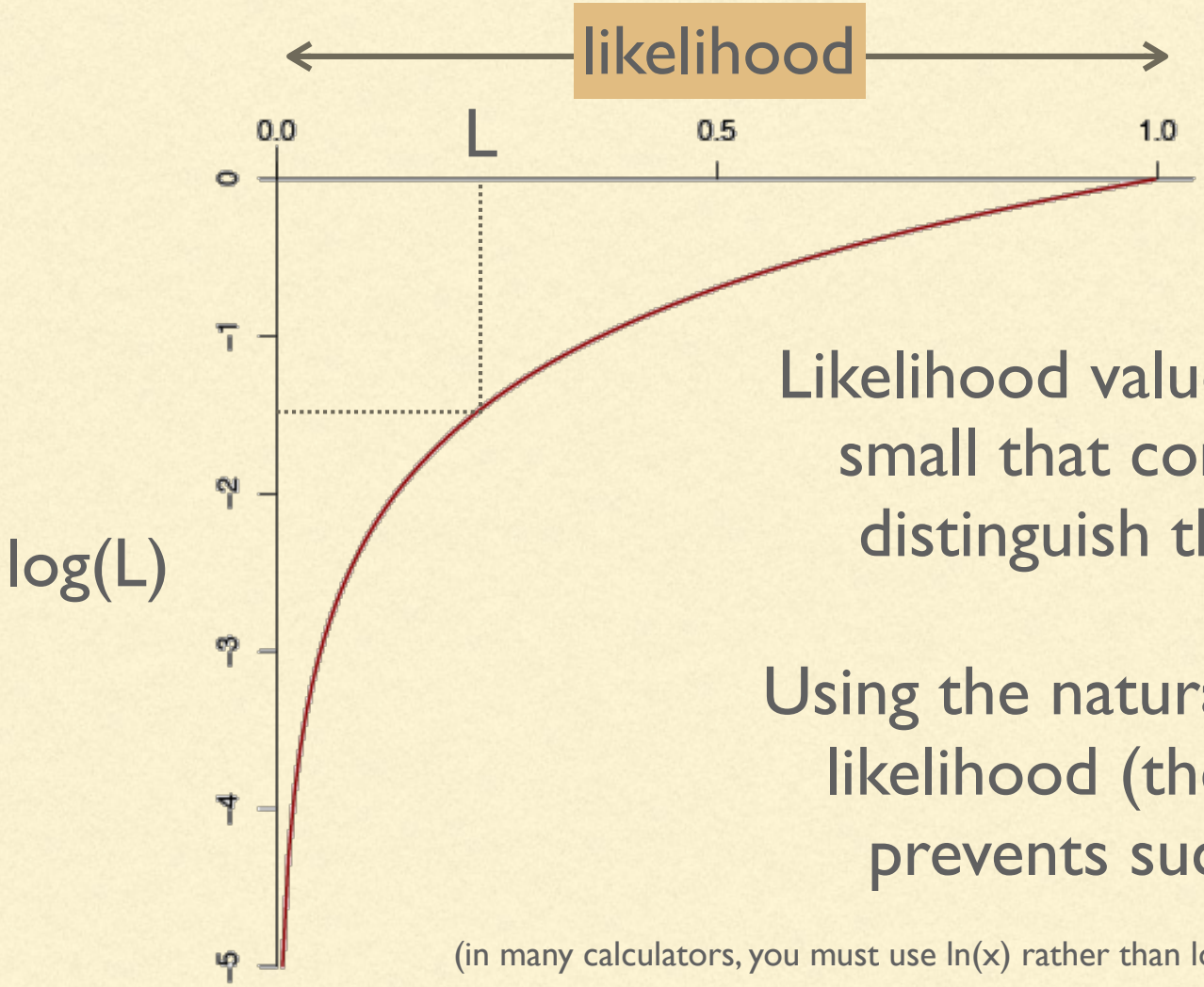
$$L = \Pr(G)\ \Pr(A)\ \Pr(A)\ \Pr(G)\ \Pr(T)\ \cdots\ \Pr(G)$$

$$L = \pi_G\ \pi_A\ \pi_A\ \pi_G\ \pi_T\ \cdots\ \pi_G$$

$$L = \pi_A^{12}\ \pi_C^{7}\ \pi_G^{7}\ \pi_T^{6}$$

$$\log L = 12\log(\pi_A) + 7\log(\pi_C) + 7\log(\pi_G) + 6\log(\pi_T)$$

Natural logarithm



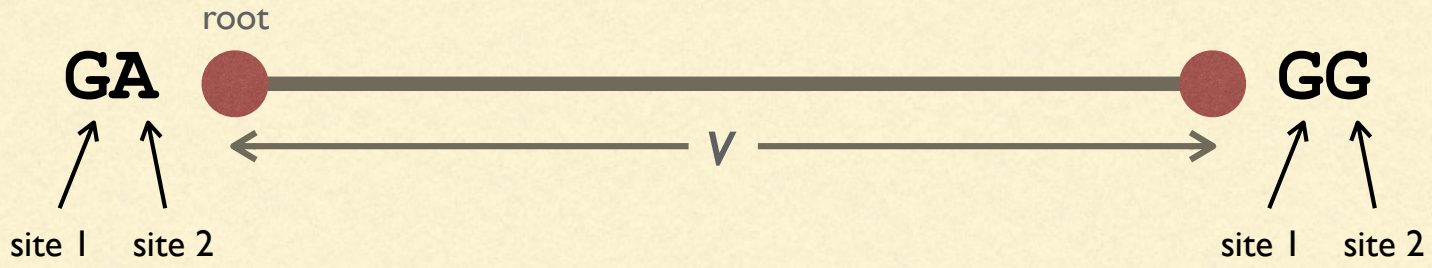
Likelihood values can become so small that computers cannot distinguish them from zero.

Using the natural logarithm of the likelihood (the log-likelihood) prevents such "underflow"

(in many calculators, you must use ln(x) rather than log(x) to take the natural log of the value x)

Likelihood of a single-edge tree

Two nodes have sequence data (but only for two sites)



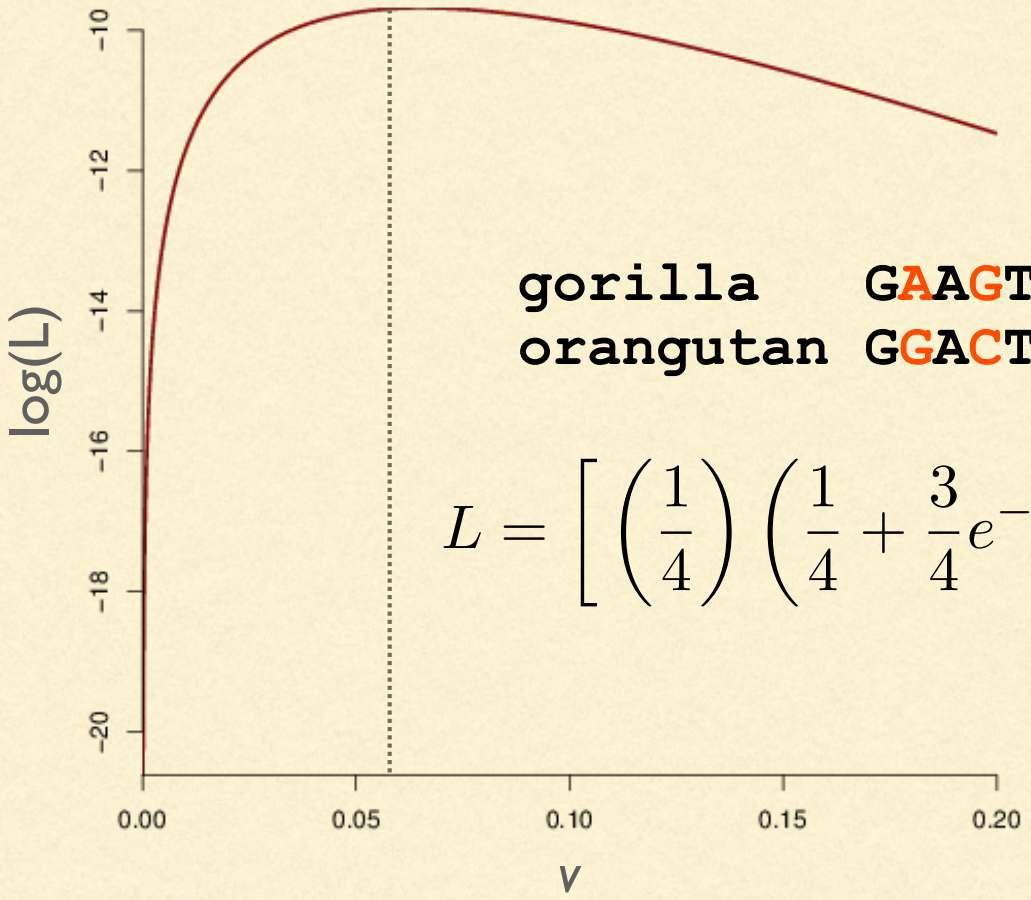
$$L = \left[\left(\frac{1}{4}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-4v/3}\right)\right] \left[\left(\frac{1}{4}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-4v/3}\right)\right]$$

site 1 site 2

Each **site likelihood** is the probability of the **starting state** at the root (1/4) times the **transition probability** (probability of the end state given the starting state)

