Module 19: Molecular Phylogenetics

MTH thanks to Paul Lewis, Tracy Heath, Joe Felsenstein, Peter Beerli, Derrick Zwickl, and Joe Bielawski for slides

Wednesday July 27: Day I

1:30PM to 3:00PM	Introduction Parsimony methods for phylogeny reconstruction Distance–based methods for phylogeny reconstruction (Mark Holder)
3:30PM to 5:00PM	Topology Searching (Mark Holder) Parsimony and distances demo in PAUP* (Mark Holder)

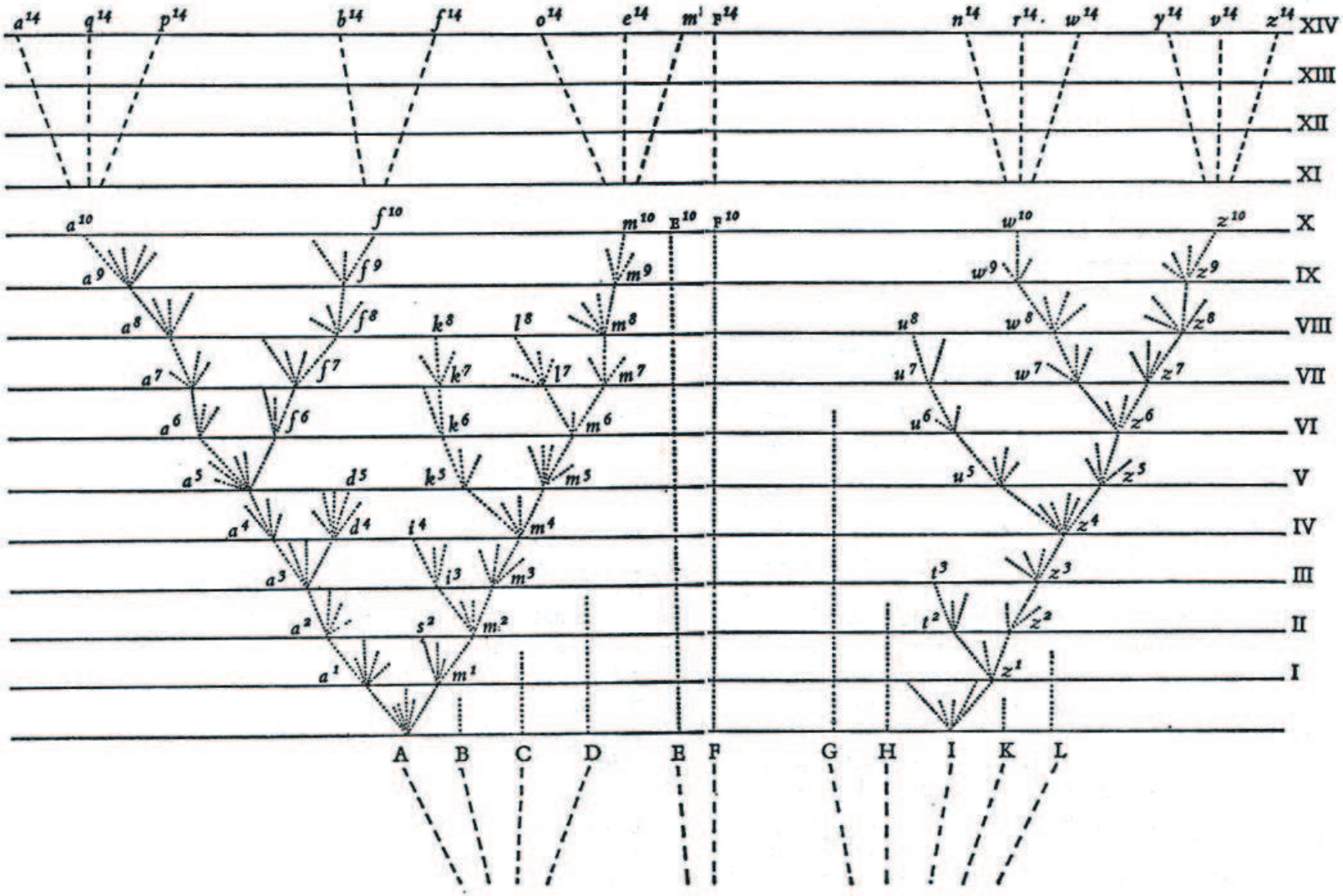
Thursday July 28: Day II

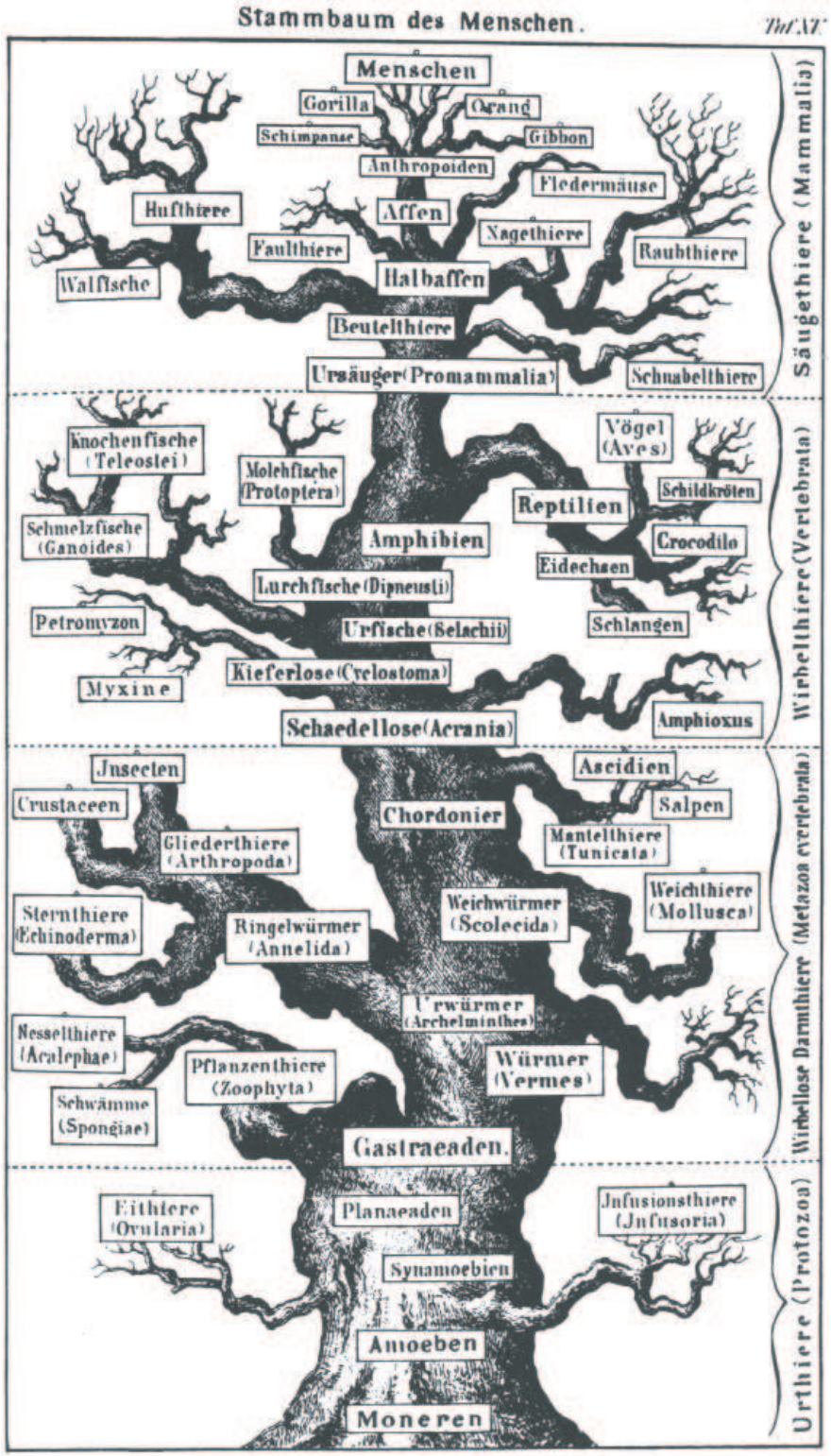
8:30AM to 10:00AM	Nucleotide Substitution Models and Transition Probabilities (Jeff Thorne) Likelihood – (Joe Felsenstein)
10:30AM to noon	PHYLIP lab: likelihood – (Joe Felsenstein) PAUP* lab (Mark Holder)
1:30PM to 3:00PM	Bootstraps and Testing Trees (Joseph Felsenstein) Bootstrapping in Phylip (Joe Felsenstein)
3:30PM to 5:00PM	More Realistic Evolutionary Models (Jeff Thorne)

Friday July 29: Day III

8:30AM to 10:00AM	Bayesian Inference and Bayesian Phylogenetics (Jeff Thorne)
10:30AM to noon	MrBayes Computer Lab – (Mark Holder) Divergence Time Estimation (Jeff Thorne)
1:30PM to 3:00PM	Divergence Time Estimation (continued) (Jeff Thorne) BEAST demo (Mark Holder)
3:30PM to 5:00PM	The Coalescent – (Joe Felsenstein) The Comparative Method – (Joe Felsenstein) Future Directions – (Joe Felsenstein)

Darwin's 1859 "On the Origin of Species" had one figure:





Human family tree

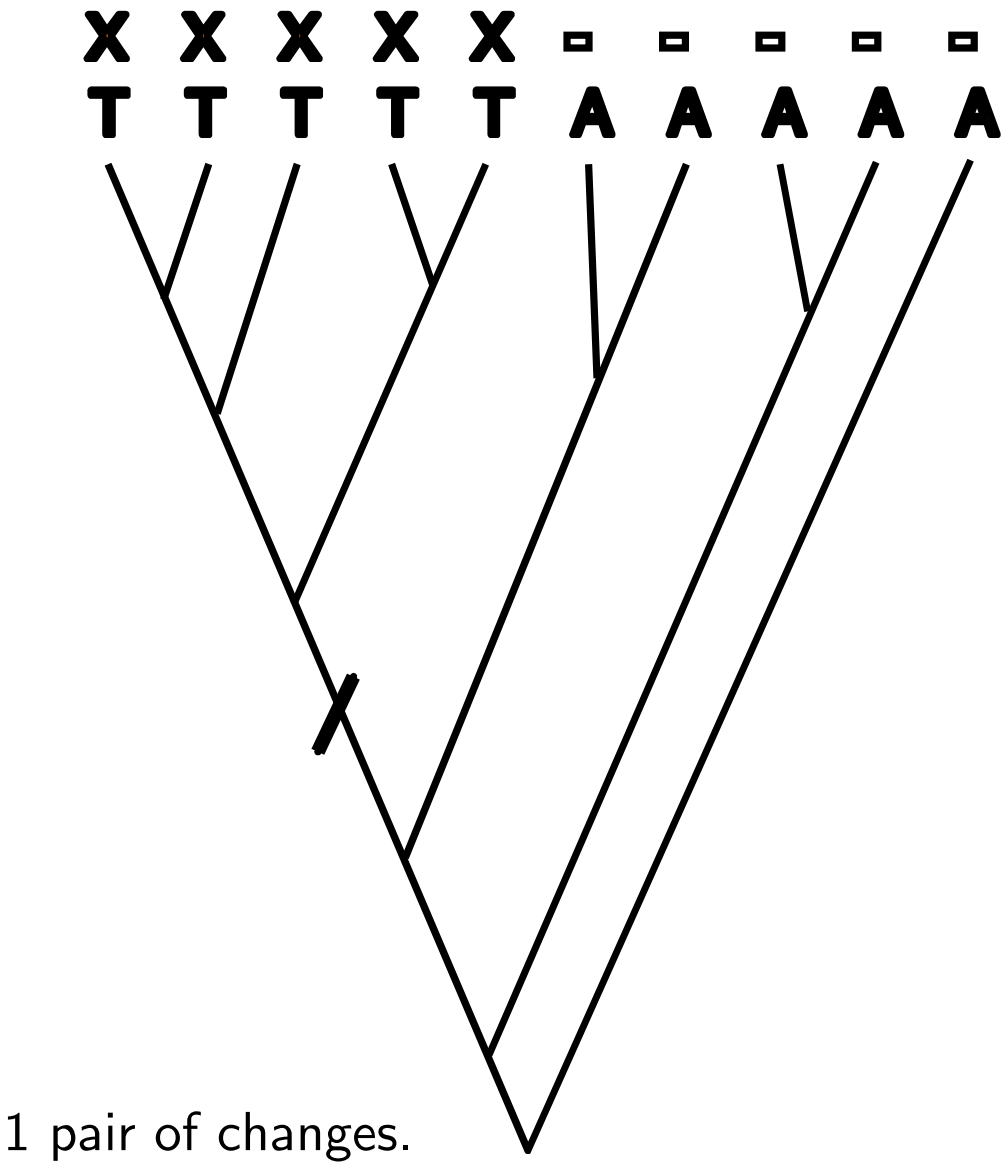
from Haeckel, 1874

Fig. 20, p. 171, in Gould, S. J. 1977.
Ontogeny and phylogeny.
Harvard University Press, Cambridge, MA

