

Introduction to Phylogenetics

Marine Biological Laboratory, Woods Hole,
Massachusetts

26 May - 5 June, 2023

Paul O. Lewis

Department of Ecology & Evolutionary Biology

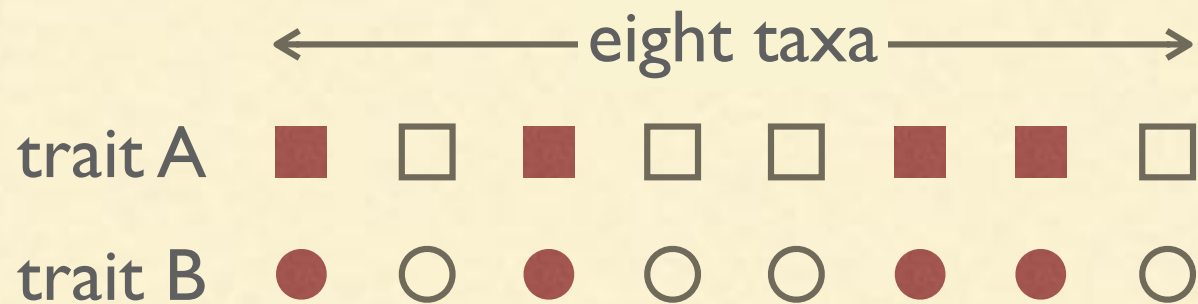


Phylogenetics is key

Dobzhansky, T. 1973. Nothing in **biology** makes sense except in the light of **evolution**. The American Biology Teacher 35:125-129.

Nothing in **evolutionary biology** makes sense except in the light of **phylogeny**. - Society of Systematic Biologists

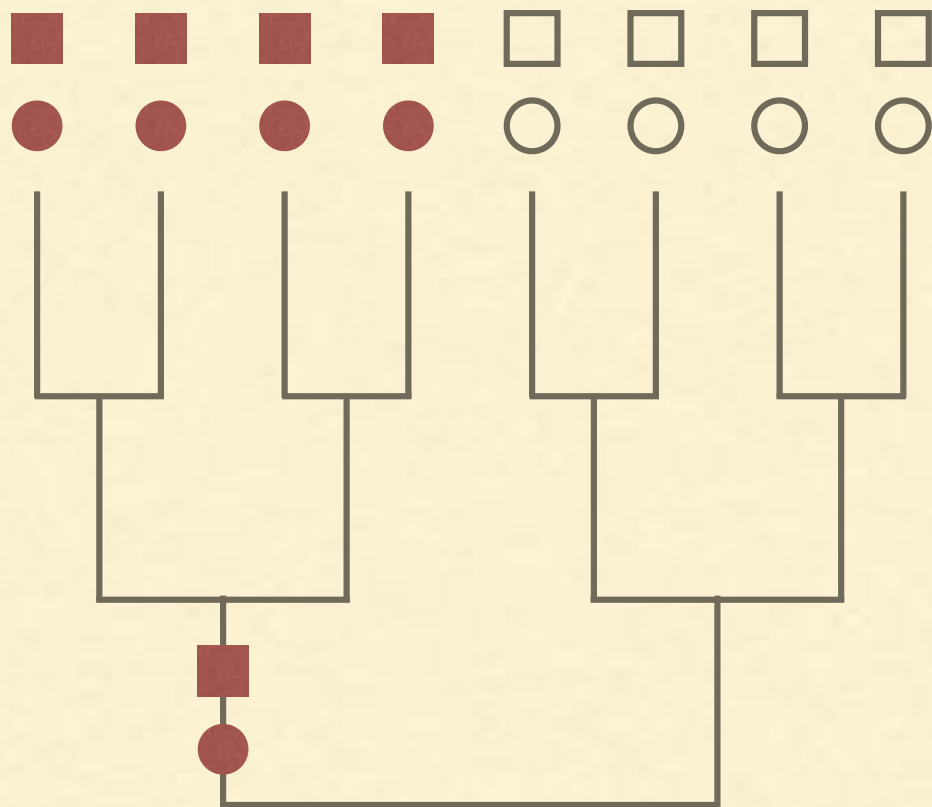
Perfect correlation



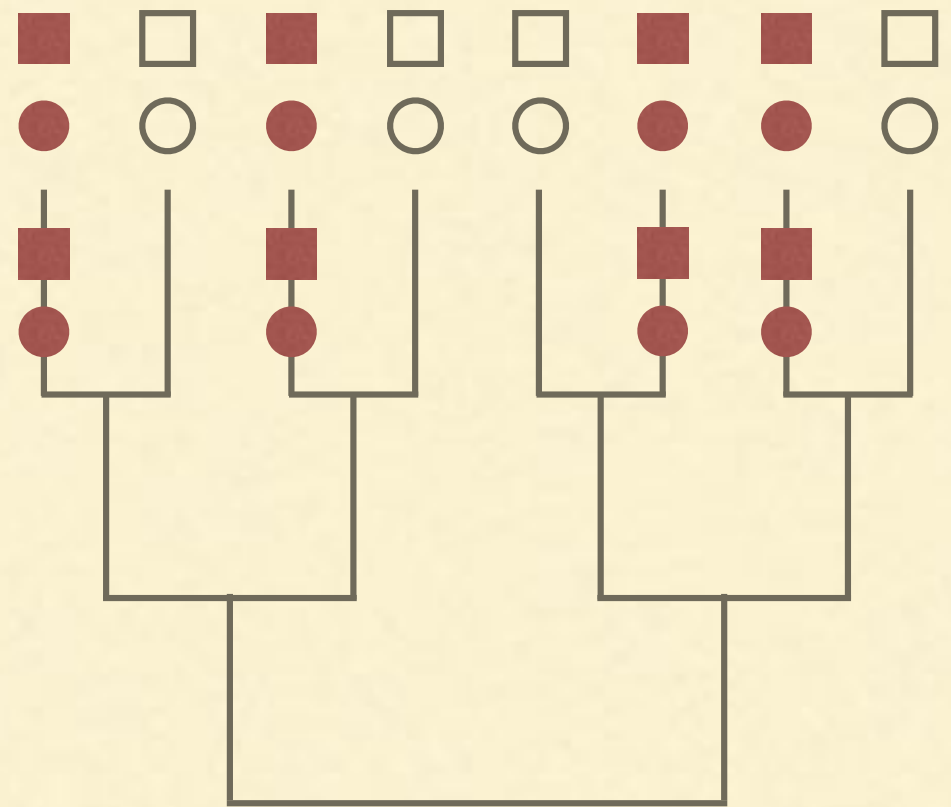
How much importance should we attach to the co-distribution of these two traits?

Two very different explanations

Simple inheritance



Correlated evolution

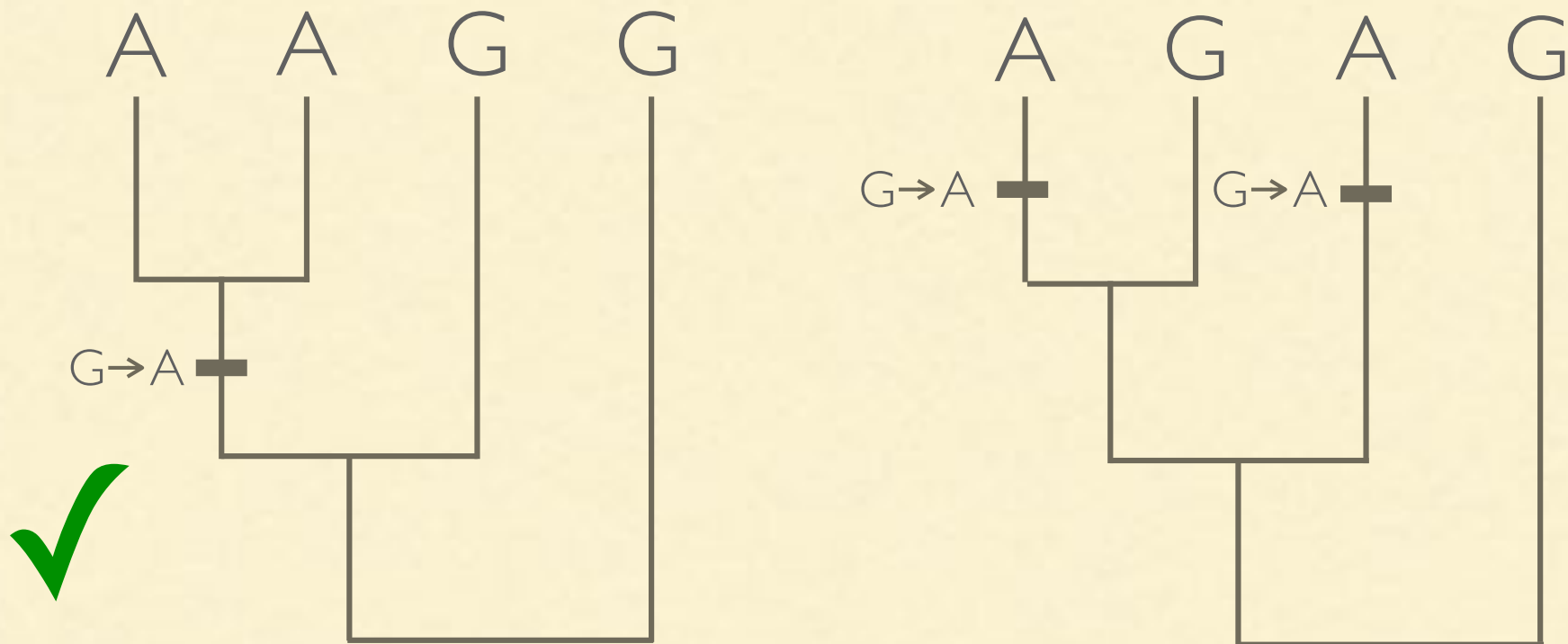


How to estimate a tree

*I think that I shall never see
A thing so awesome as the Tree
That links us all in paths of genes
Down into depths of time unseen
--- DAVID MADDISON (2013)*

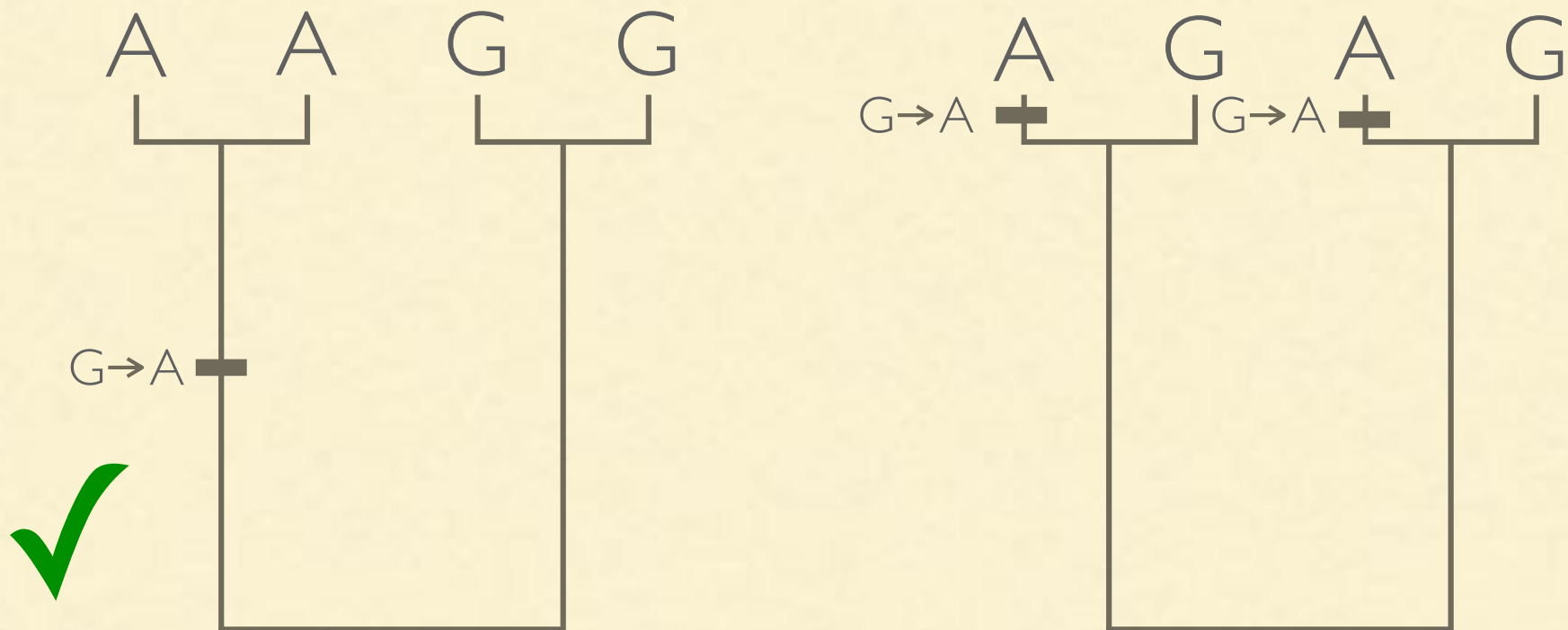
Maddison, D. 2013. The Tree of Life. Systematic Biology 62:179

Which tree is better?



Parsimony criterion says tree requiring fewer changes is better

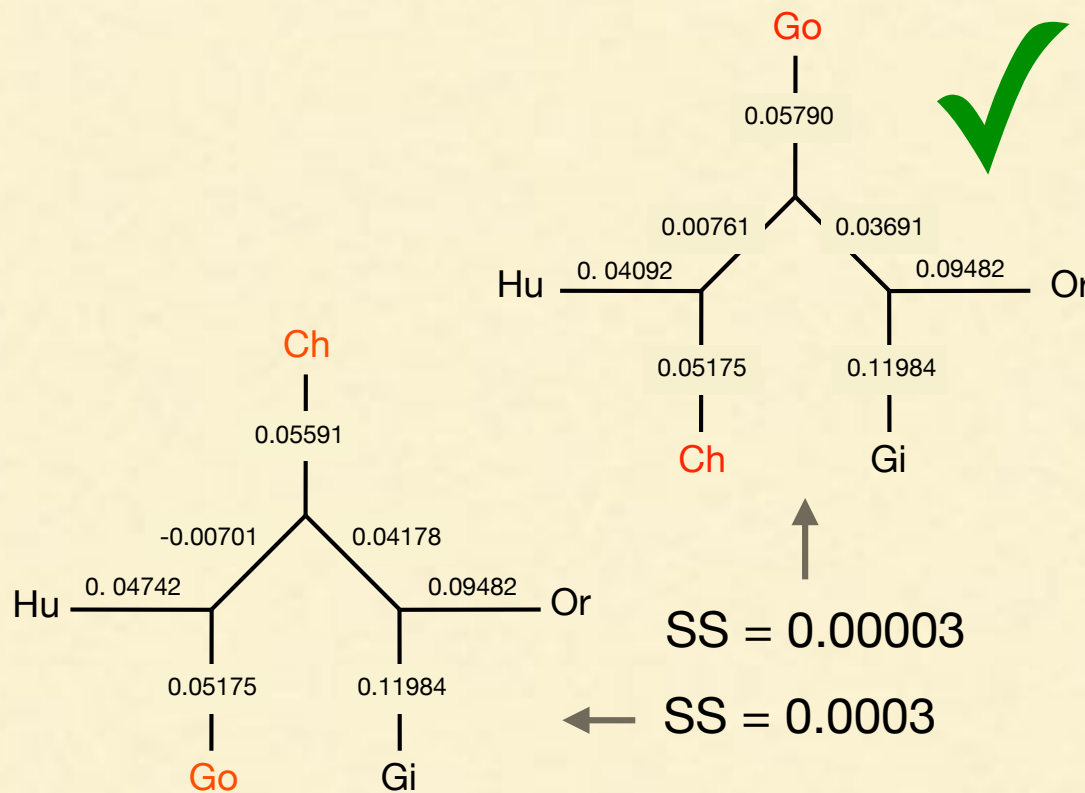
Which tree is better?