Maximum likelihood estimation

0.065 is the maximum likelihood estimate (MLE) of *v*




```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```

$$L = \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} + \frac{3}{4}e^{-4\nu/3}\right) \right]^{30} \left[\left(\frac{1}{4}\right) \left(\frac{1}{4} - \frac{1}{4}e^{-4\nu/3}\right) \right]^{2}$$

Transition probabilities

$$\frac{1}{4} + \frac{3}{4}e^{-4\nu/3}$$

same state

$$\frac{1}{4} - \frac{1}{4}e^{-4\nu/3}$$

different states

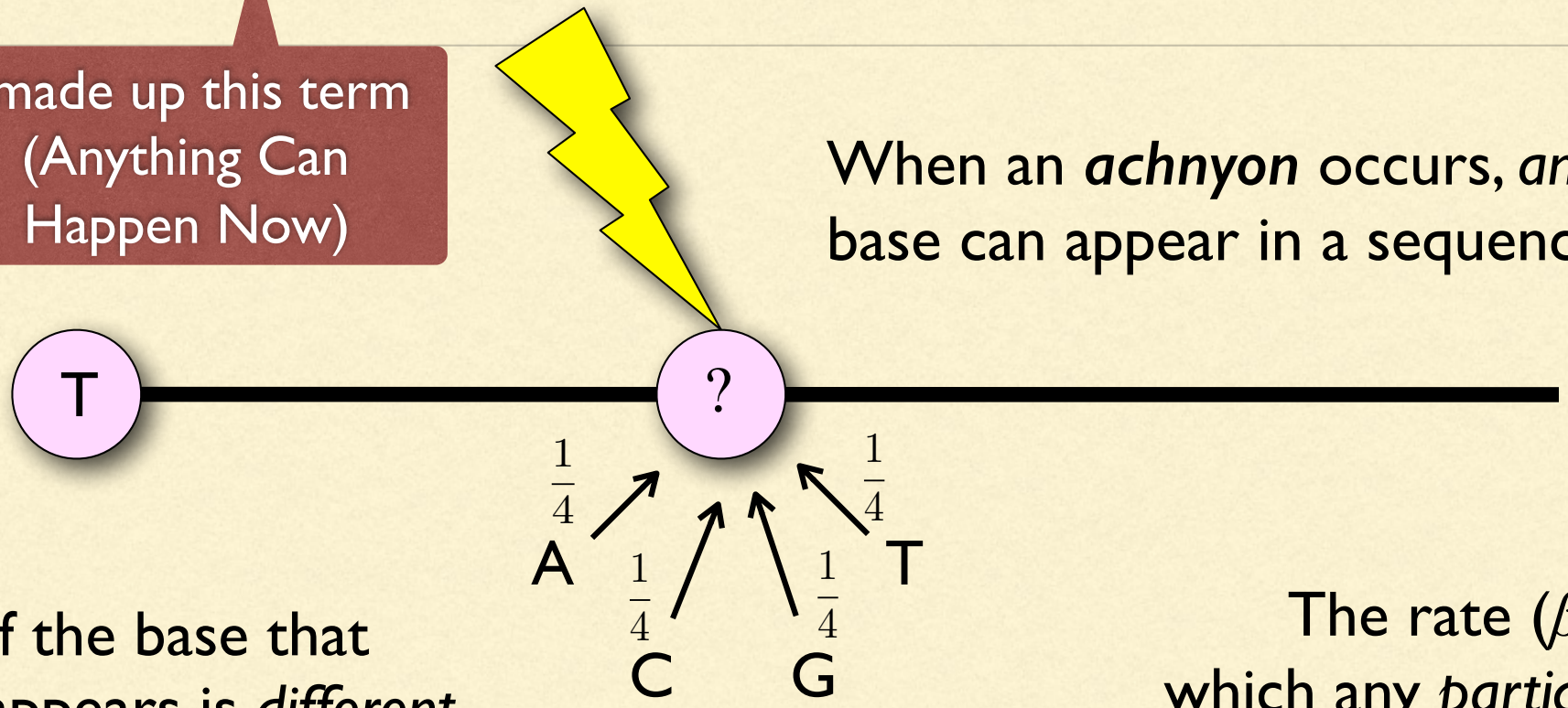
Conditional probability of end state given starting state and edge length

How do these formulas arise?

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 in H. N. Munro (ed.), Mammalian Protein Metabolism. Academic Press, New York.

"ACHNyons" vs. substitutions

I made up this term (Anything Can Happen Now)



When an ***achnyon*** occurs, *any* base can appear in a sequence.

If the base that appears is *different* from the base that was already there, then a **substitution** event has occurred.

The rate (β) at which any *particular* substitution occurs will be 1/4 the achnyon rate (μ). That is, $\beta = \mu/4$ (or $\mu = 4\beta$)

Deriving a transition probability

Calculate the probability that a site currently T will change to G over time t when the rate of this particular substitution is β:

Pr(0 achnyons) = $e^{-\mu t}$ (Poisson probability of zero events)

Pr(at least 1 achnyon) = $1 - e^{-\mu t}$