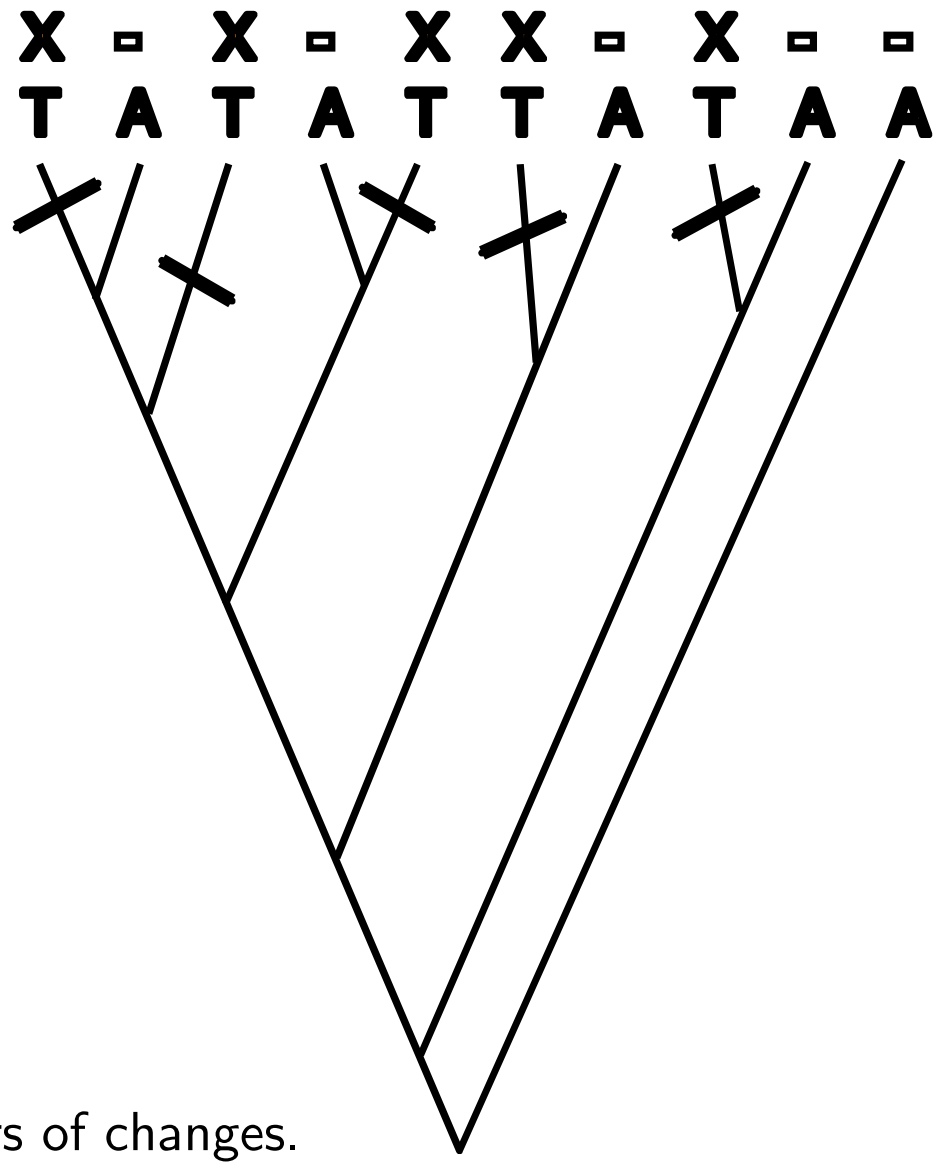
Are desert green algae adapted to high light intensities?

Species	Habitat	Photoprotection
1	terrestrial	xanthophyll
2	terrestrial	xanthophyll
3	terrestrial	xanthophyll
4	terrestrial	xanthophyll
5	terrestrial	xanthophyll
6	aquatic	none
7	aquatic	none
8	aquatic	none
9	aquatic	none
10	aquatic	none

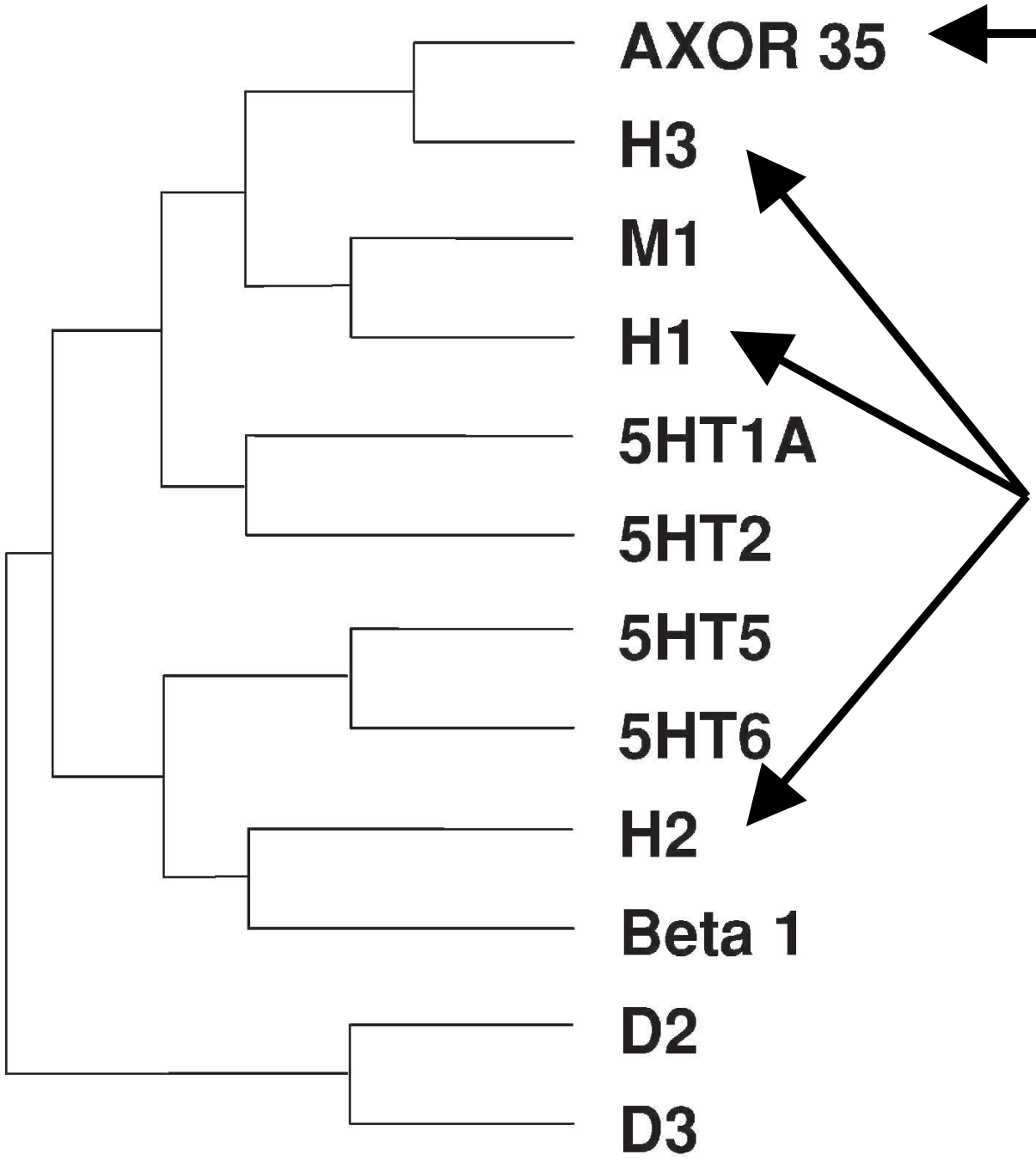
Phylogeny reveals the events that generate the pattern



1 pair of changes.
Coincidence?



5 pairs of changes.
Much more convincing



GPCR with unknown ligand

Natural ligand known
to be histamine

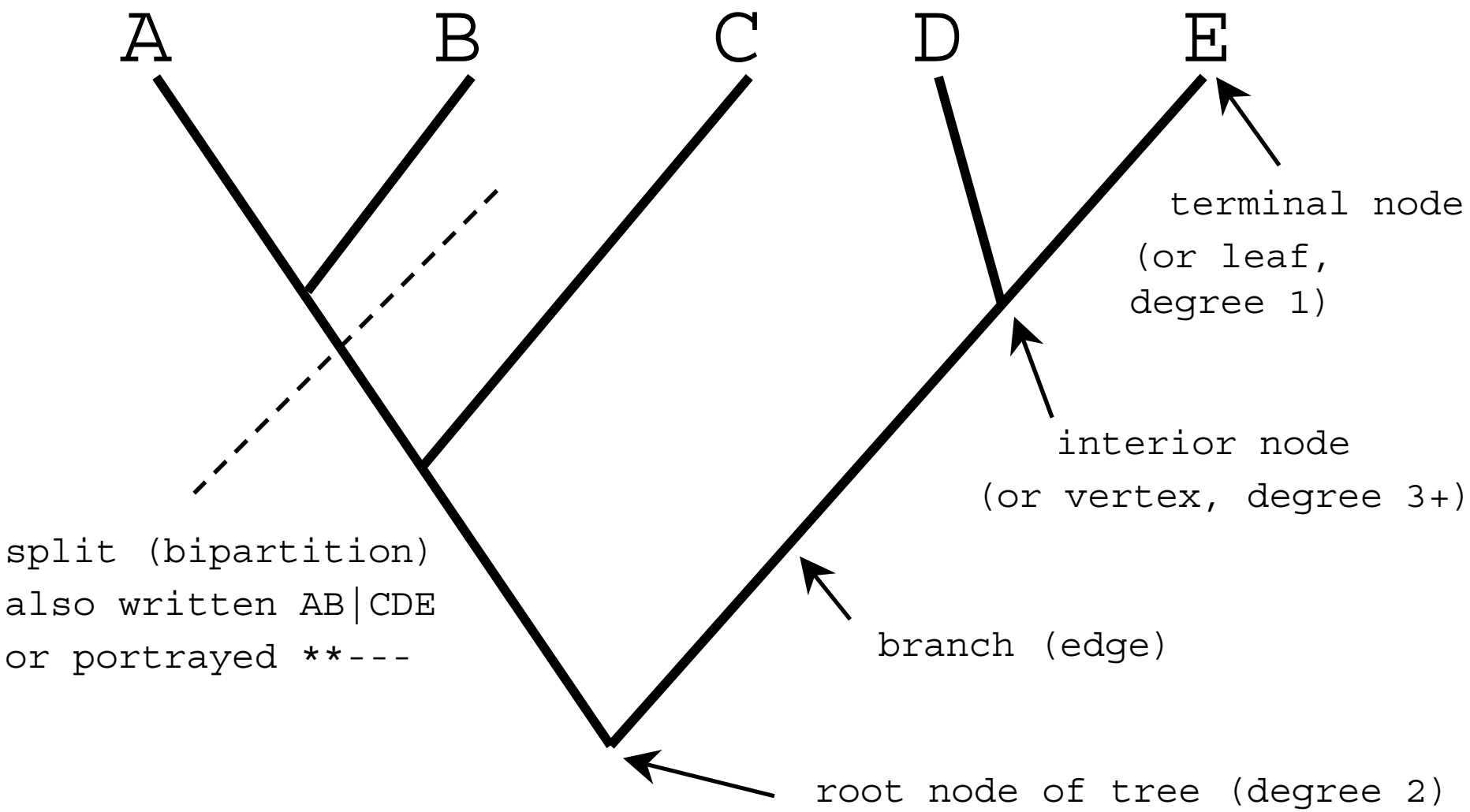
Which ligand for AXOR35
would you test first?