Likelihood criterion says tree that makes us less surprised at the observed data is better

Which tree is better?

$$(0.10928 - 0.10643)^2$$

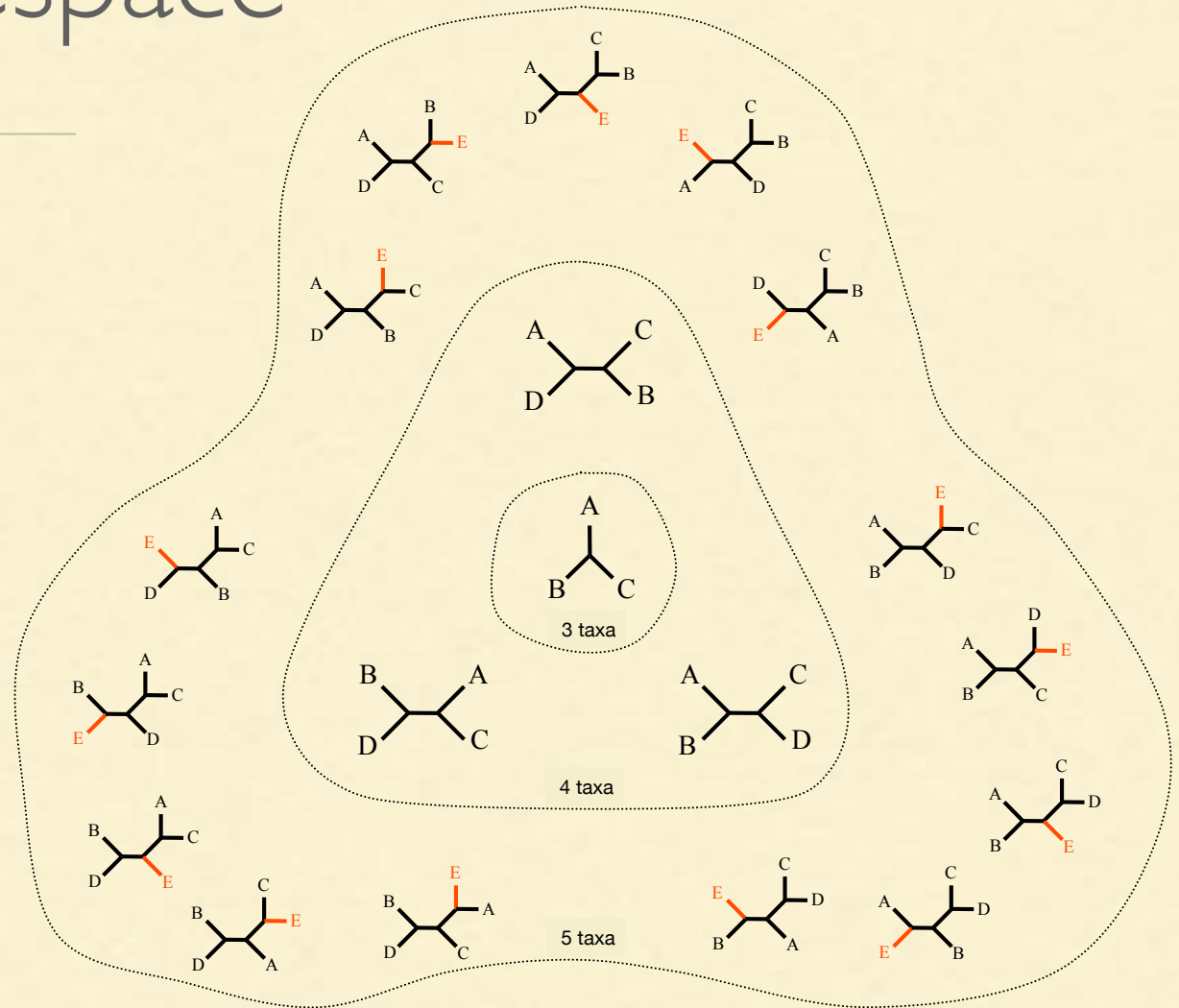


Taxon Pair	distance (data)	distance (tree)	squared differences
Hu-Ch	0.09267	0.09267	0
Hu-Go	0.10928	0.10643	0.000008123
Hu-Or	0.17848	0.18026	0.000003168
Hu-Gi	0.2042	0.20528	0.000001166
Ch-Go	0.1144	0.11726	0.00000818
Ch-Or	0.19413	0.19109	0.000009242
Ch-Gi	0.21591	0.21611	0.00000004
Go-Or	0.18836	0.18963	0.000001613
Go-Gi	0.21592	0.21465	0.000001613
Or-Gi	0.21466	0.21466	0
			0.000033144

Least squares criterion says tree that better matches pairwise distances is better

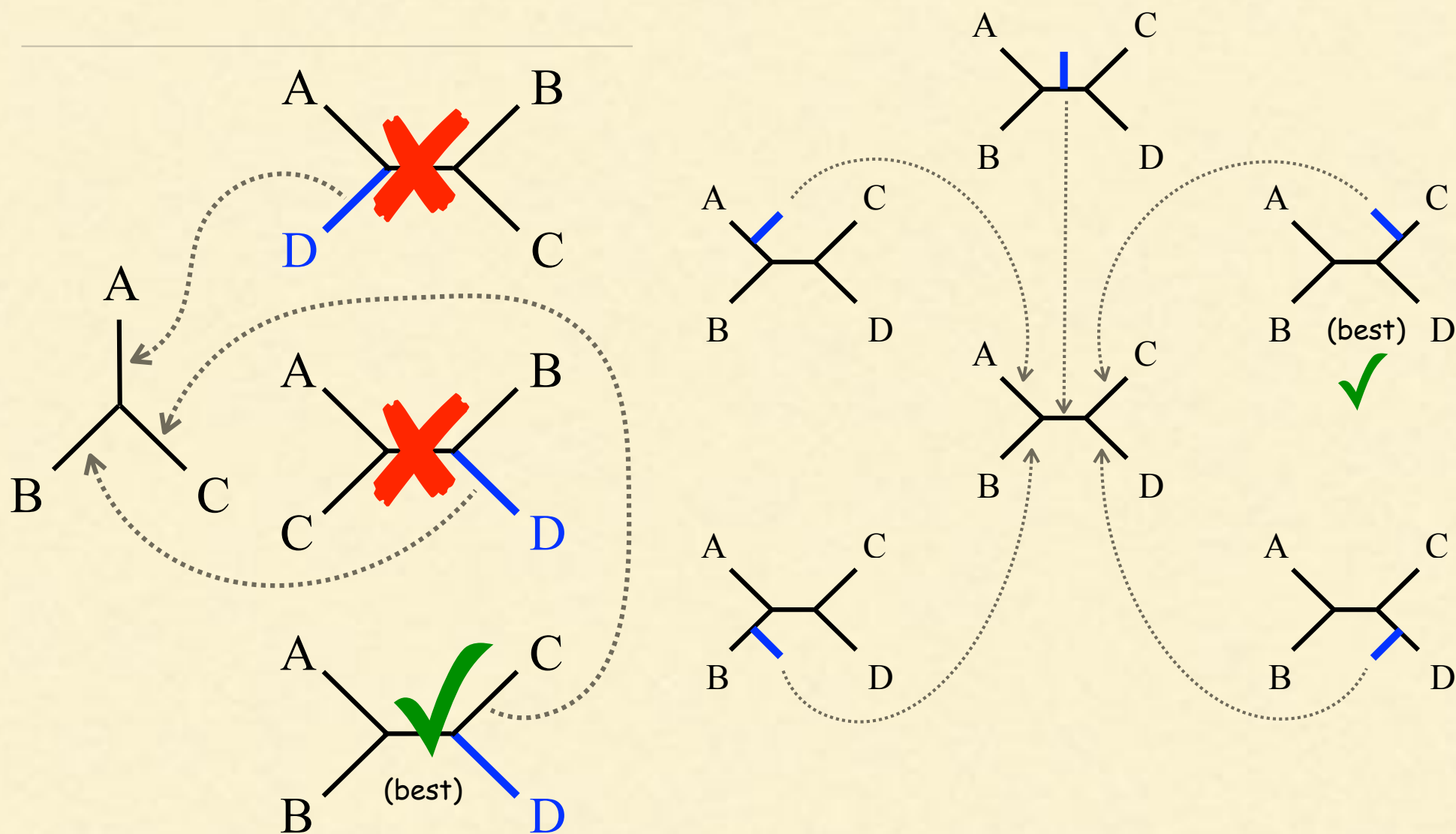
Searching treespace

Taxa	Number of unrooted trees
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625
16	213,458,046,676,875
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	221,643,095,476,699,771,875
21	8,200,794,532,637,891,559,375
22	319,830,986,772,877,770,815,625
23	13,113,070,457,687,988,603,440,625

