Thanks to Paul Lewis and Joe Felsenstein for the use of slides

Review

- Hennigian logic reconstructs the tree if we know **polarity** of characters and there is **no homoplasy**

- UPGMA infers a tree from a distance matrix:
  - groups based on **similarity**
  - fails to give the correct tree if rates of character evolution vary much

- modern distance-based approaches:
  - extend the ideas behind Buneman's (1971) approach
  - try to find trees and branch lengths, such that the path lengths implied by the tree match the pairwise distances that are estimated from character data.
  - are **not** sensitive to variation in rate.
  - do not use all of the information in the data.

Imperfections of distance base approaches

- Summarizing the character data into a distance matrix loses information.
- We cannot tell which characters evolve quickly and which evolve slowly from pairwise comparisons.

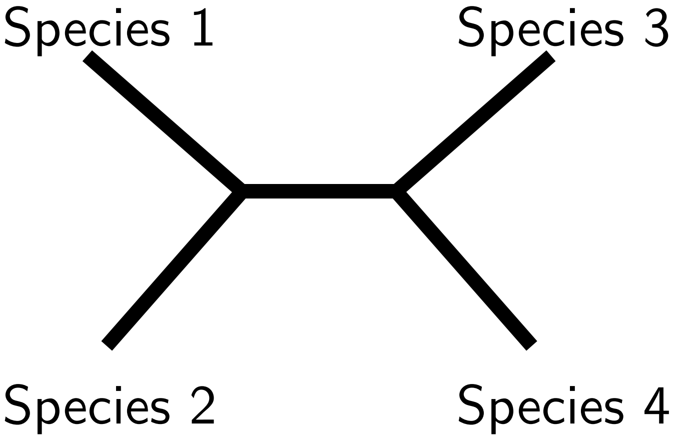
	1	2	3	4
Species 1	C	G	A	C
Species 2	C	G	A	T
Species 3	C	G	A	C
Species 4	C	G	A	T
Species 5	C	G	A	C
Species 6	C	G	A	T
Species 7	C	G	A	C
Species 8	C	G	A	T
Species 9	C	G	G	C
Species 10	C	G	G	T
Species 11	C	G	G	C
Species 12	C	G	G	T
Species 13	C	G	G	C
Species 14	C	G	G	T
Species 15	C	G	G	C
Species 16	C	G	G	T

Imperfections of distance base approaches

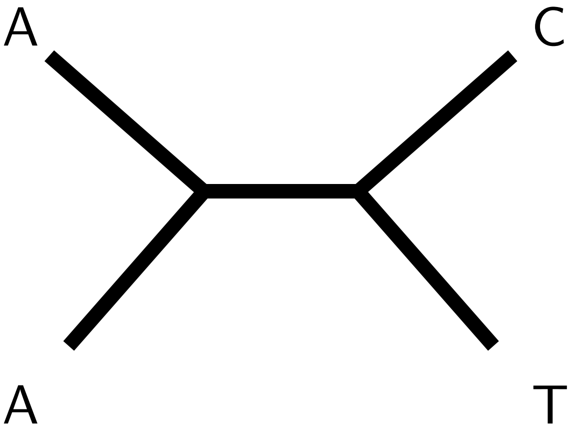
- Summarizing the character data into a distance matrix loses information.
- We cannot tell which characters evolve quickly and which evolve slowly from pairwise comparisons.
- Reconstructing character histories on a tree can reveal homoplasy – this means we can't condense the character data before we start looking for trees.

	1	2	3	4	5	6	7	8	9	.	.	.
Species 1	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 2	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 3	C	G	G	T	C	**C**	G	G	T	.	.	.
Species 4	C	G	G	C	C	**T**	G	G	T	.	.	.

