Workshop on Molecular Evolution

Marine Biological Laboratory, Woods Hole, Massachusetts

27 May - 6 June, 2022

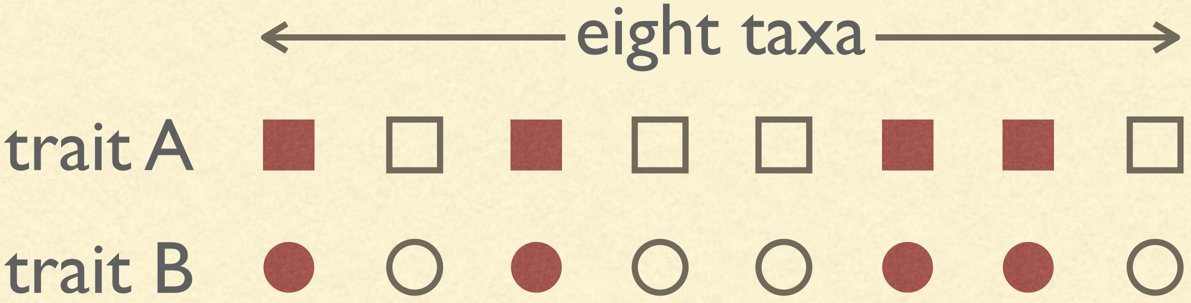
Paul O. Lewis
Department of Ecology & Evolutionary Biology
UCONN
UNIVERSITY OF CONNECTICUT

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



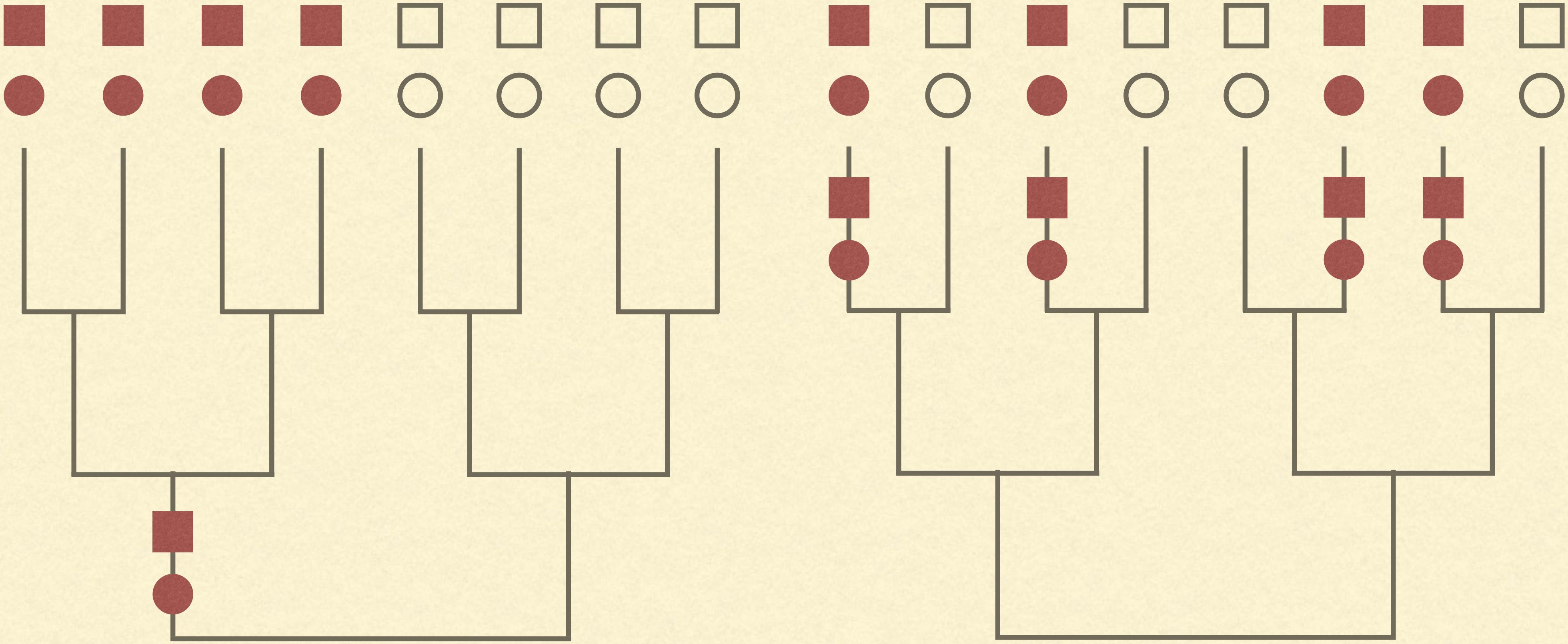
eight taxa

trait A ■ □ ■ □ □ ■ ■ □

trait B ● ○ ● ○ ○ ● ● ○

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

Two very different explanations



Simple inheritance

Correlated evolution

Overview

					27	28
29	30	31	1	2	3	4
5	6					

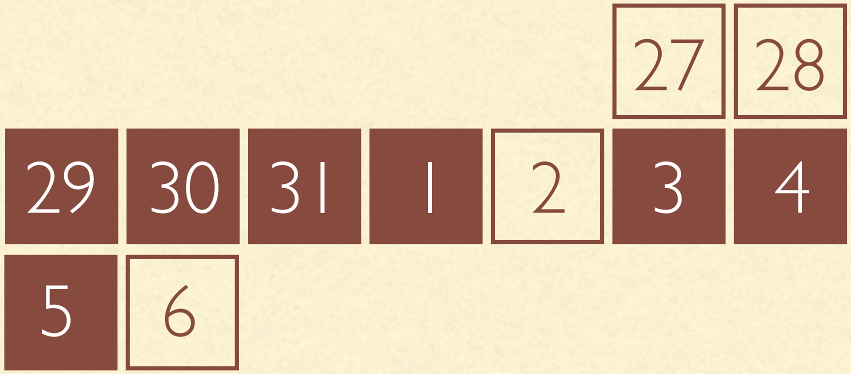
Intro to phylogenetics, likelihood and likelihood models:

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

Computing introduction, sequence alignment:

Tonight: **Láruson**, **Gonçalves**, **Taylor**, **Satler**

Under the hood



27 28

29 30 31 1 2 3 4

5 6

C++ Programming subworkshop (optional):

Mornings 8-9am: **Huelsenbeck**

Model selection

					27	28
29	30	31	1	2	3	4
5	6					

Model selection:

Sunday morning: **Lewis, Swofford**

PAUP* lab:

Sunday afternoon: **Swofford**

Bayesian statistics:

Sunday evening: **Lewis**

RevBayes

					27	28
29	**30**	31	1	2	3	4
5	6					

Graphical models, tree estimation:

Monday morning: **Brown**

Divergence time estimation:

Monday afternoon lecture/evening lab: **Heath**

Coalescence, phylogenomics

					27	28
29	30	31	1	2	3	4
5	6					

Introduction to coalescent theory:

Tuesday morning: **Beerli**

Open Tree of Life, gene tree updating lab:

Tuesday afternoon: **McTavish**

Phylogenomics:

Tuesday evening: **McTavish**

Phylogeography

					27	28
29	30	31	1	2	3	4
5	6					

Phylogeography, species trees vs. gene trees:

Wednesday June 1: **Edwards**, **Yoder**

Course **Dinner Party**

Free day: Thursday, June 2

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

Species trees, networks, migration

					27	28
29	30	31	1	2	**3**	4
5	6					

Species tree estimation lab:

Friday morning: **Swofford** (**Kubatko**)

Networks and hybridization lab:

Friday afternoon: **Solís-Lemus**

MIGRATE lab:

Friday evening: **Beerli**

Selection

					27	28
29	30	31	1	2	3	4
5	6					

Selection and codon models:

Saturday morning: **Bielawski**

Adaptive protein evolution:

Saturday afternoon: **Chang**

PAML lab:

Saturday evening: **Bielawski**

Species trees, networks, migration

					27	28
29	30	31	1	2	3	4
5	6					

Large tree maximum likelihood inference lab:

Sunday morning: **Bui**

Amino acid models, topology tests:

Sunday afternoon: **Susko**

Capstone: Evolutionary applications of genomics

Sunday evening: **Knowles**

Species trees, networks, migration

					27	28
29	30	31	1	2	3	4
5	**6**					

Scientific ethics:

Monday morning: **Swofford**, **Bielawski**

Open lab:

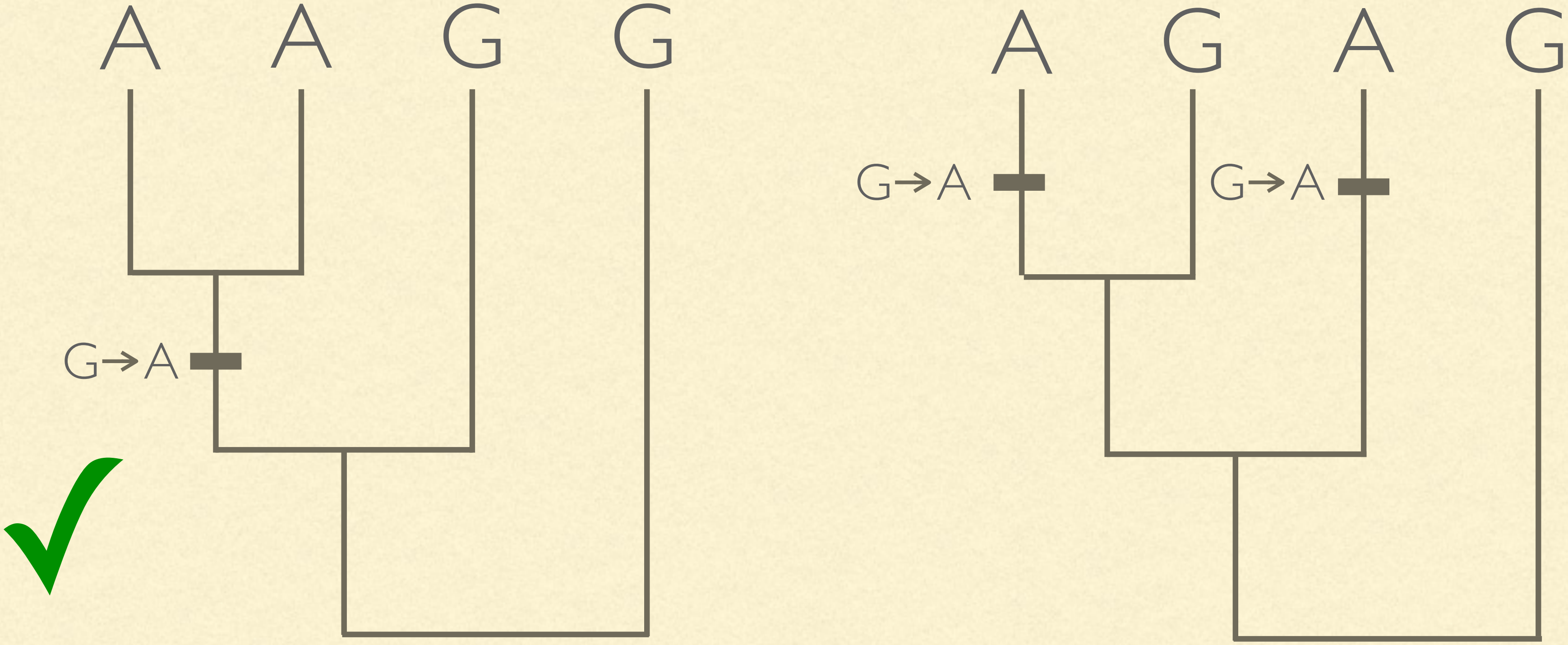
Your last chance to ask questions

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?

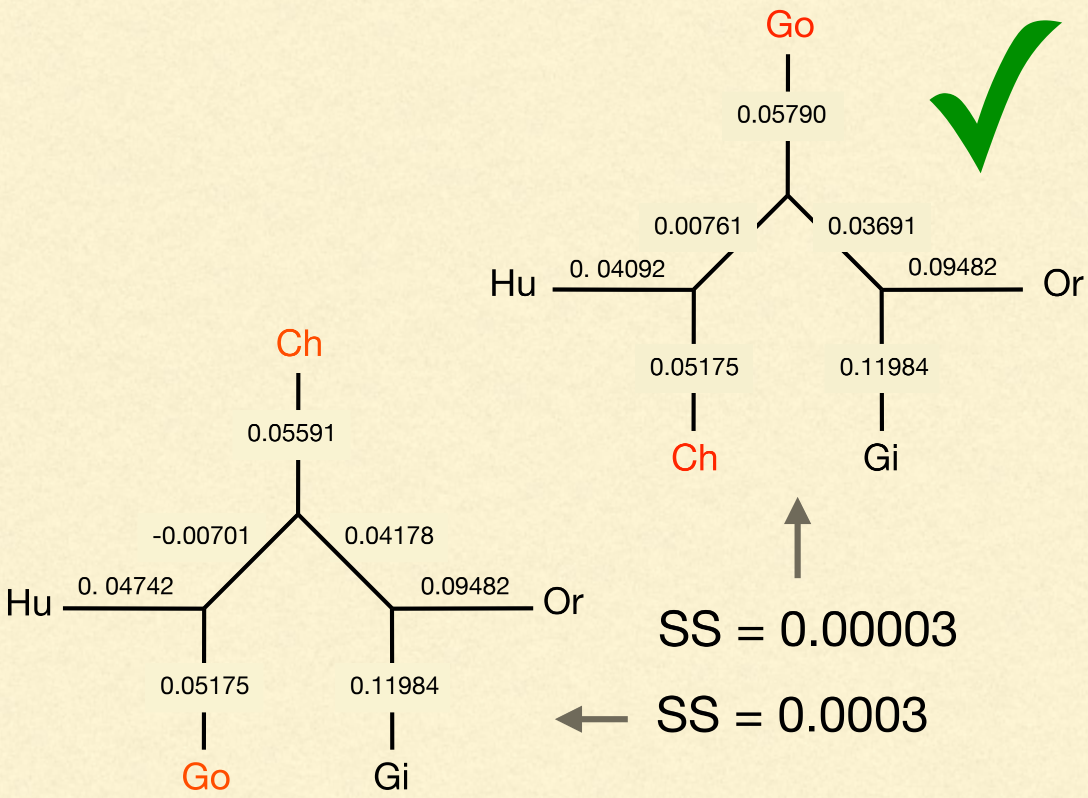


Parsimony criterion says tree requiring fewer changes is better

Which tree is better?