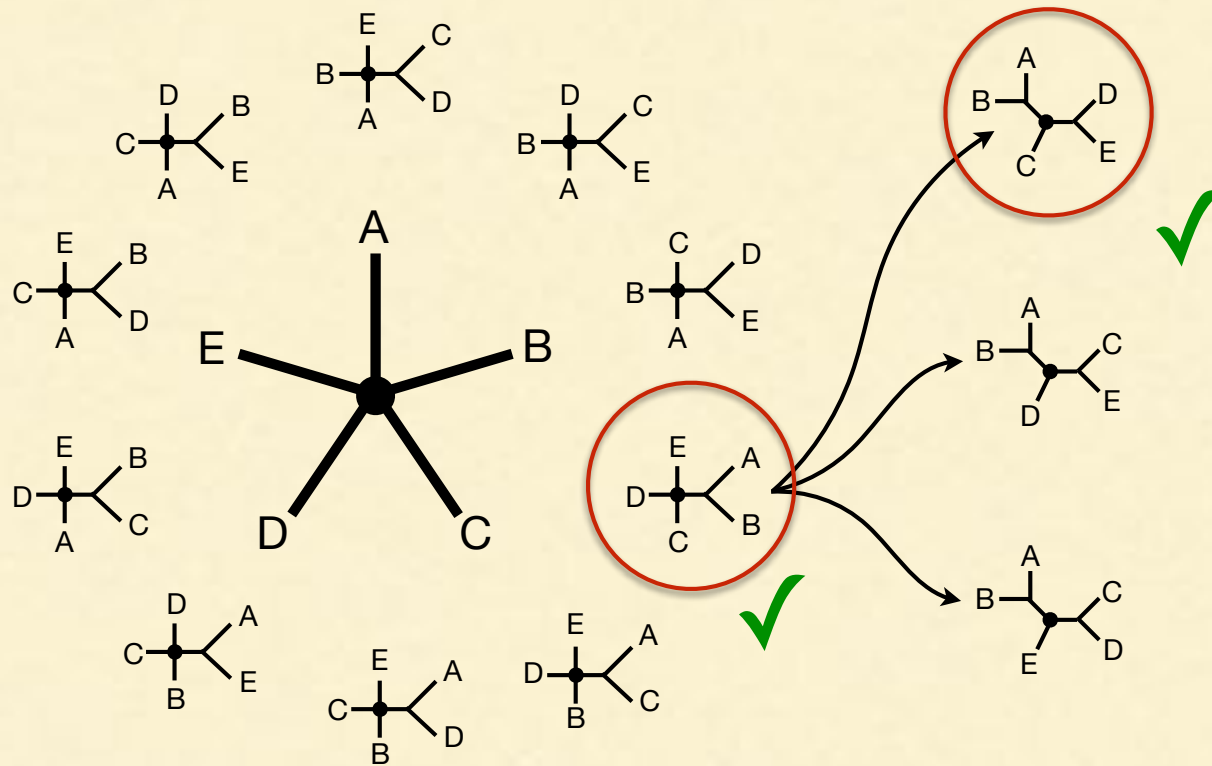


← 83.2 billion years @ 5 million trees/sec

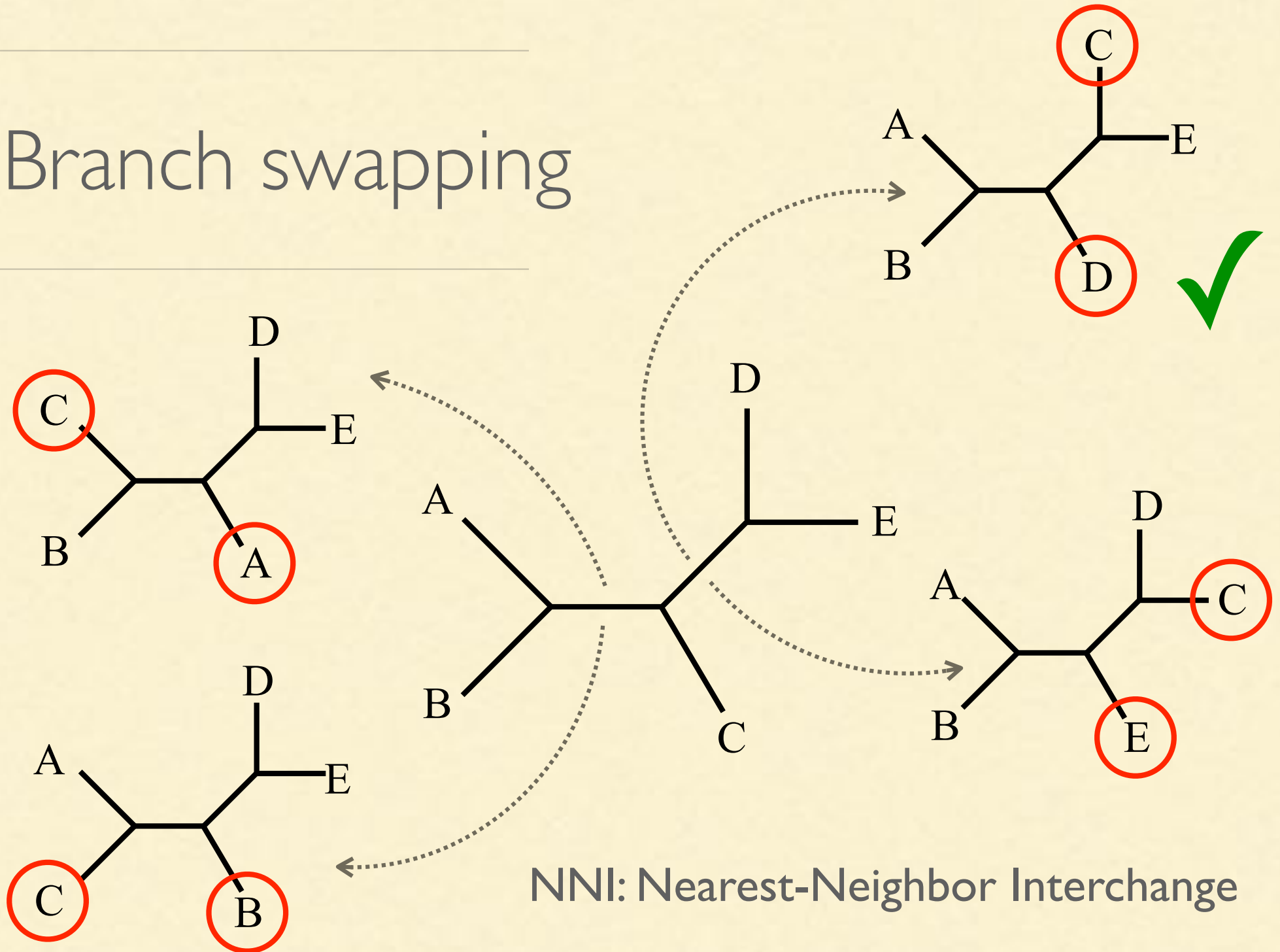
Stepwise addition



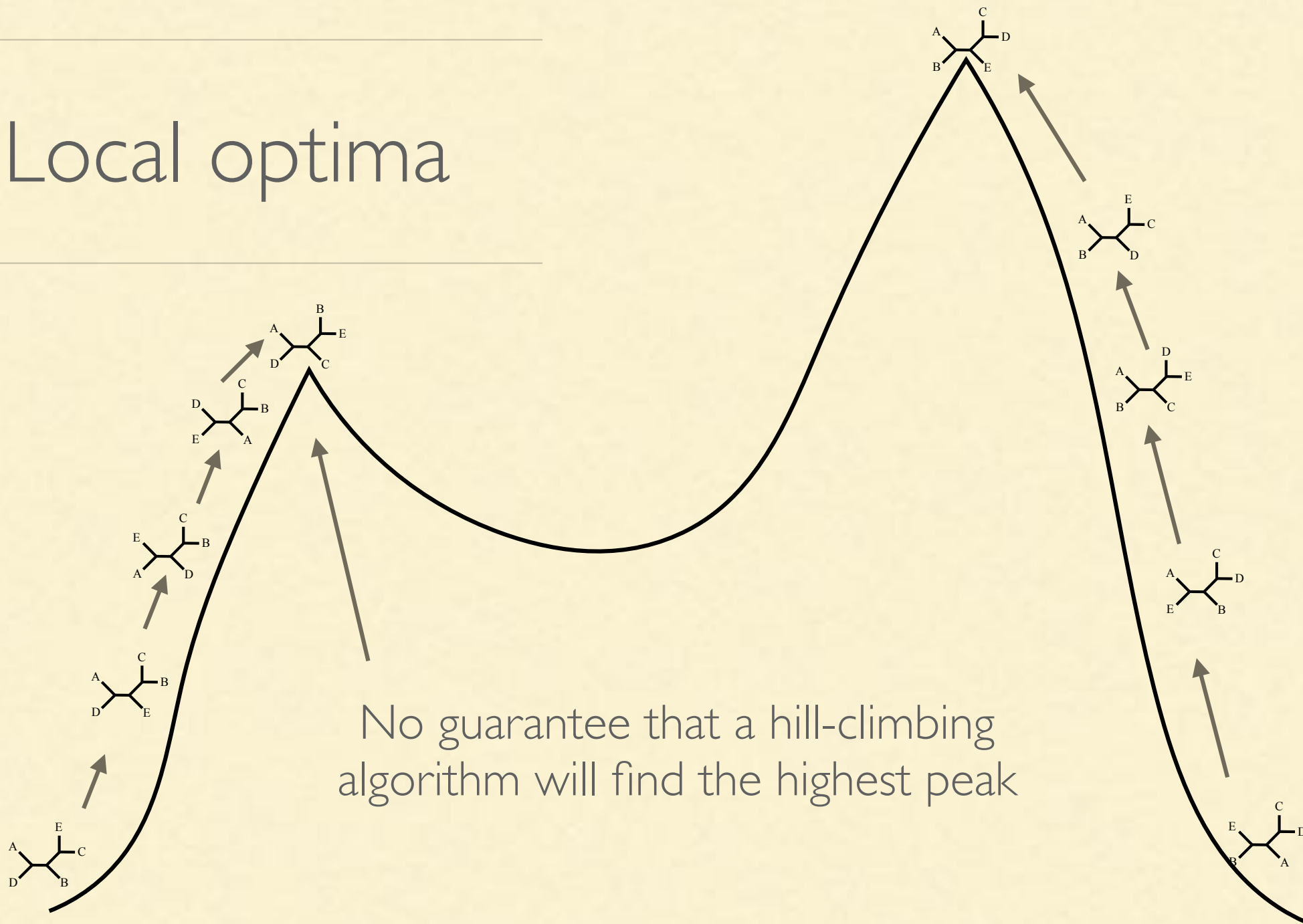
Star decomposition (e.g. Neighbor Joining)



Branch swapping



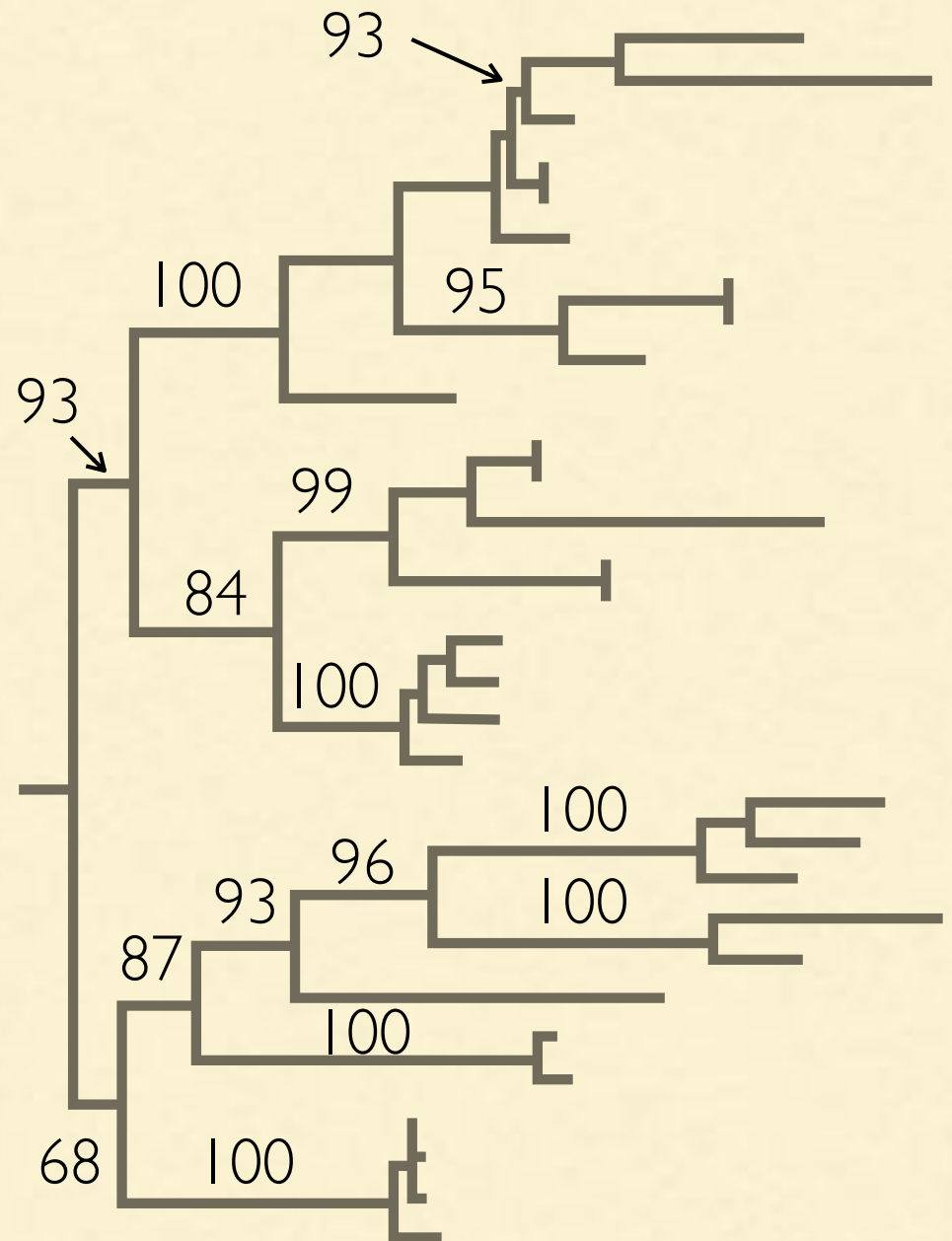
Local optima



Support

Not all parts of a tree are equally well supported by the data.

Support values on the branches tell us how confident we can be in the clade defined by that branch.