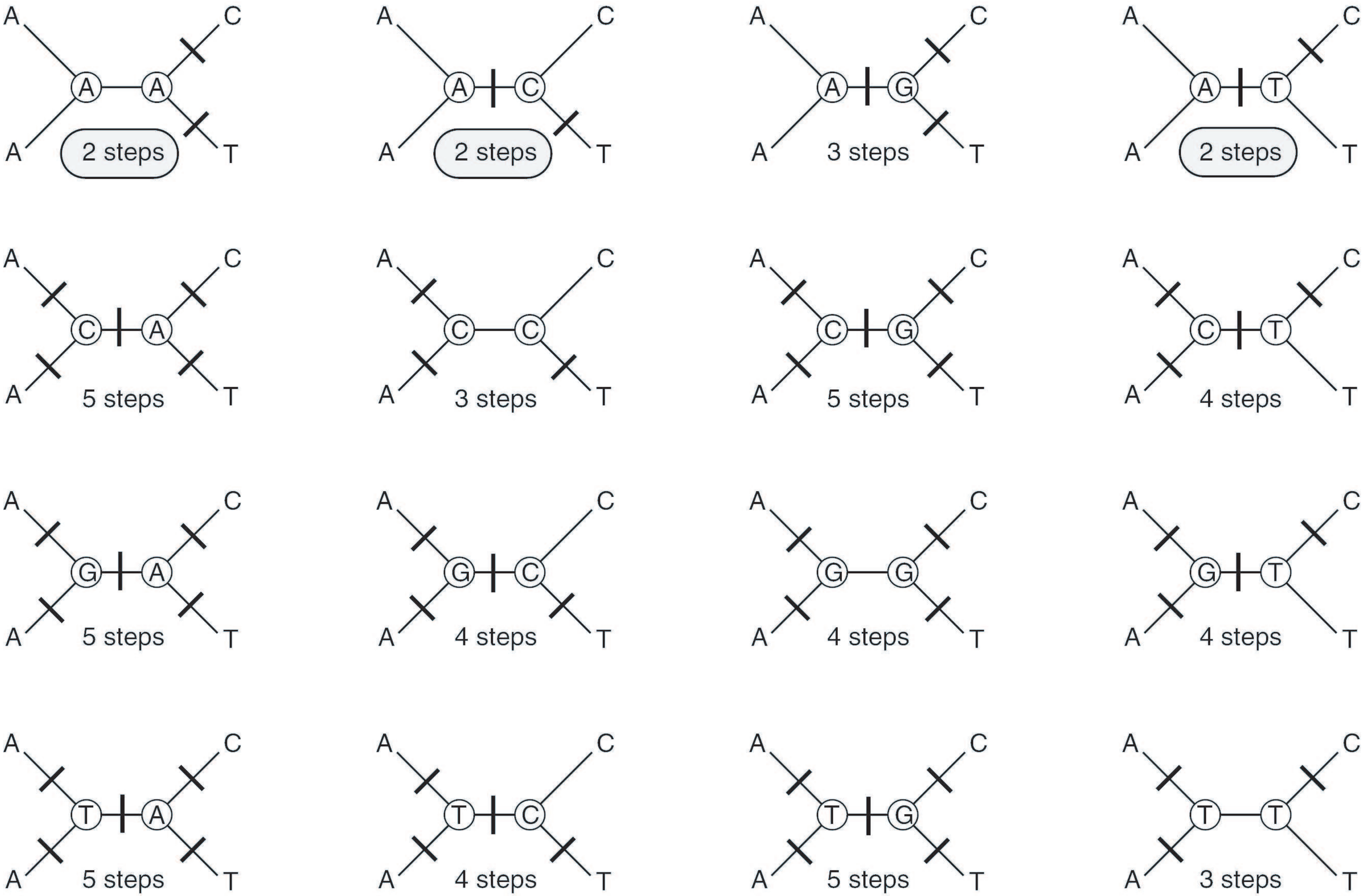
One of the 3 possible trees:



Same tree with states at character 6 instead of species names



“Standard” Parsimony



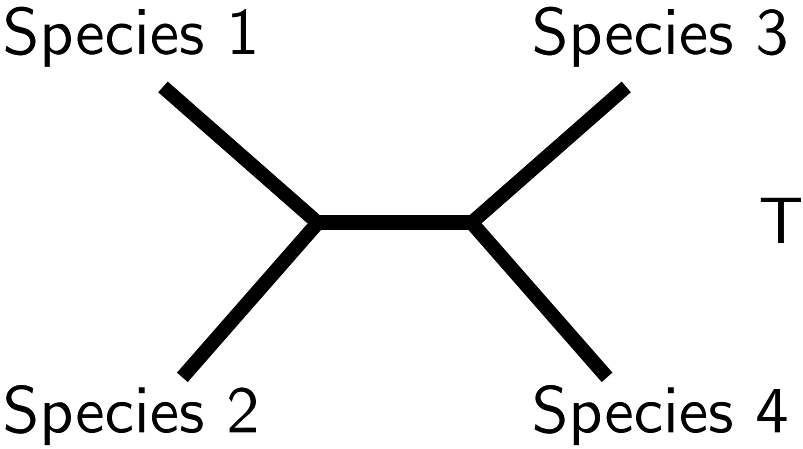
Things to note about the last slide

- 2 steps was the minimum score attainable.
- Multiple ancestral character state reconstructions gave a score of 2.
- Enumeration of all possible ancestral character states is **not** the most efficient algorithm.