$(0.10928 - 0.10643)^2$

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



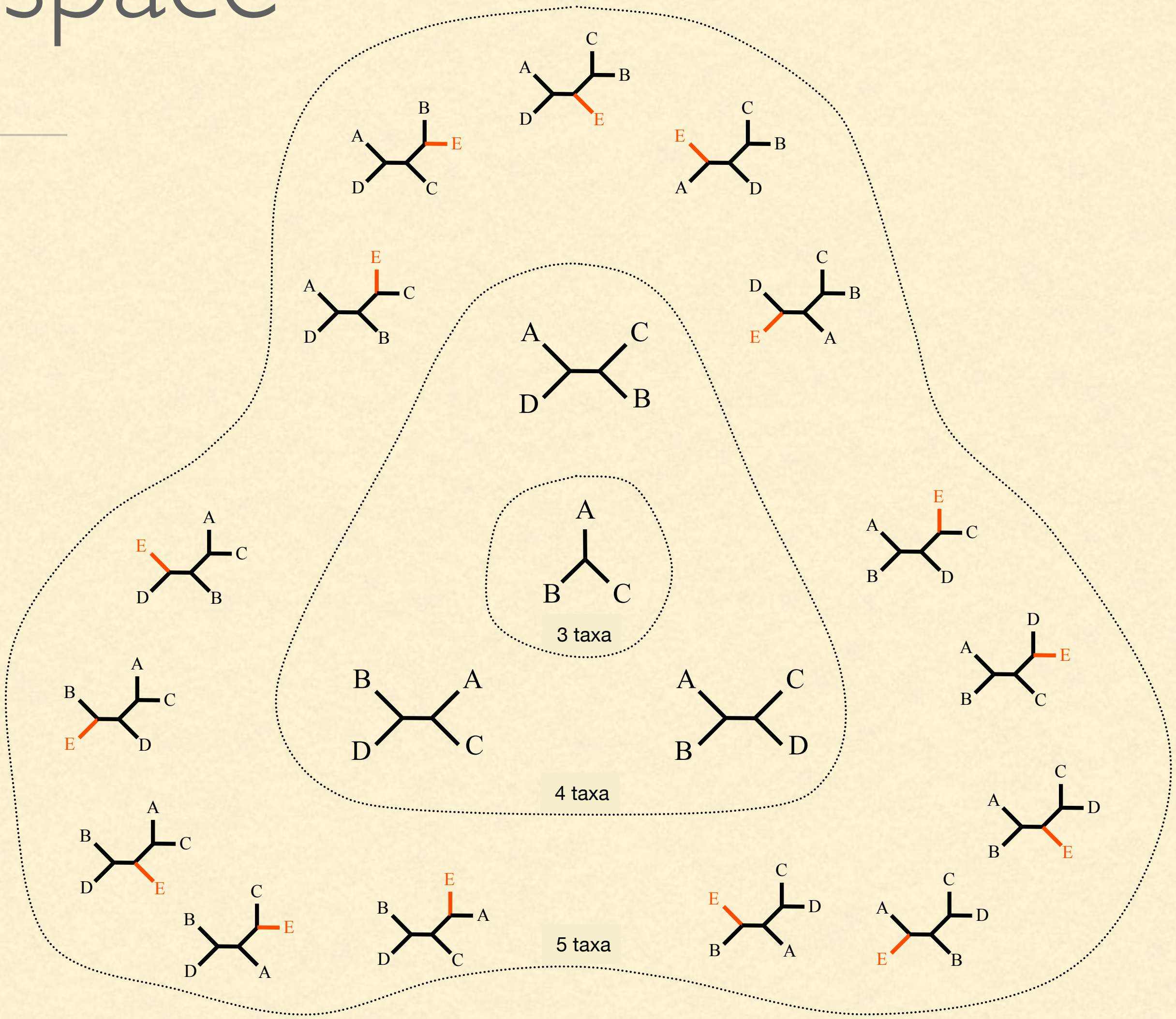
SS = 0.00003

SS = 0.0003

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

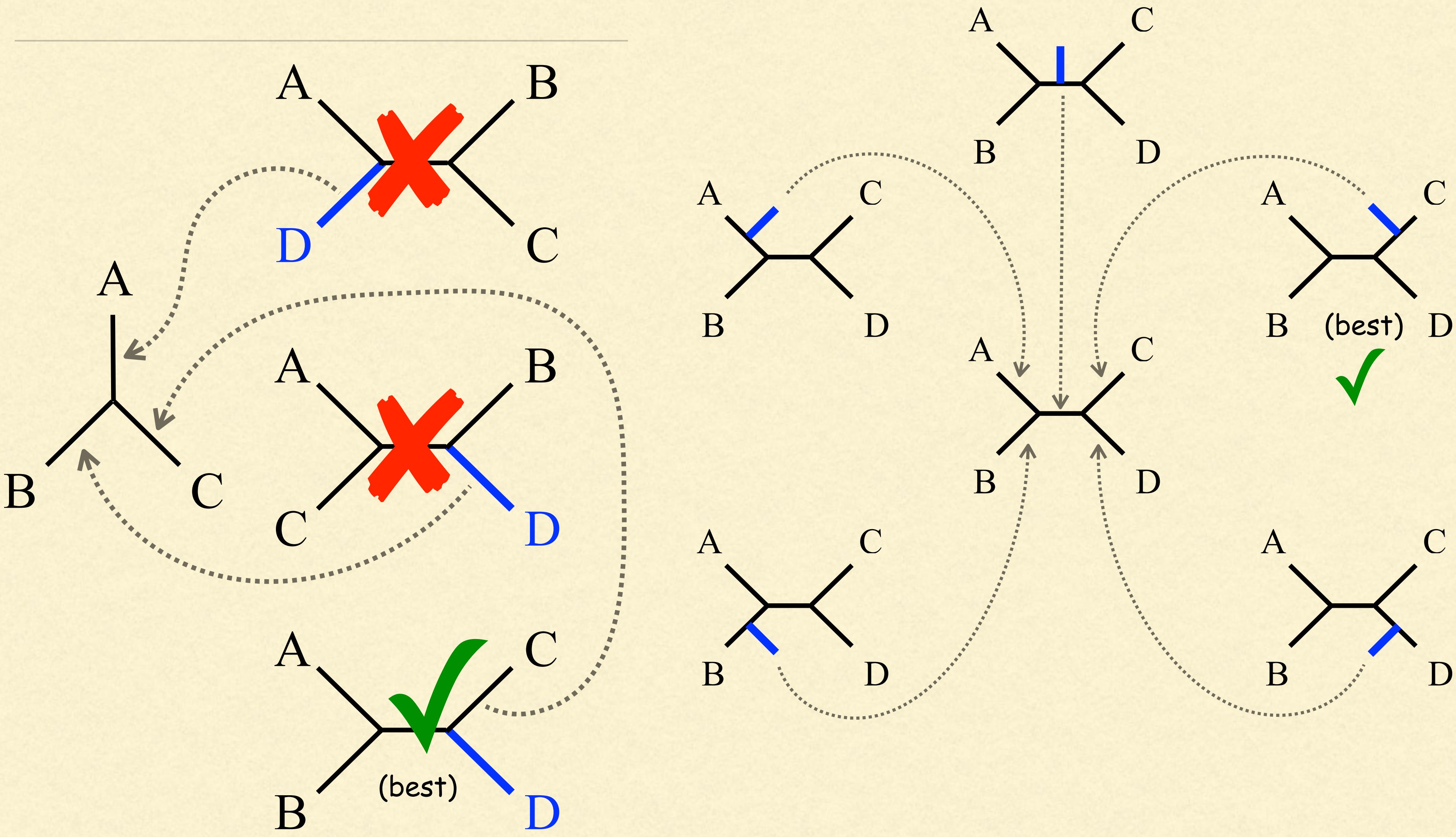
Searching treespace

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

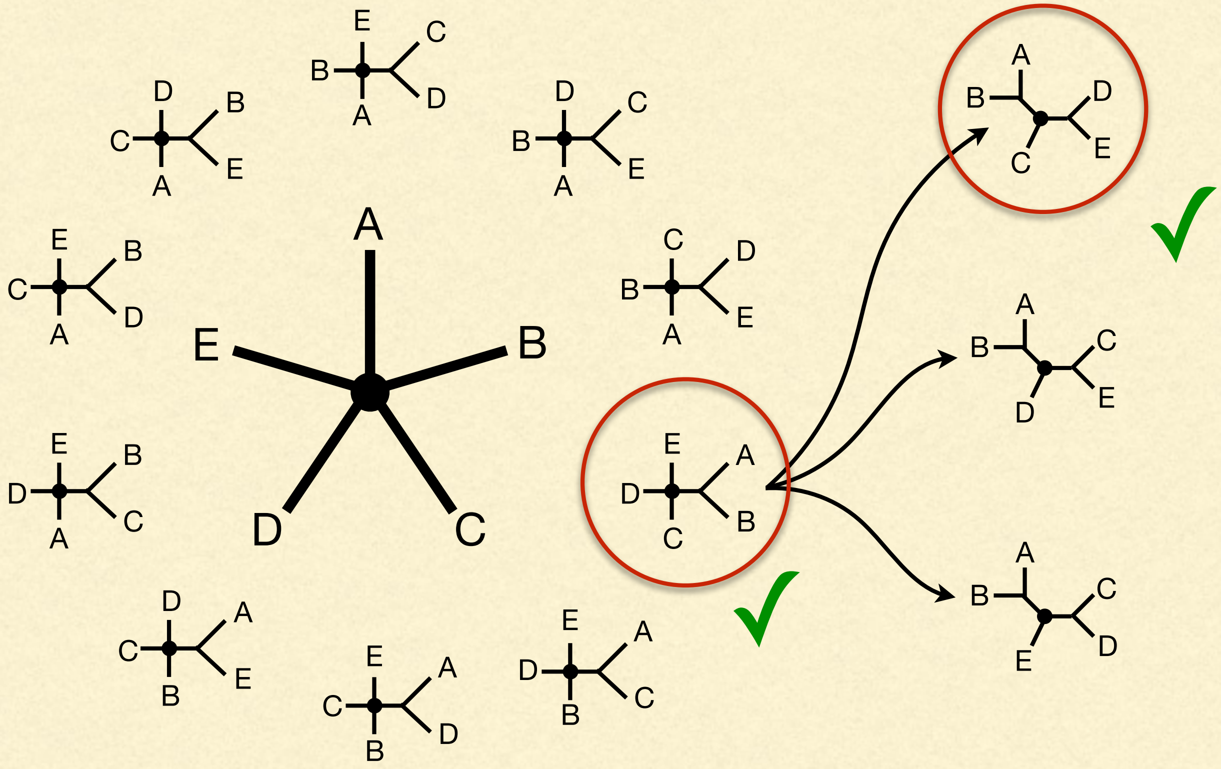


← 83.2 billion years @ 5 million trees/sec

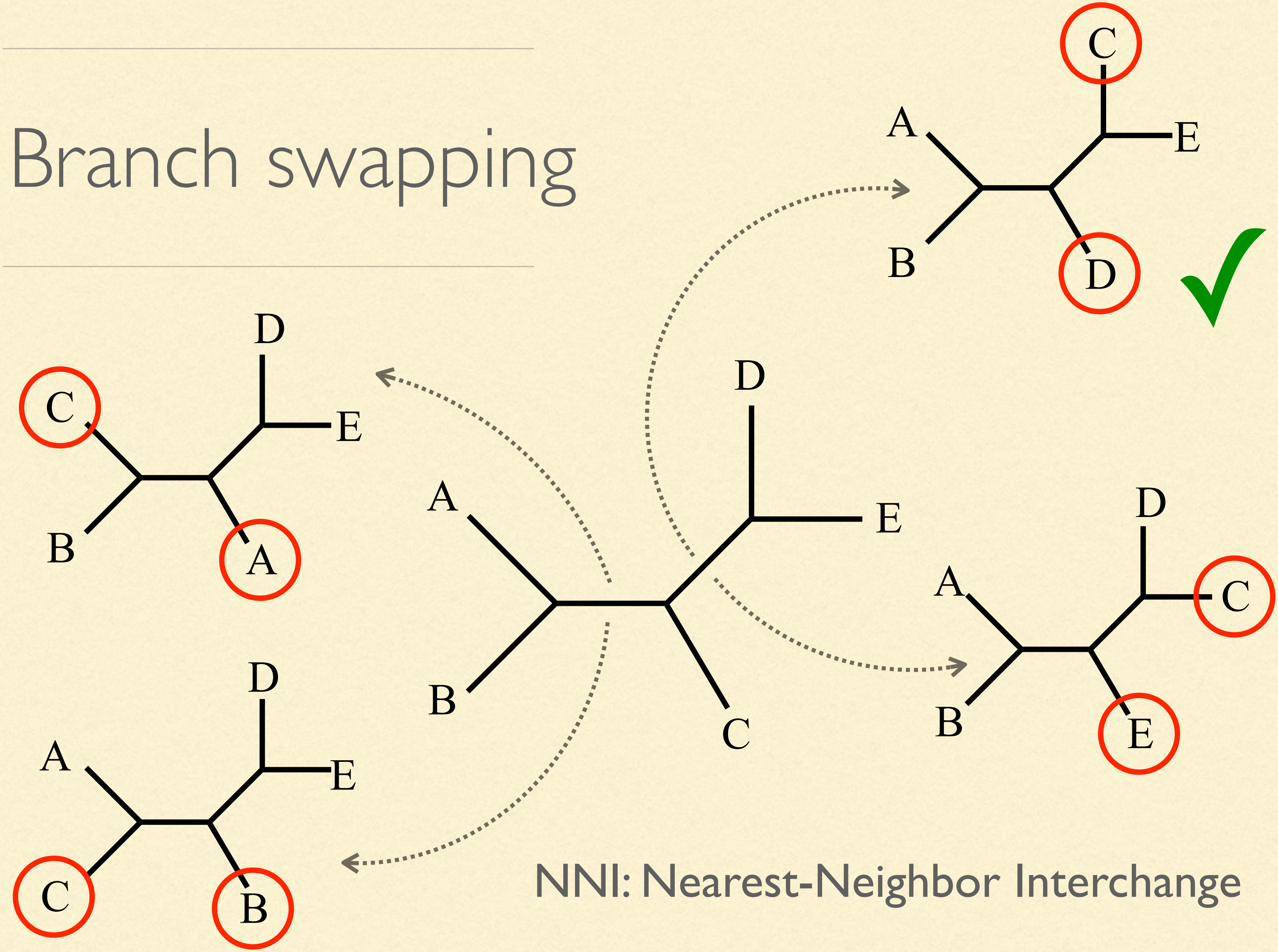
Stepwise addition



Star decomposition (e.g. Neighbor Joining)