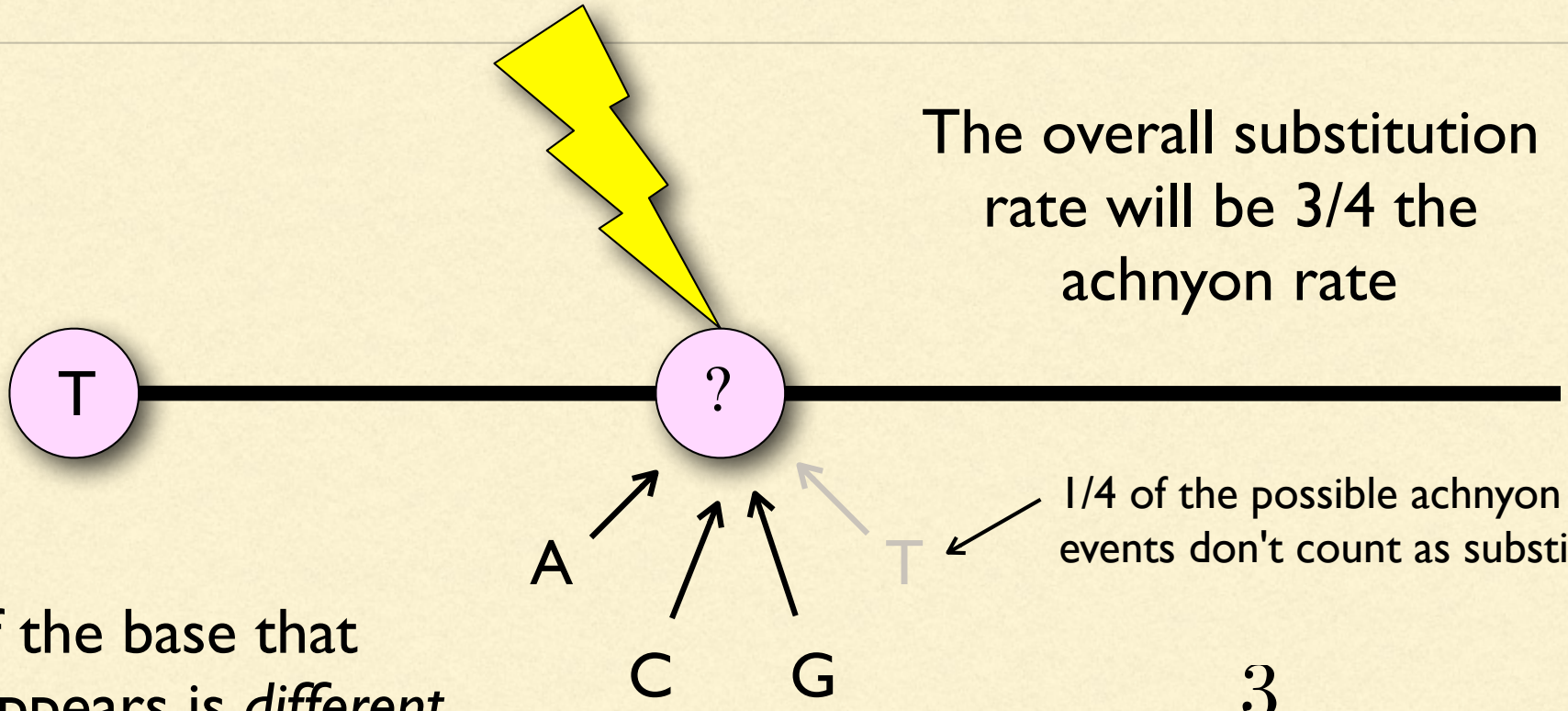
Pr(last achnyon results in base G) = $\frac{1}{4}$

Pr(end in G | start in T) = $\frac{1}{4}\left(1 - e^{-\mu t}\right)$

Remember that the achnyon rate (μ) is 4 times the rate (β) of any *particular* substitution:

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

Expected number of substitutions



The overall substitution rate will be 3/4 the achnyon rate

1/4 of the possible achnyon events don't count as substitutions

If the base that appears is *different* from the base that was already there, then a **substitution** event has occurred.

$$\nu = \frac{3}{4}\mu t = 3\beta t$$

$$\frac{\nu}{3} = \beta t$$

Deriving a transition probability

Calculate the probability that a site currently T will change to G over time t when the rate of this particular substitution is β:

$$\vdots$$

Remember that the achnyon rate (μ) is 4 times the rate (β) of any *particular* substitution:

$$P_{GT}(t) = \tfrac{1}{4}\left(1 - e^{-4\beta t}\right)$$

Substitute $\nu/3$ for βt:

$$P_{GT}(t) = \tfrac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

Transition Probabilities: Remarks

$$P_{TA}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$
$$P_{TC}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$
$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$
$$P_{TT}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$1 - e^{-4\nu/3}$$