Each character (site) is assumed to be independene

To calculate the parsimony score for a tree we simply sum the scores for every site.

	1	2	3	4	5	6	7	8	9
Species 1	C	G	**A**	C	C	A	G	G	T
Species 2	C	G	**A**	C	C	A	G	G	T
Species 3	C	G	**G**	T	C	C	G	G	T
Species 4	C	G	**G**	C	C	T	G	G	T
Score	0	0	**1**	1	0	2	0	0	0

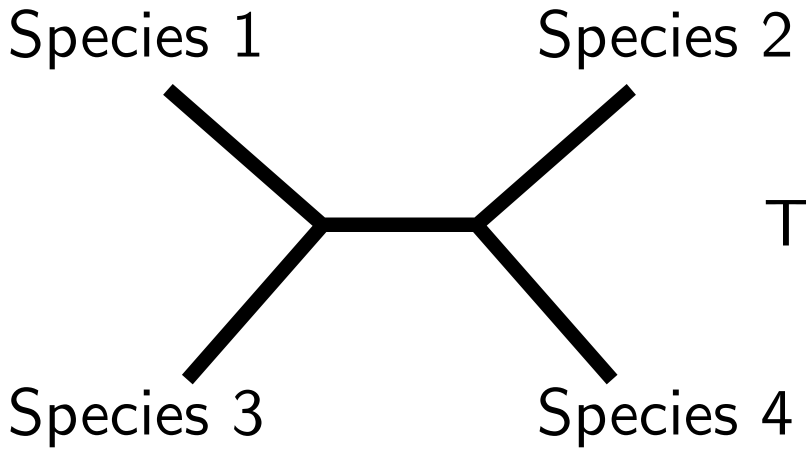


Tree 1 has a score of **4**

Considering a different tree

We can repeat the scoring for each tree.

	1	2	3	4	5	6	7	8	9
Species 1	C	G	**A**	C	C	A	G	G	T
Species 2	C	G	**A**	C	C	A	G	G	T
Species 3	C	G	**G**	T	C	C	G	G	T
Species 4	C	G	**G**	C	C	T	G	G	T
Score	0	0	**2**	1	0	2	0	0	0

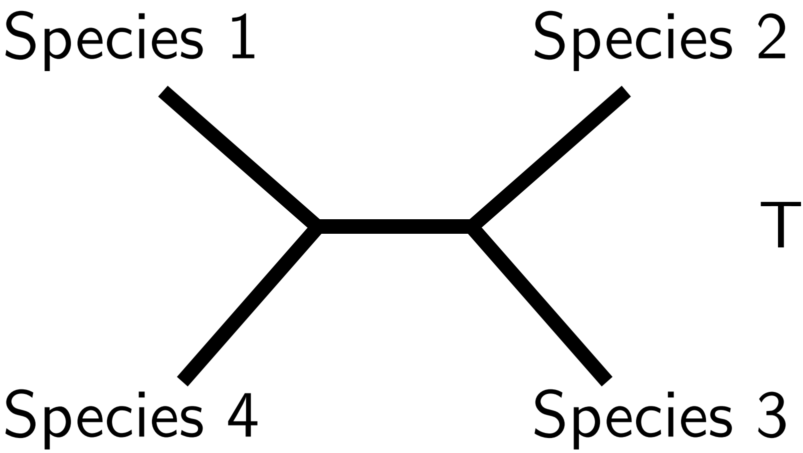


Tree 2 has a score of **5**

One more tree

Tree 3 has the same score as tree 2

	1	2	3	4	5	6	7	8	9
Species 1	C	G	**A**	C	C	A	G	G	T
Species 2	C	G	**A**	C	C	A	G	G	T
Species 3	C	G	**G**	T	C	C	G	G	T
Species 4	C	G	**G**	C	C	T	G	G	T
Score	0	0	**2**	1	0	2	0	0	0



Tree 3 has a score of **5**

Parsimony criterion prefers tree 1

Tree 1 required the *fewest* number of state changes (DNA substitutions) to explain the data.