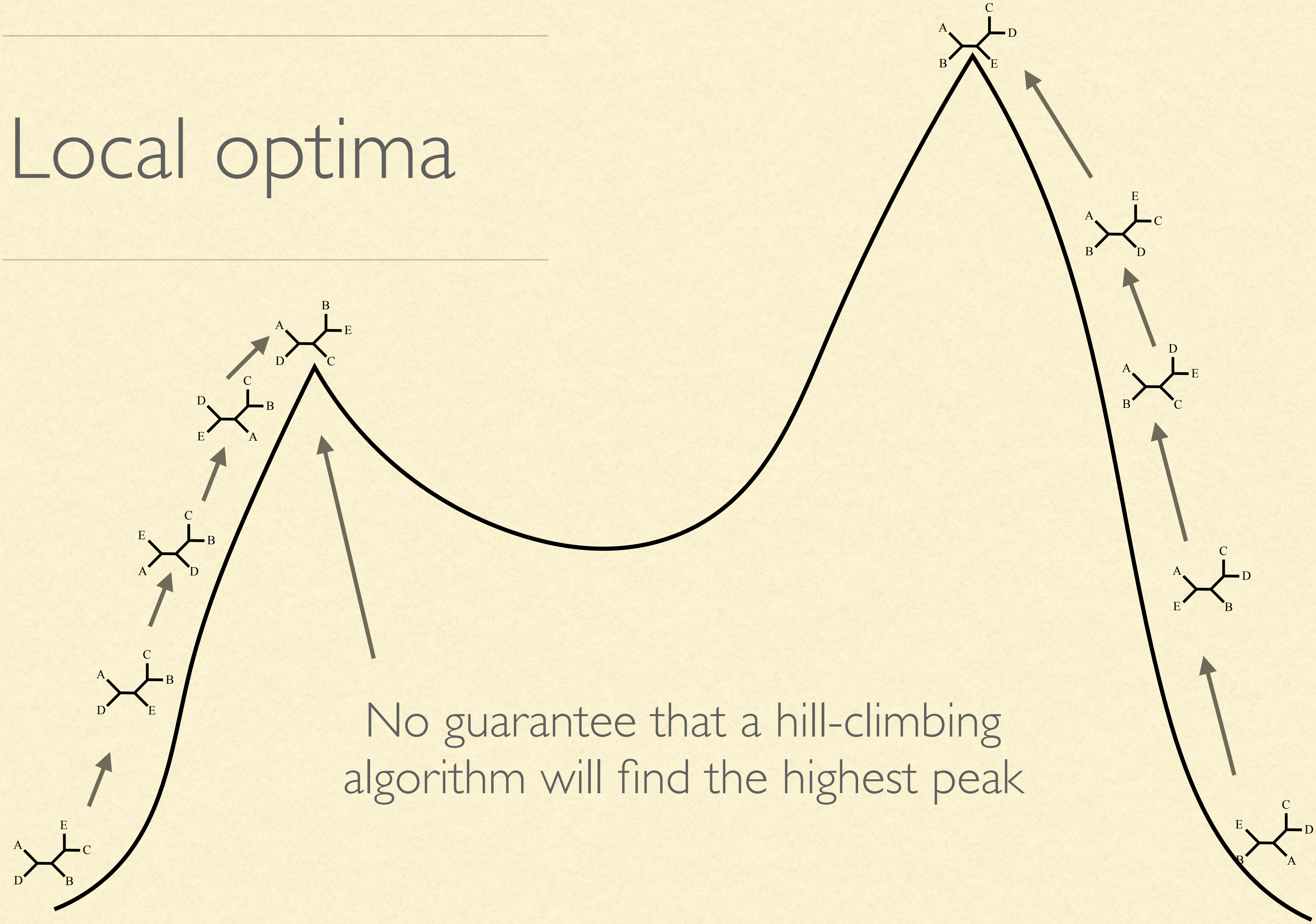


Branch swapping



NNI: Nearest-Neighbor Interchange

Local optima

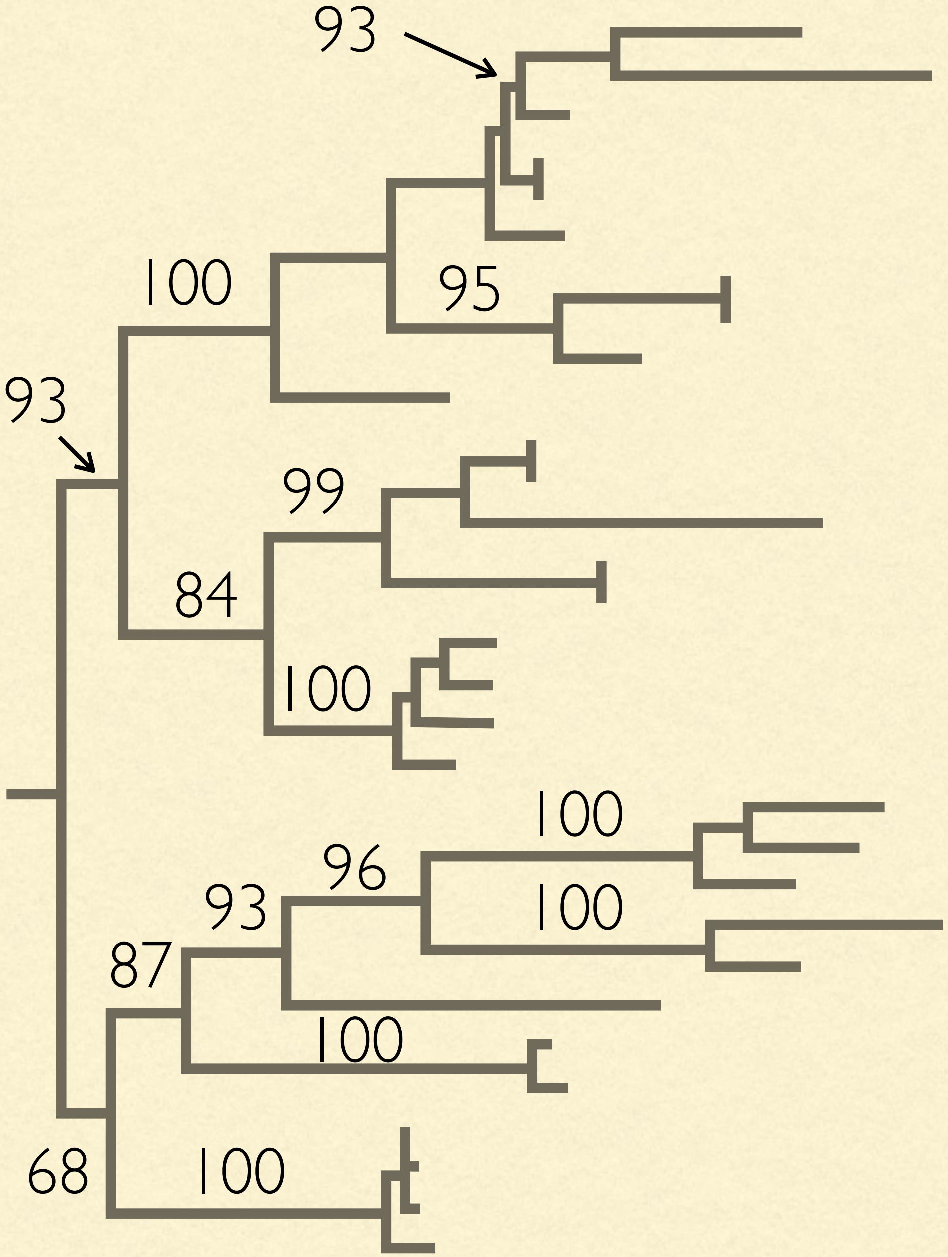


No guarantee that a hill-climbing algorithm will find the highest peak

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.