These should add to 1.0 because T *must* change to something!

Doh! Something must be wrong here...

Transition Probabilities: Remarks

$$P_{TA}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TC}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TG}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right)$$

$$P_{TT}(t) = \frac{1}{4}\left(1 - e^{-4\nu/3}\right) + e^{-4\nu/3}$$

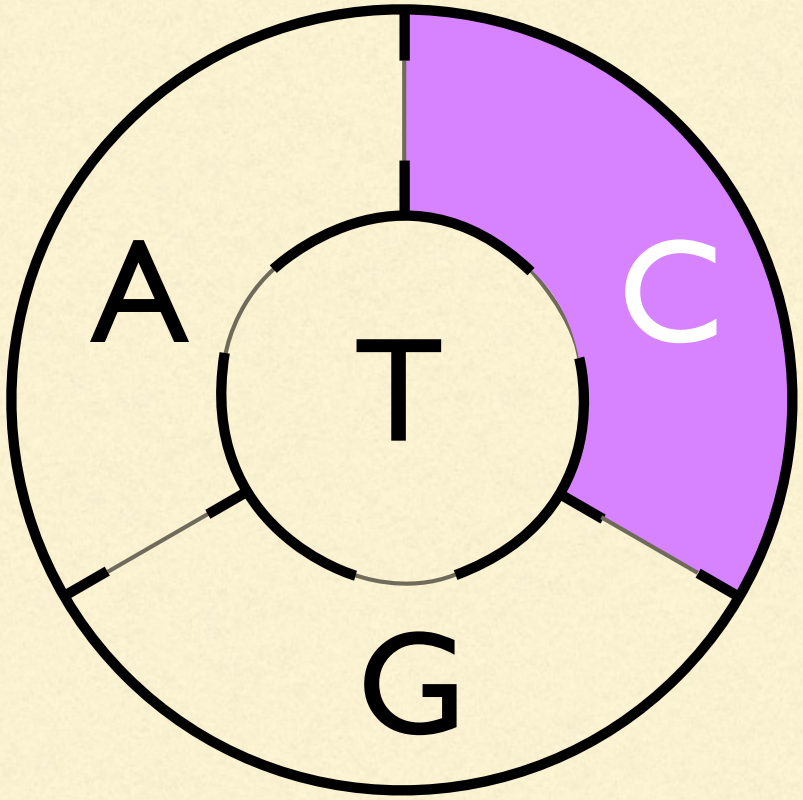
I forgot to account for the possibility of *no* acnyons over time t

Equilibrium Frequencies

Imagine a bottle of perfume has been spilled in room C.

The doors to the other rooms are closed, so the perfume has, thus far, not been able to spread.

What would happen if we opened all the doors?



Equilibrium Frequencies

At the **instant the doors open**, perfume molecules...

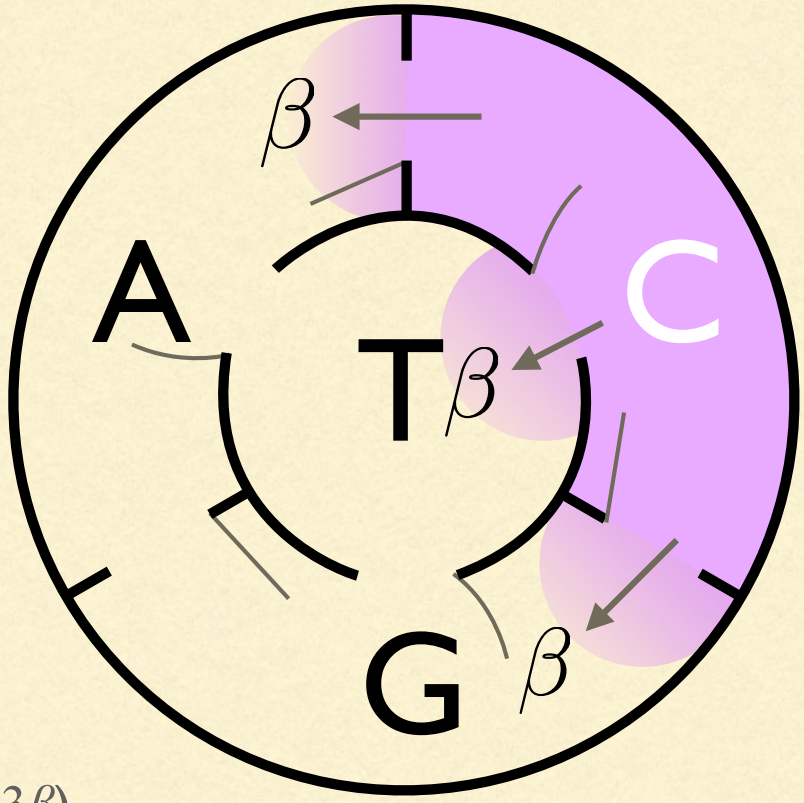
enter room A at rate β

enter room T at rate β

enter room G at rate β

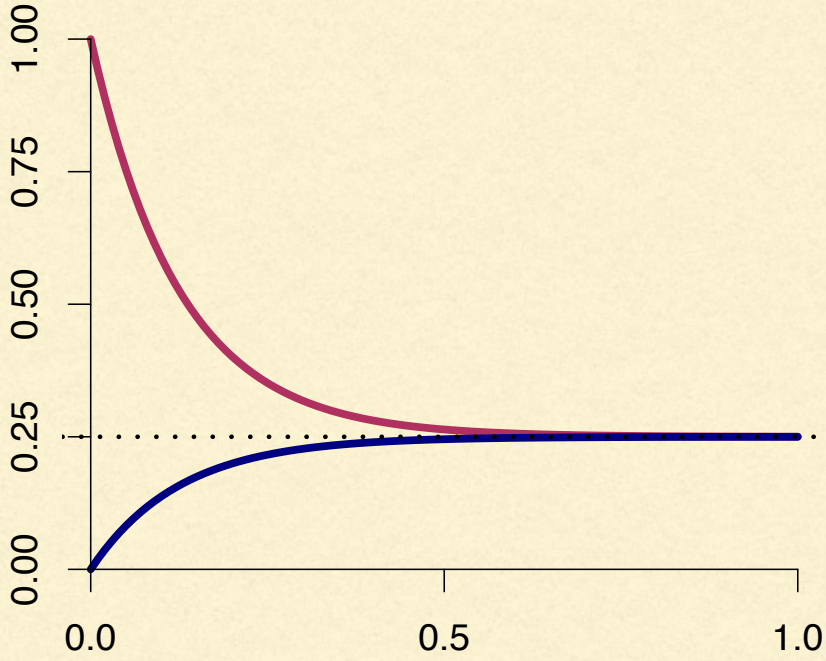
enter room C at rate -3β

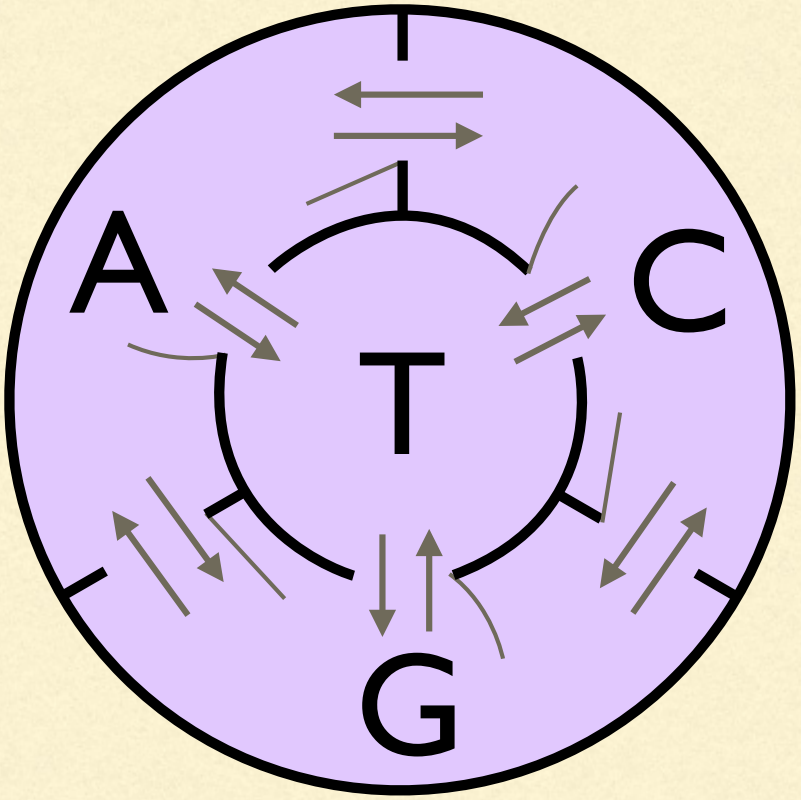
(you could also say they *leave* C at rate 3β)



Equilibrium Frequencies

At **equilibrium**, the relative concentration of perfume is **equal** in all rooms





$$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$$

Transition probability demo

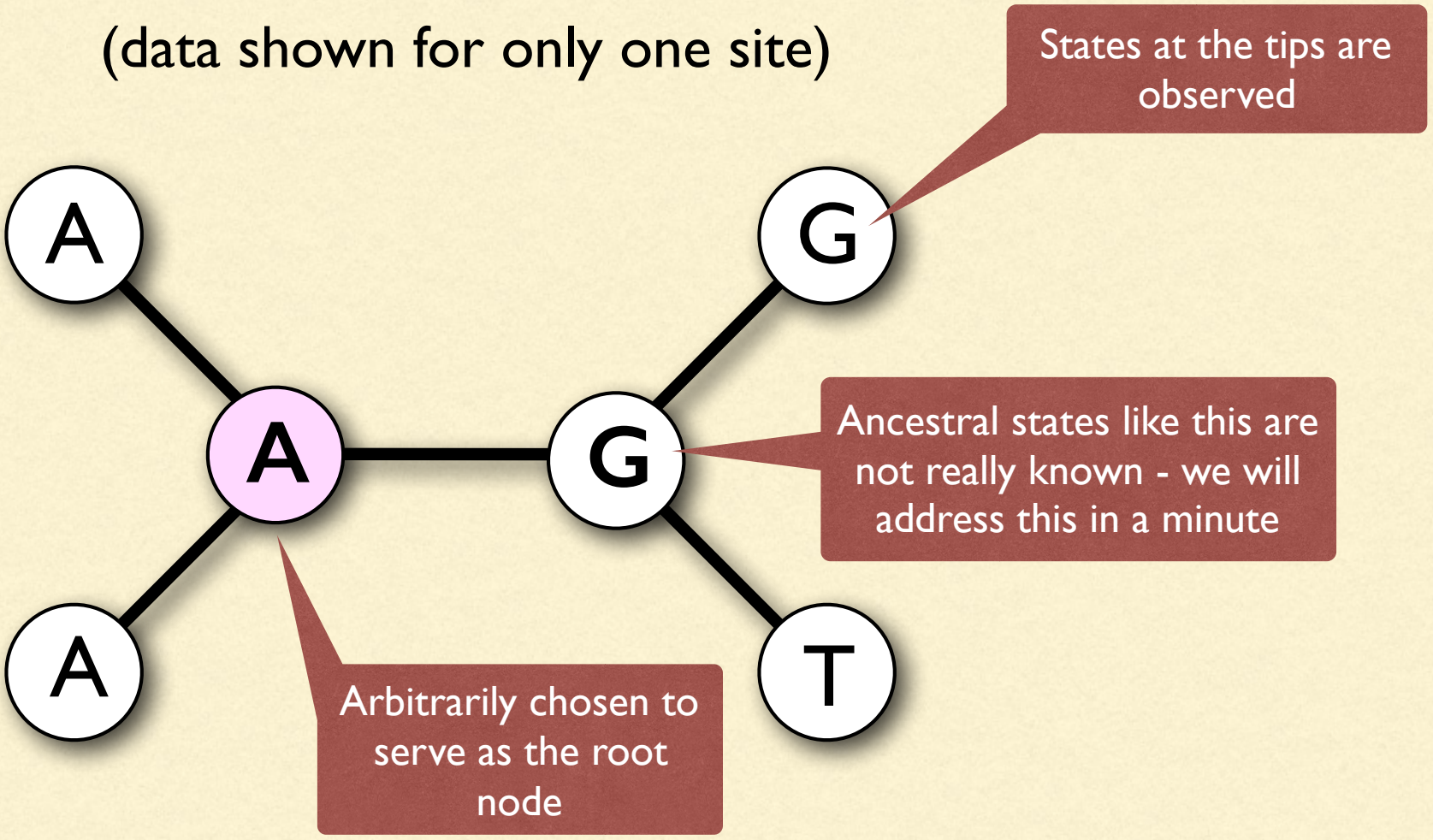
https://phylogeny.uconn.edu/transition-probability/

Sequence data for four taxa

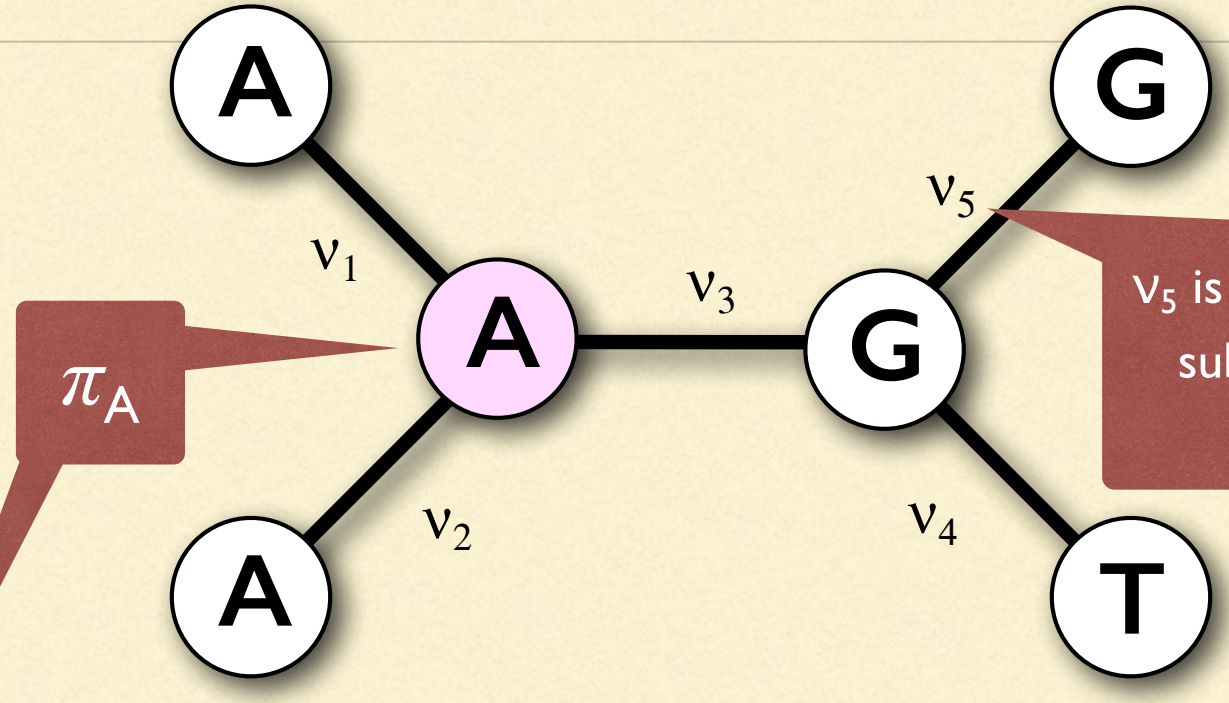
one site