Wise, A., Jupe, S. C., and Rees, S. 2004. The identification of ligands at orphan G-protein coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* 44:43-66.

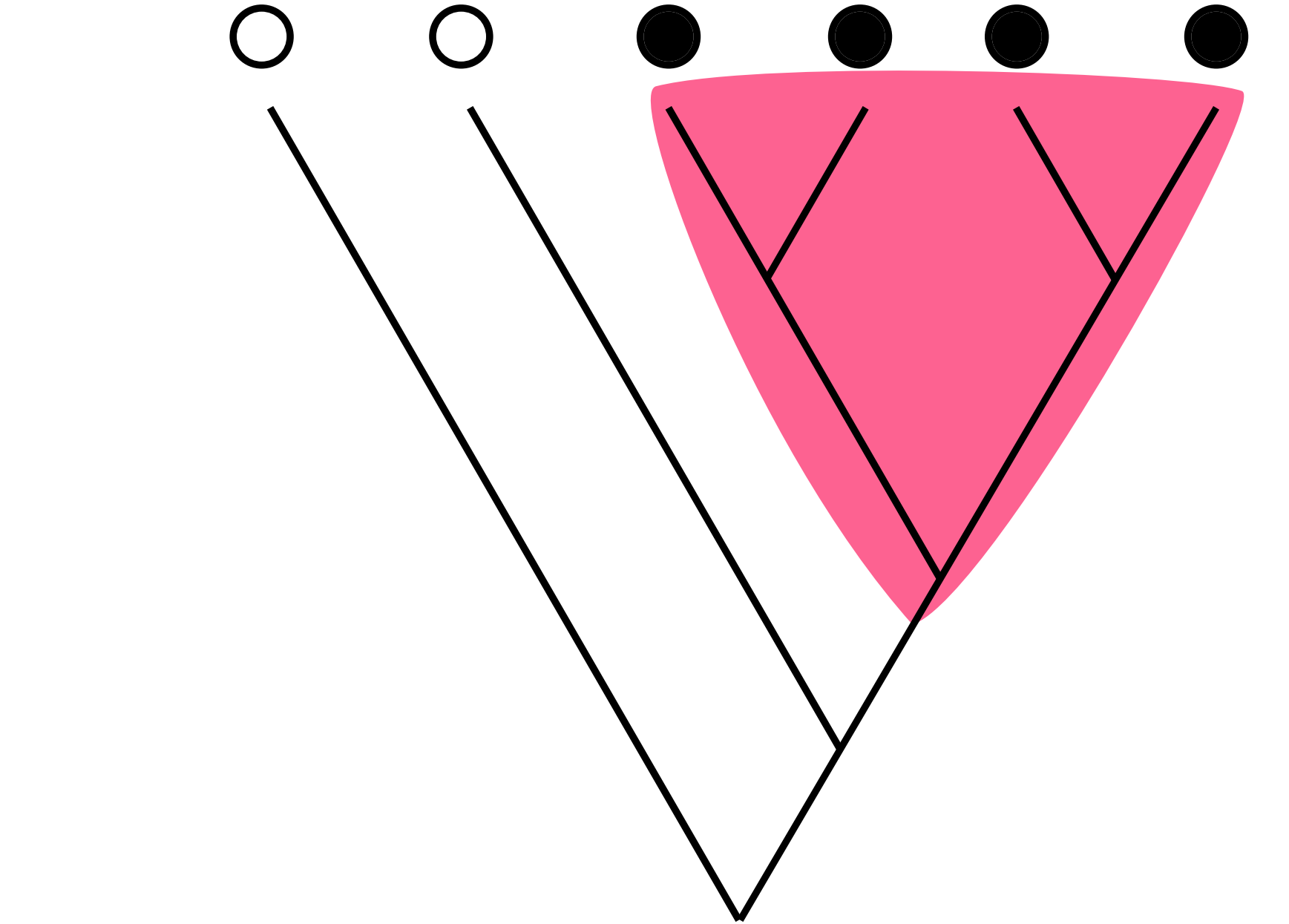
Many evolutionary questions require a phylogeny

- Estimating the number of times a trait evolved
- Determining whether a trait tends to be lost more often than gained, or vice versa
- Estimating divergence times
- Distinguishing homology from analogy
- Inferring parts of a gene under strong positive selection

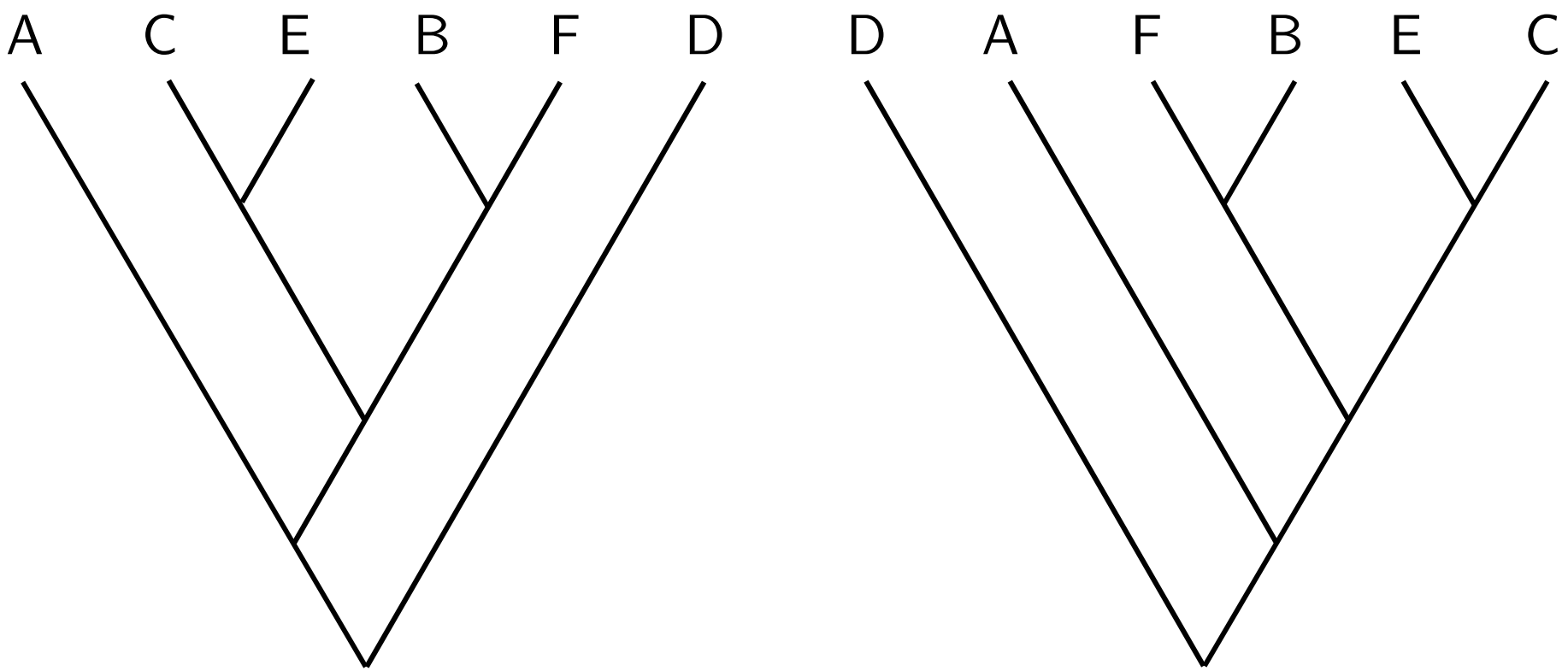
Tree terminology



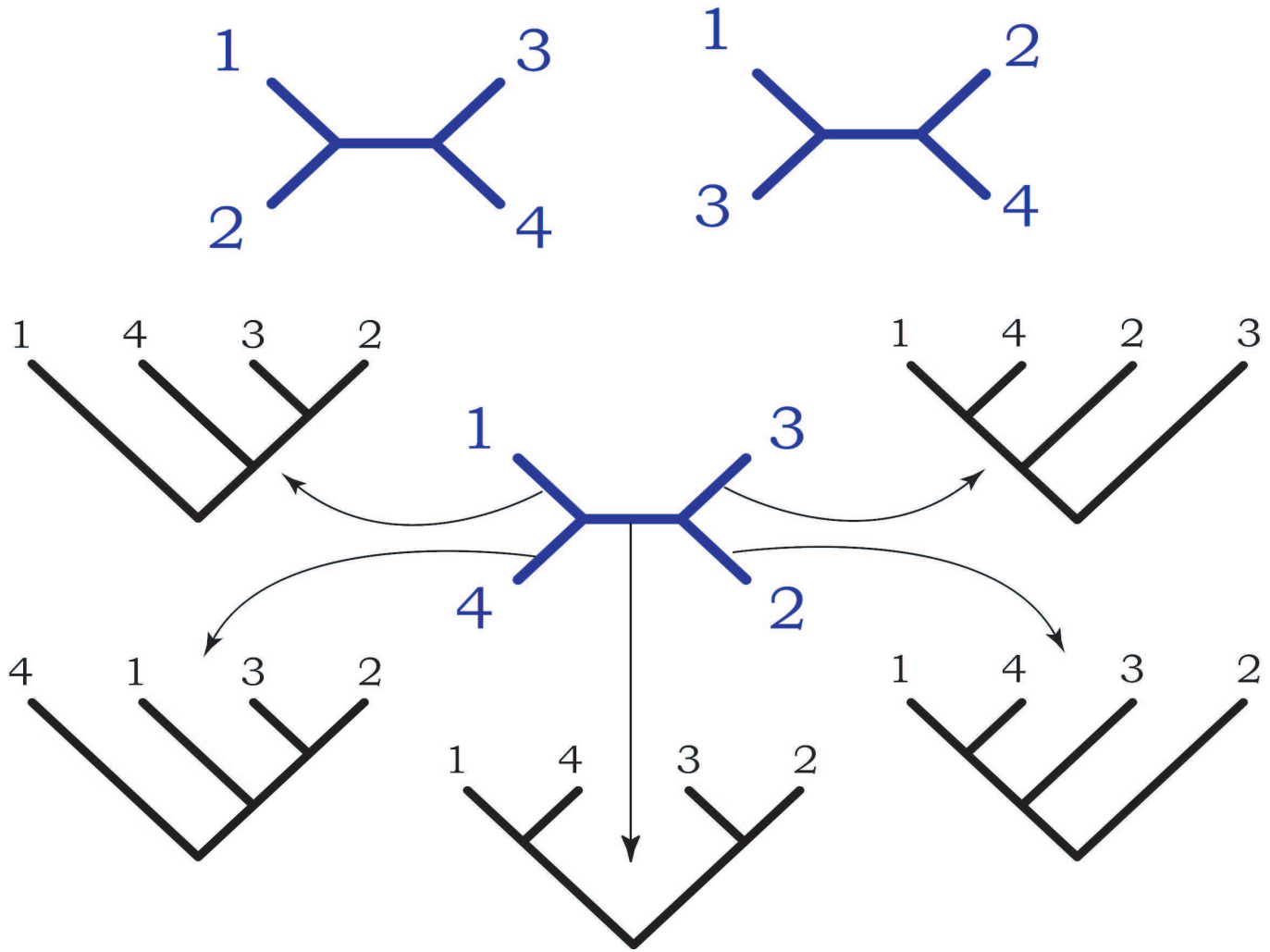
Monophyletic groups ("clades"): the basis of phylogenetic classification