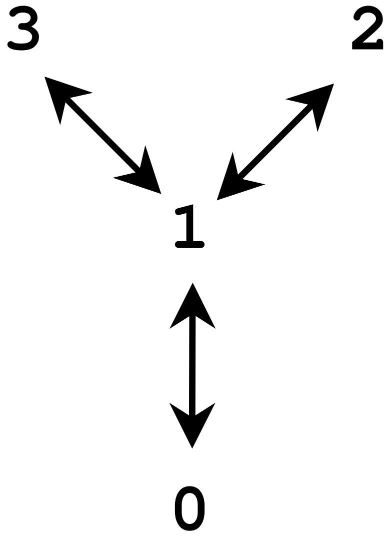
Some parsimony advocates equate the preference for the fewest number of changes to the general scientific principle of preferring the simplest explanation (Ockham's Razor), but this connection has not been made in a rigorous manner.

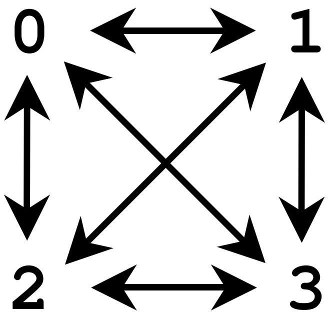
The parsimony criterion is equivalent to minimizing homoplasy.

In the example matrix at the beginning of these slides, only character 3 is parsimony informative.

	1	2	3	4	5	6	7	8	9
Species 1	C	G	**A**	C	C	A	G	G	T
Species 2	C	G	**A**	C	C	A	G	G	T
Species 3	C	G	**G**	T	C	C	G	G	T
Species 4	C	G	**G**	C	C	T	G	G	T
Max score	0	0	**2**	1	0	2	0	0	0
Min score	0	0	**1**	1	0	2	0	0	0

Assumptions about the evolutionary process can be incorporated using different step costs





Fitch Parsimony
“unordered”

Stepmatrices

Fitch Parsimony Stepmatrix

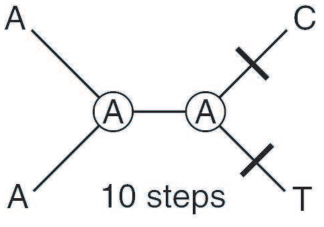
		To			
		A	C	G	T
From	A	0	1	1	1
	C	1	0	1	1
	G	1	1	0	1
	T	1	1	1	0

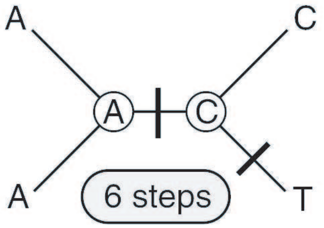
Stepmatrices

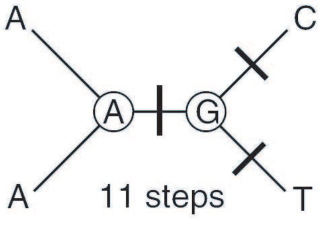
Transversion-Transition 5:1 Stepmatrix

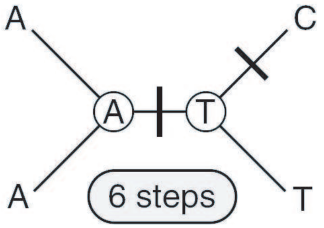
		To			
		A	C	G	T
	A	0	5	1	5
From	C	5	0	5	1
	G	1	5	0	5
	T	5	1	5	0