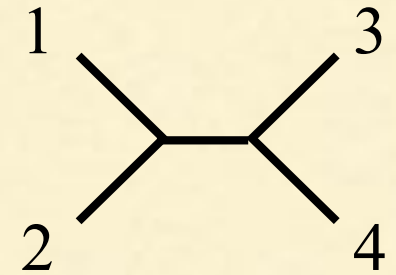
Bootstrap support

sites sampled with replacement

	1	2	3	4	5	6	7	8
1	A	G	G	C	G	T	A	C
2	A	A	G	C	G	T	A	T
3	A	G	T	C	A	C	G	G
4	A	A	T	C	G	C	G	G

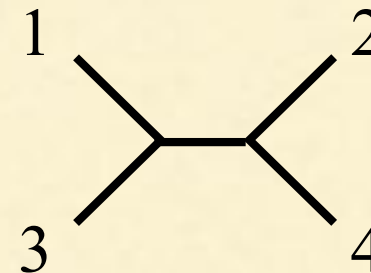
X **X** **X**

original data

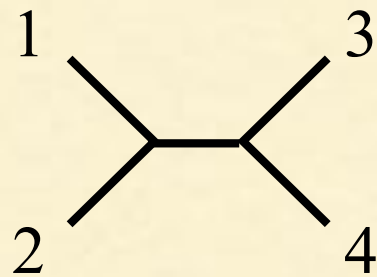


	1	2	3	4	5	6	7	8
1	G	G	C	G	G	C	G	G
2	G	A	C	A	G	T	A	G
3	T	G	C	G	A	G	G	A
4	T	A	C	A	G	G	A	G

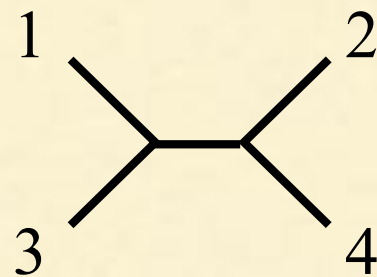
bootstrap replicate



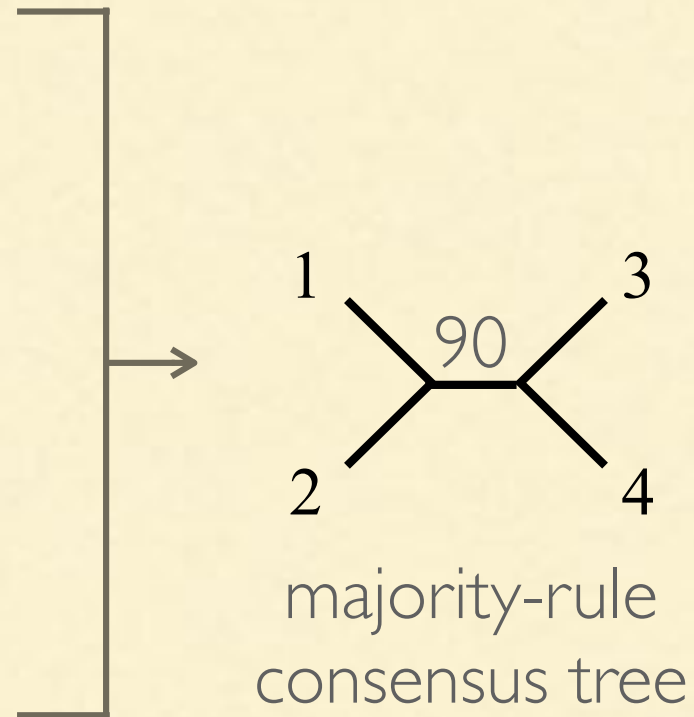
Consensus trees



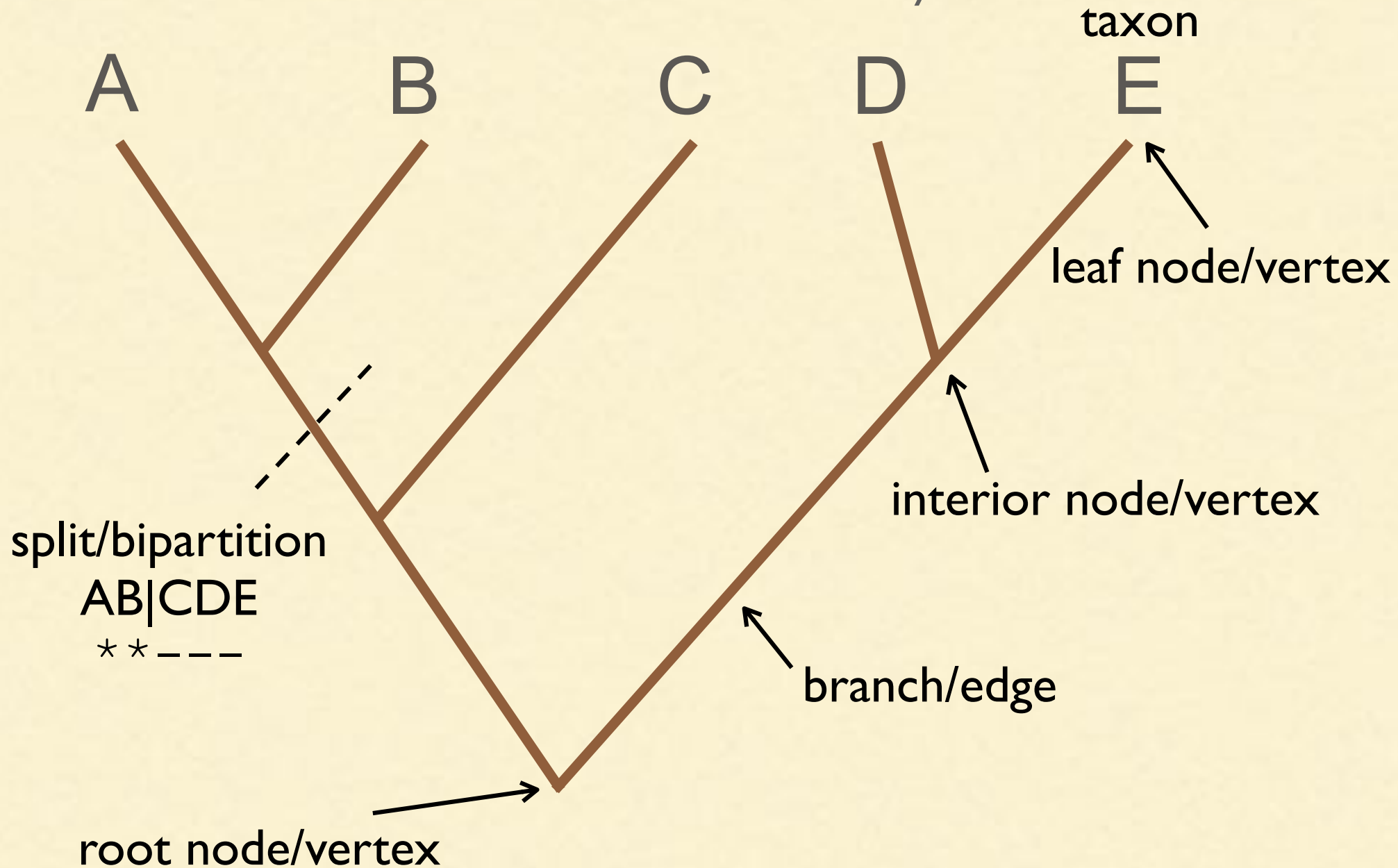
90% of
bootstrap
replicates



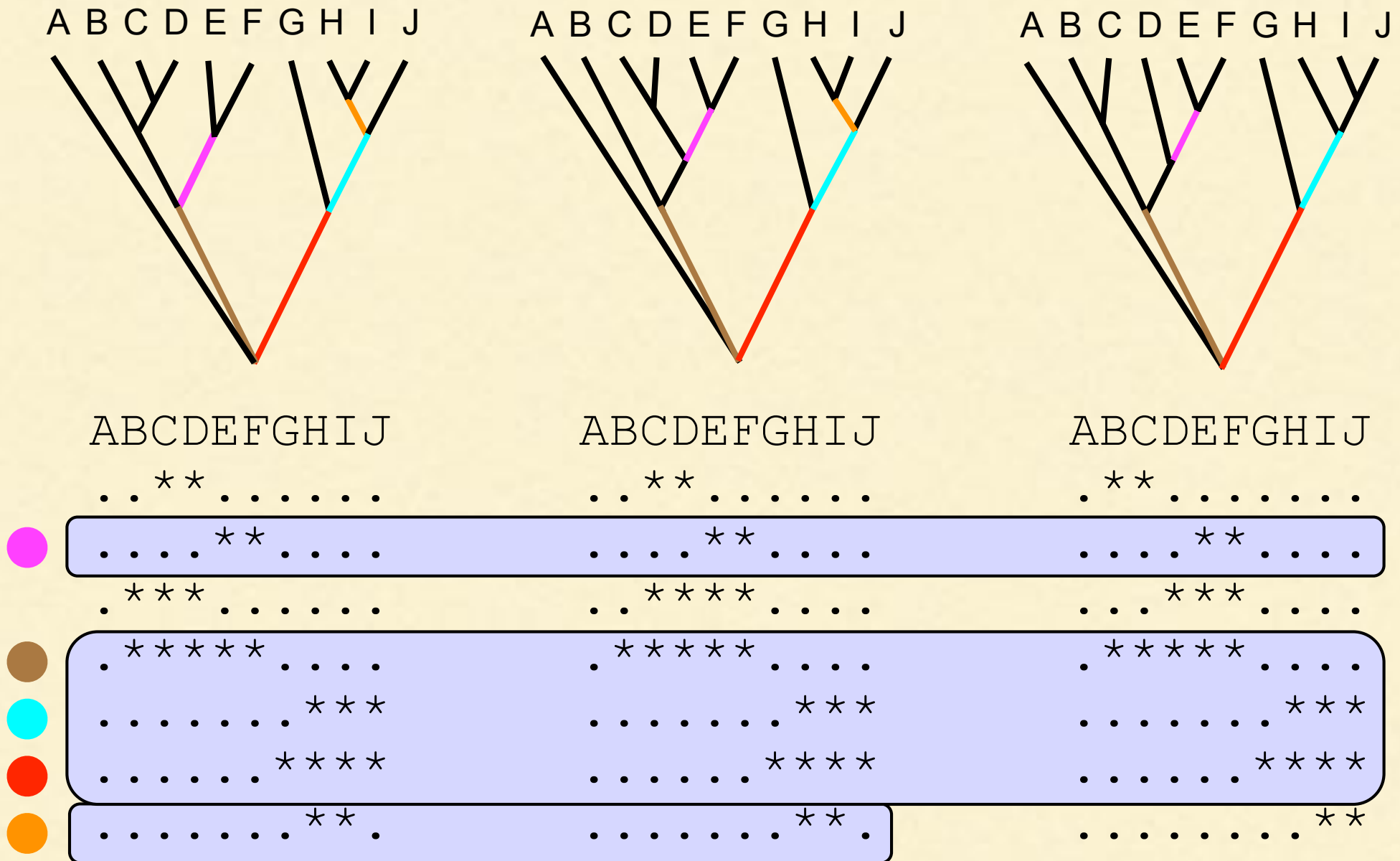
10% of
bootstrap
replicates

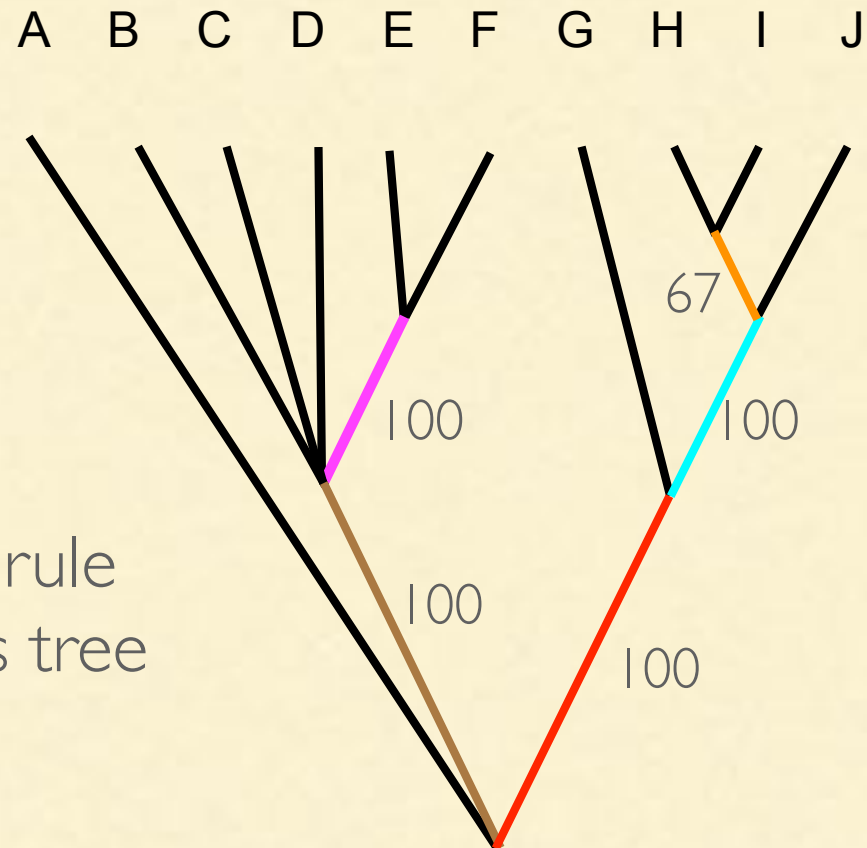
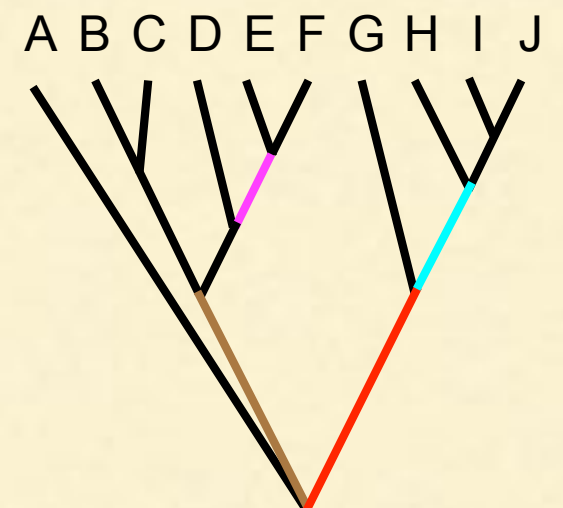


Tree anatomy



Consensus trees

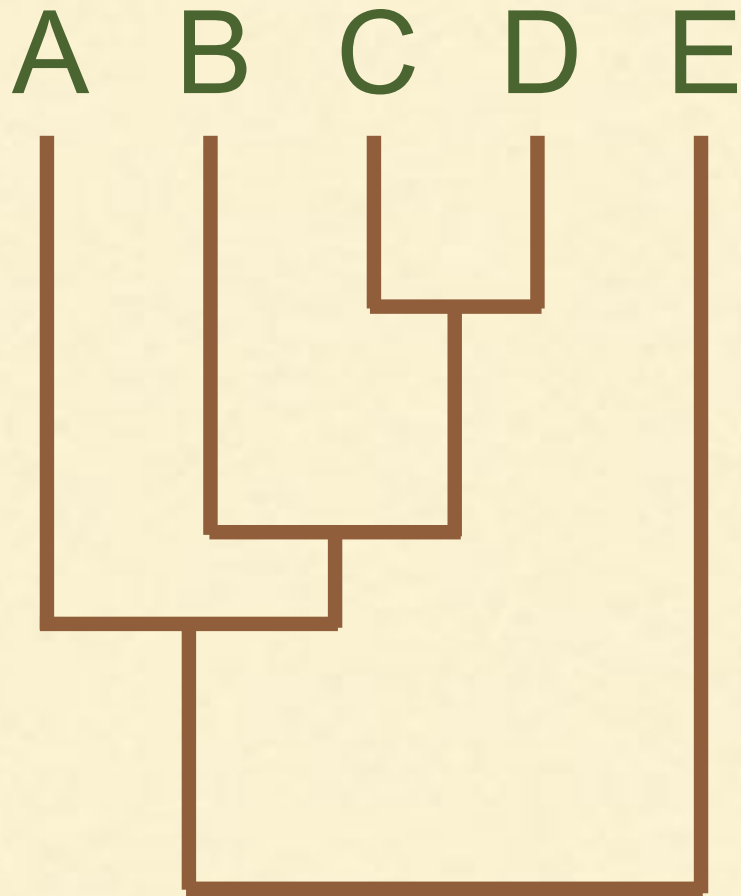




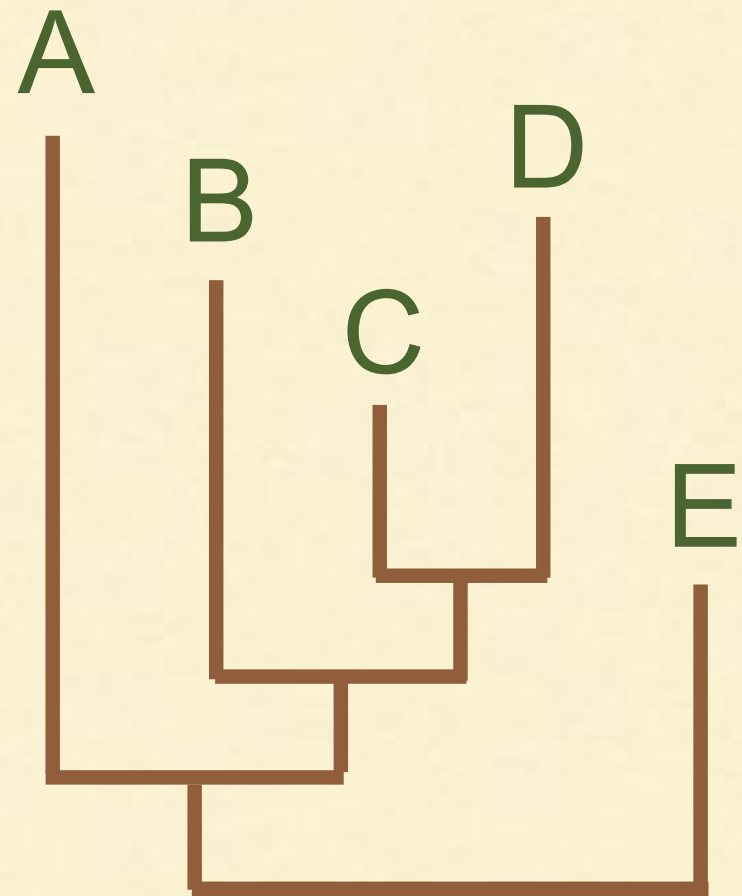
majority rule
consensus tree

% of input trees

Edge lengths



edge lengths are
time only



edge lengths are
rate x time

Newick descriptions

#NEXUS

Begin trees;

Translate

```
1 Chlamydomodium_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
```

;