```
Sphagnum   CGACATTTTGGCAGCATTTCGAATGACTCCTCAACCTGGAGTACCACCCG...
Asplenium  CGATATCTTGGCAGCTTTCCGGATGACCCCACAACCCGGAGTACCAGCTG...
Picea      GGATATTTTGGCAGCATTCCGAGTAACTCCTCAACCAGGGGTGCCGCCCG...
Avena      TGATATCTTGGCAGCATTCCGAGTAACTCCTCAACCTGGGGTTCCGCCGG...
```

Likelihood for tree (one site)

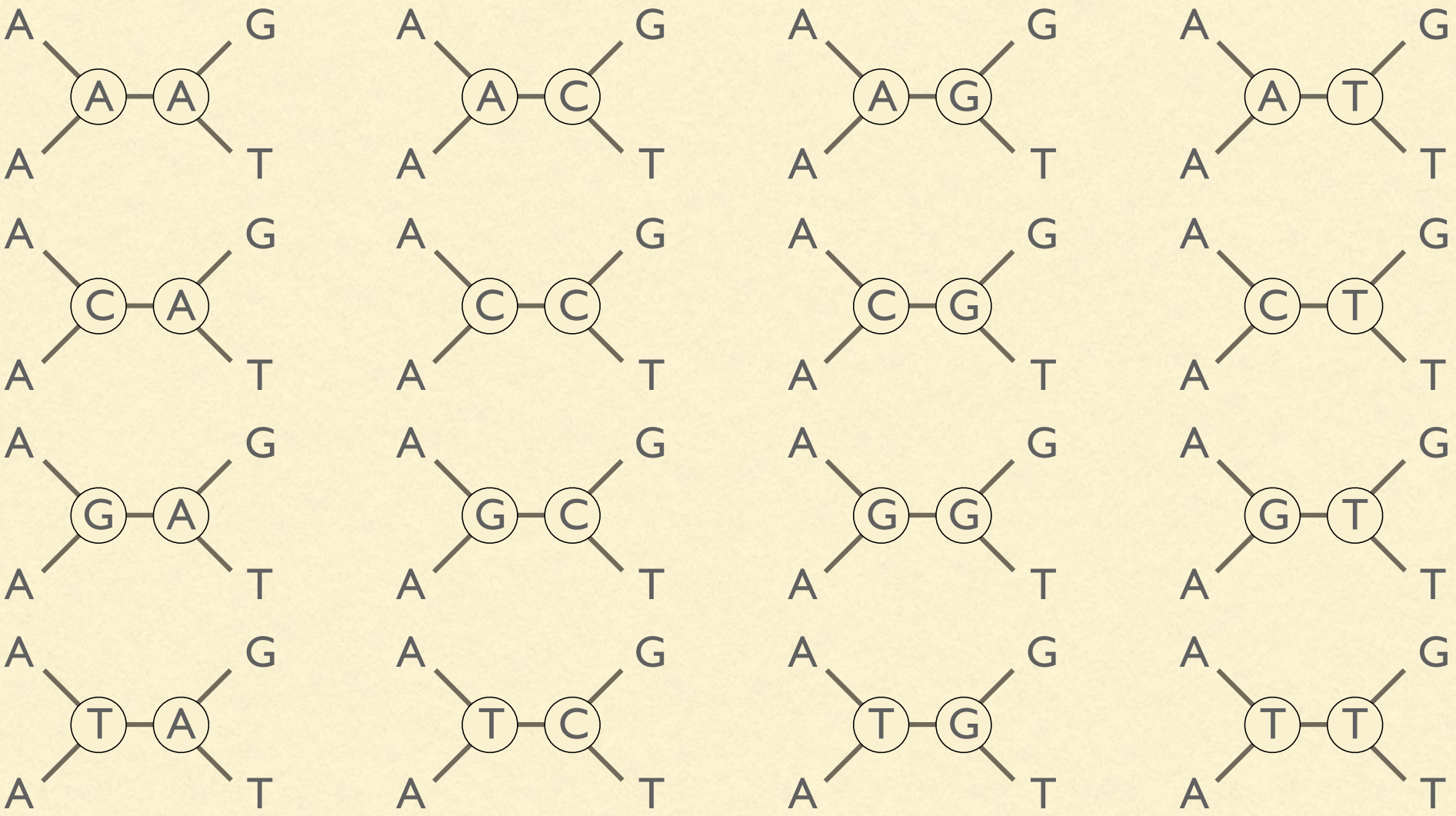


Likelihood for tree (one site)