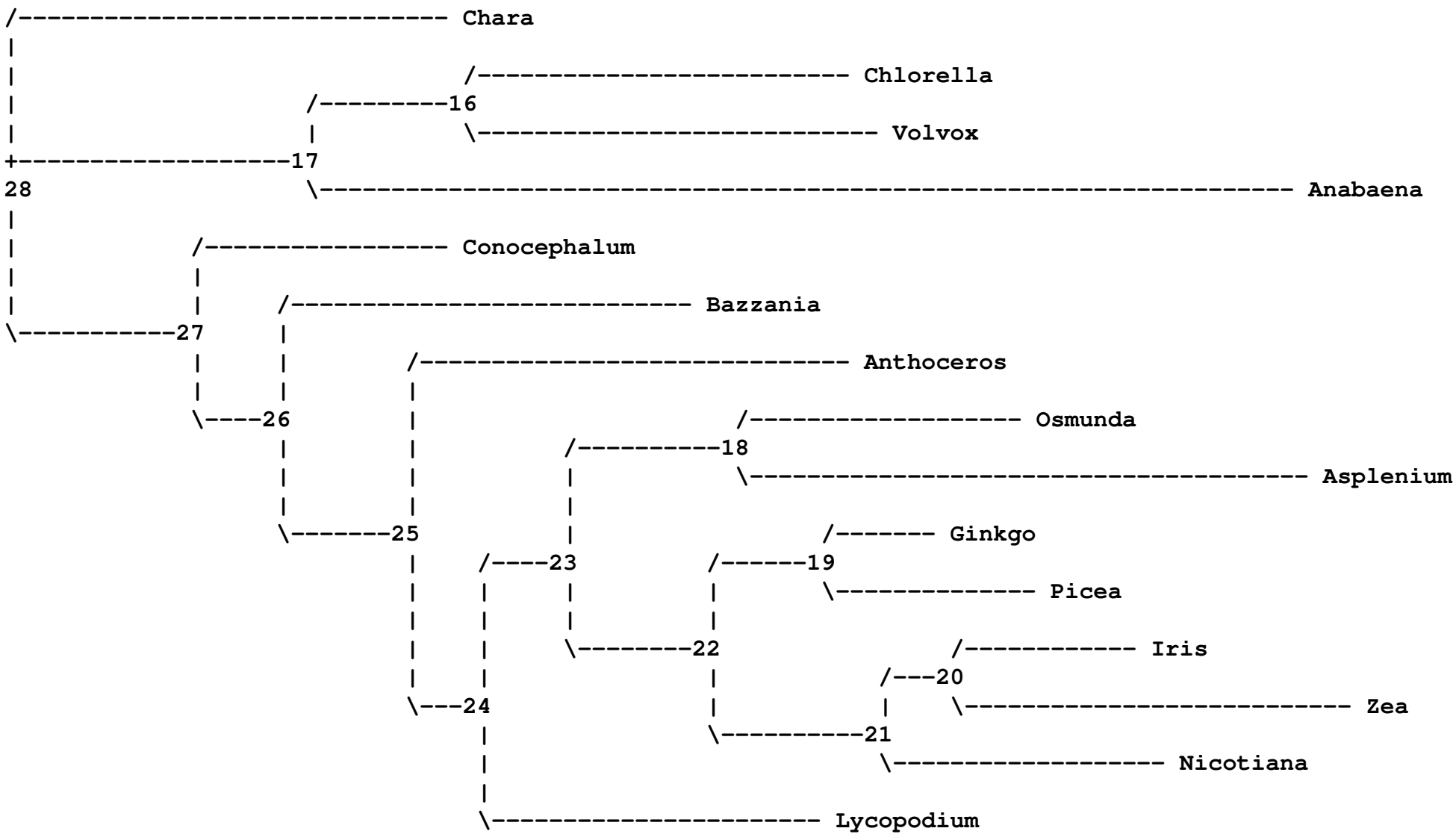
Branch rotation does not matter



Rooted vs unrooted trees



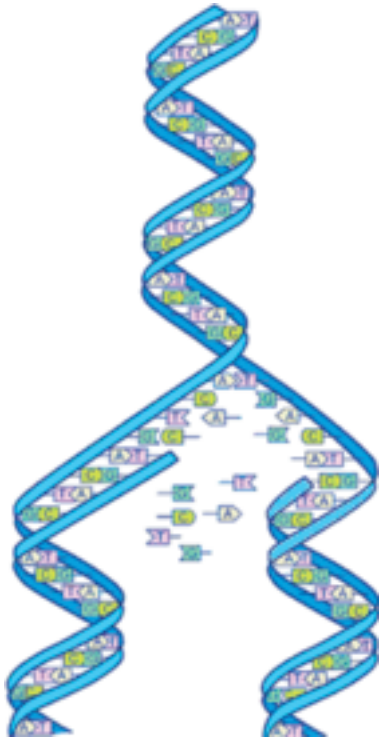
Warning: software often displays unrooted trees like this:

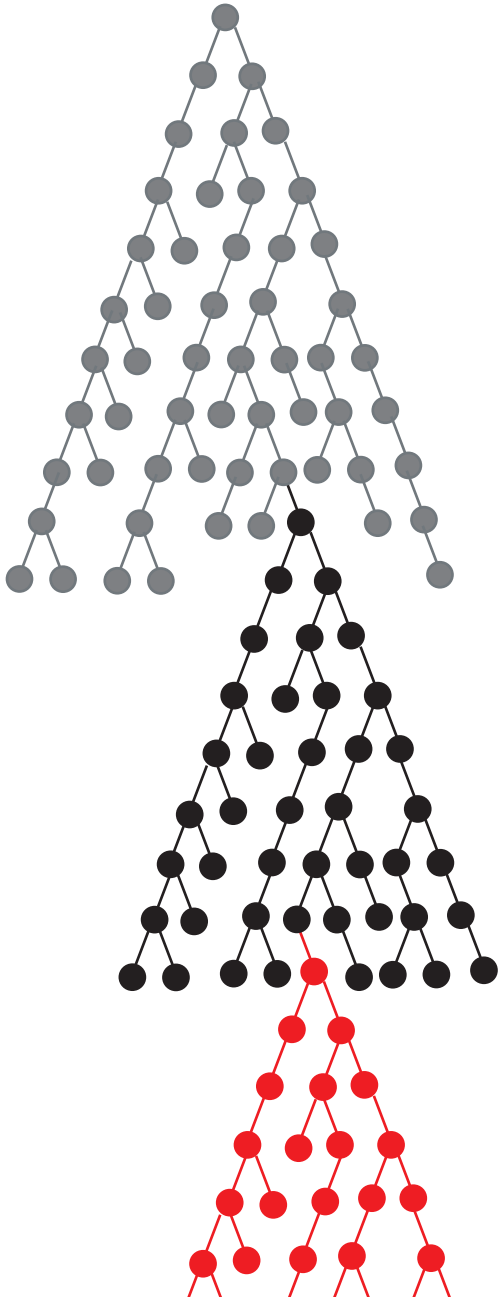


We use trees to represent genealogical relationships in several contexts.

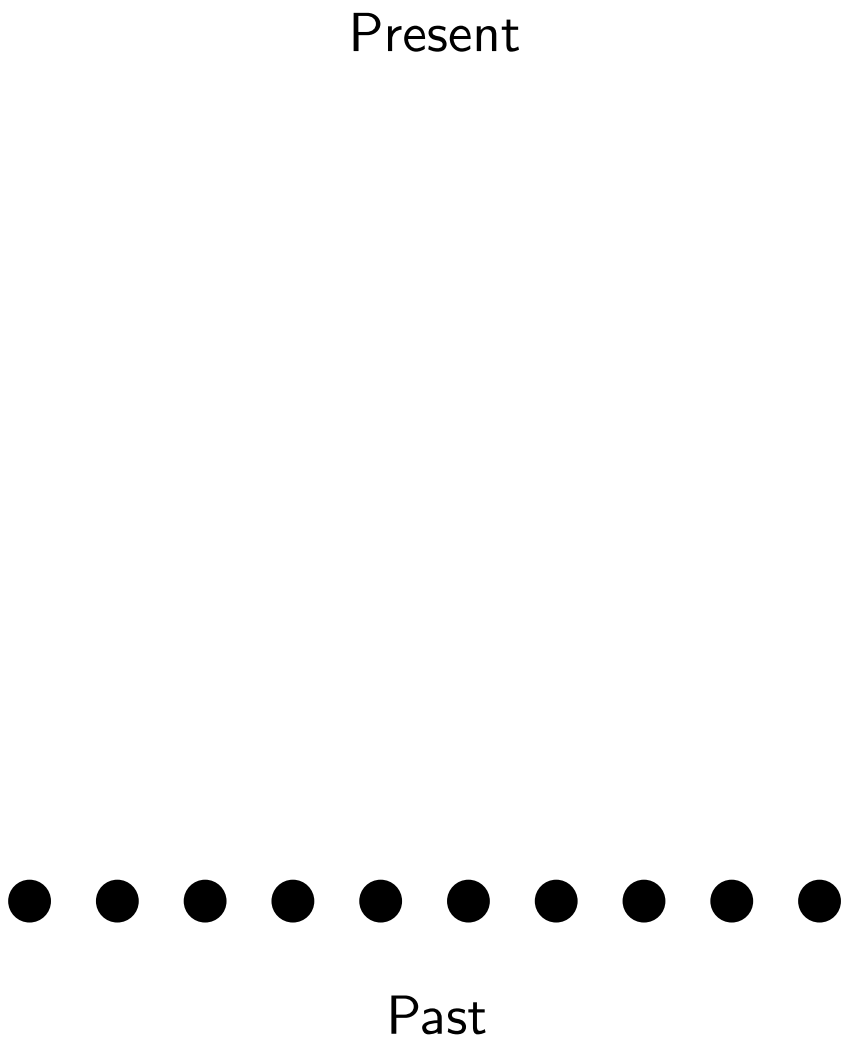
Domain	Sampling	tree	The cause of splitting
Population Genetics	> 1 indiv/sp. Few species	Gene tree	> 1 descendants of a single gene copy
Phylogenetics	Few indiv/sp. Many species	Phylogeny	speciation
Molecular Evolution	> 1 locus/sp. > 1 species	Gene tree. Gene family tree	speciation or duplication