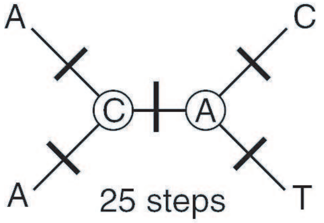
5:1 Transversion:Transition parsimony

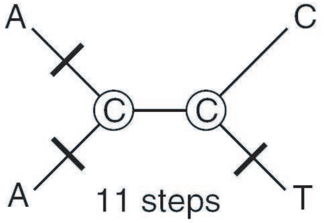


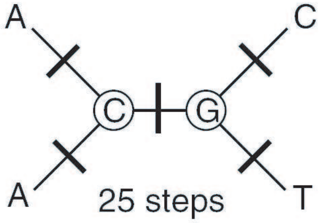


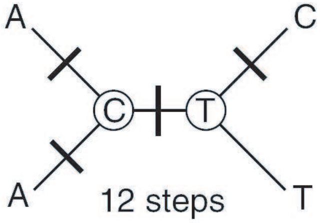


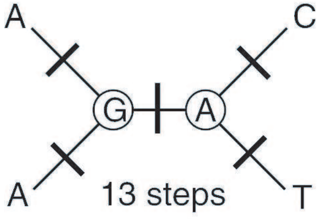


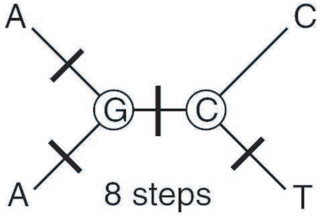


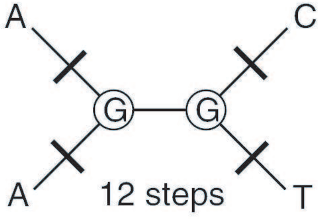


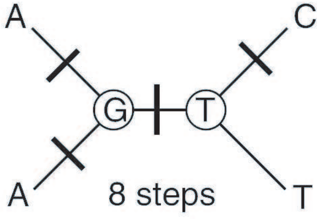


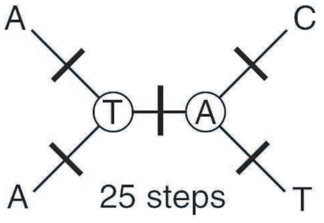


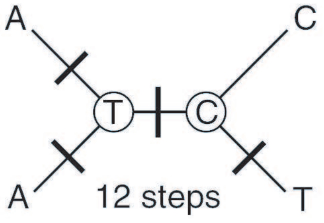


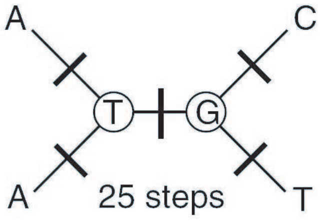


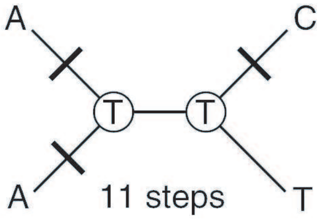












Stepmatrix considerations

- Parsimony scores from different stepmatrices cannot be meaningfully compared (31 under Fitch is not “better” than 45 under a transversion:transition stepmatrix)
- Parsimony cannot be used to infer the stepmatrix weights