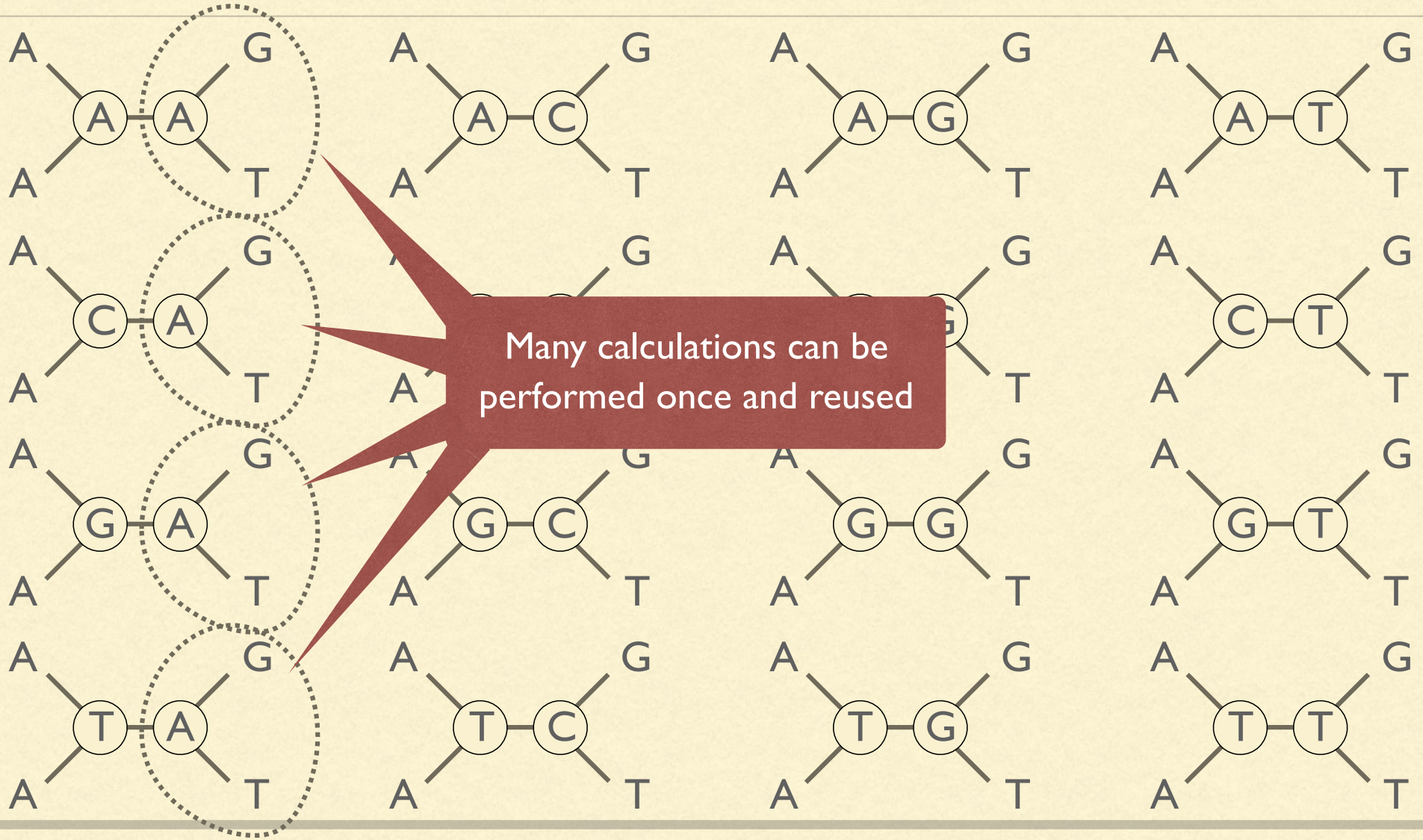
$$L = \frac{1}{4}\left[\frac{1}{4} + \frac{3}{4}e^{-4\nu_1/3}\right]\left[\frac{1}{4} + \frac{3}{4}e^{-4\nu_2/3}\right]\left[\frac{1}{4} - \frac{1}{4}e^{-4\nu_3/3}\right]\left[\frac{1}{4} - \frac{1}{4}e^{-4\nu_4/3}\right]\left[\frac{1}{4} + \frac{3}{4}e^{-4\nu_5/3}\right]$$

Brute force approach

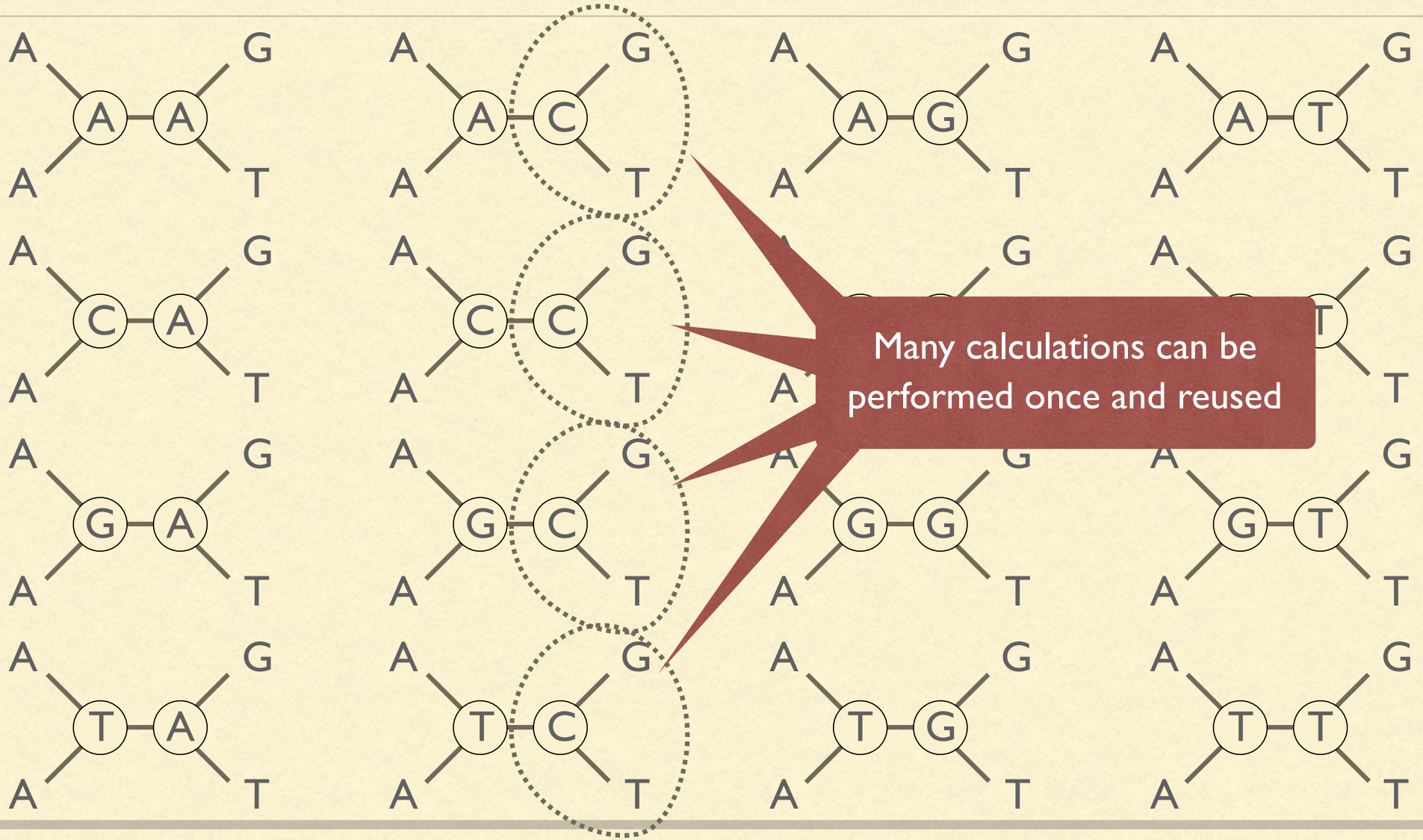