Phylogenies are an inevitable result of molecular genetics



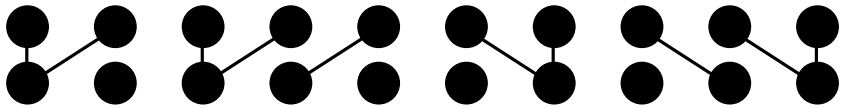


Genealogies within a population



Genealogies within a population

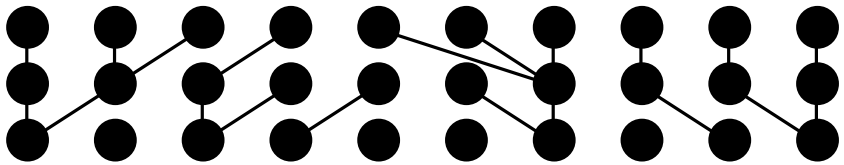
Present



Past

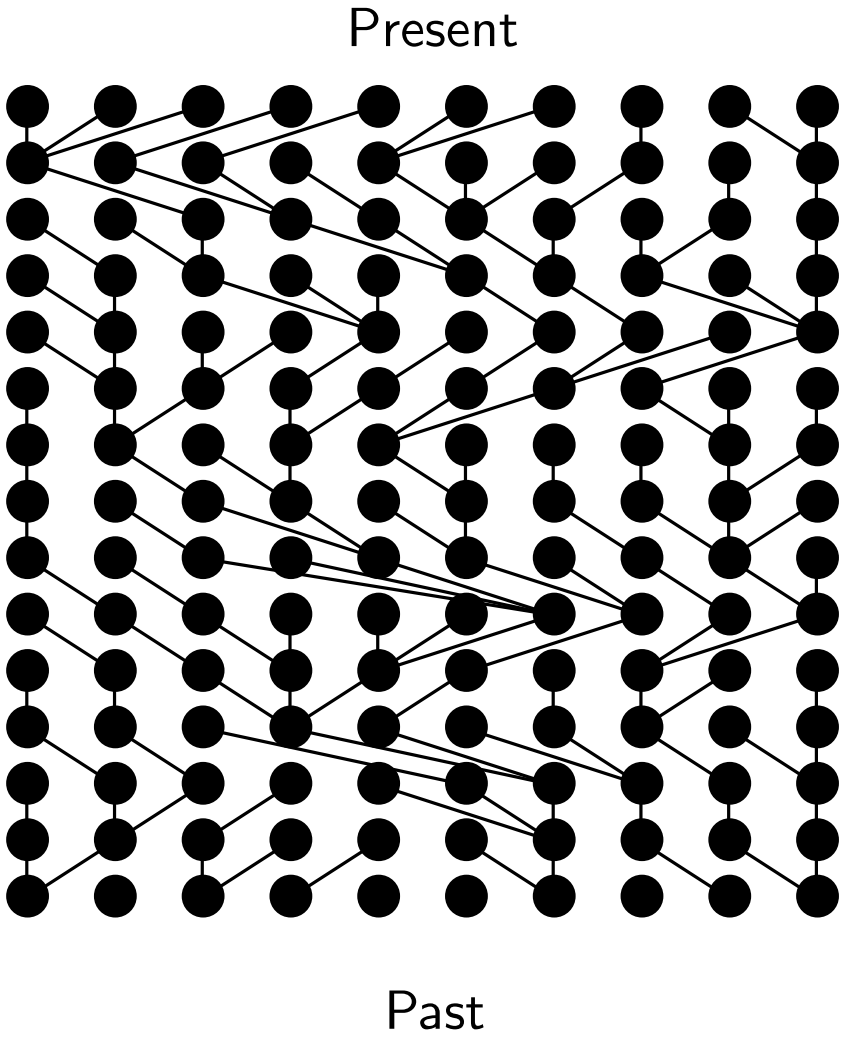
Genealogies within a population

Present

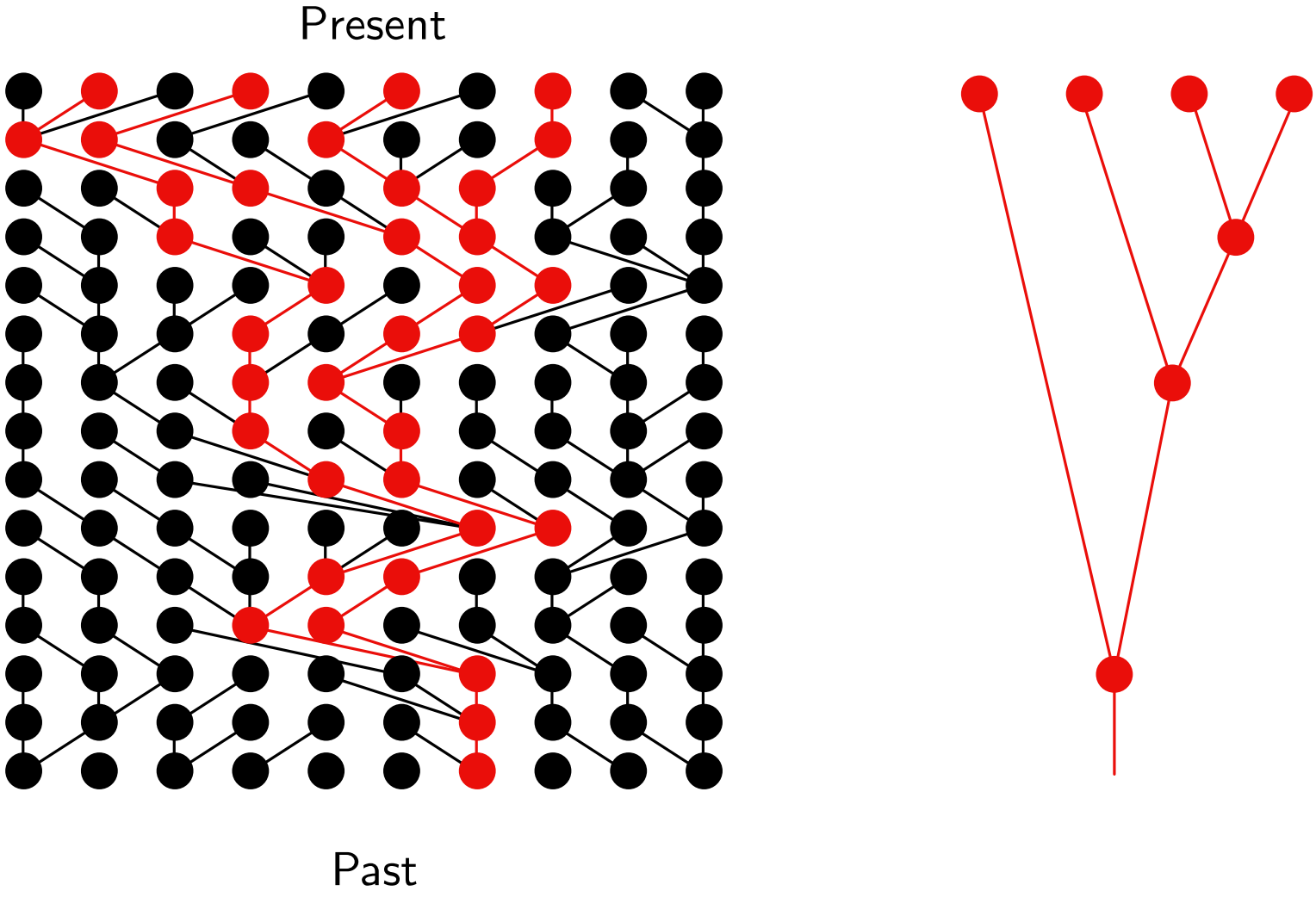


Past

Genealogies within a population



Genealogies within a population



Biparental inheritance would make the picture messier, but the genealogy of the gene copies would still form a tree (if there is no recombination).

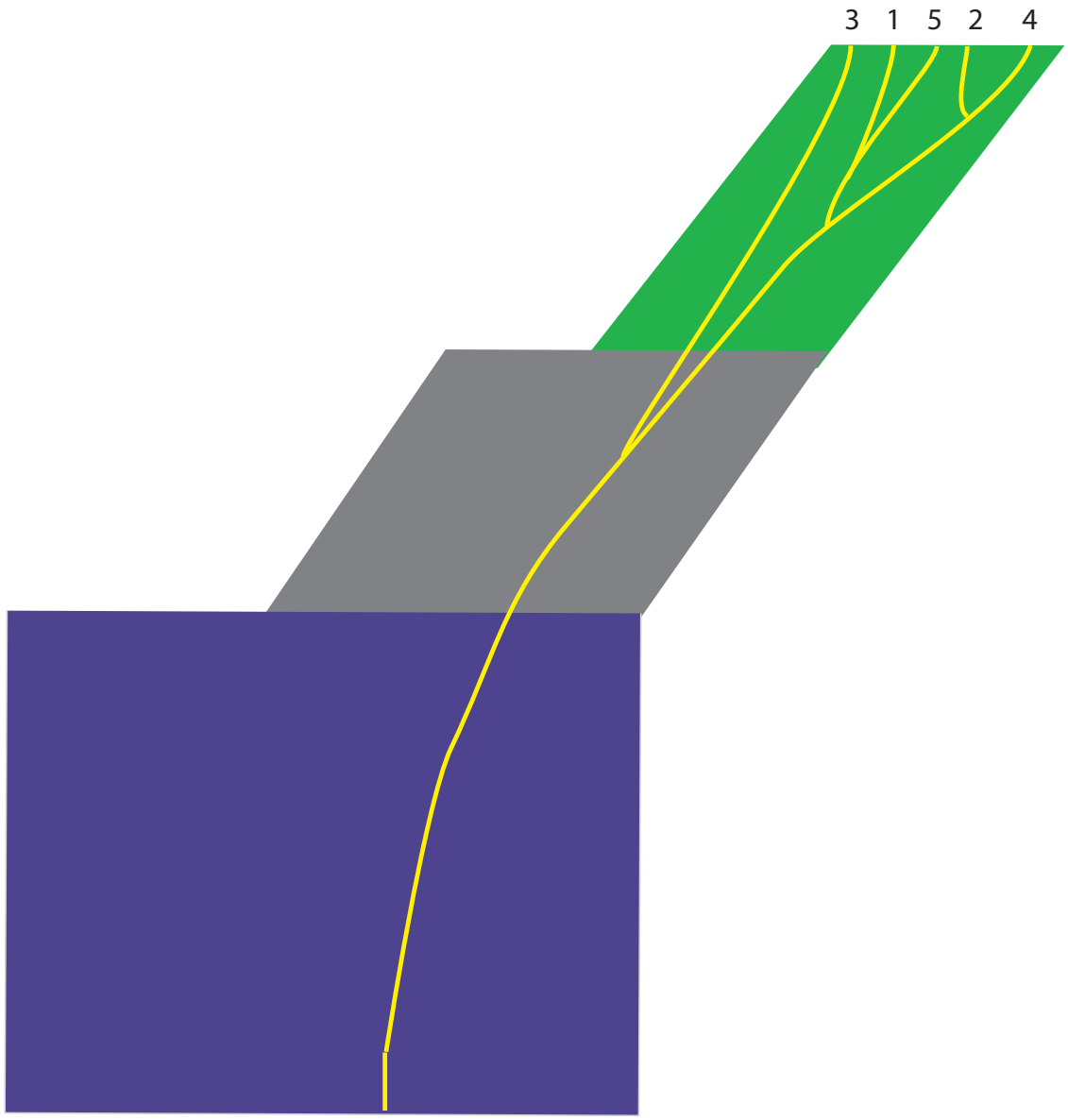
terminology: genealogical trees within population or species trees