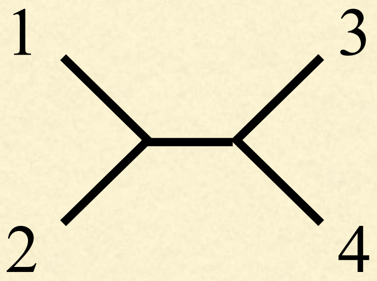


Bootstrap support

sites sampled with replacement

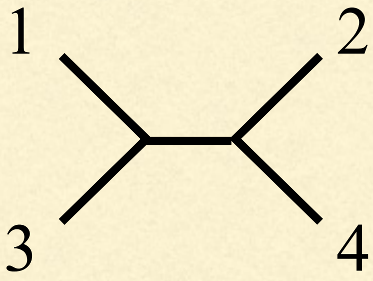
original data

	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
	✗					✗	✗	

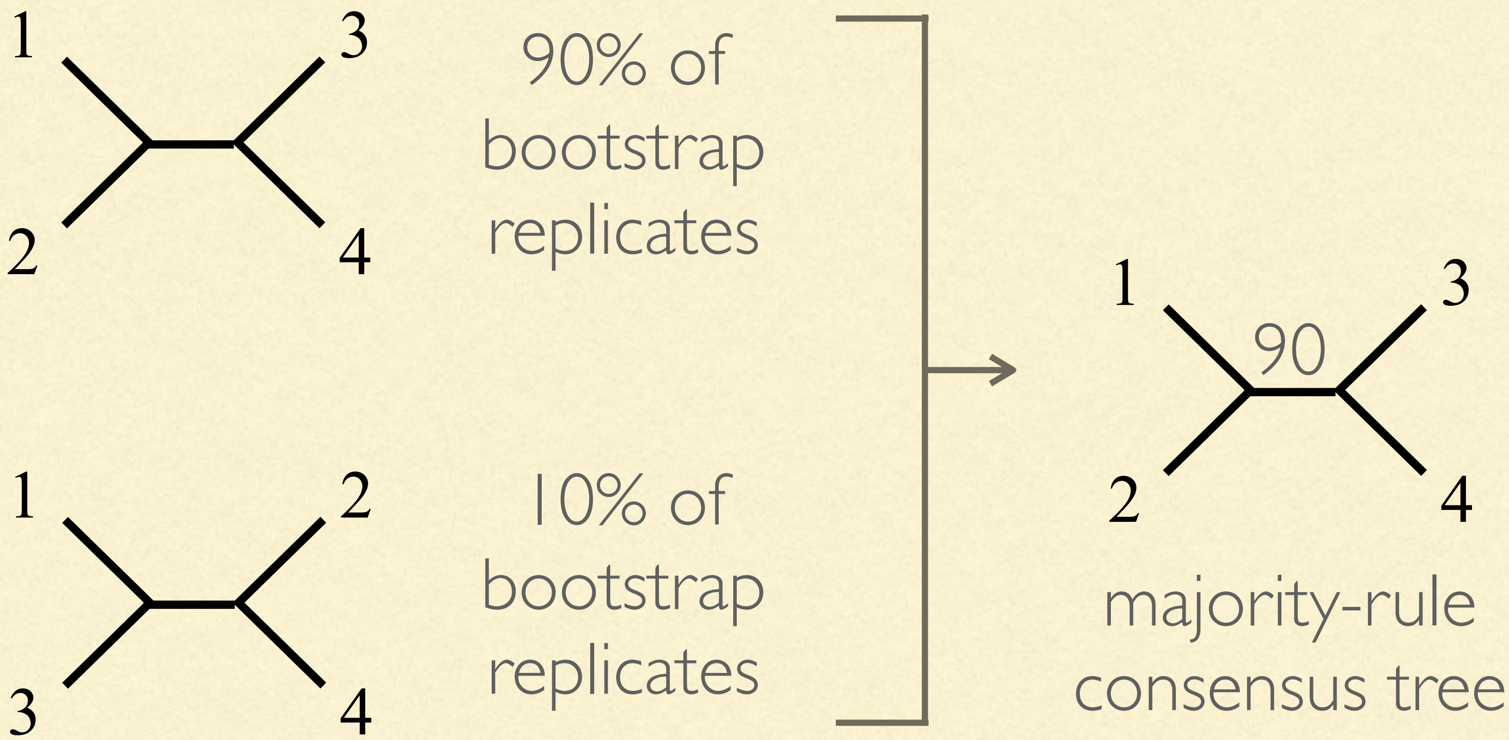


bootstrap replicate

	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



Consensus trees

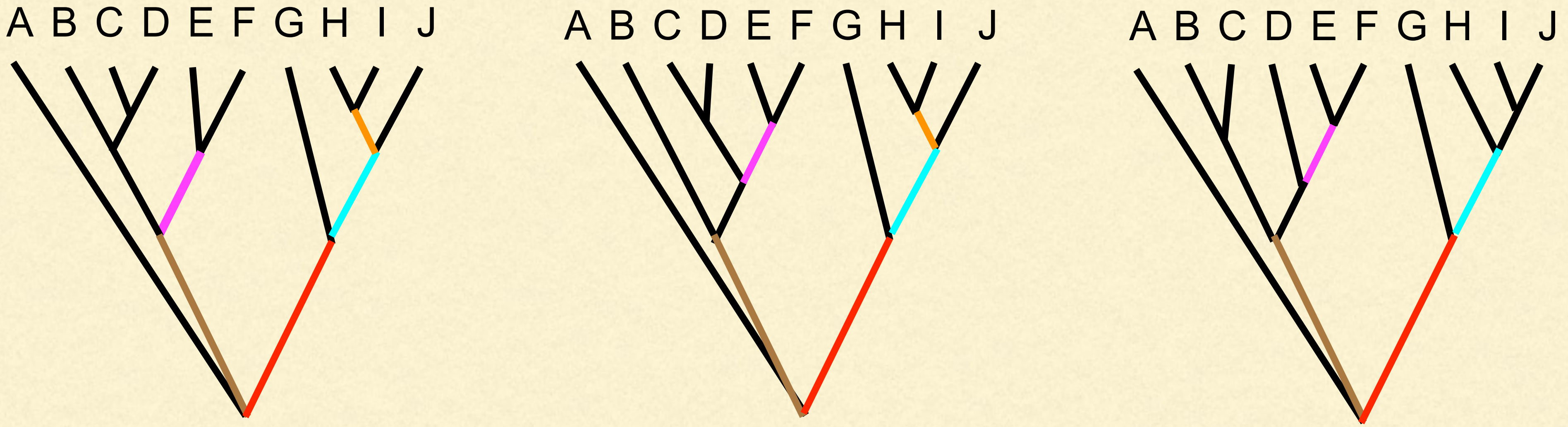


90% of bootstrap replicates

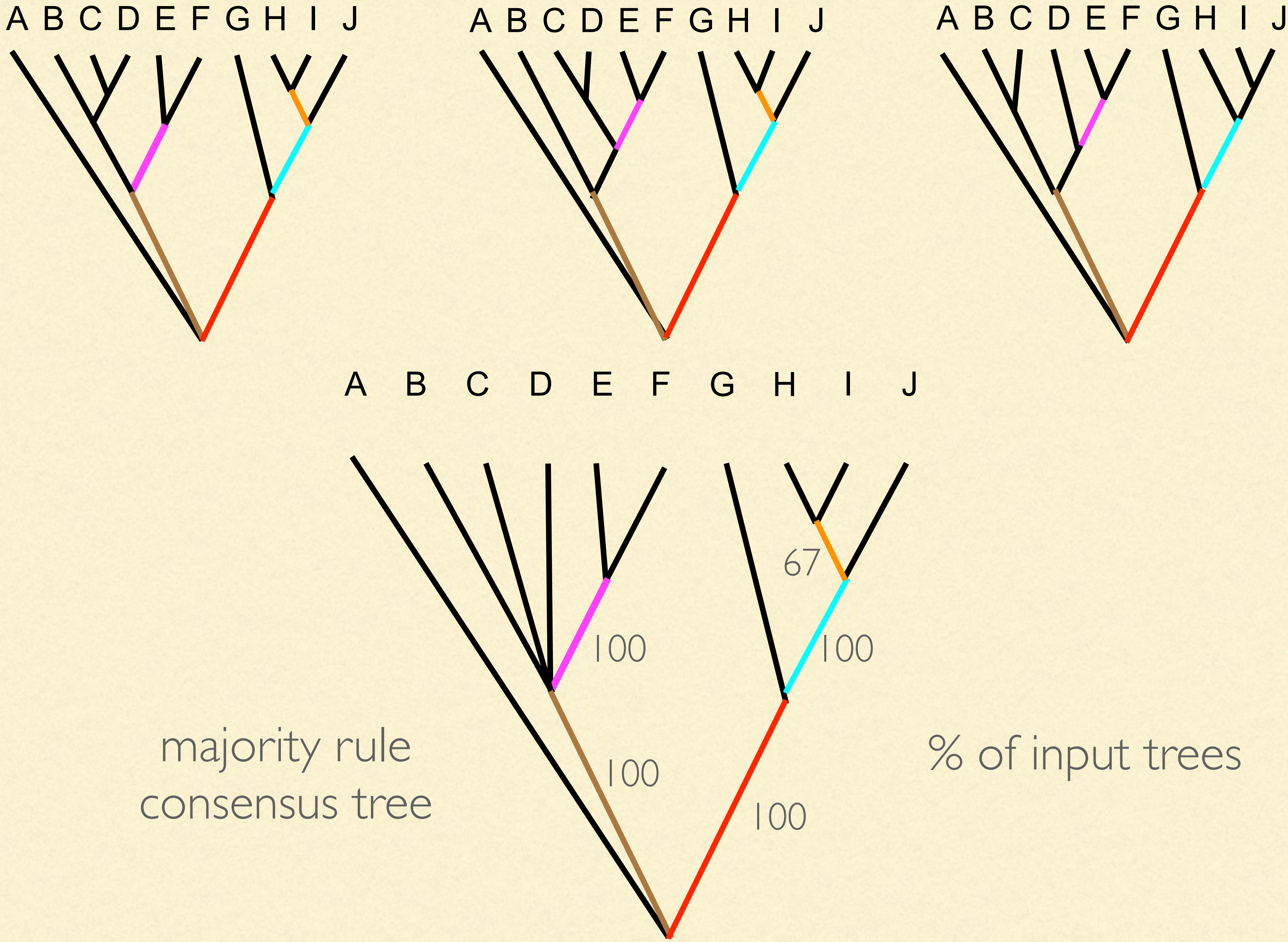
10% of bootstrap replicates

majority-rule consensus tree

Consensus trees



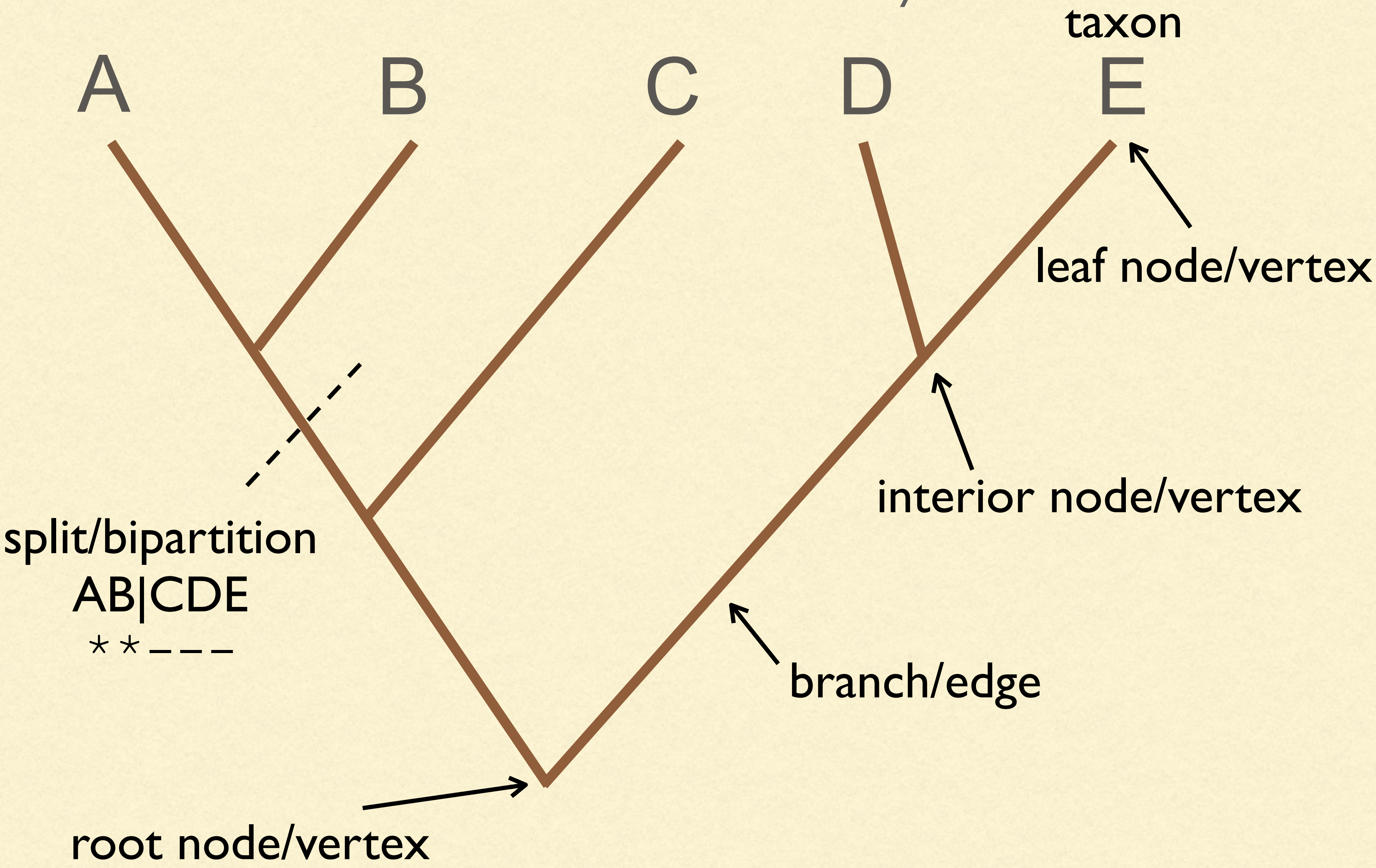
	ABCDEFGHIJ	ABCDEFGHIJ	ABCDEFGHIJ
	..**......	..**......	.**.......
●	....**....	....**....	....**....
	.***......	..****....	...***....
●	.*****....	.*****....	.*****....
●	.......***	.......***	.......***
●	......****	......****	......****
●	........**.	........**.	........**