Pruning algorithm

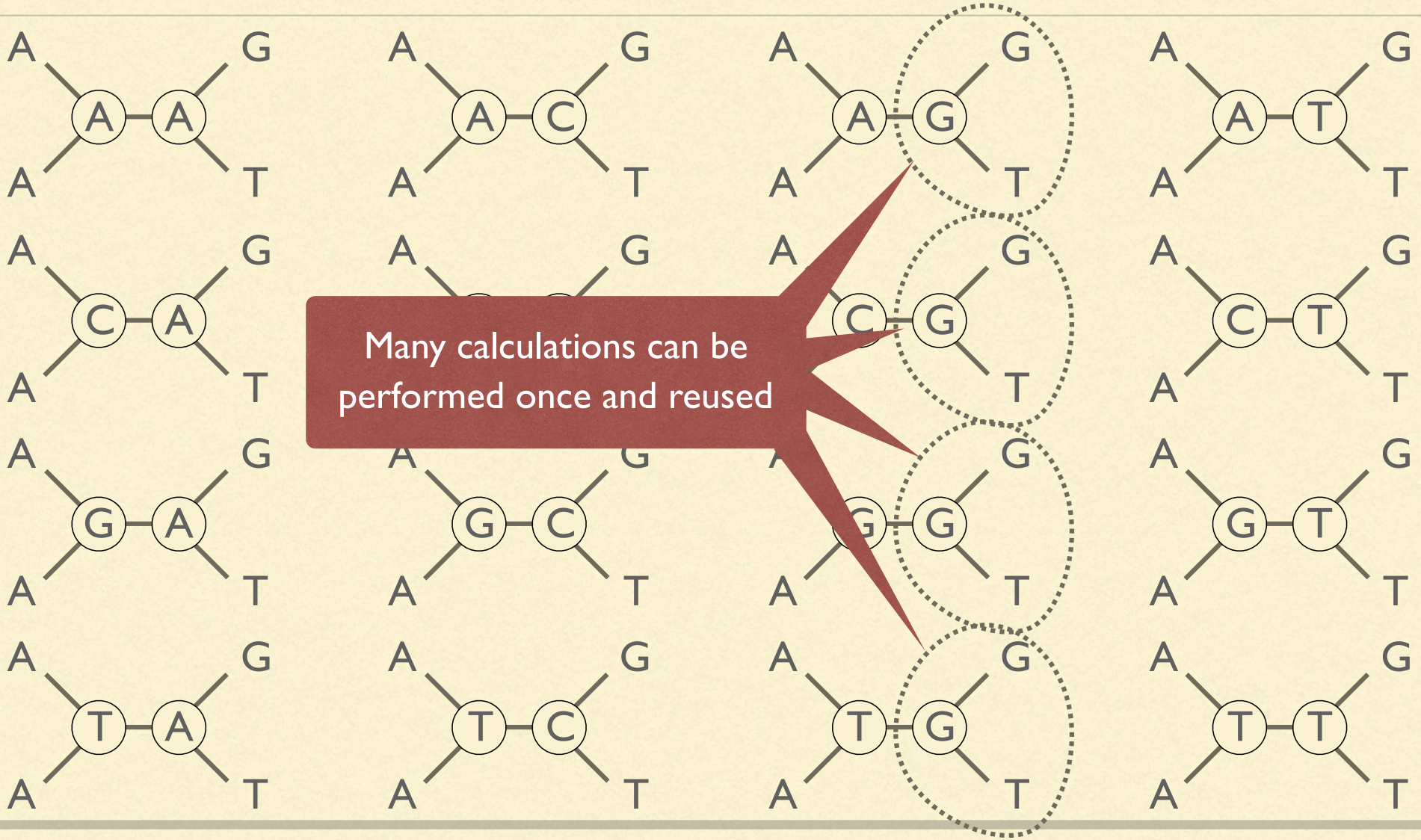
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376.



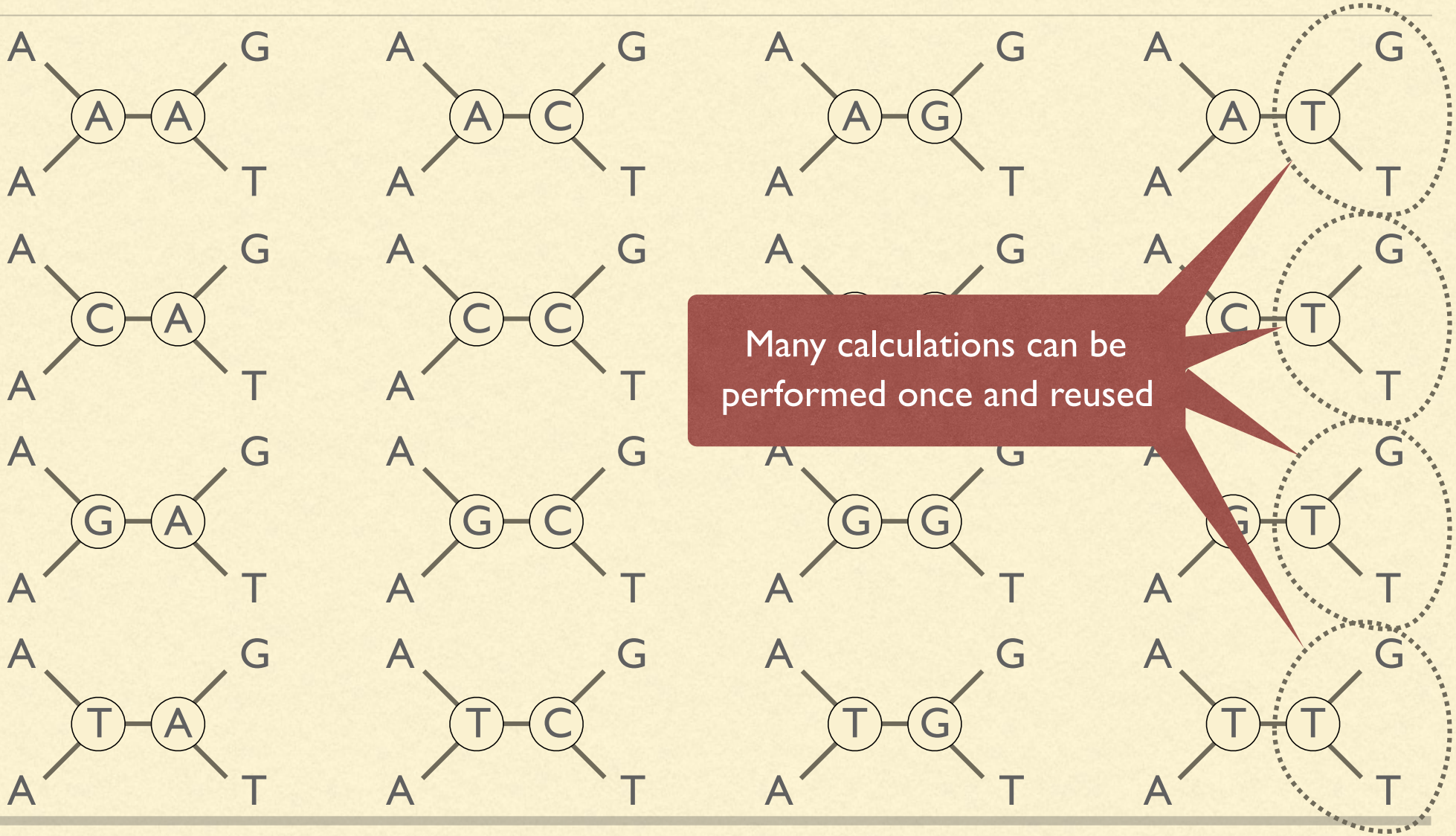
Pruning algorithm



Pruning algorithm

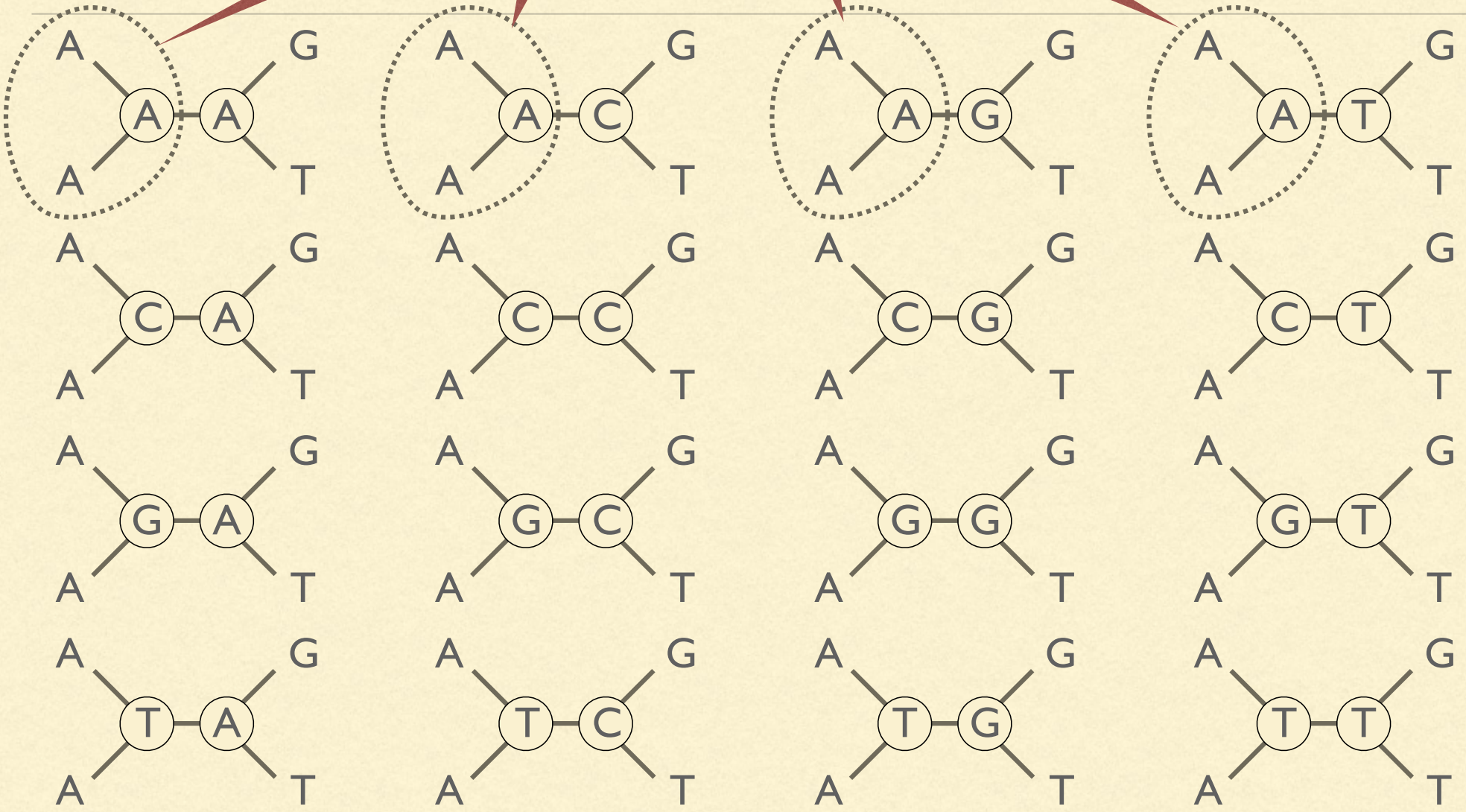


Pruning algorithm

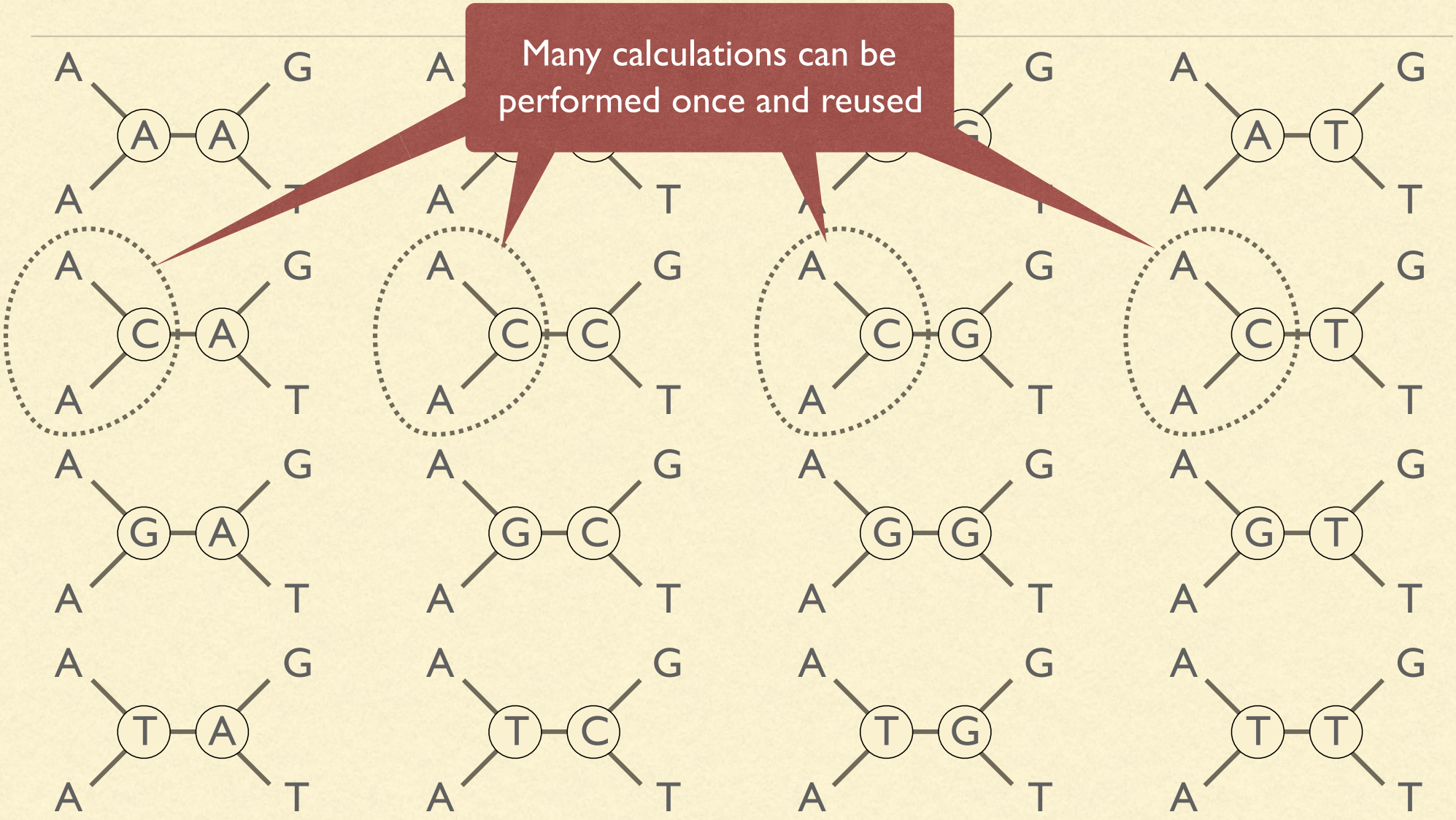


Pruning alg

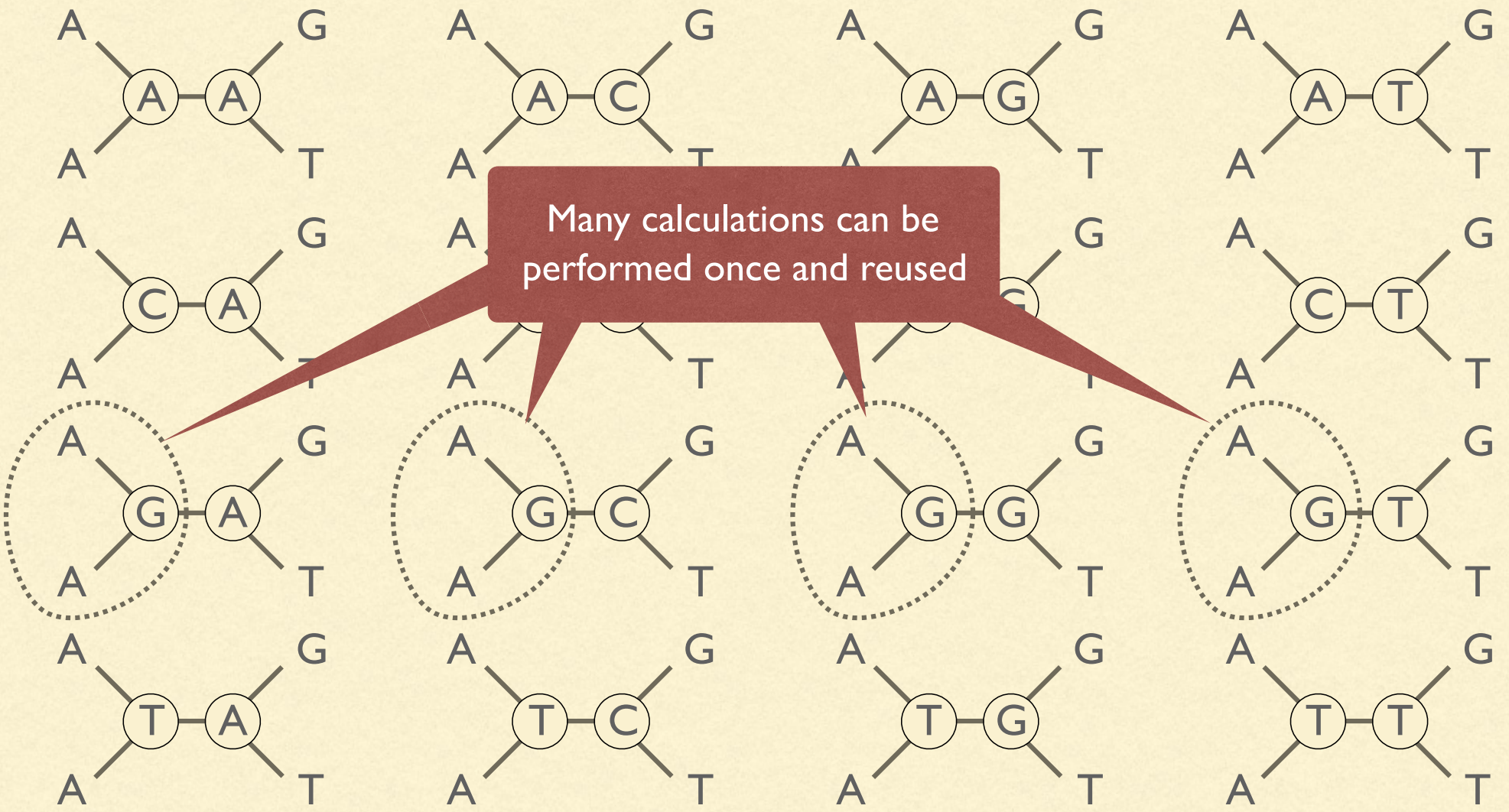