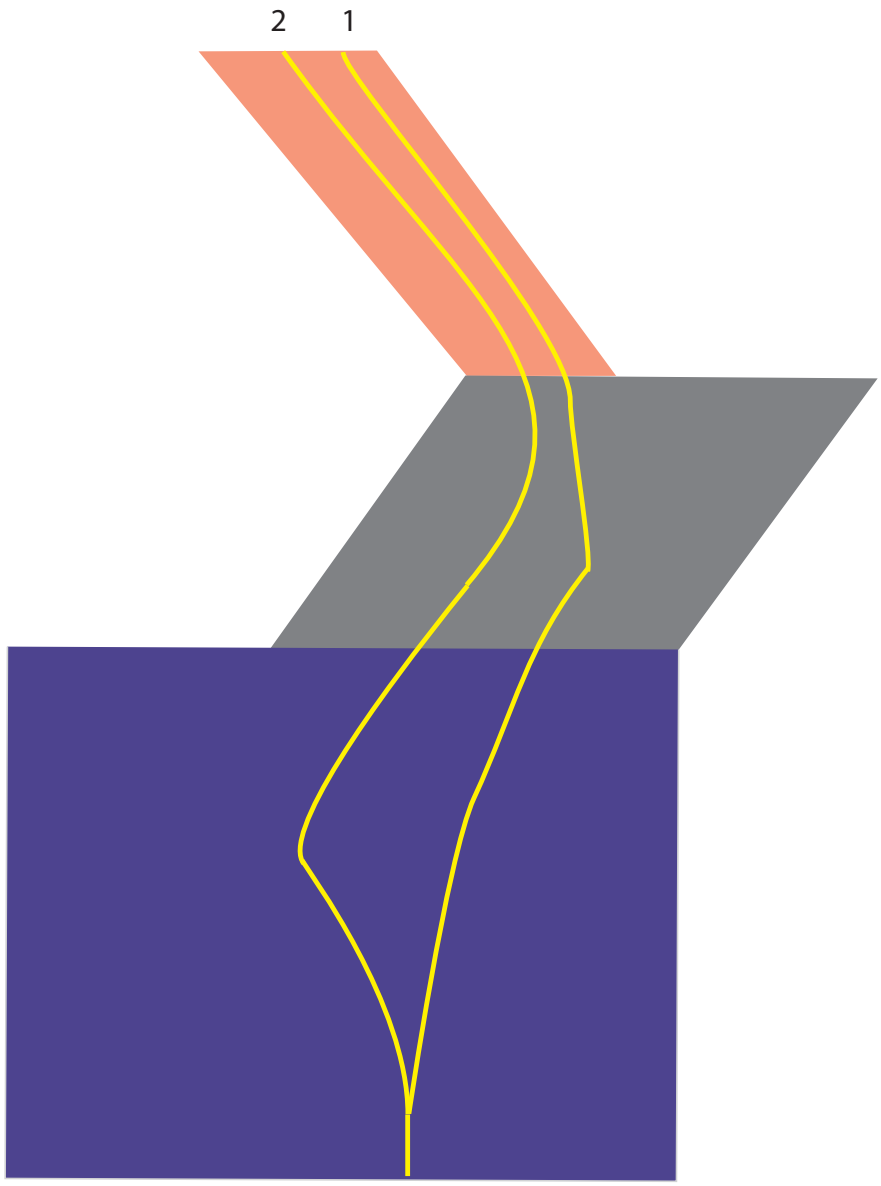
It is tempting to refer to the tips of these gene trees as alleles or haplotypes.

- allele – an alternative form a gene.
- haplotype – a linked set of alleles

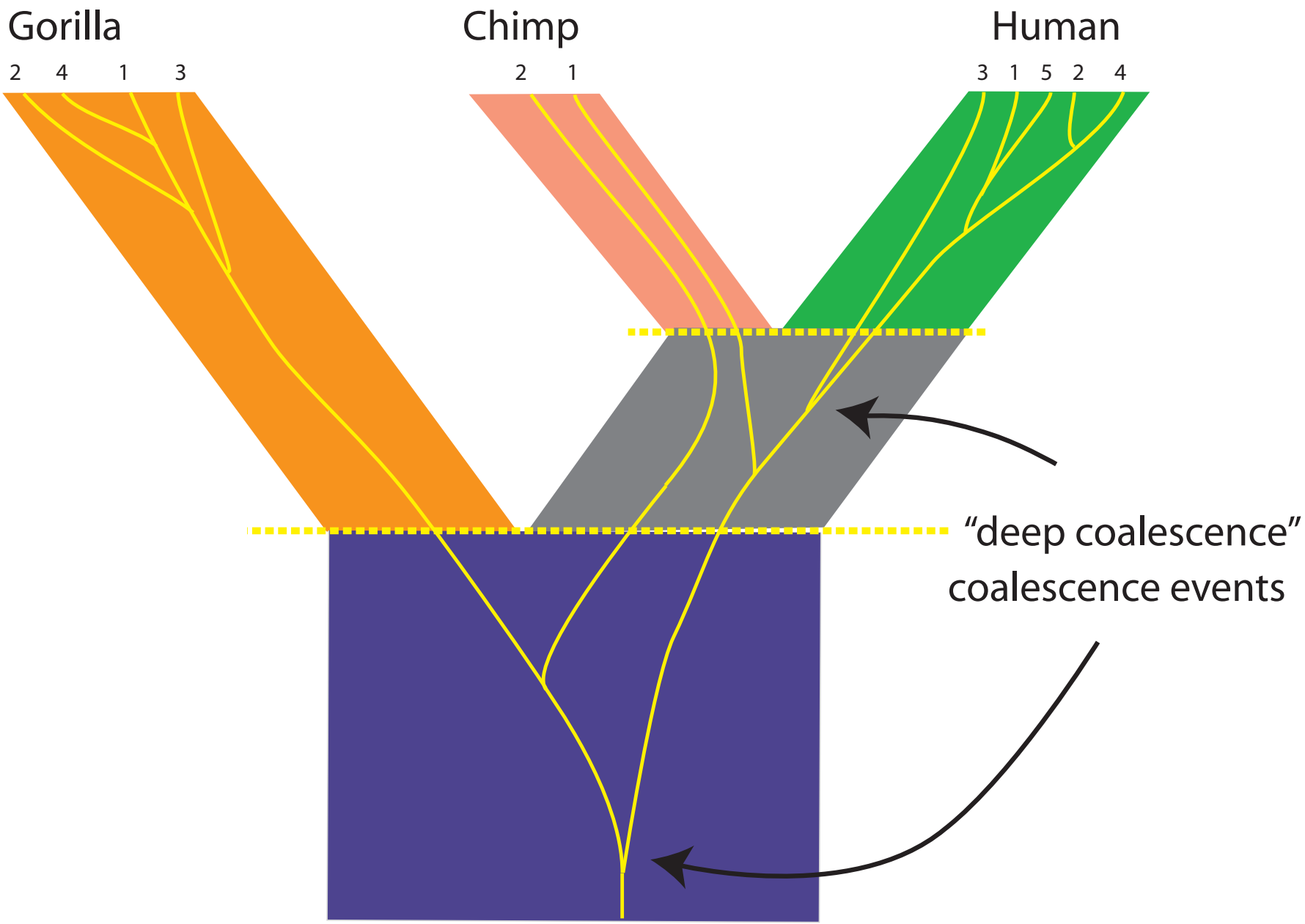
But both of these terms require a differences in sequence.

The gene trees that we draw depict genealogical relationships – regardless of whether or not nucleotide differences distinguish the "gene copies" at the tips of the tree.


3 1 5 2 4


2
1

A “gene tree” within a species tree



terminology: genealogical trees within population or species trees

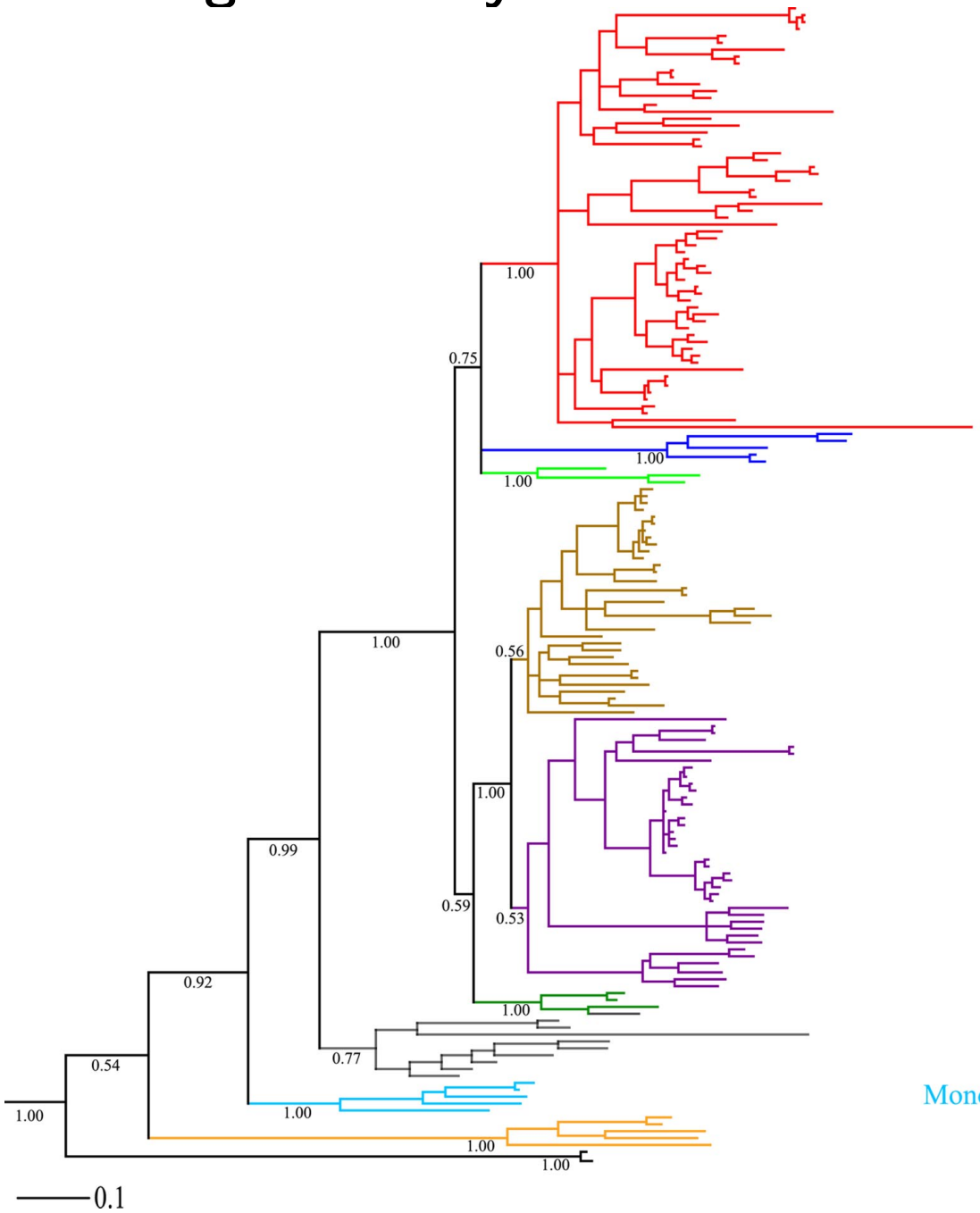
- coalescence – merging of the genealogy of multiple gene copies into their common ancestor. "Merging" only makes sense when viewed *backwards in time*.