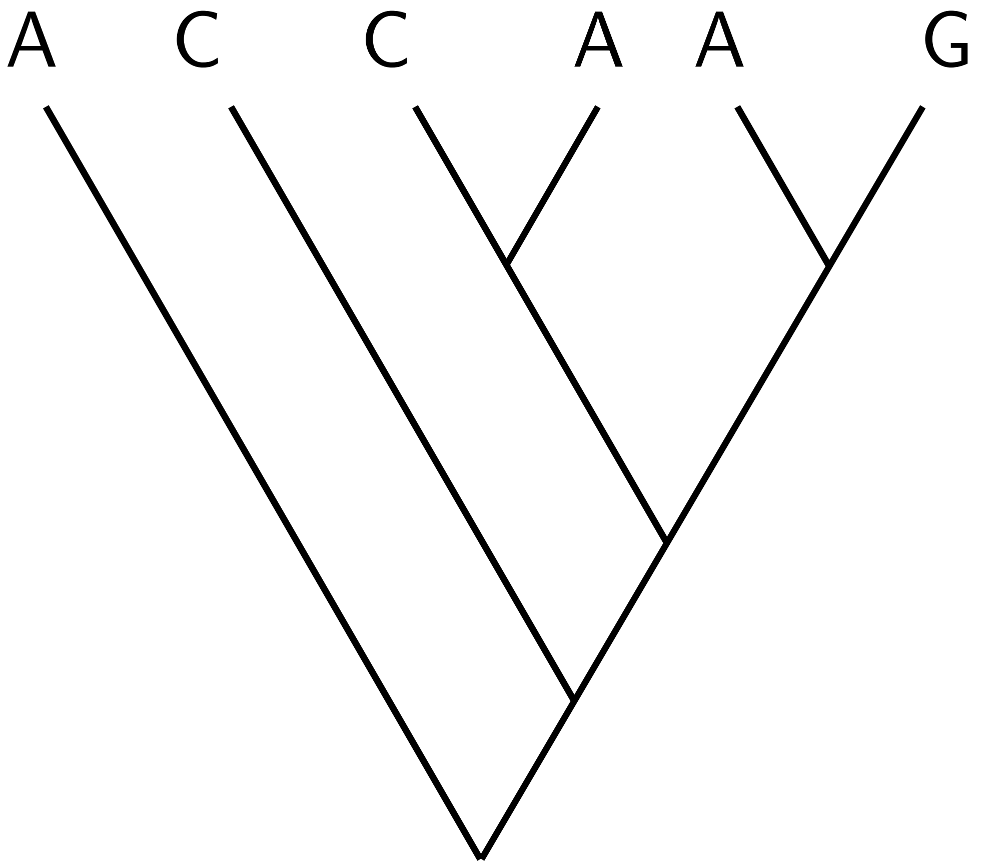
Other Parsimony variants

- *Dollo* derived state can only arise once, but reversals can be frequent (*e.g.* restriction enzyme sites).

- “weighted” - usually means that different characters are weighted differently (slower, more reliable characters usually given higher weights).