majority rule
consensus tree

% of input trees

Tree anatomy



taxon

A B C D E

leaf node/vertex

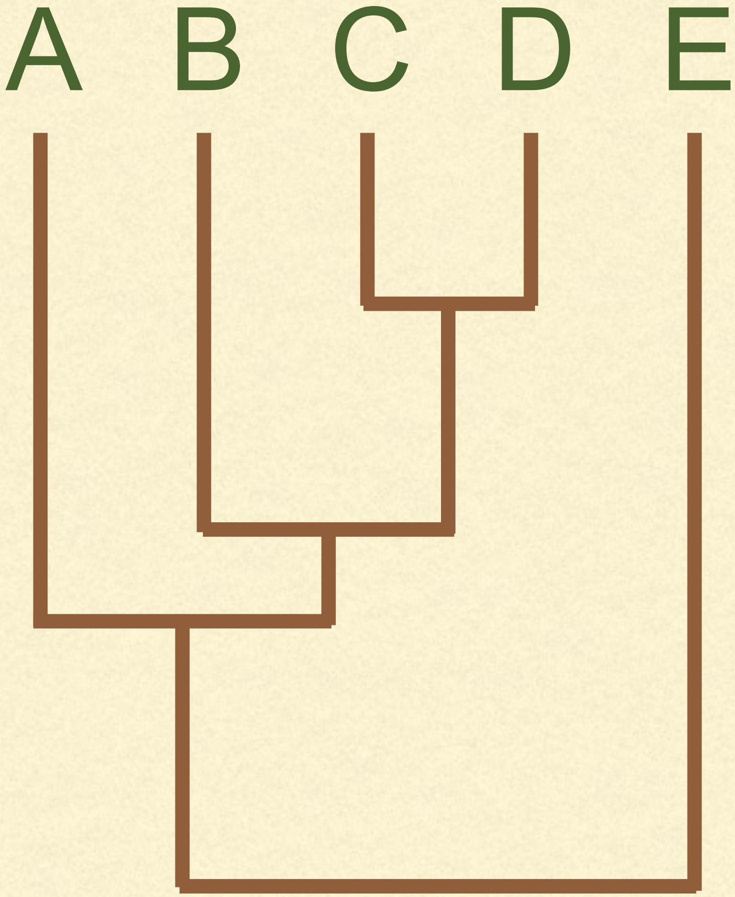
interior node/vertex

branch/edge

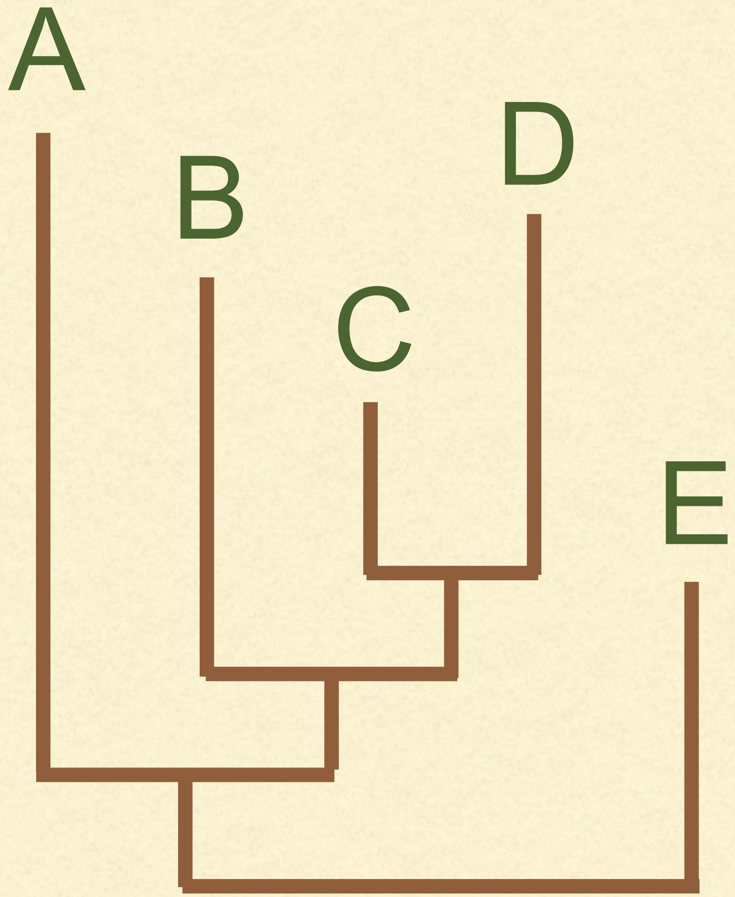
split/bipartition
AB|CDE
* * – – –

root node/vertex

Edge lengths



edge lengths are
time only



edge lengths are
rate x time

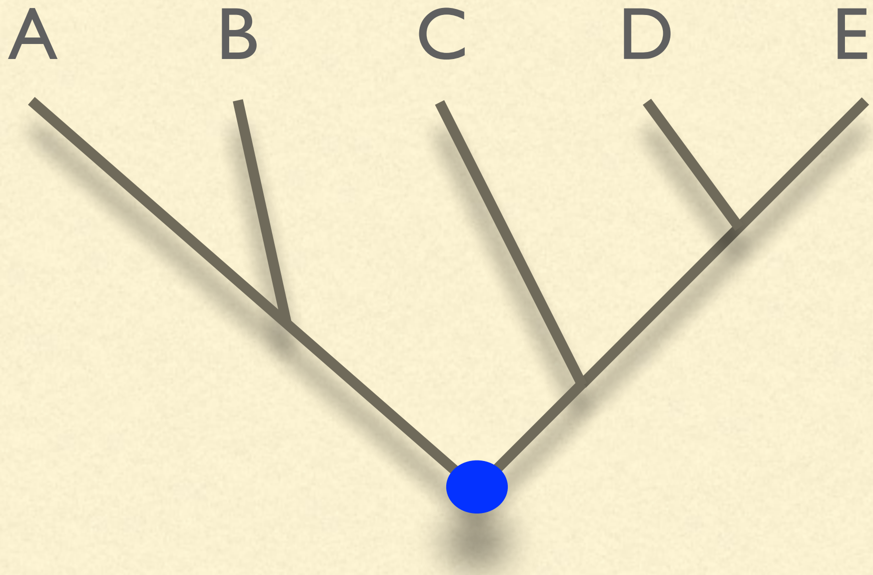
Newick descriptions

```
#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;
```

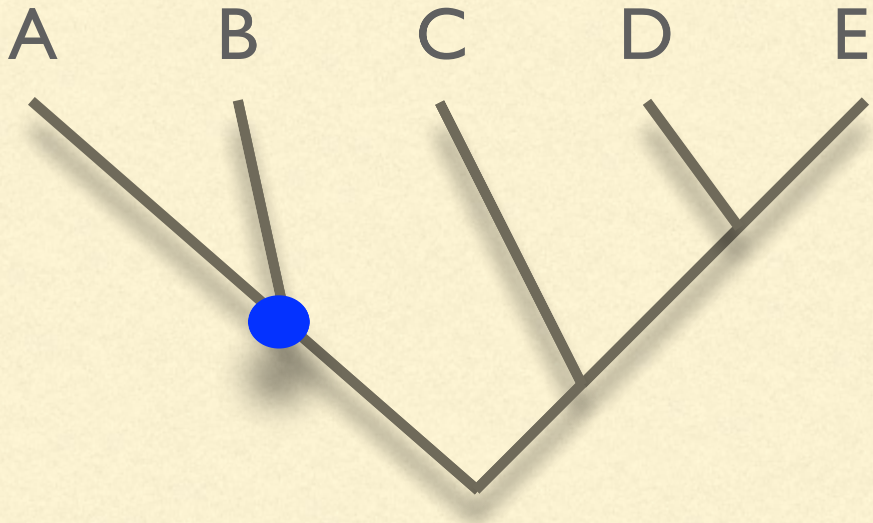
https://en.wikipedia.org/wiki/Newick_format

Newick tree descriptions



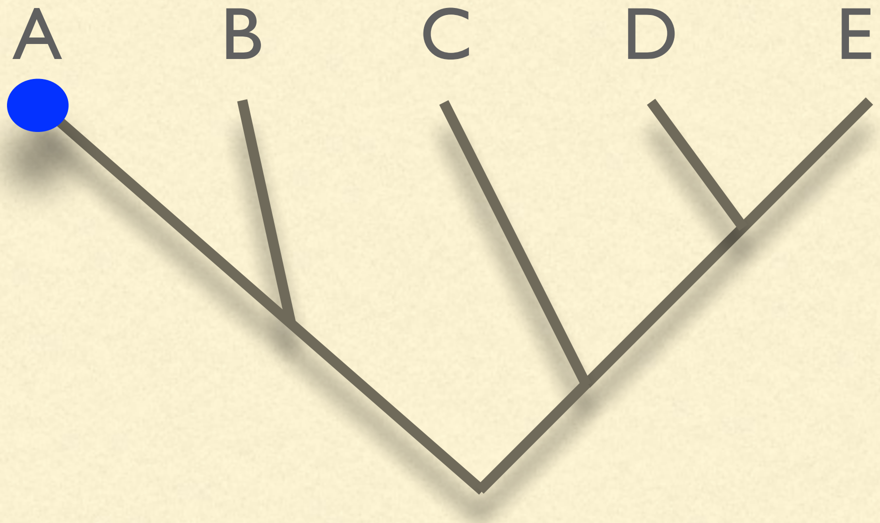
((A,B),(C,(D,E)))

Newick tree descriptions



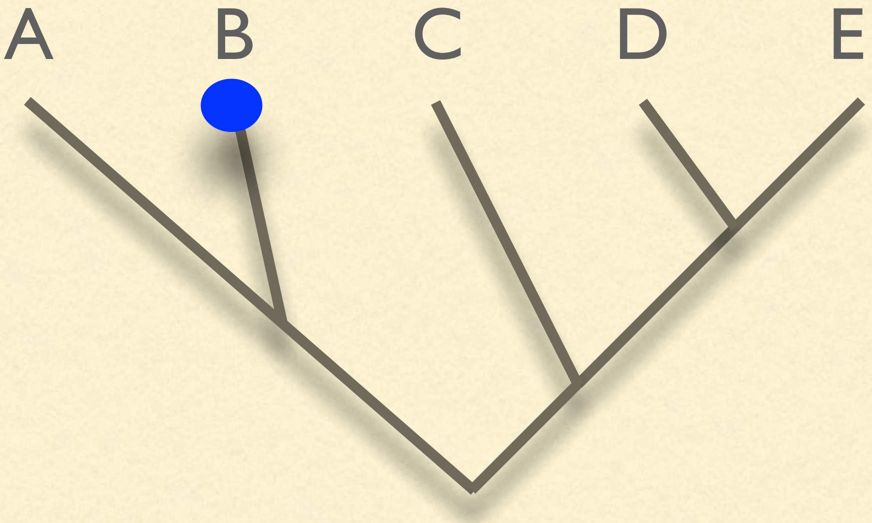
((A,B),(C,(D,E)))

Newick tree descriptions



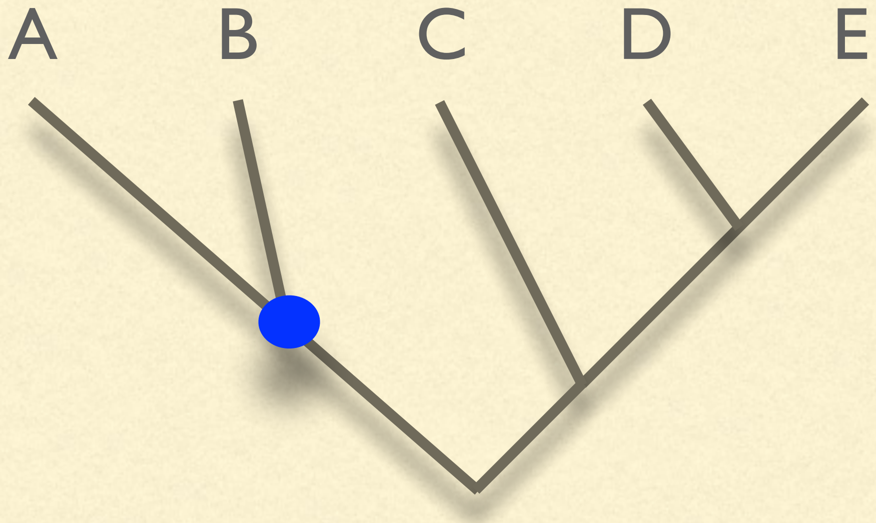
((**A**,B),(C,(D,E)))

Newick tree descriptions



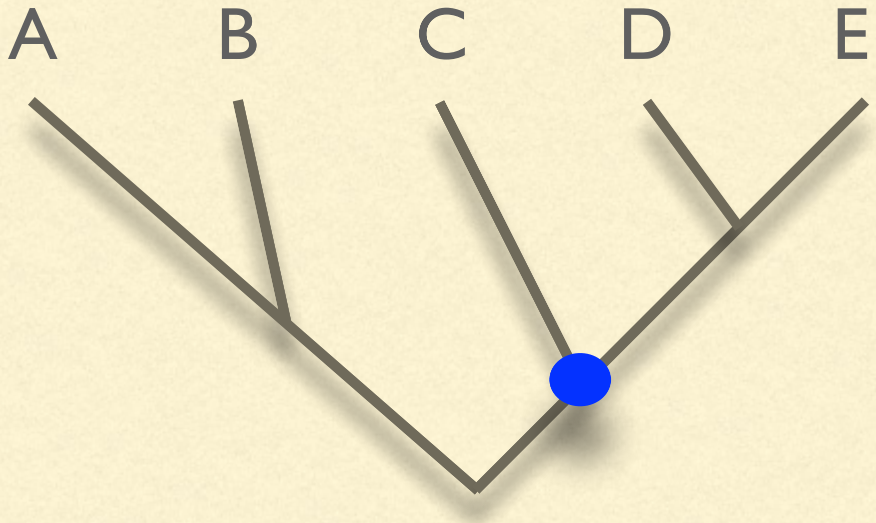
((A,**B**),(C,(D,E)))

Newick tree descriptions



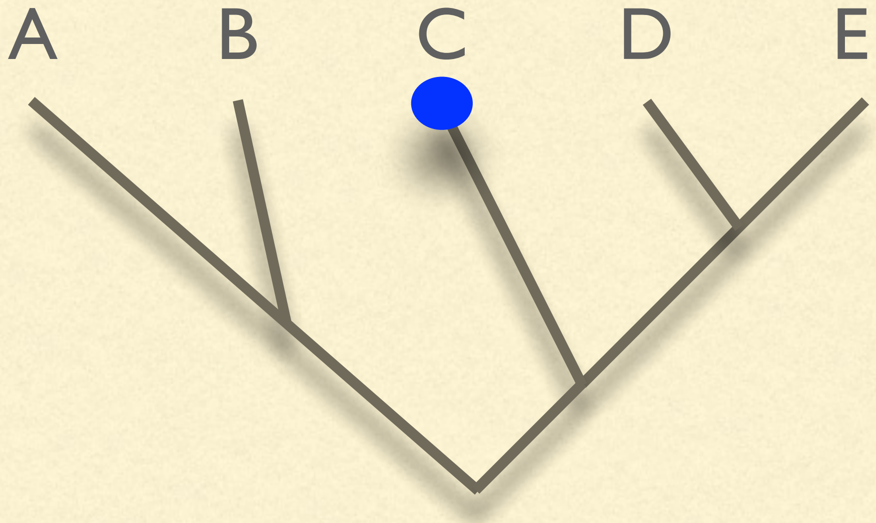
((A,B),(C,(D,E)))

Newick tree descriptions



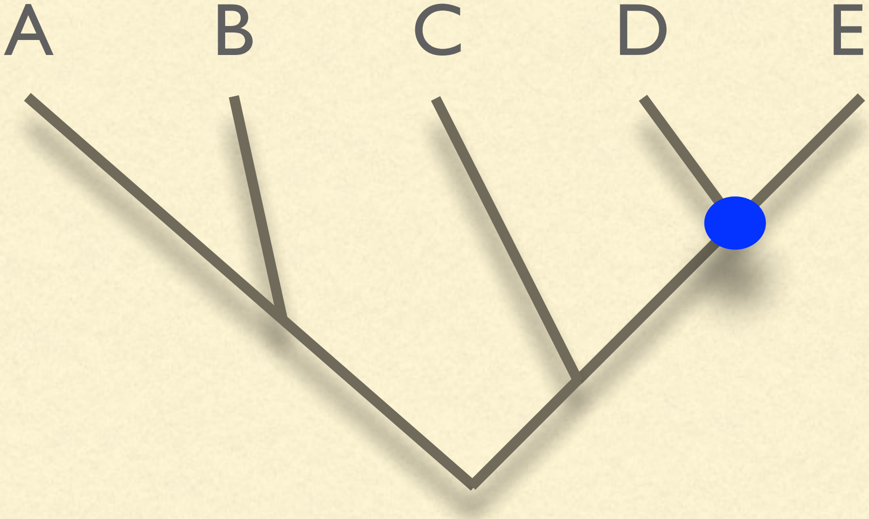
((A,B),(C,(D,E)))

Newick tree descriptions



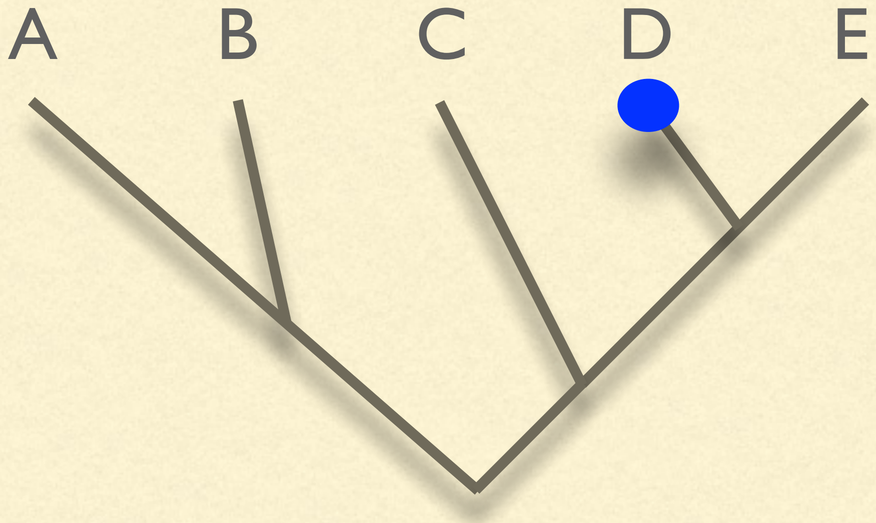
((A,B),**(C**,(D,E)))