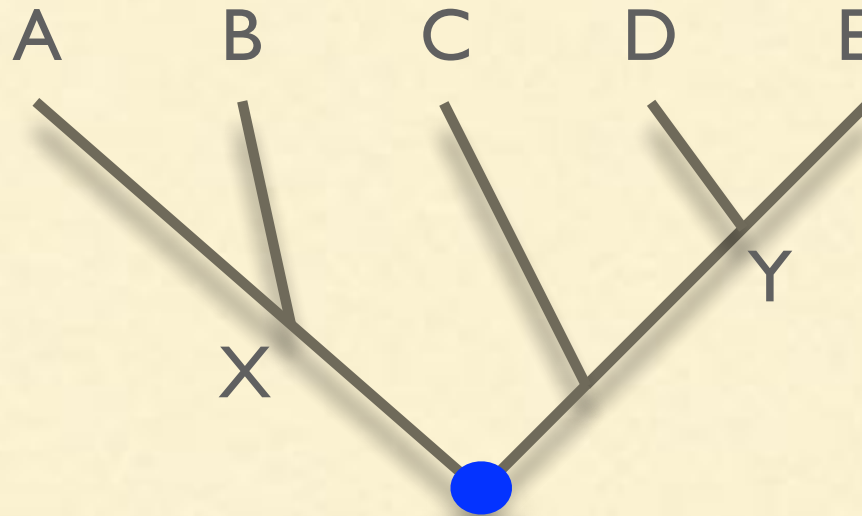
```
tree 'PAUP_1' = [&U] (1:0.104899,((2:0.009446,  
(4:0.001635,6:7.29892e-07):0.030410):0.005612,3:0.007100):0.002552,5:0.001416);
```

End;



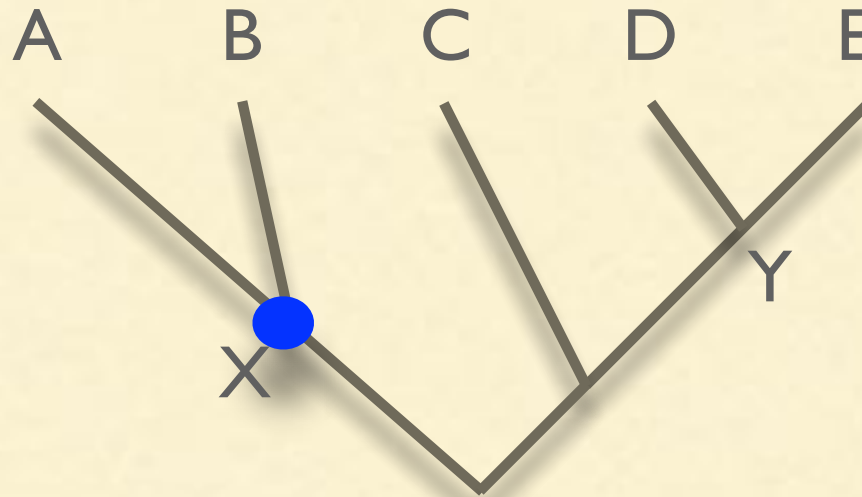
https://en.wikipedia.org/wiki/Newick_format

Newick tree descriptions



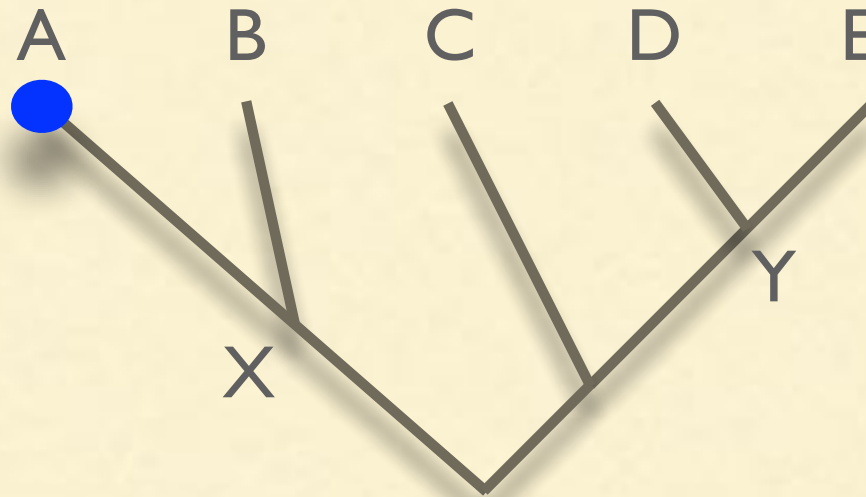
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



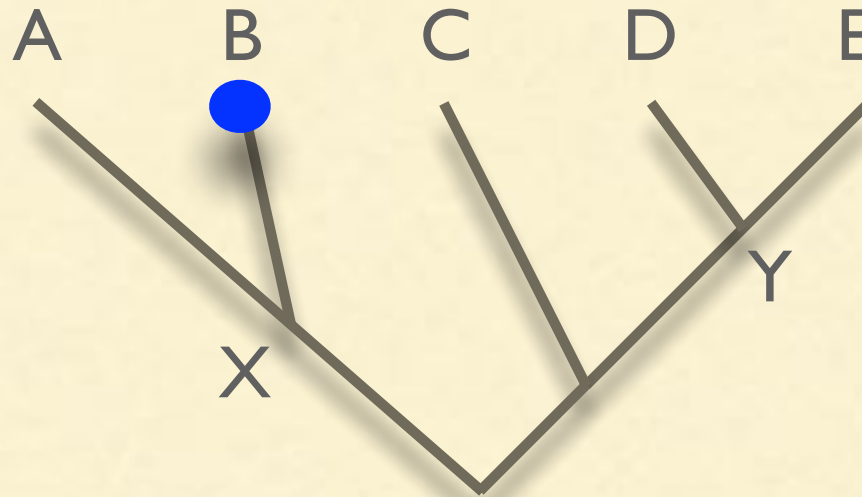
((A,B)X,(C,(D,E)Y)**)**

Newick tree descriptions



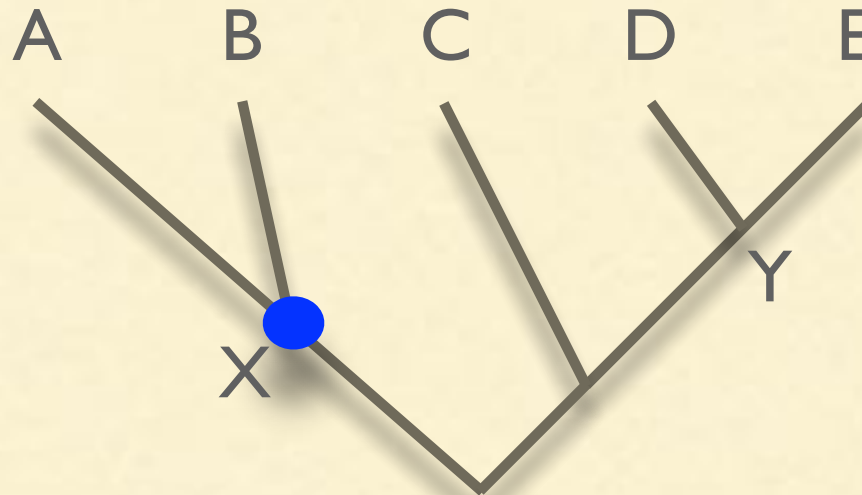
((A,B)X,(C,(D,E)Y))

Newick tree descriptions



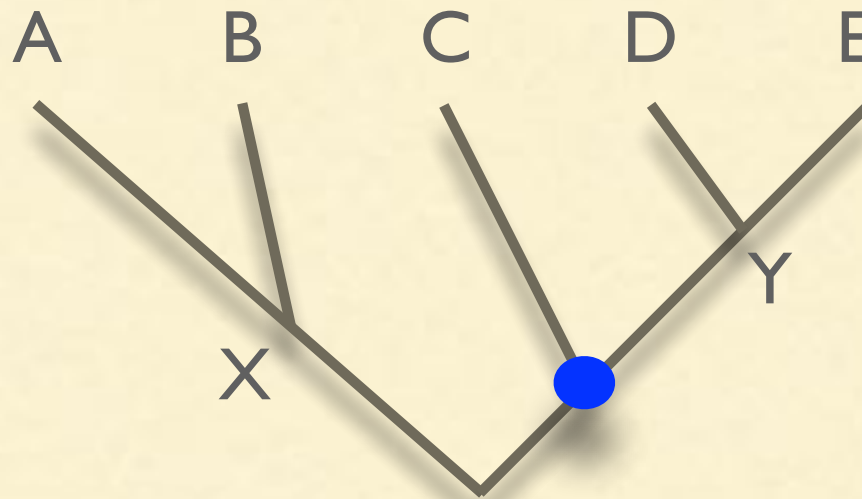
$((A, \mathbf{B})X, (C, (D, E)Y))$

Newick tree descriptions



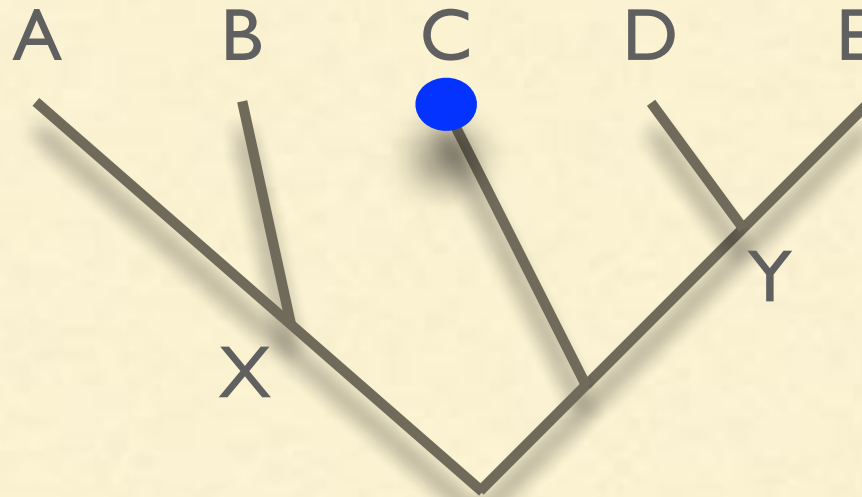
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



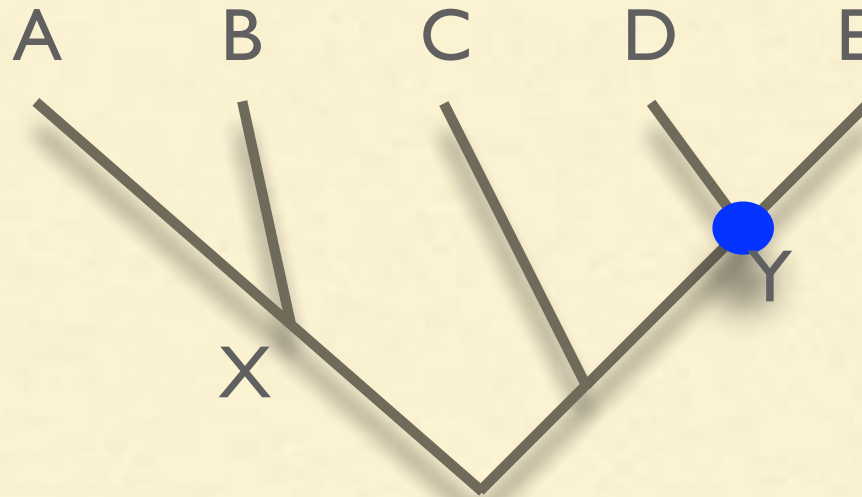
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



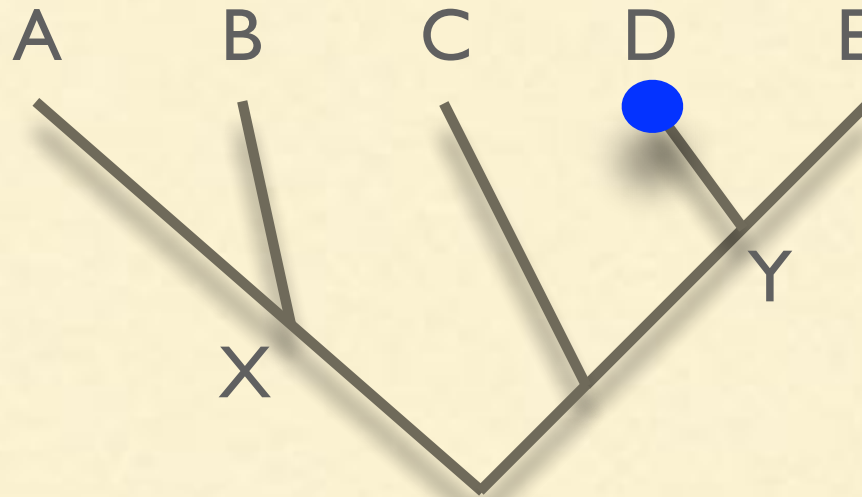
$((A,B)X, (C,(D,E)Y))$

Newick tree descriptions



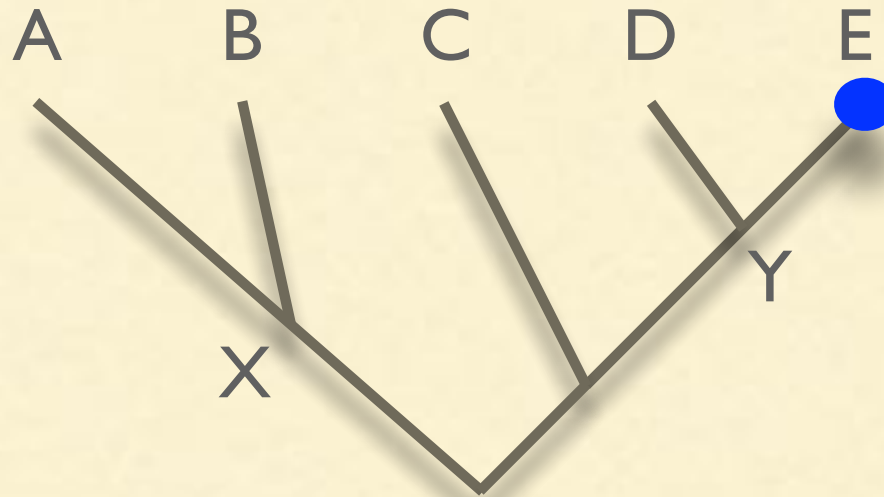
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



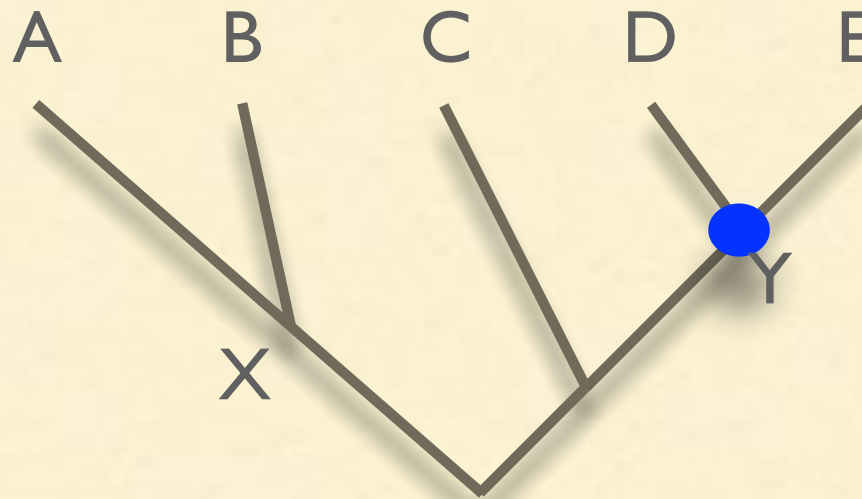
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