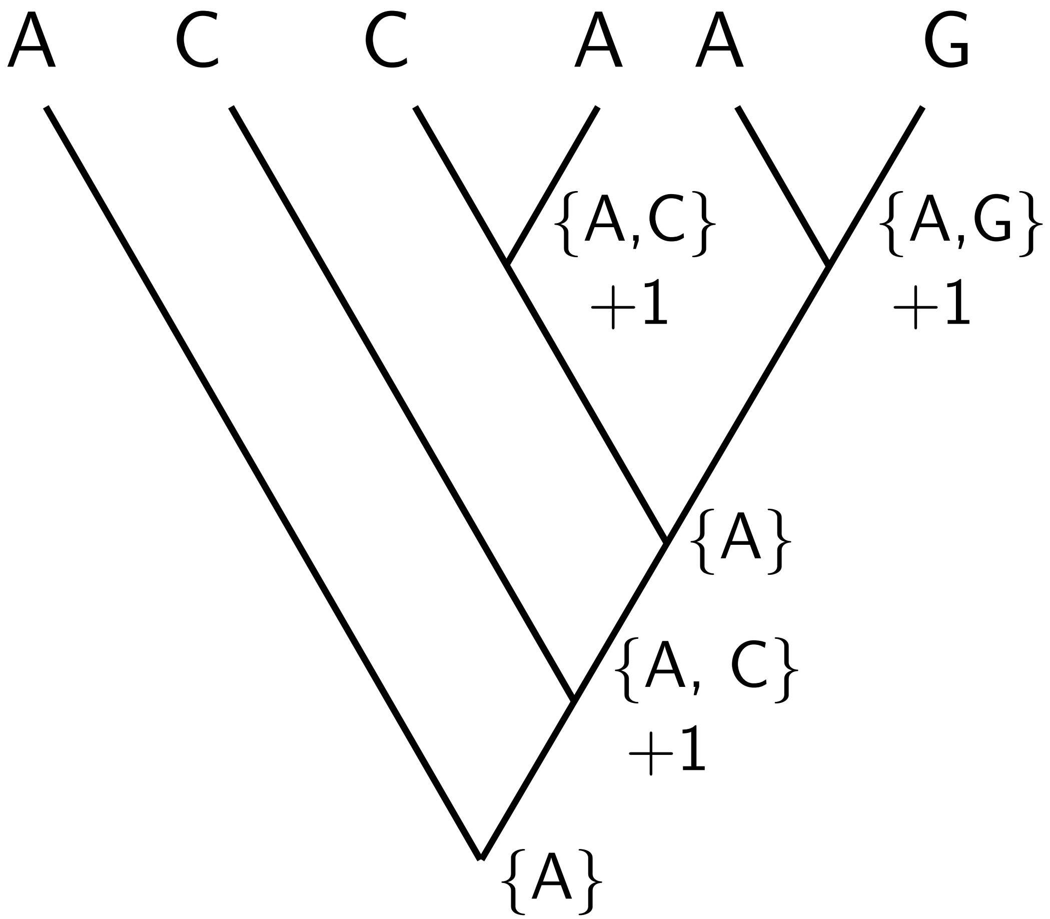
- implied weights **?**

Scoring trees under parsimony is fast



Scoring trees under parsimony is fast – Fitch algorithm

3 steps



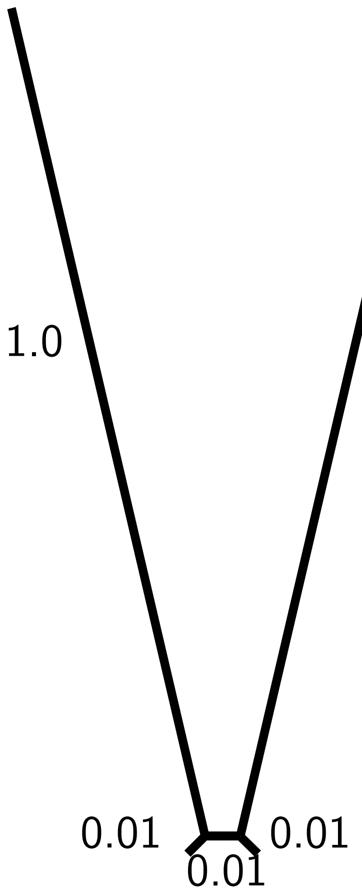
Scoring a trees under parsimony

- Fitch and Sankoff algorithms are fast (# of calculations increases linearly with the # of leaves, despite the fact that the # of possible ancestral state reconstructions increases exponentially).
- Fitch and Sankoff algorithms are guaranteed to succeed (return the minimum # of changes).
- Elaborations of these algorithms allow us to reconstruct ancestral states.

Qualitative description of parsimony

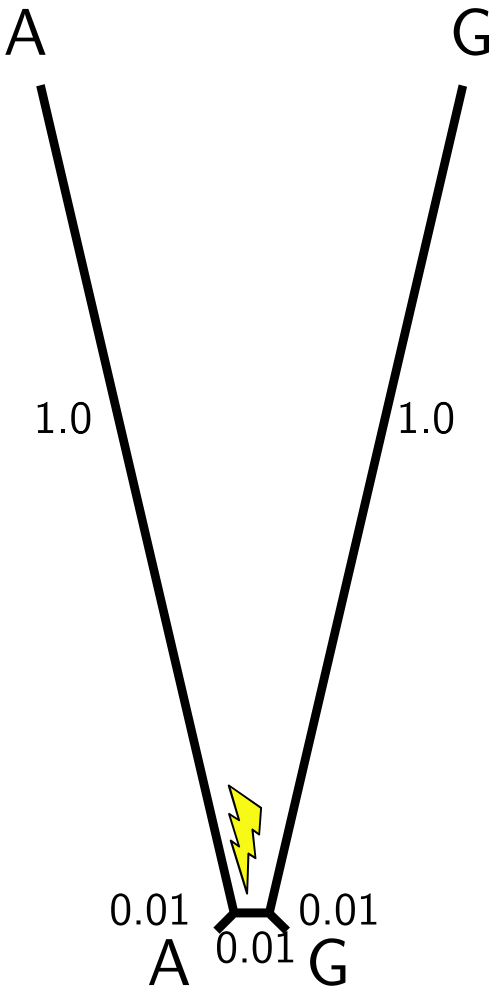
- It can perform well even when changes are not rare.
- Does not "prefer" to put changes on one branch over another
- Hard to characterize statistically
  - the set of conditions in which parsimony is guaranteed to work well is very restrictive (low probability of change and not too much branch length heterogeneity);
  - Parsimony often performs well in simulation studies (even when outside the zones in which it is guaranteed to work);
  - Estimates of the tree can be extremely biased.