Newick tree descriptions



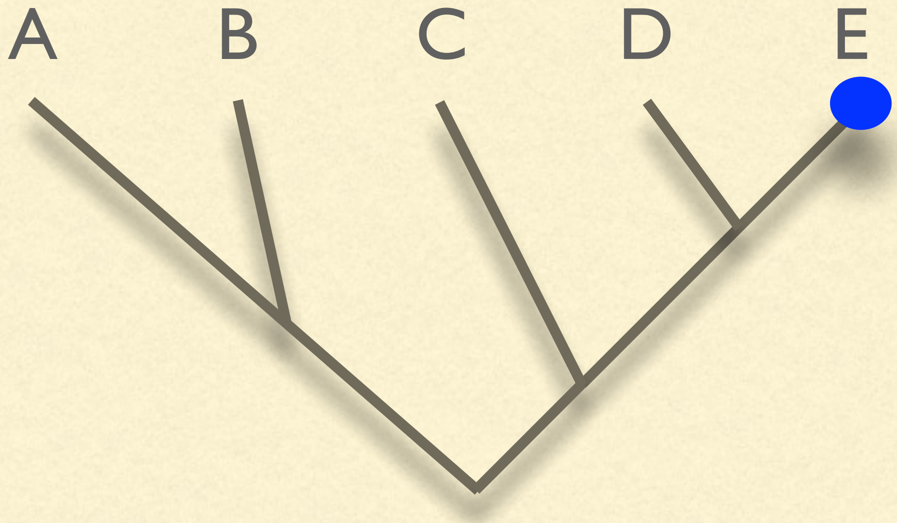
((A,B),(C,(D,E)))

Newick tree descriptions



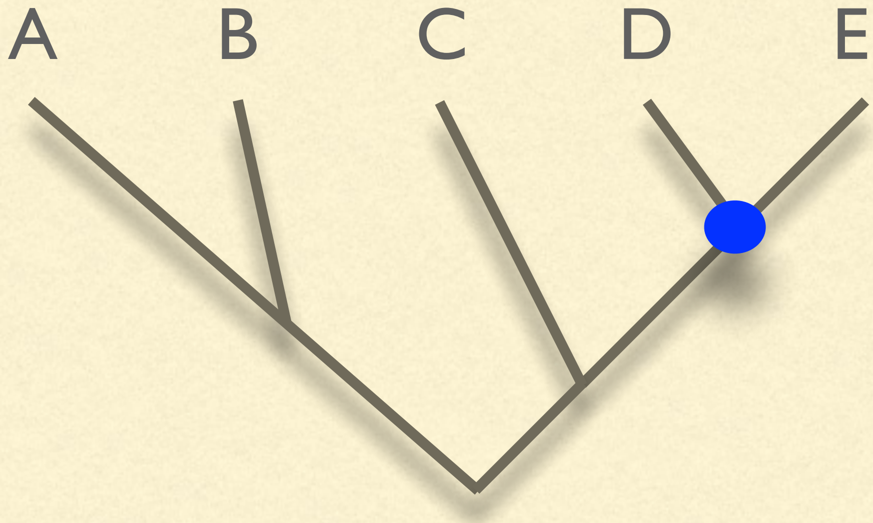
((A,B),(C,**(D**,E)))

Newick tree descriptions



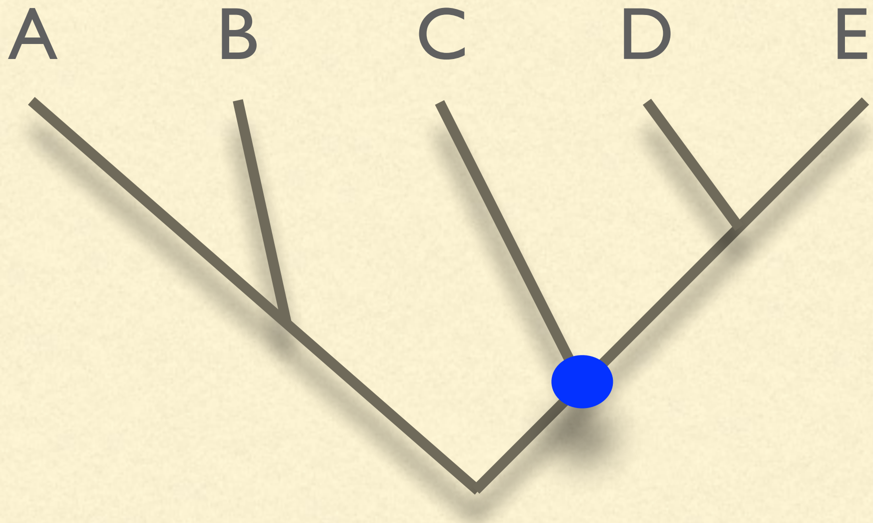
((A,B),(C,(D,**E**)))

Newick tree descriptions



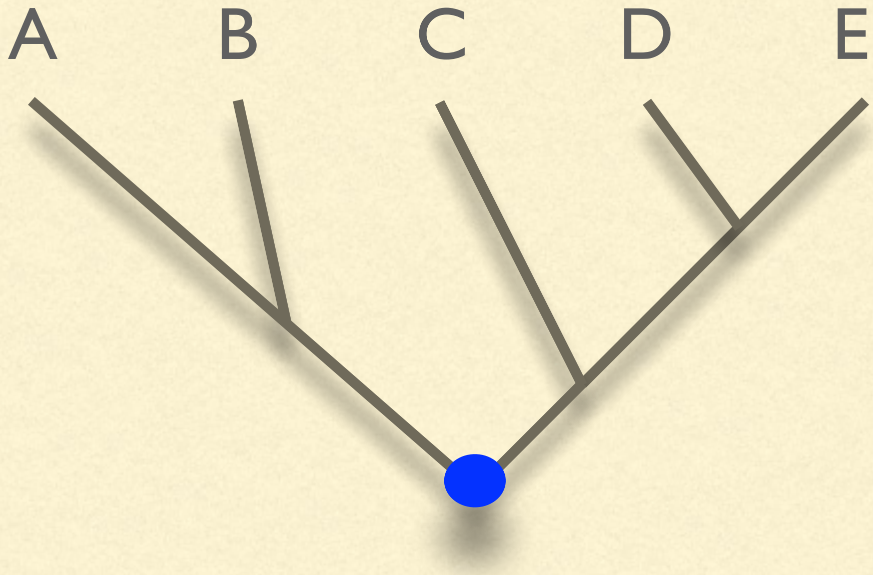
((A,B),(C,(D,E)))

Newick tree descriptions



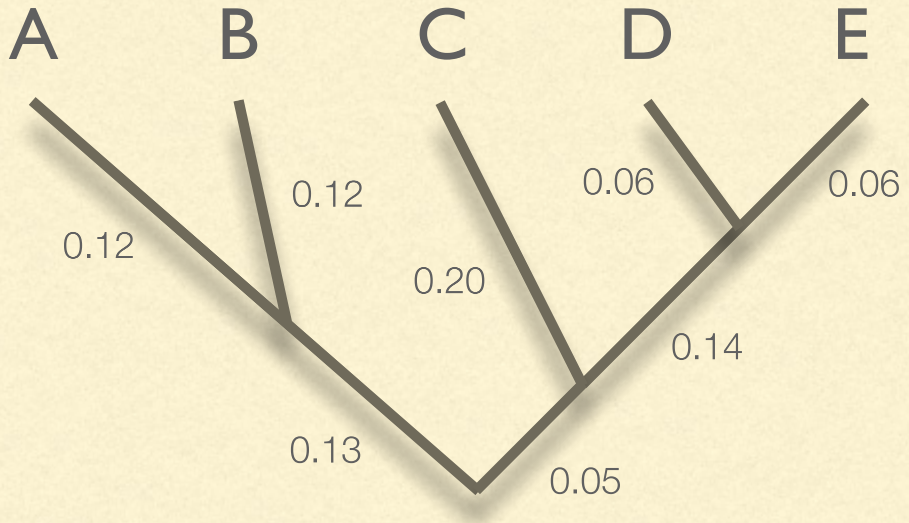
((A,B),(C,(D,E)))

Newick tree descriptions



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

Newick tree descriptions

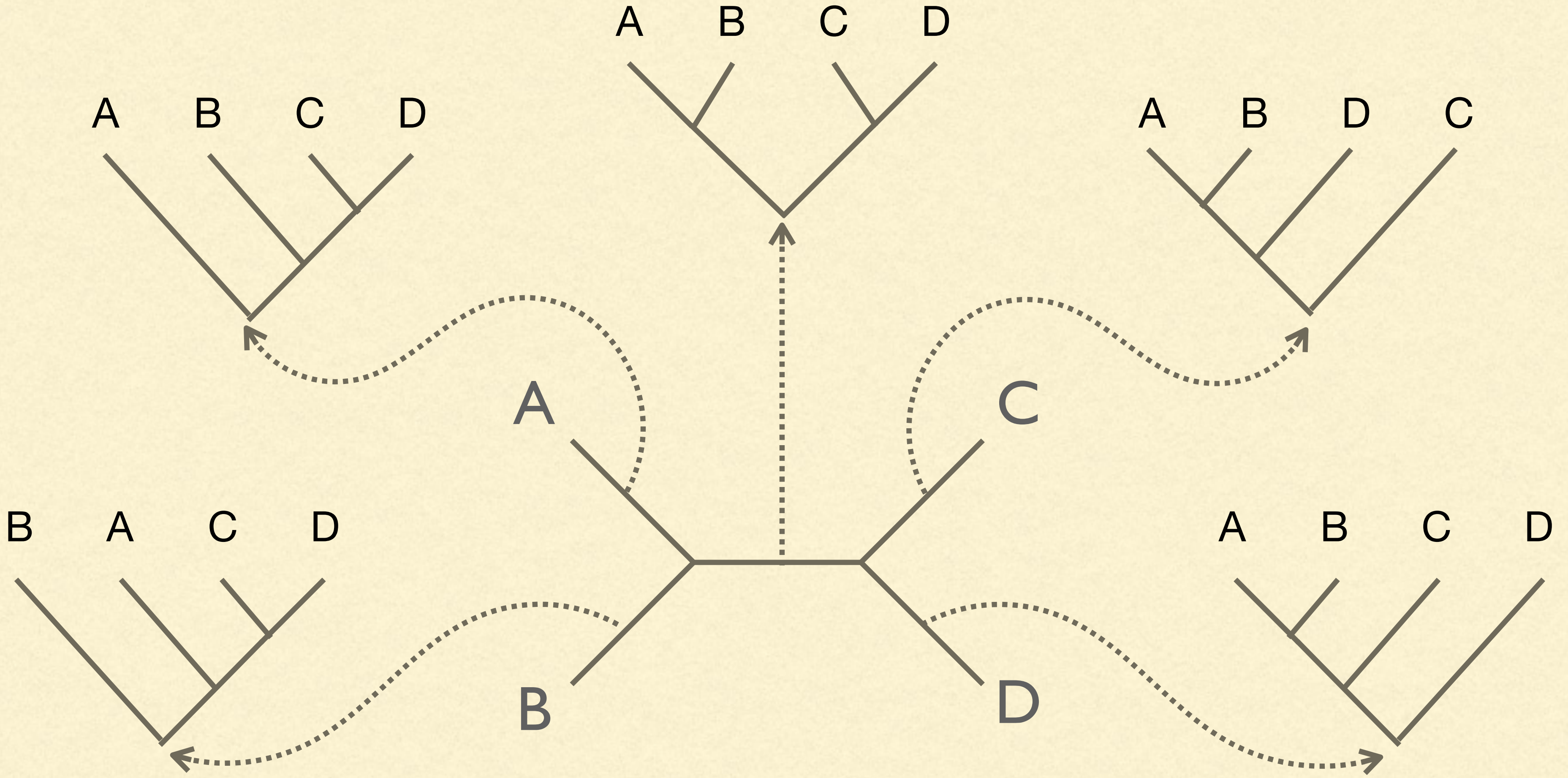


((A:.12,B:.12):.13,(C:.2,(D:.06,E:.06):.14):.05)

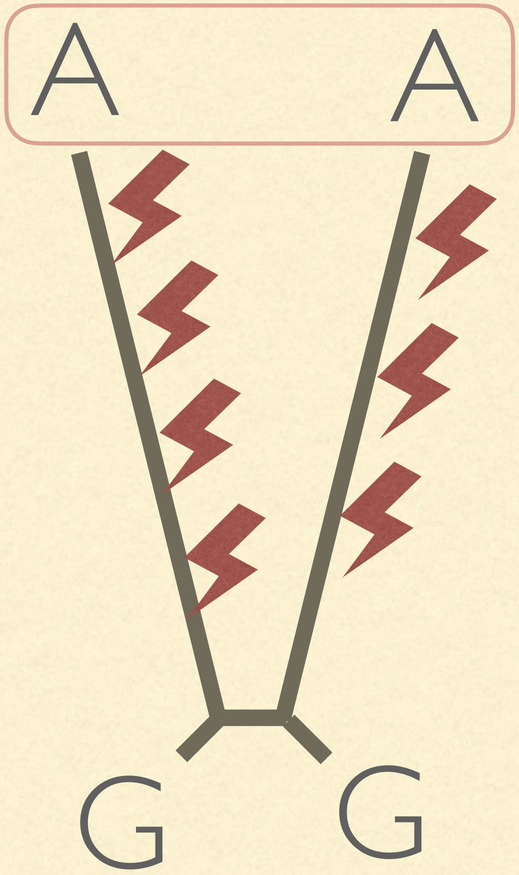
edge lengths follow colon after node name (if present)