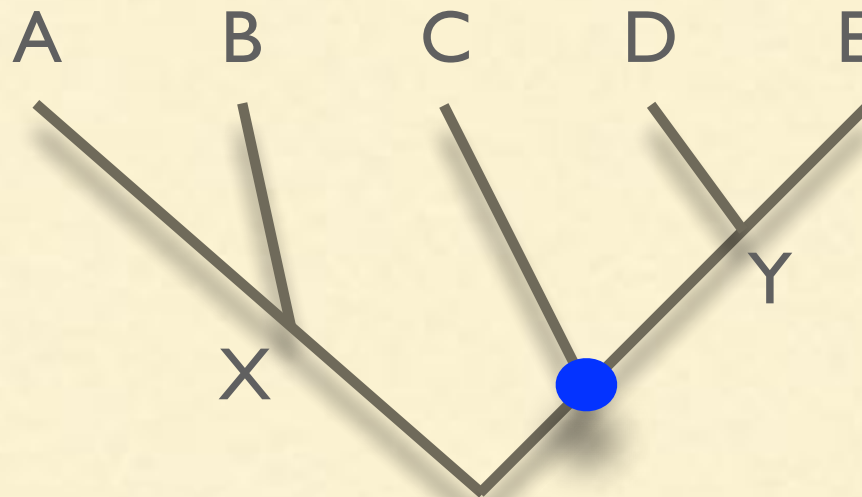
$((A,B)X,(C,(D,\textcolor{blue}{E})Y))$

Newick tree descriptions



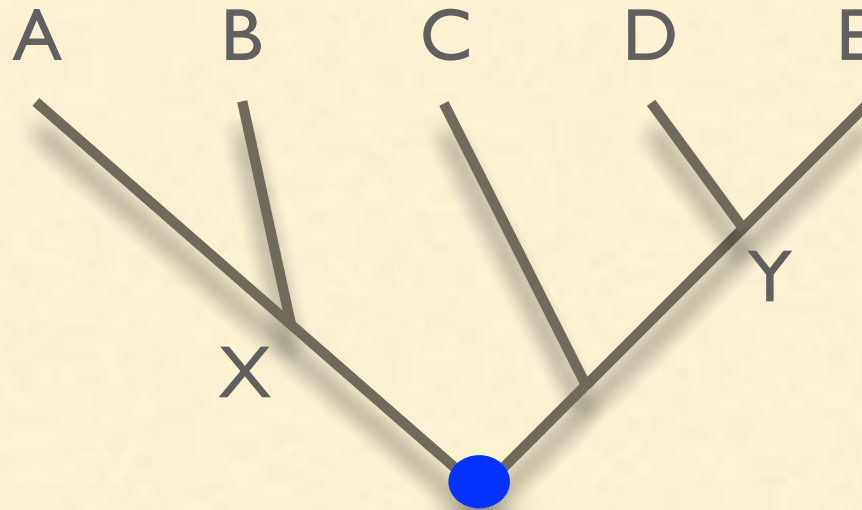
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



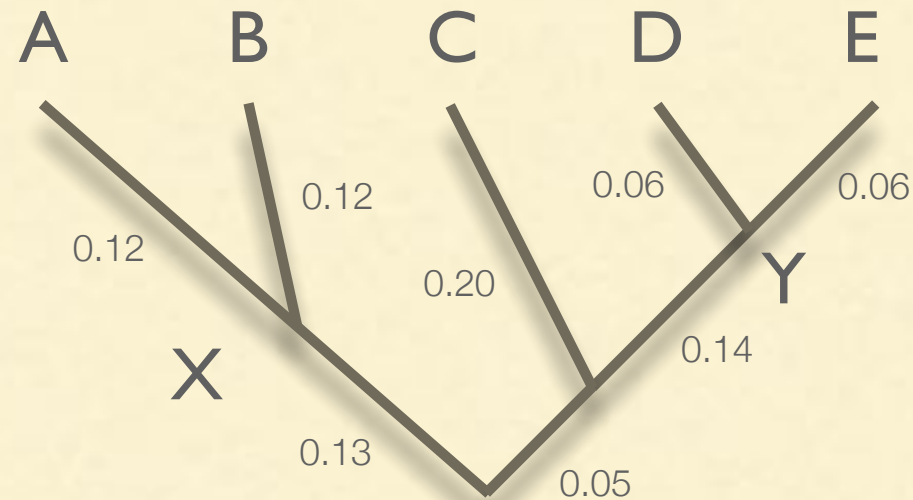
$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions



$((A,B)X,(C,(D,E)Y))$

Newick tree descriptions

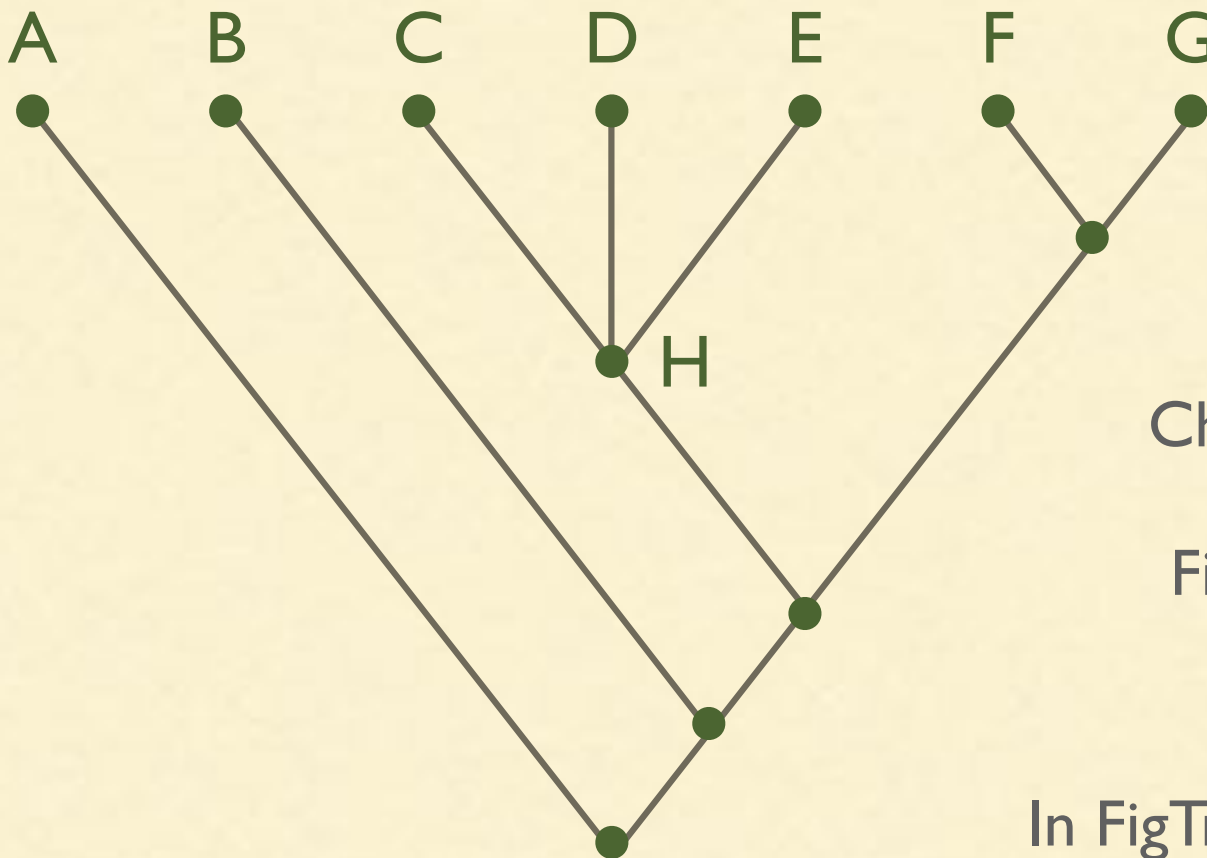


`((A:.12,B:.12)X:.13,(C:.2,(D:.06,E:.06)Y:.14):.05)`

edge lengths follow colon after node name (if present)

Newick challenge

Create a newick tree description for this tree (just the topology, no branch lengths)

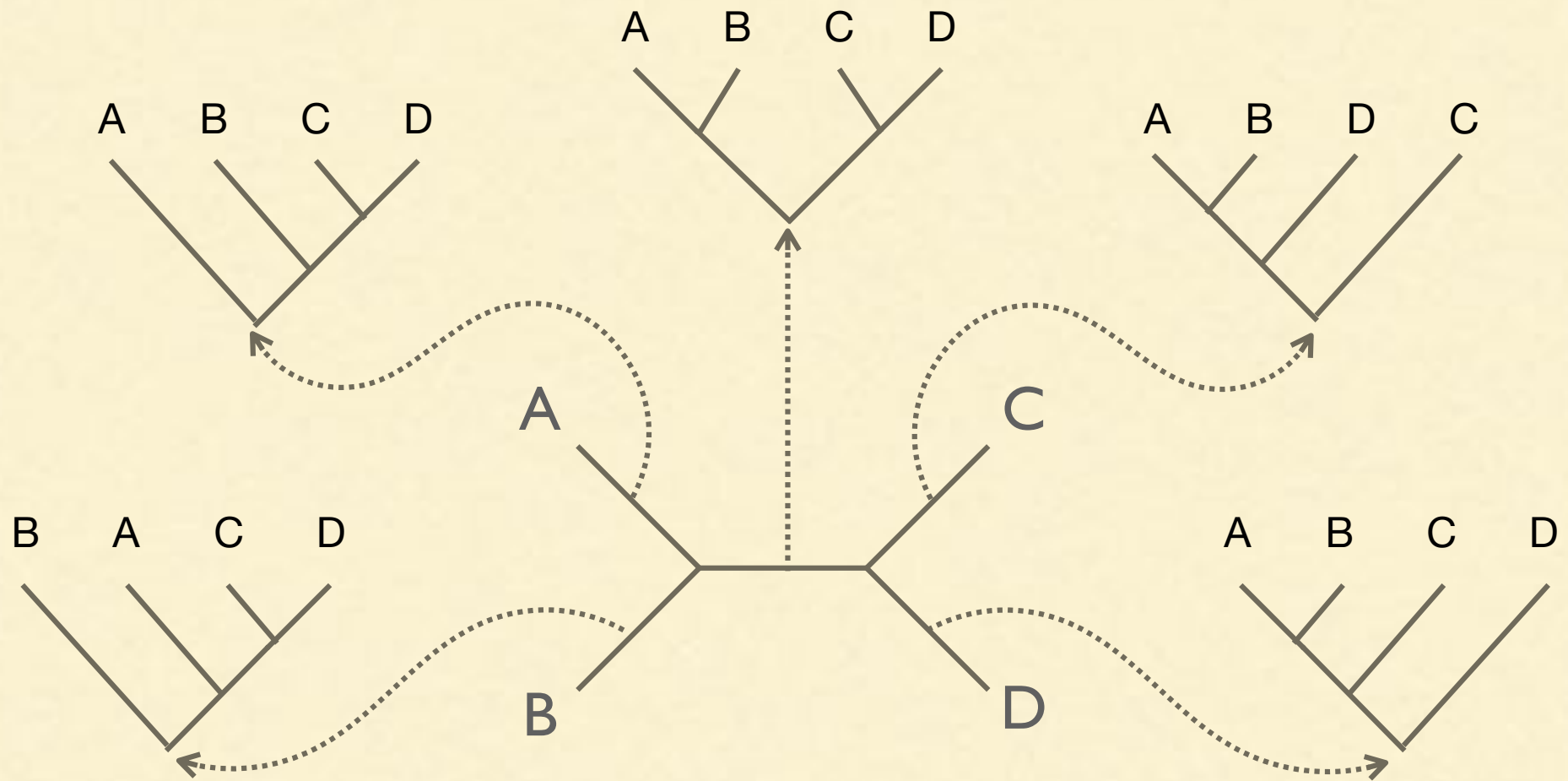


Check your work by pasting
your newick string into
FigTree ([https://github.com/
rambaut/figtree/releases](https://github.com/rambaut/figtree/releases))

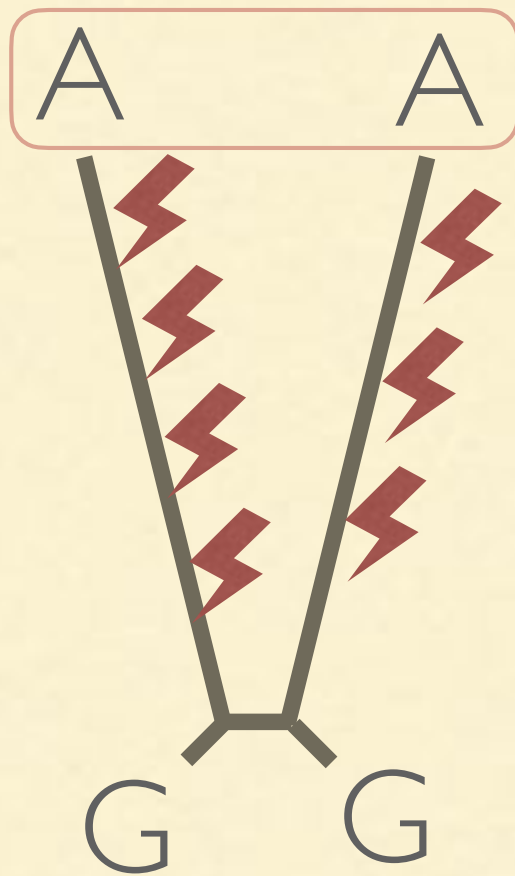
In FigTree, check Node labels and
choose display = label" to see the "H"