Long branch attraction



Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

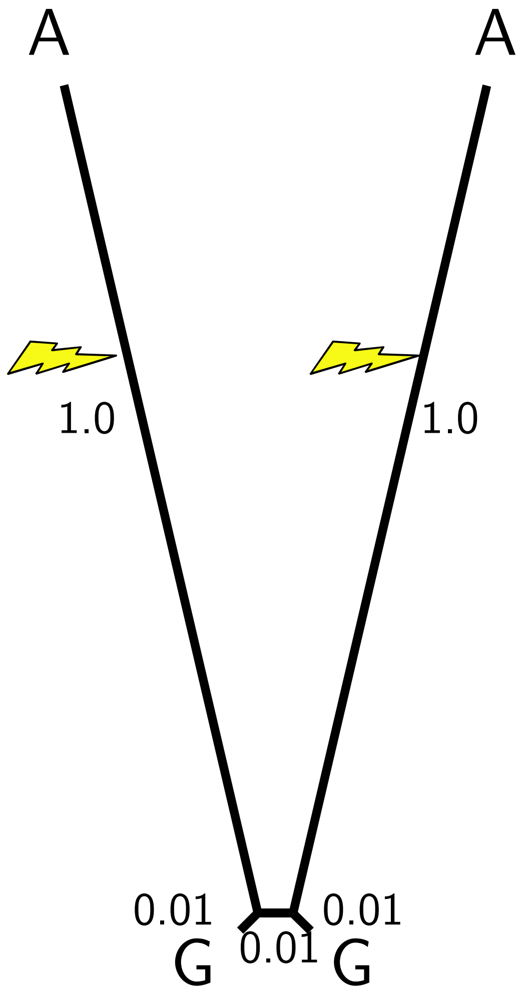
Long branch attraction



Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

Long branch attraction



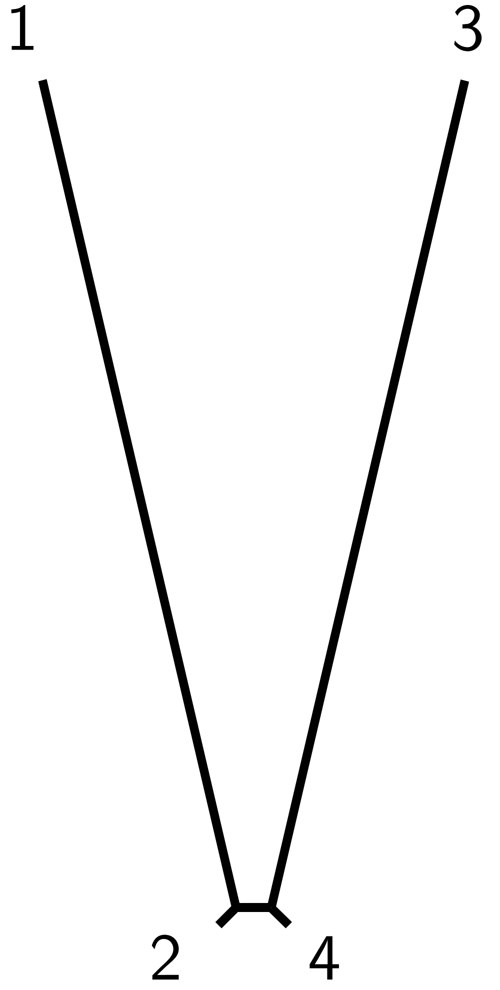
Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

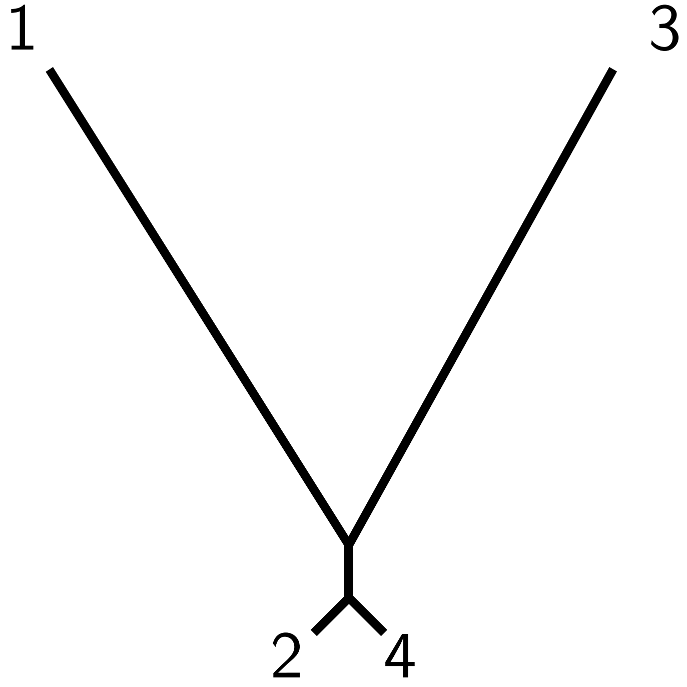
The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).

Long branch attraction

Parsimony is almost guaranteed to get this tree wrong.



True



Inferred

Inconsistency

- Statistical Consistency (roughly speaking) is converging to the true answer as the amount of data goes to ∞.
- Parsimony based tree inference is *not* consistent for some tree shapes. In fact it can be "positively misleading":
  - "Felsenstein zone" tree
  - Many clocklike trees with short internal branch lengths and long terminal branches (Penny *et al.*, 1989, Huelsenbeck and Lander, 2003).
- Methods for assessing confidence (e.g. bootstrapping) will indicate that you should be very confident in the wrong answer.