Rooted vs unrooted

rooting and **adding a taxon** increase treespace by the same amount

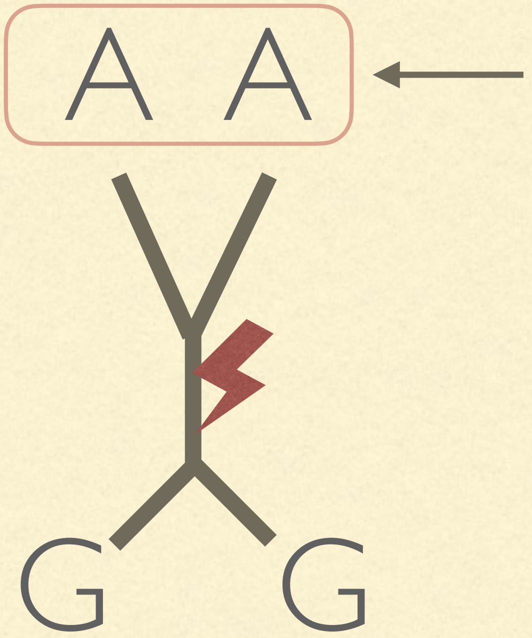


Challenges: model violations



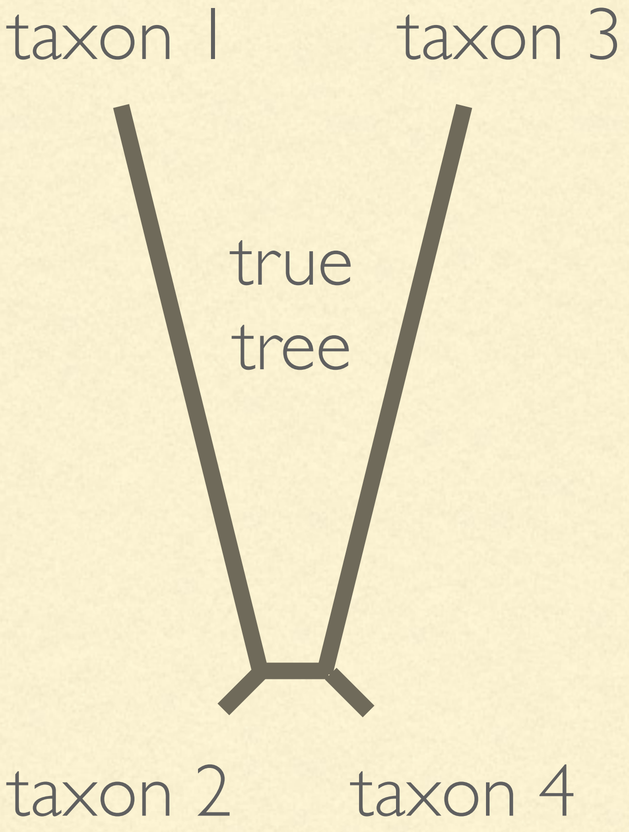
Long external branches favor a **convergence explanation** of this similarity

Challenges: model violations



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

Challenges: model violations



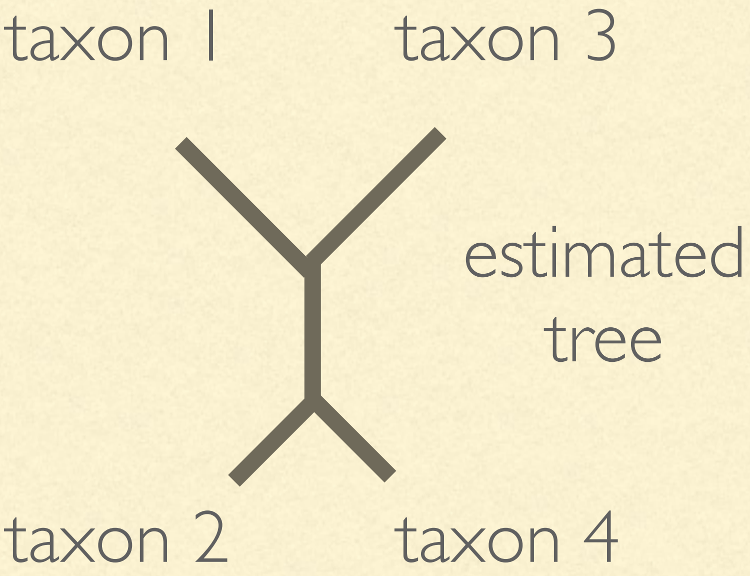
taxon 1

taxon 3

true tree

taxon 2

taxon 4



taxon 1

taxon 3

estimated tree

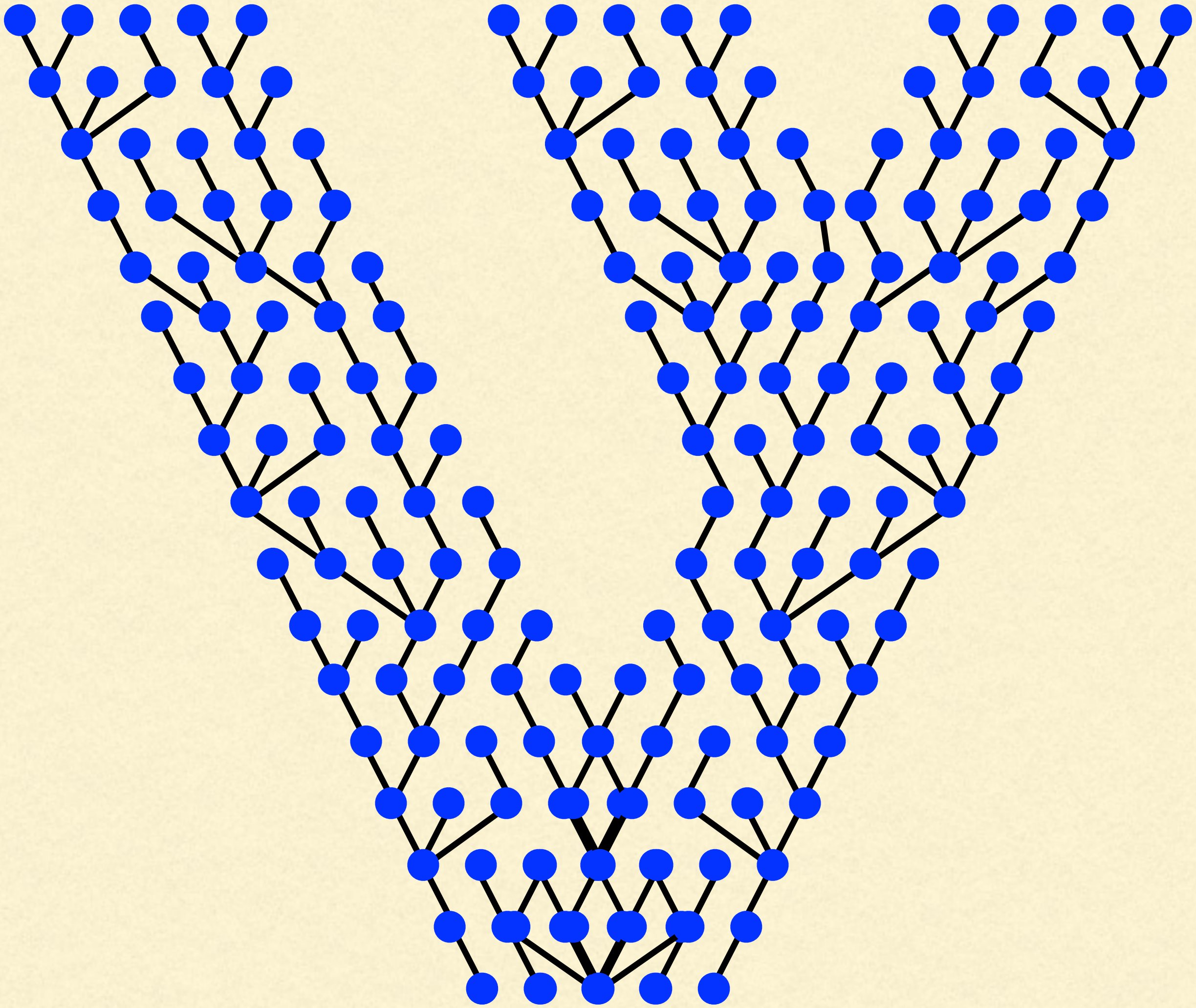
taxon 2

taxon 4

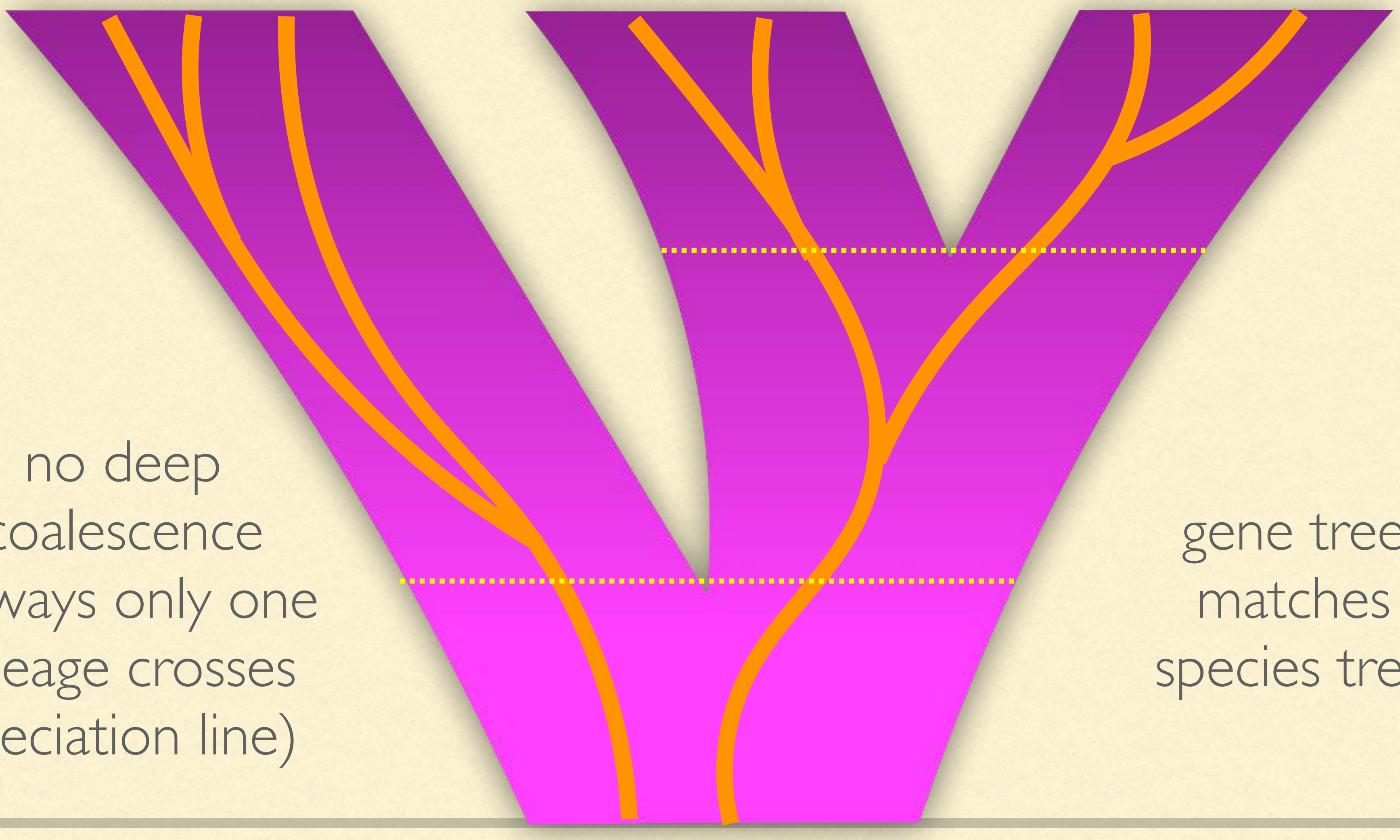
Models that are too simple often underestimate branch lengths