Rooted vs unrooted

rooting and adding a
taxon increase
treespace by the same
amount



Challenges: model violations



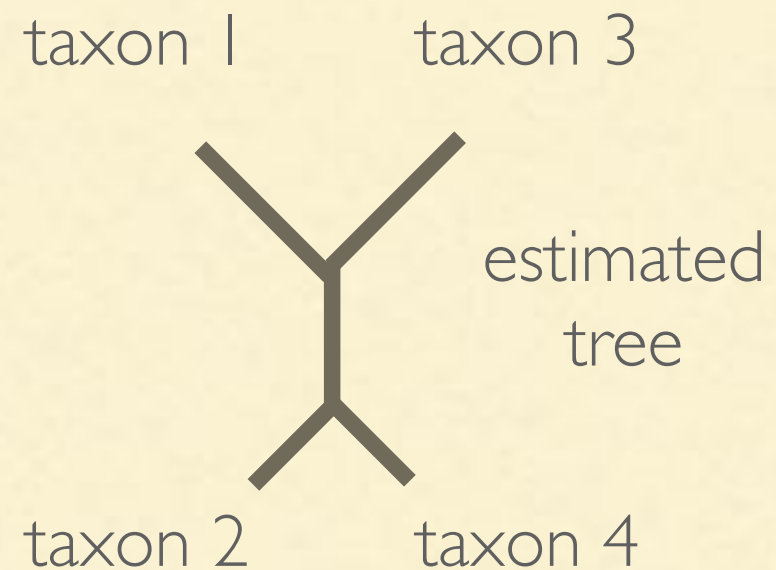
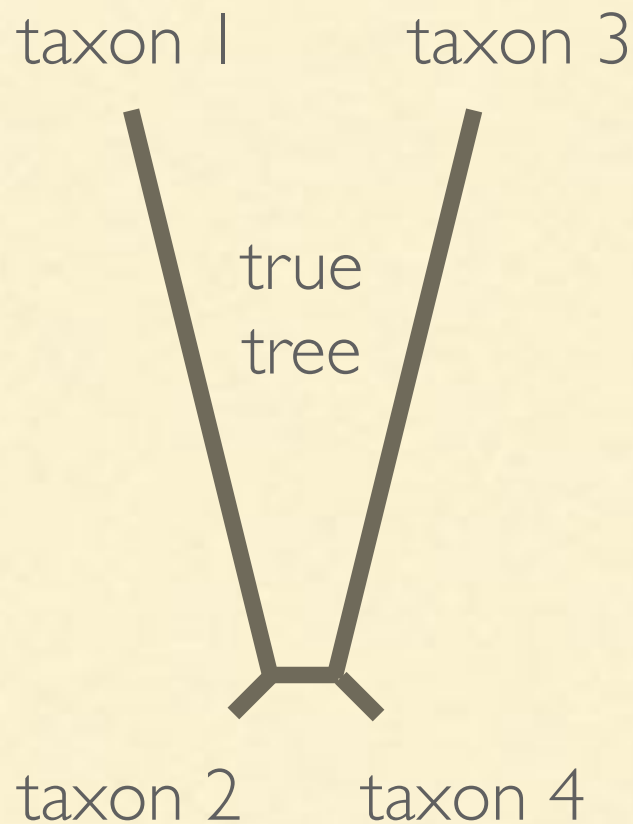
Long external branches favor
a **convergence explanation**
of this similarity

Challenges: model violations



← Short external branches favor
an **inheritance explanation**
of this similarity

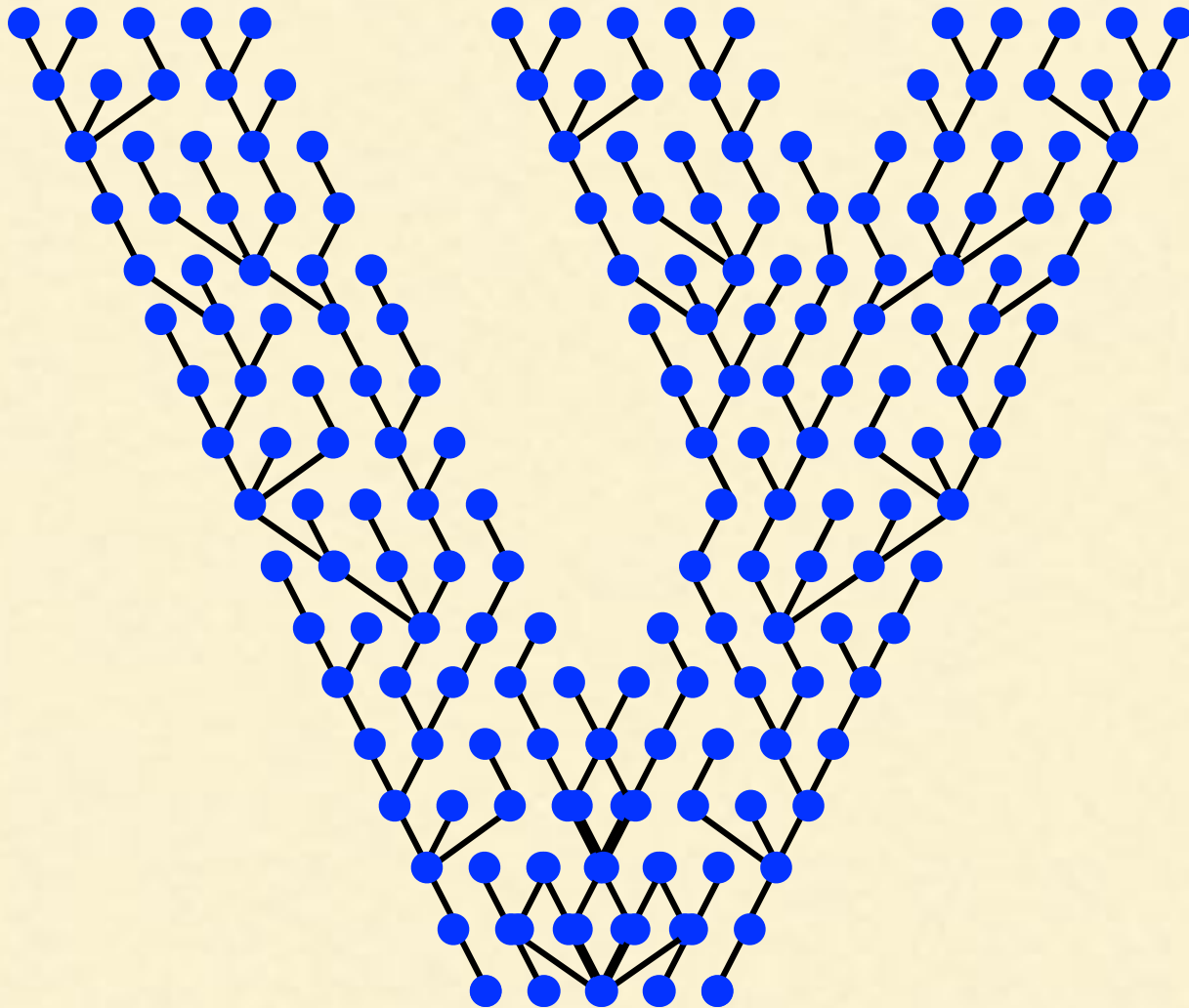
Challenges: model violations



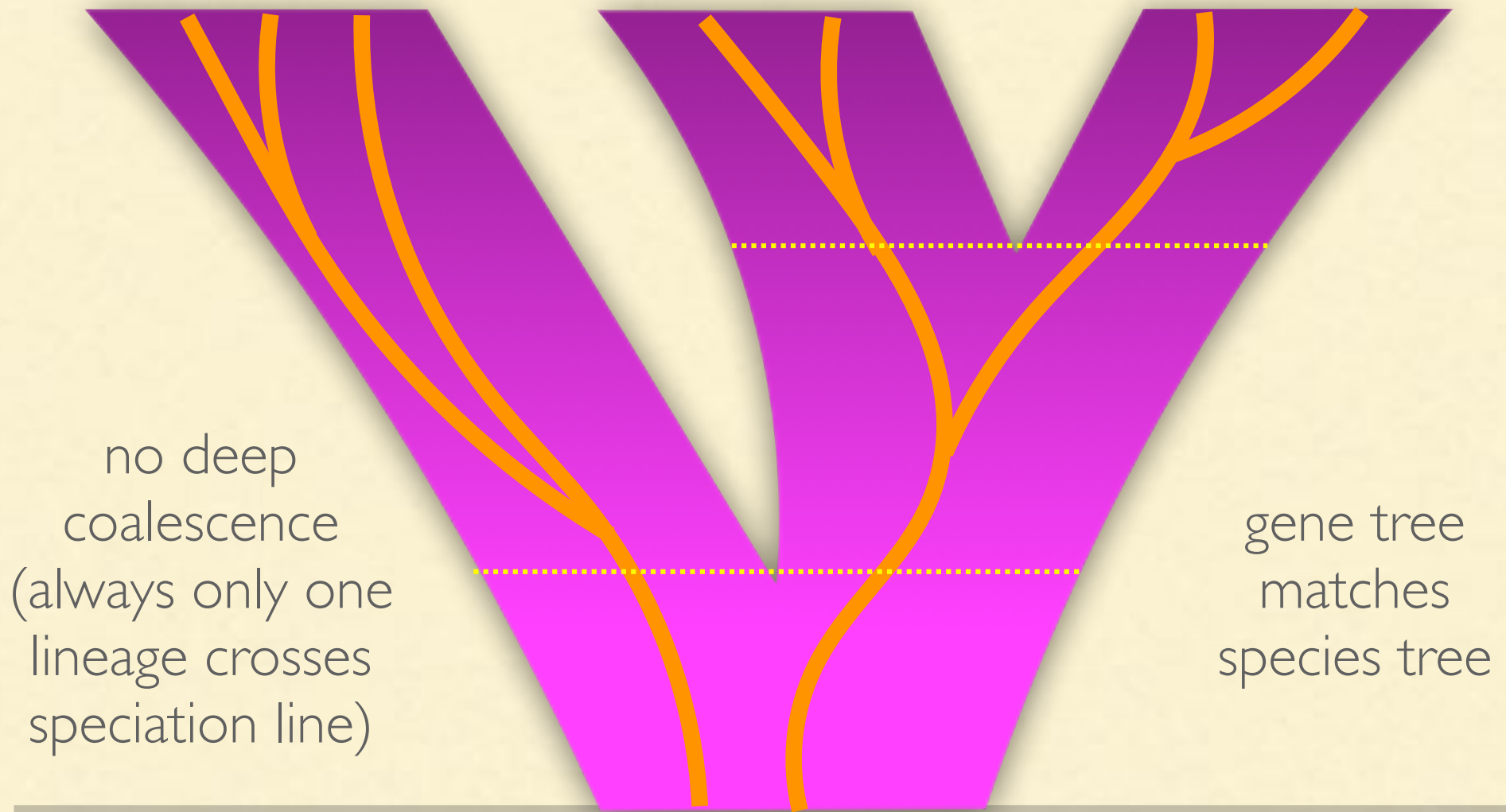
Models that are too simple often underestimate branch lengths