- "deep coalescence" or "incomplete lineage sorting" refer to the *failure* of gene copies to coalesce within the duration of the species – the lineages coalesce in an ancestral species

Inferring a species tree while accounting for the coalescent

Figure 2 from Heled and Drummond (2010)

A "gene family tree"



Eutherian β-δ globin

Marsupial β-globin

Monotreme ε^P-β^Pglobin

Eutherian ε-globin

Eutherian γ-globin

Marsupial ε-globin

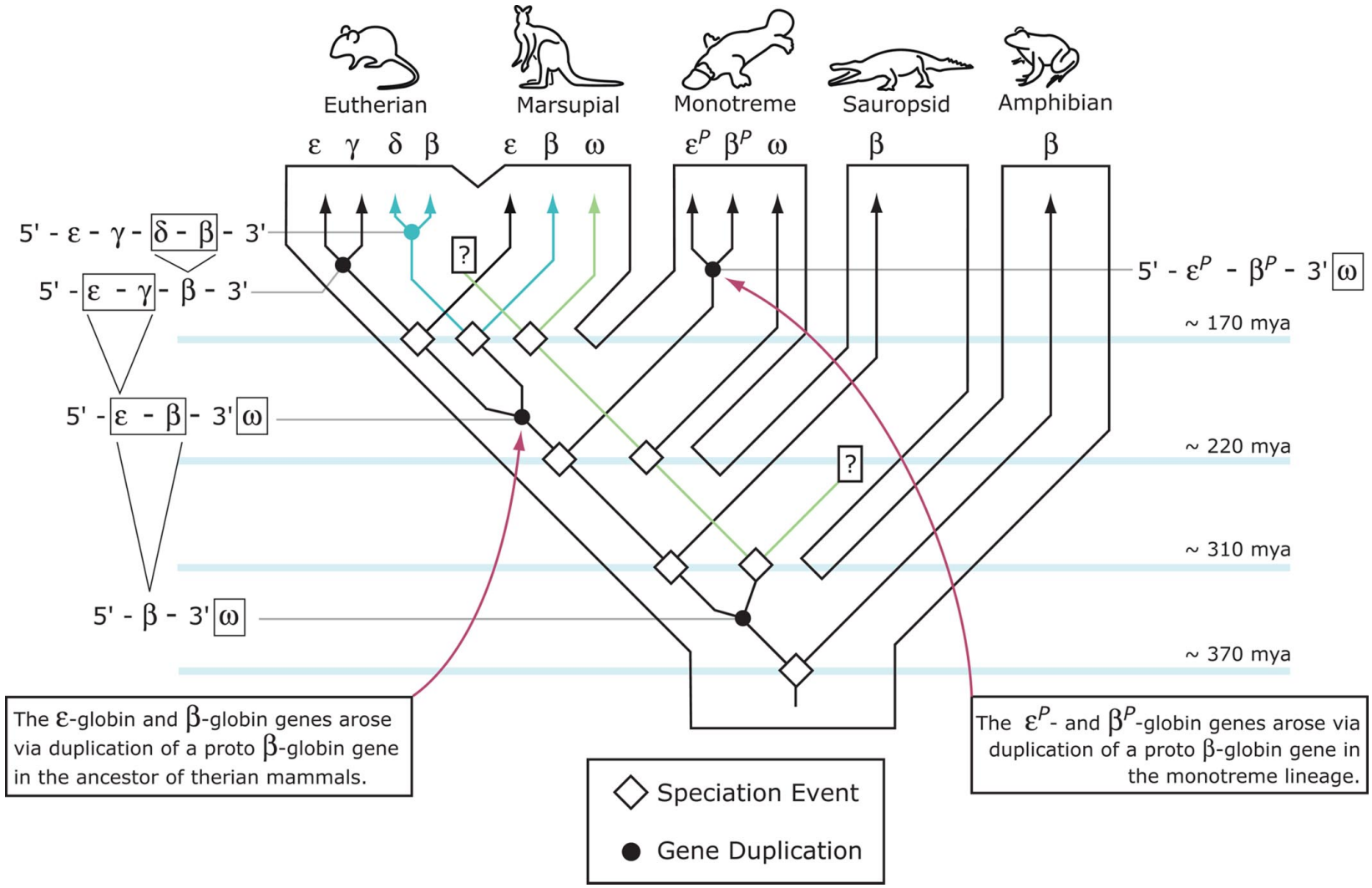
Sauropsida β-like globins

Monotreme and marsupial ω-globin

Amphibian β-like globins

Fish β-like globins

Opazo, Hoffmann and Storz "Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals" PNAS **105(5)** 2008



Opazo, Hoffmann and Storz "Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals" PNAS **105(5)** 2008

terminology: trees of gene families

- duplication – the creation of a new copy of a gene within the same genome.
- homologous – descended from a common ancestor.
- paralogous – homologous, but resulting from a gene duplication in the common ancestor.
- orthologous – homologous, and resulting from a speciation event at the common ancestor.

Joint estimation of gene duplication, loss, and species trees using PHYLDOG

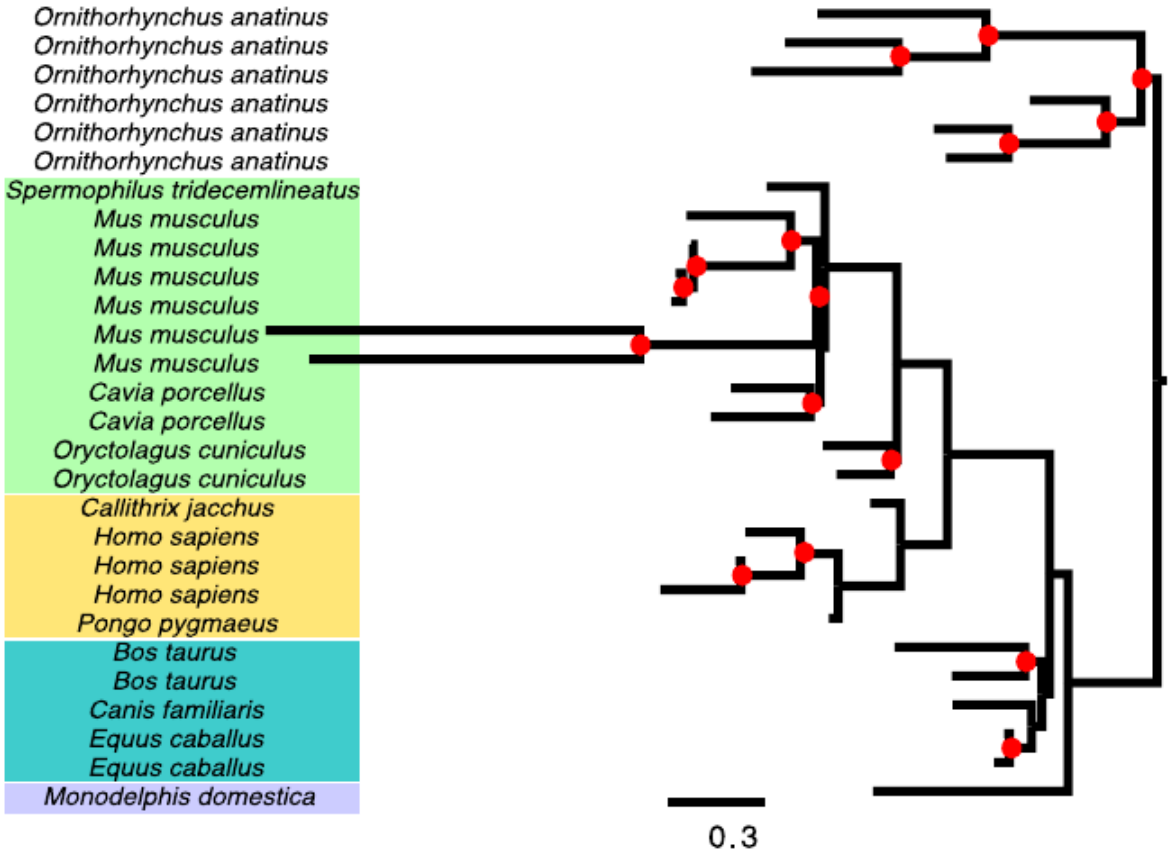


Figure 2A from Boussau et al. (2013)

Multiple contexts for tree estimation (again):

	The cause of splitting	**Important caveats**
"Gene tree" or "a coalescent"	DNA replication	recombination is usually ignored
Species tree Phylogeny	speciation	recombination, hybridization, lateral gene transfer, and deep coalescence cause conflict in the data we use to estimate phylogenies
Gene family tree	speciation or duplication	recombination (eg. domain swapping) is not tree-like

Joint estimation of gene duplication, loss, and coalescence with DLCoalRecon

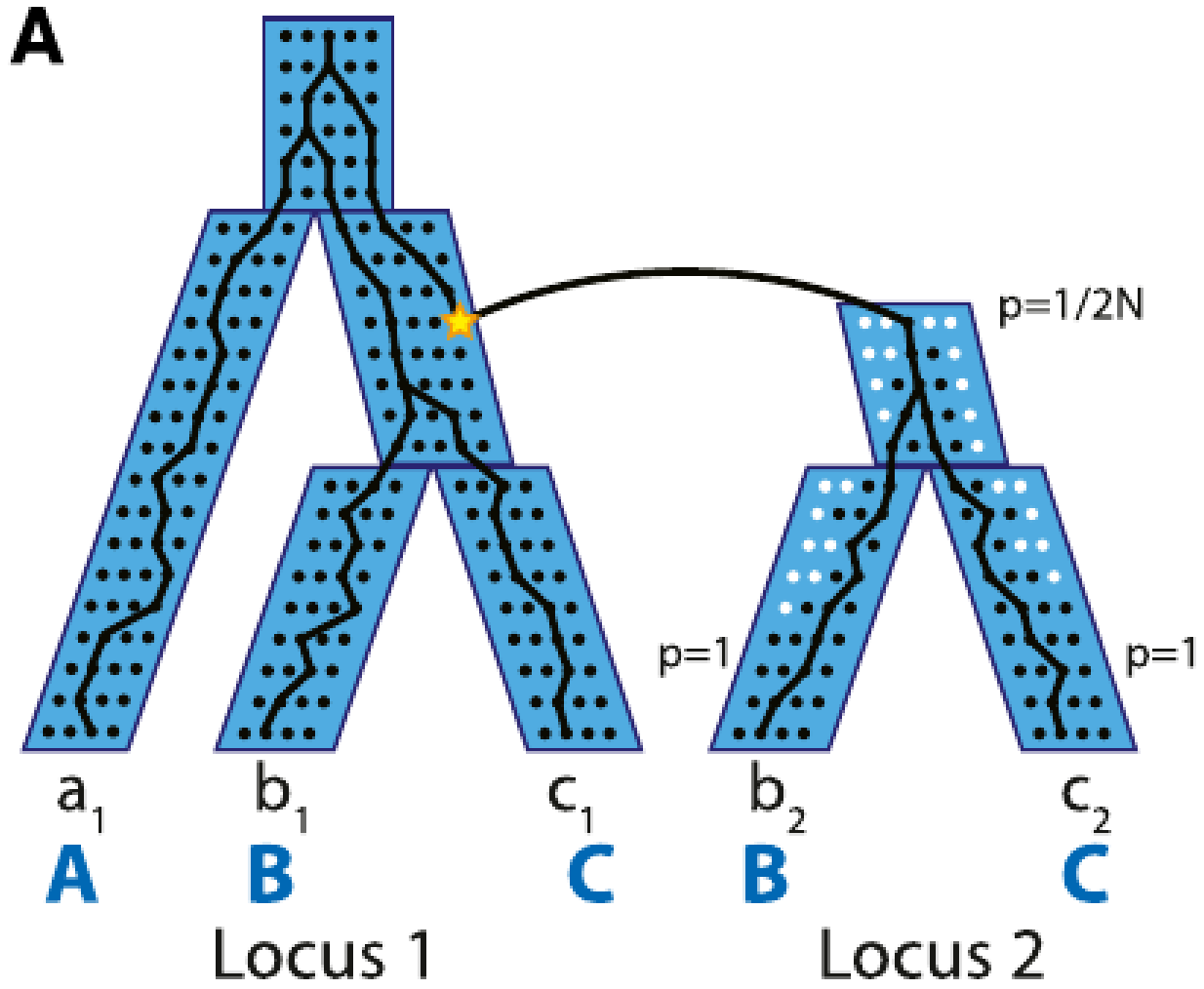


Figure 2A from Rasmussen and Kellis (2012)