Many calculations can be performed once and reused



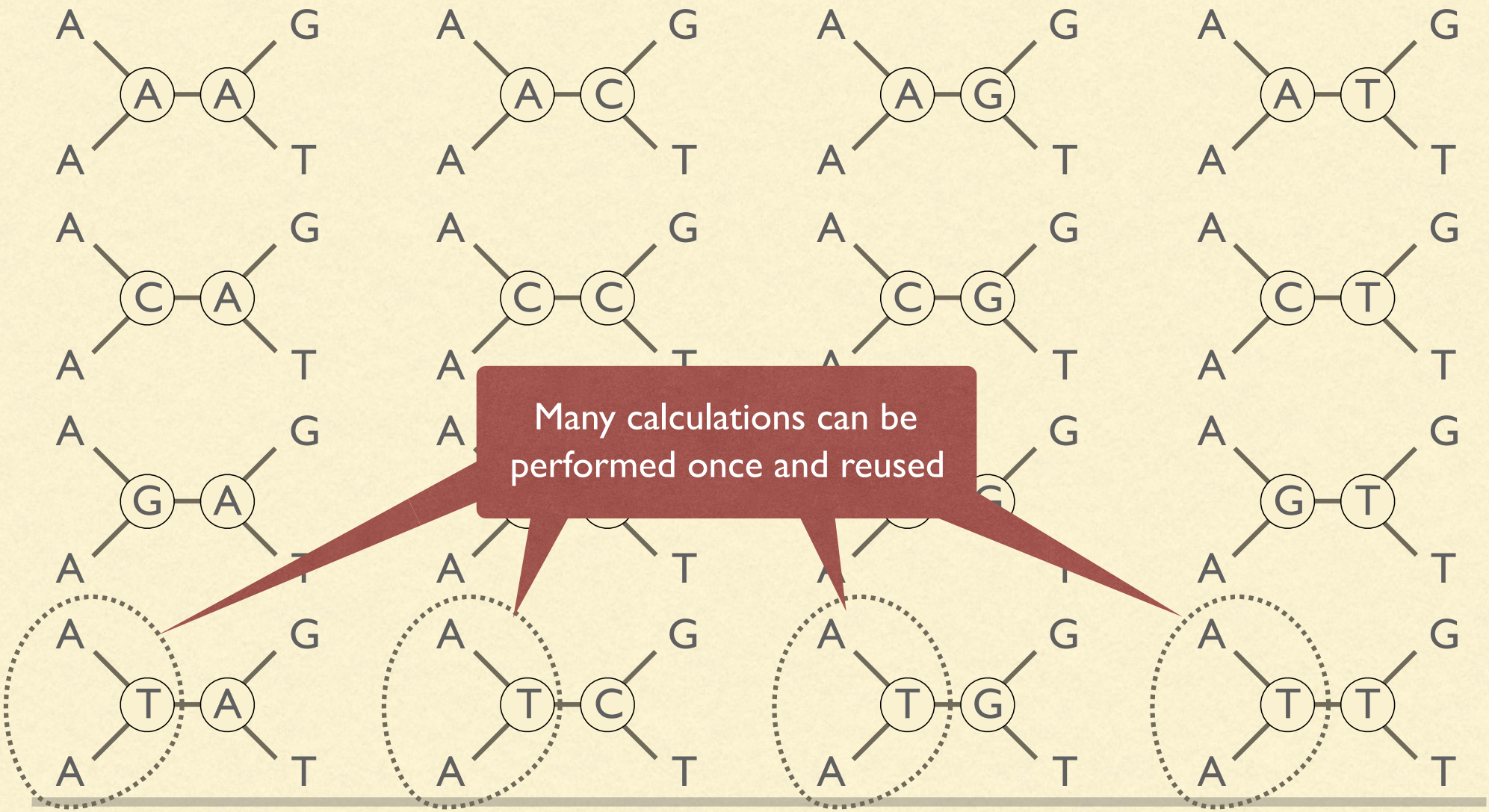
Pruning algorithm



Pruning algorithm



Pruning algorithm

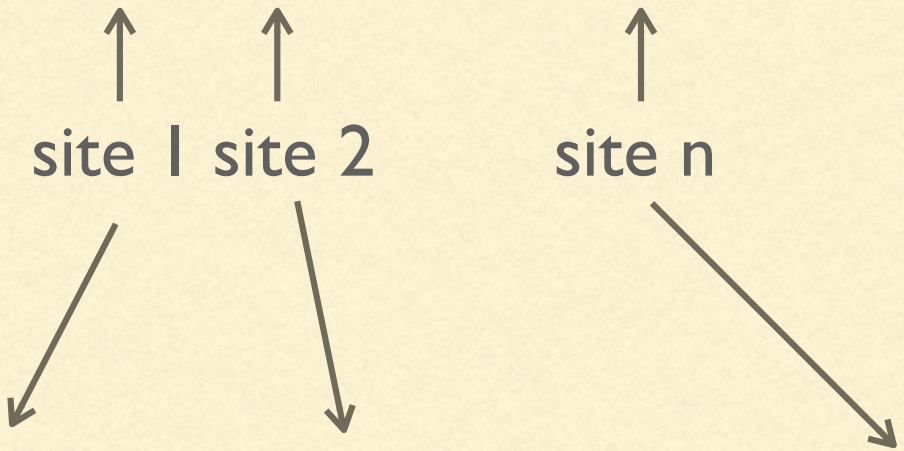


Total likelihood

$$L = L_1 L_2 \cdots L_n$$

Total likelihood

$$L = L_1 L_2 \cdots L_n$$



$$\log L = \log L_1 + \log L_2 + \cdots + \log L_n$$

~End of Part 1~