Long branch attraction

Challenges: deep coalescence



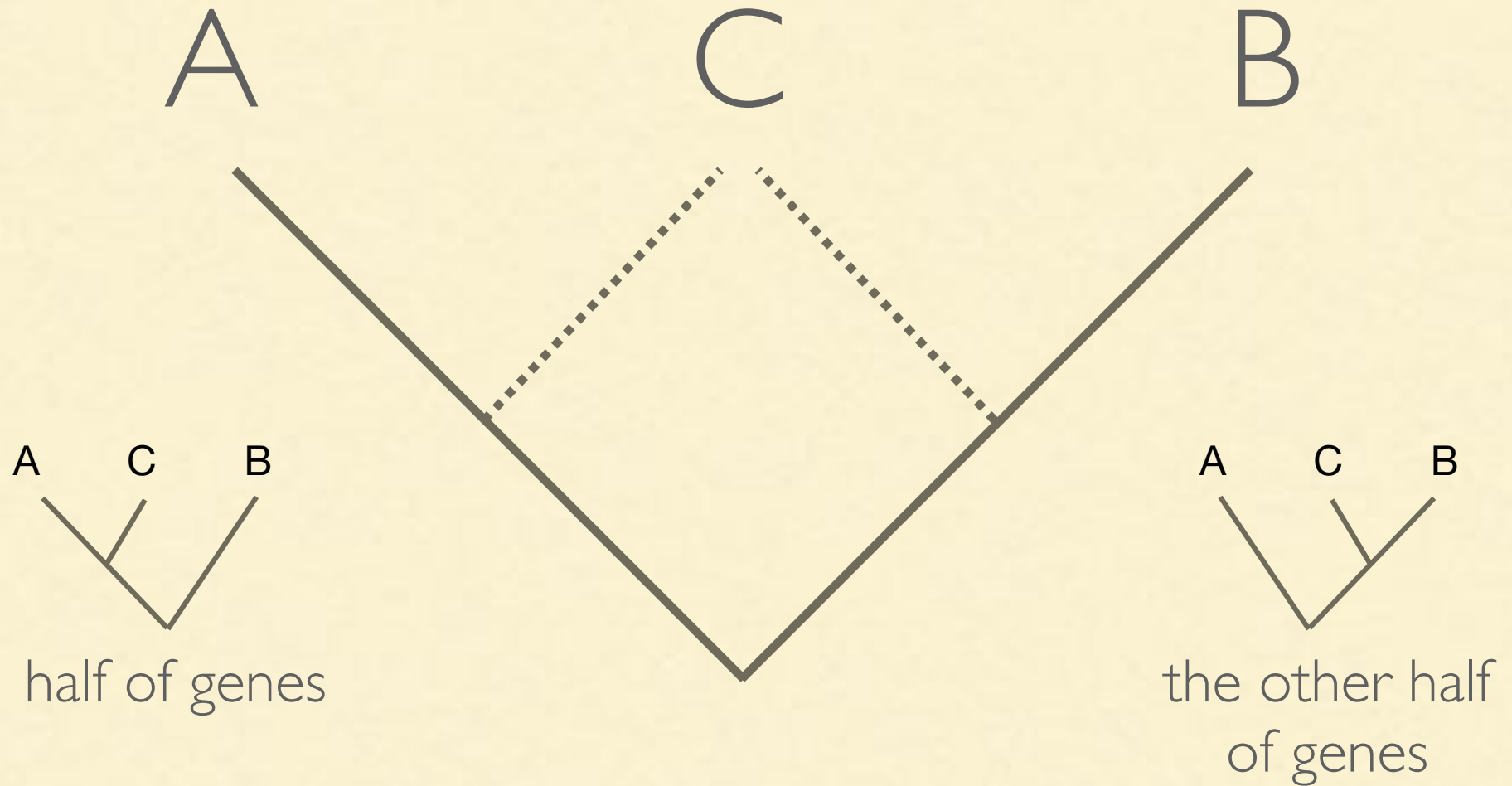
Challenges: deep coalescence



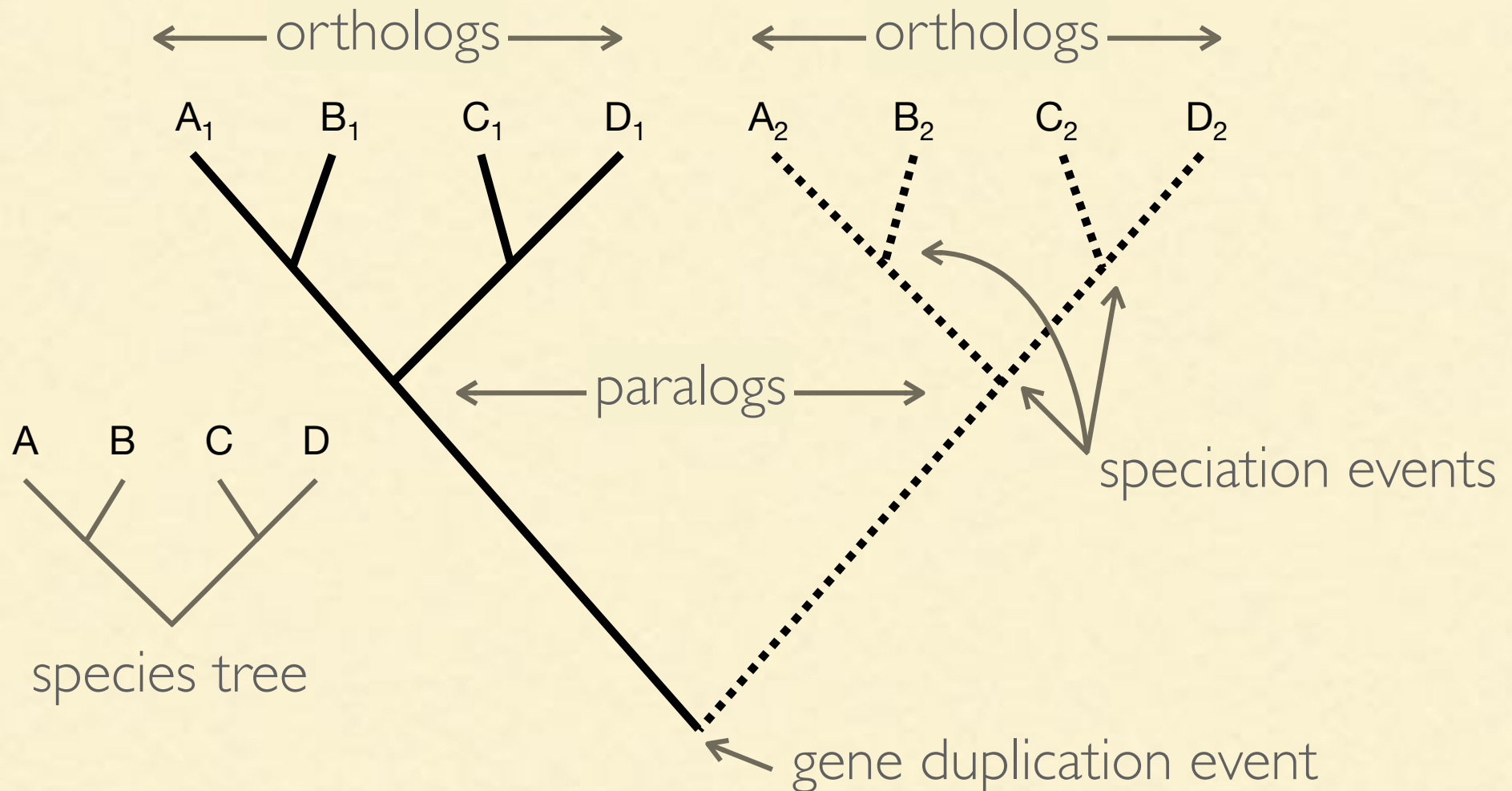
Challenges: deep coalescence



Challenges: hybridization

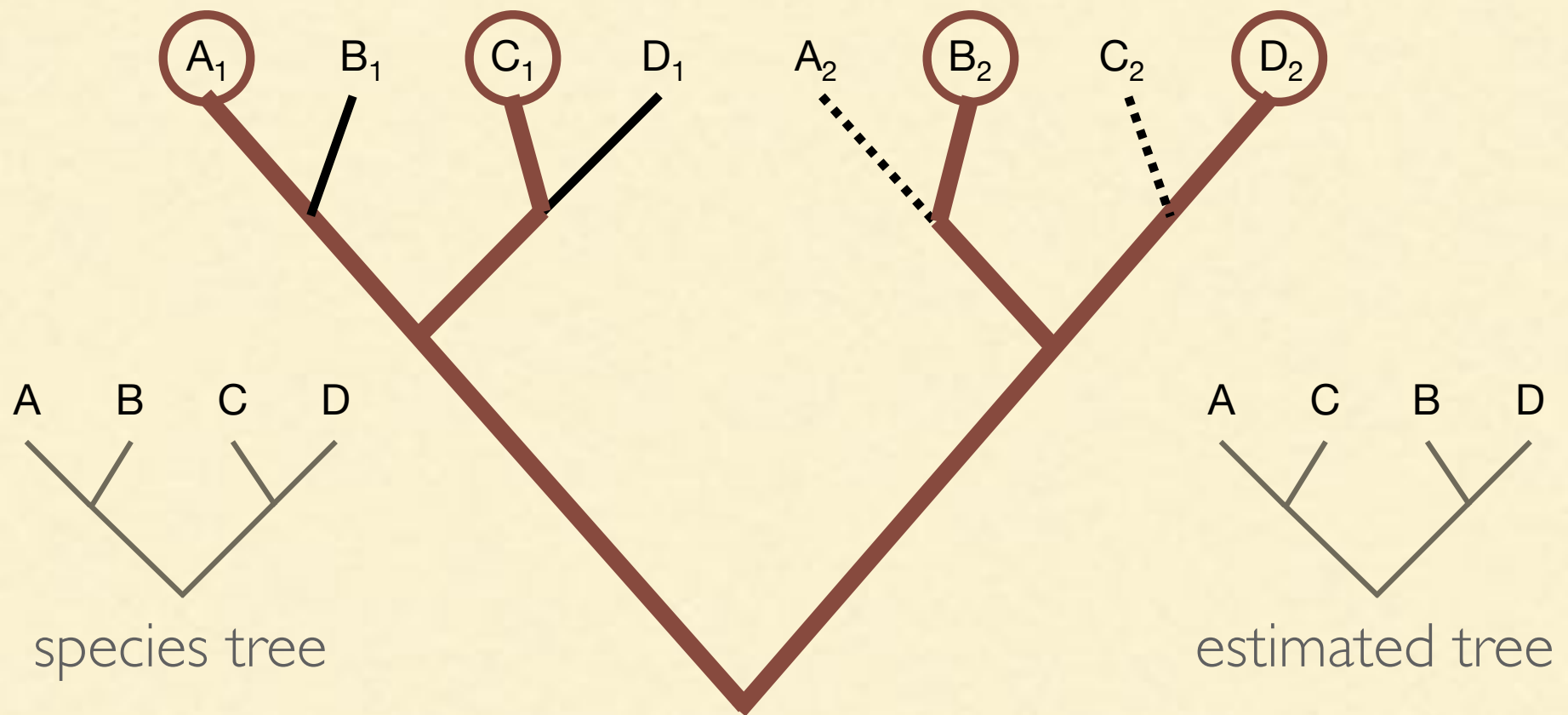


Challenges: paralogy

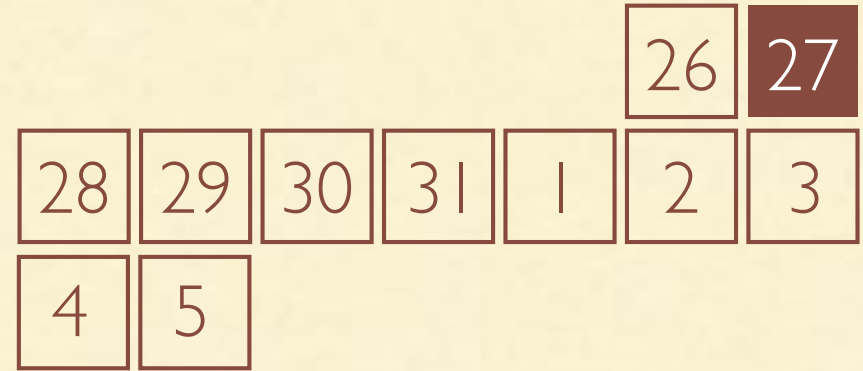


Challenges: paralogy

sampled sequences are a mixture of orthologs and paralogs



Overview of the Workshop



Intro to phylogenetics, likelihood and likelihood models:

Today (Saturday): **Lewis, Huelsenbeck**

Computing introduction, sequence alignment:

Tonight: **Satler, Taylor, Fauskee, Milkey, Kong**



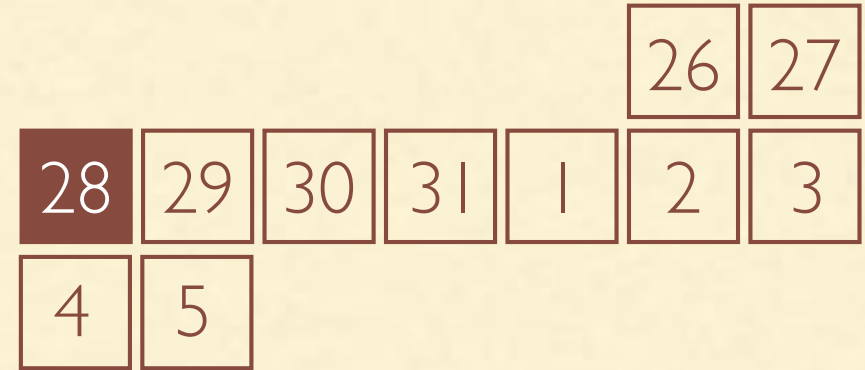
Under the hood

					26	27
28	29	30	31	1	2	3
4	5					

C++ Programming subworkshop (optional):

Mornings 8-9am: **Huelsenbeck**

Model selection and maximum likelihood



Model selection:

Sunday morning: **Lewis, Swofford**

PAUP* lab:

Sunday afternoon: **Swofford**

IQ-TREE: ML inference on a large scale

Sunday evening: **Bui**

RevBayes

					26	27
28	29	30	31	1	2	3
4	5					

Introduction to Bayesian statistics

Monday morning: **Lewis**

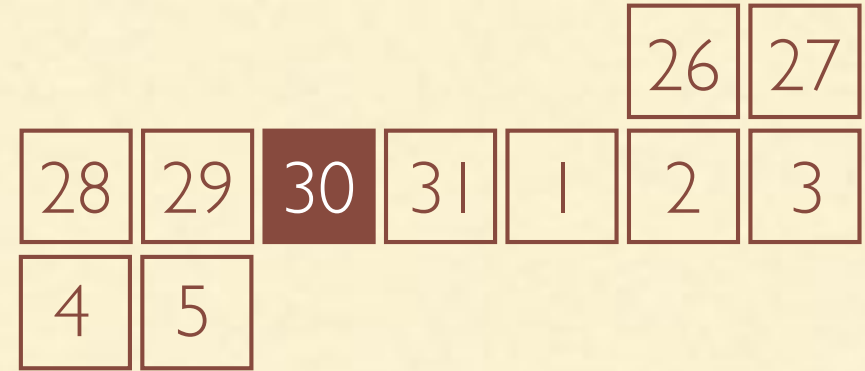
RevBayes: Graphical models, tree estimation:

Monday afternoon: **Brown**

RevBayes: Divergence time estimation:

Monday evening lecture/lab: **Heath**

Coalescence, species trees, and networks



Introduction to coalescent theory:

Tuesday morning: **Beerli**

Species tree estimation lab:

Tuesday afternoon: **Swofford (Kubatko)**

Networks and hybridization lab:

Tuesday evening: **Solís-Lemus**

Phylogeography, dinner
party, and the free day!



Phylogeography, species trees vs. gene trees:

Wednesday morning/afternoon: **Edwards, Yoder**

Course **Dinner Party**

Wednesday evening

Free day: Thursday all day

Sleep, visit Martha's Vineyard, whale watching...

Selection

					26	27
28	29	30	31	1	2	3
4	5					

Selection and codon models:

Friday morning: **Bielawski**

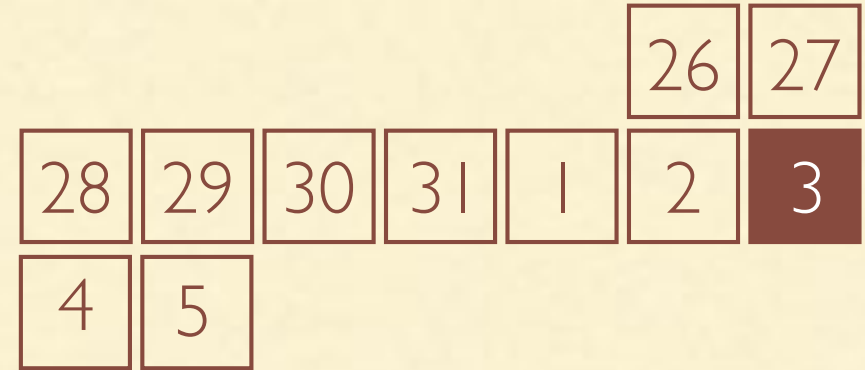
Adaptive protein evolution:

Friday afternoon: **Chang**

PAML lab:

Friday evening: **Bielawski**

Machine learning, amino
acid models, migration



Machine learning:

Saturday morning: **Smith**

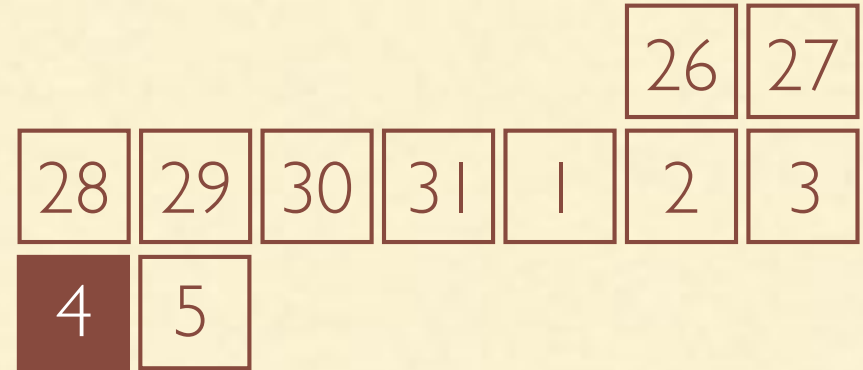
Amino acid models, topology tests:

Saturday afternoon: **Susko**

MIGRATE: population structure and migration

Saturday evening: **Beerli**

Deep time, updating, and capstone



Amino acid models and deep time phylogenetics:

Sunday morning: **Eme**

Open Tree of Life, phylogenomics, gene tree updating:

Sunday afternoon: **McTavish**

Capstone: Evolutionary applications of genomics

Sunday evening: **Knowles**

Ethics, open lab

					26	27
28	29	30	31	1	2	3
4	5					

Scientific ethics:

Monday morning: **Swofford, Bielawski**

Open lab:

Your last chance to ask questions