Long branch attraction

Challenges: deep coalescence



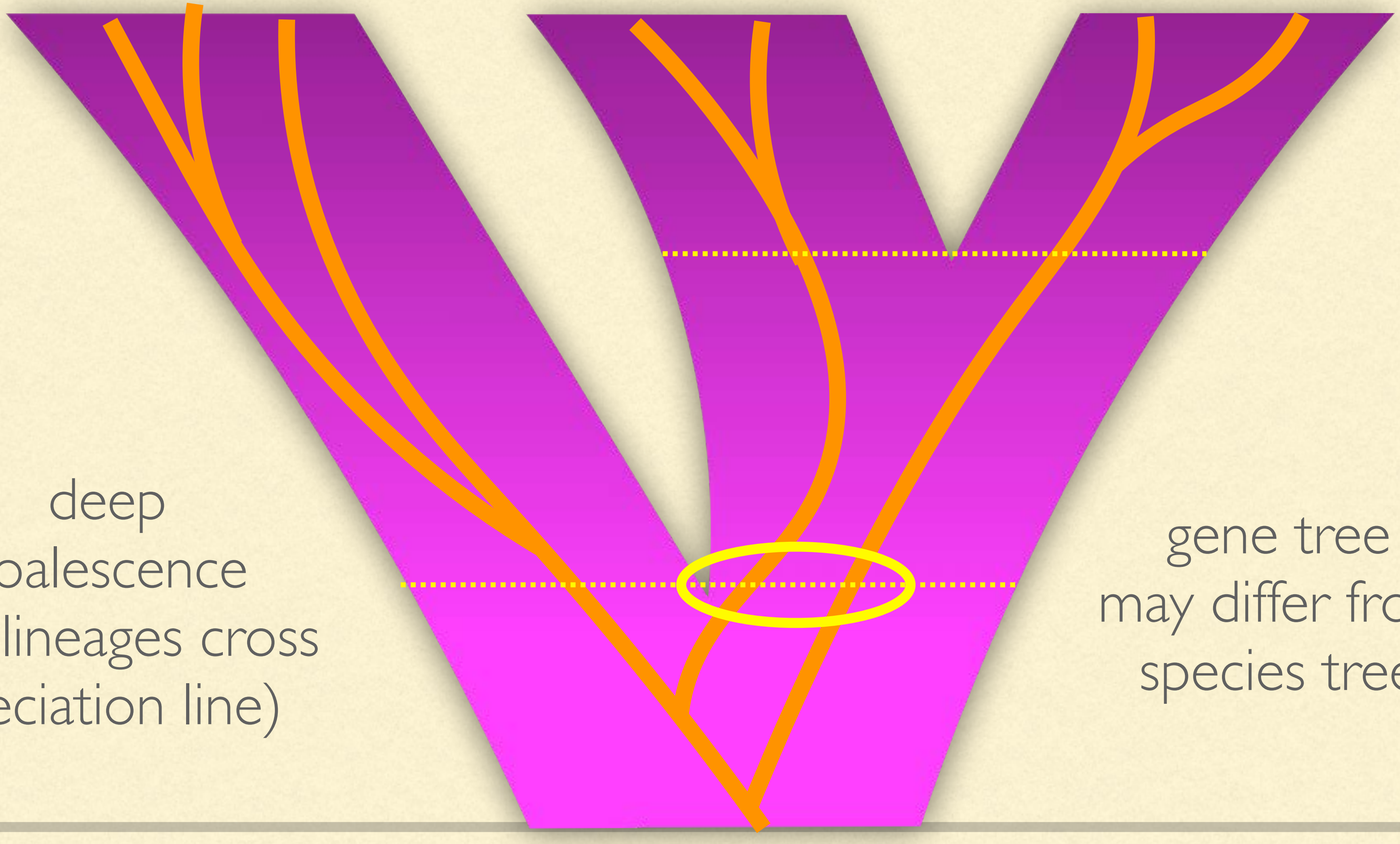
Challenges: deep coalescence



no deep coalescence (always only one lineage crosses speciation line)

gene tree matches species tree

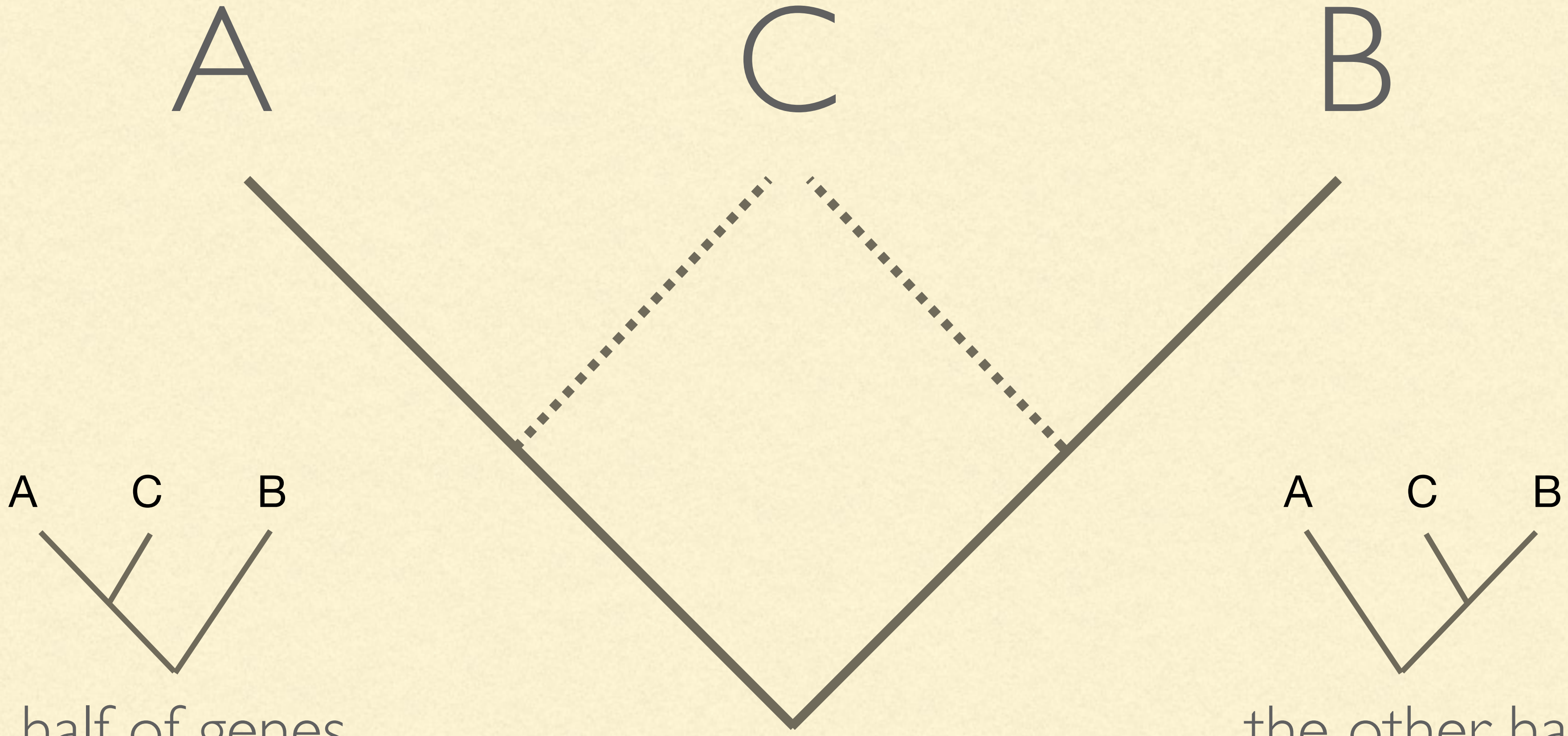
Challenges: deep coalescence



deep
coalescence
(2+ lineages cross
speciation line)

gene tree
may differ from
species tree

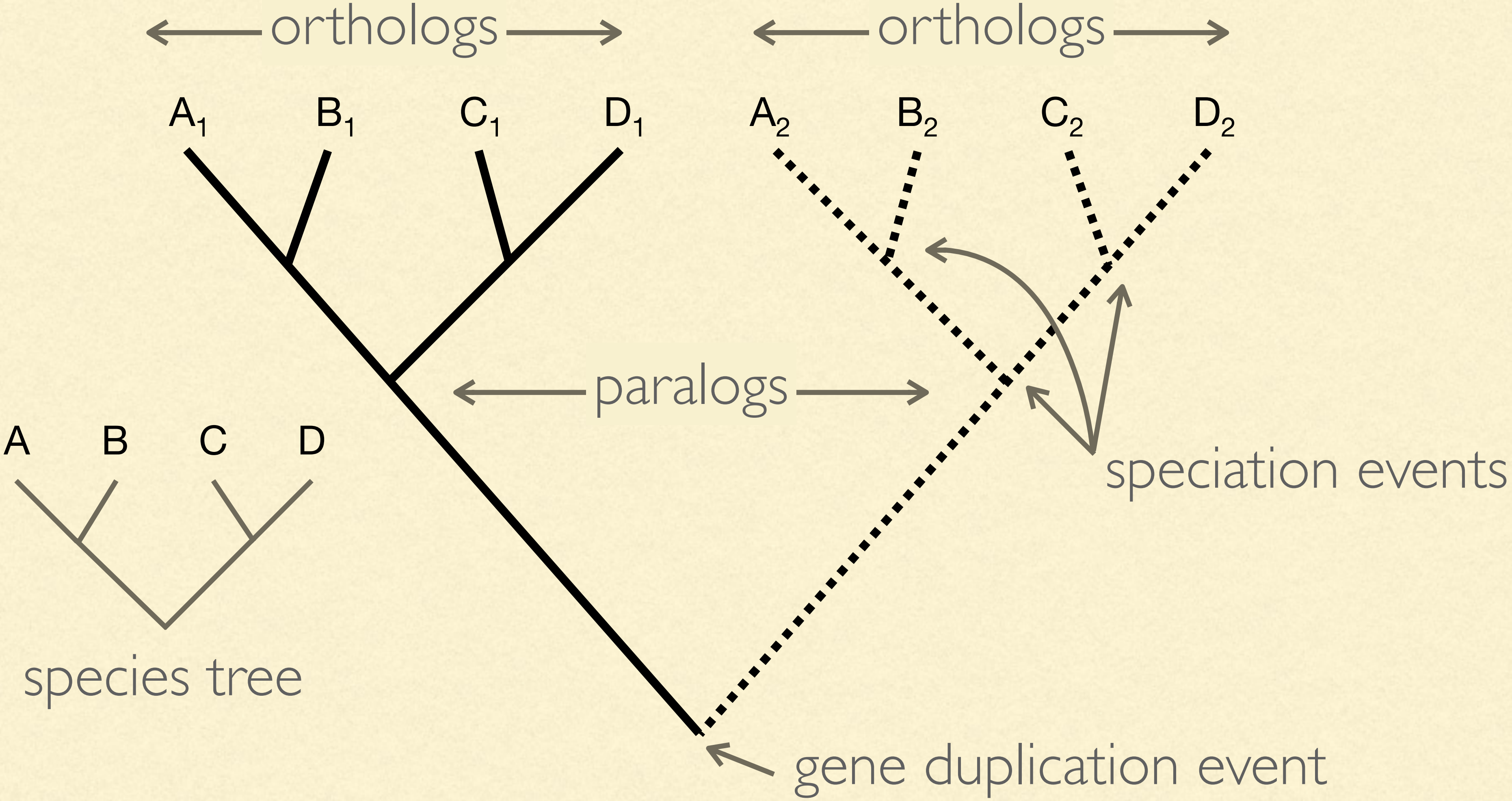
Challenges: hybridization



half of genes

the other half of genes

Challenges: paralogy



orthologs

orthologs

paralogs

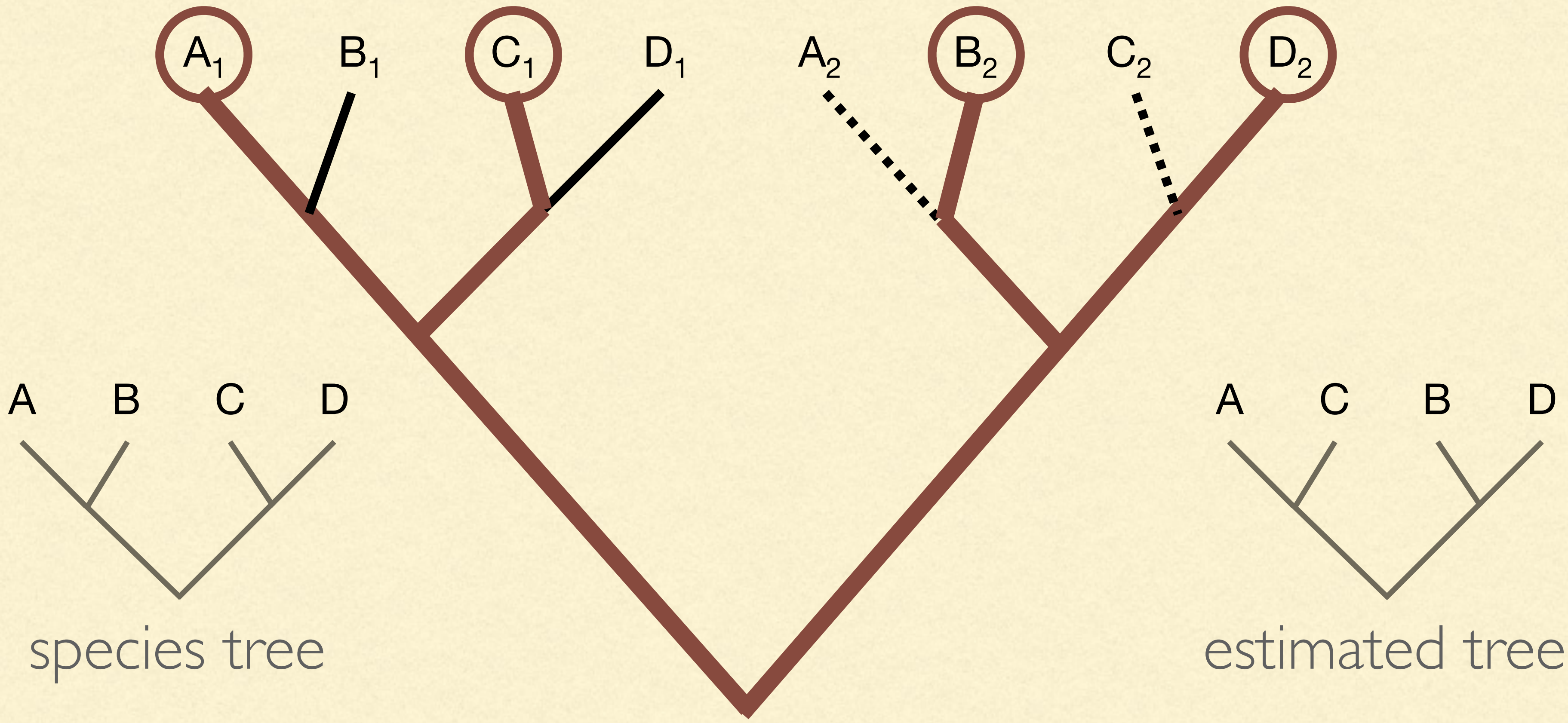
speciation events

species tree

gene duplication event

Challenges: paralogy

sampled sequences are a mixture of orthologs and paralogs


A_1
B_1
C_1
D_1
A_2
B_2
C_2
D_2
A
B
C
D
A
C
B
D
species tree
estimated tree