Future: improved integration of DL models and coalescence

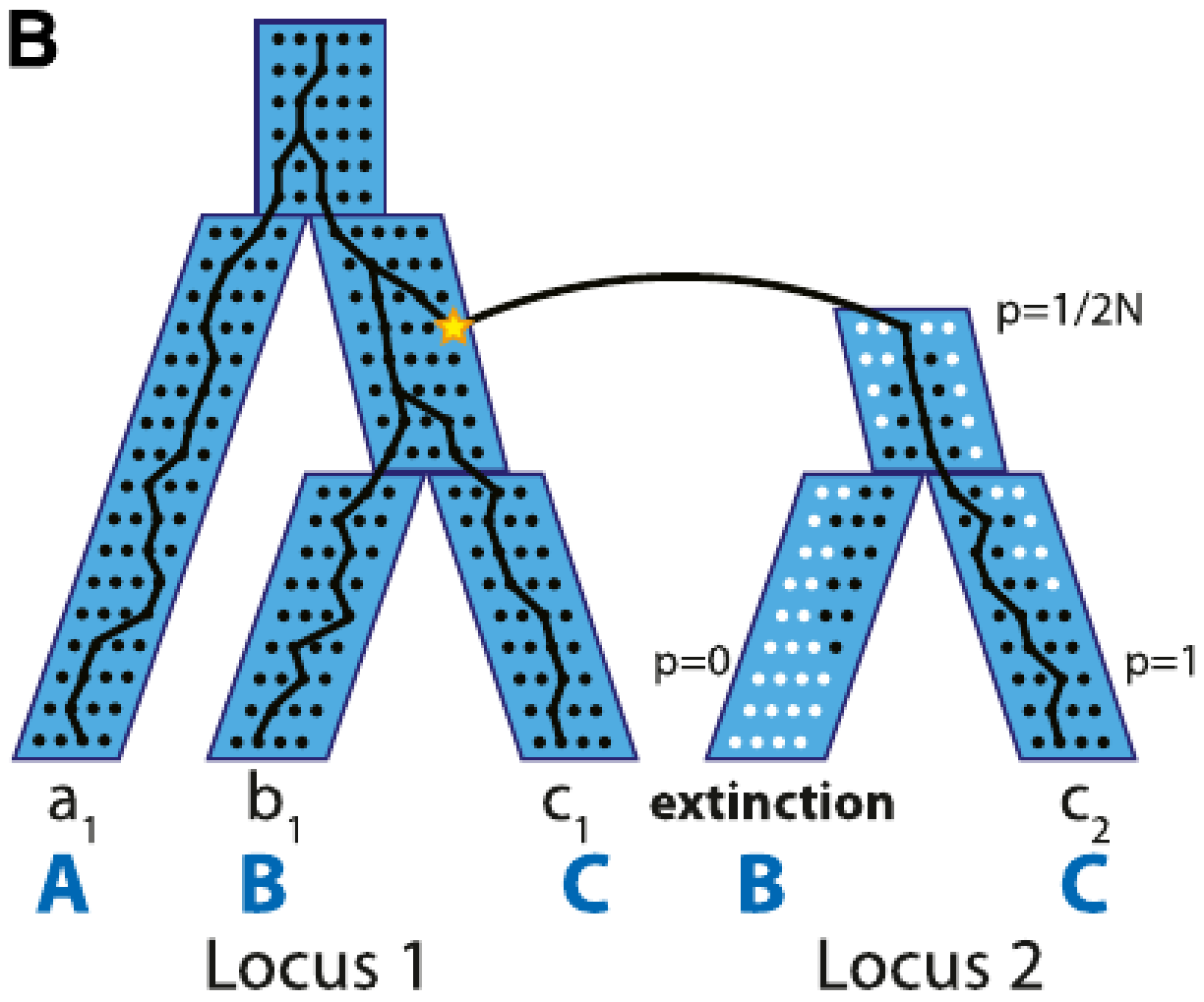


Figure 2B from Rasmussen and Kellis (2012)

Lateral Gene Transfer

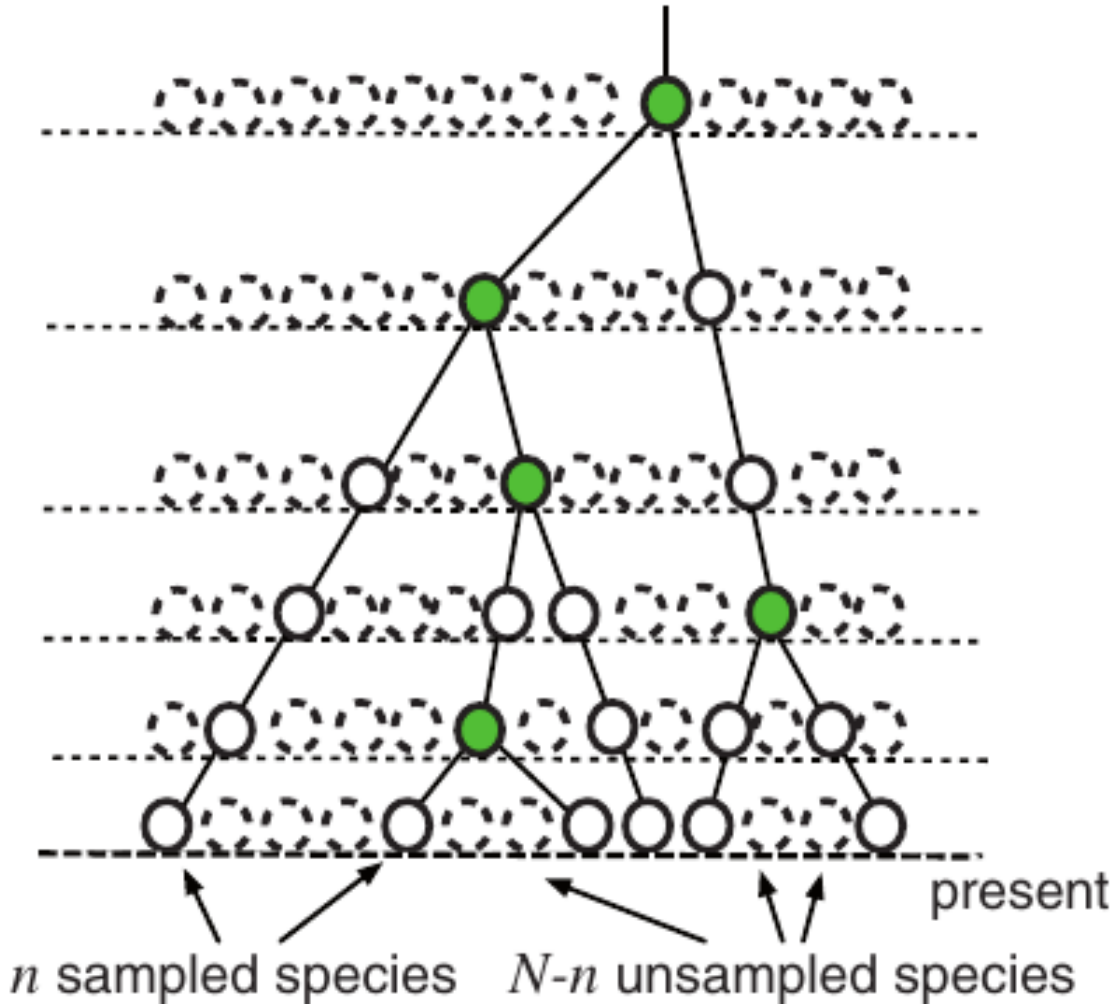


Figure 2c from Szöllősi et al. (2013)

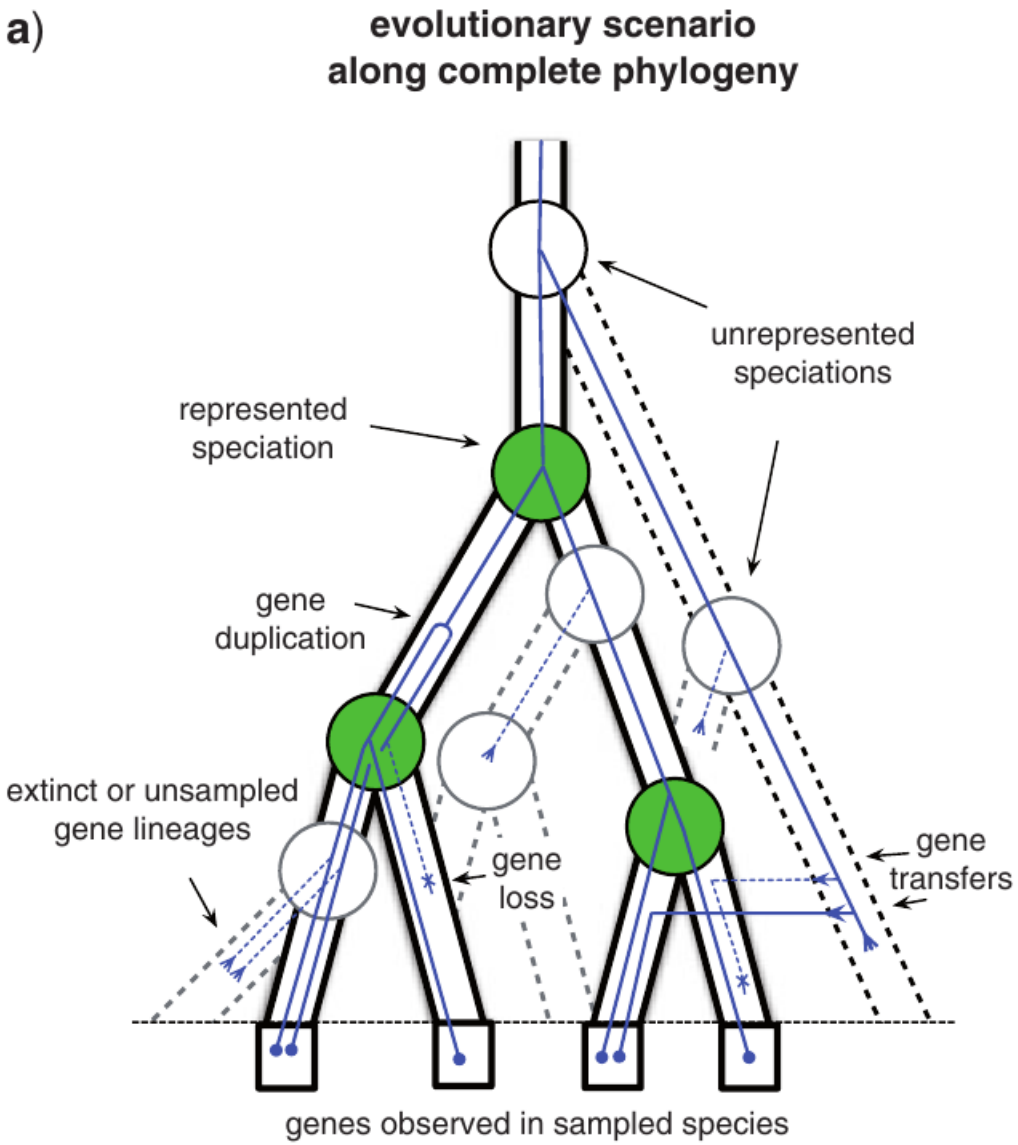


Figure 3 from Szöllősi et al. (2013)

They used 423 single-copy genes

in ≥ 34 of 36 cyanobacteria

They estimate:

2.56 losses/family

2.15 transfers/family

$\approx 28\%$ of transfers between

non-overlapping branches

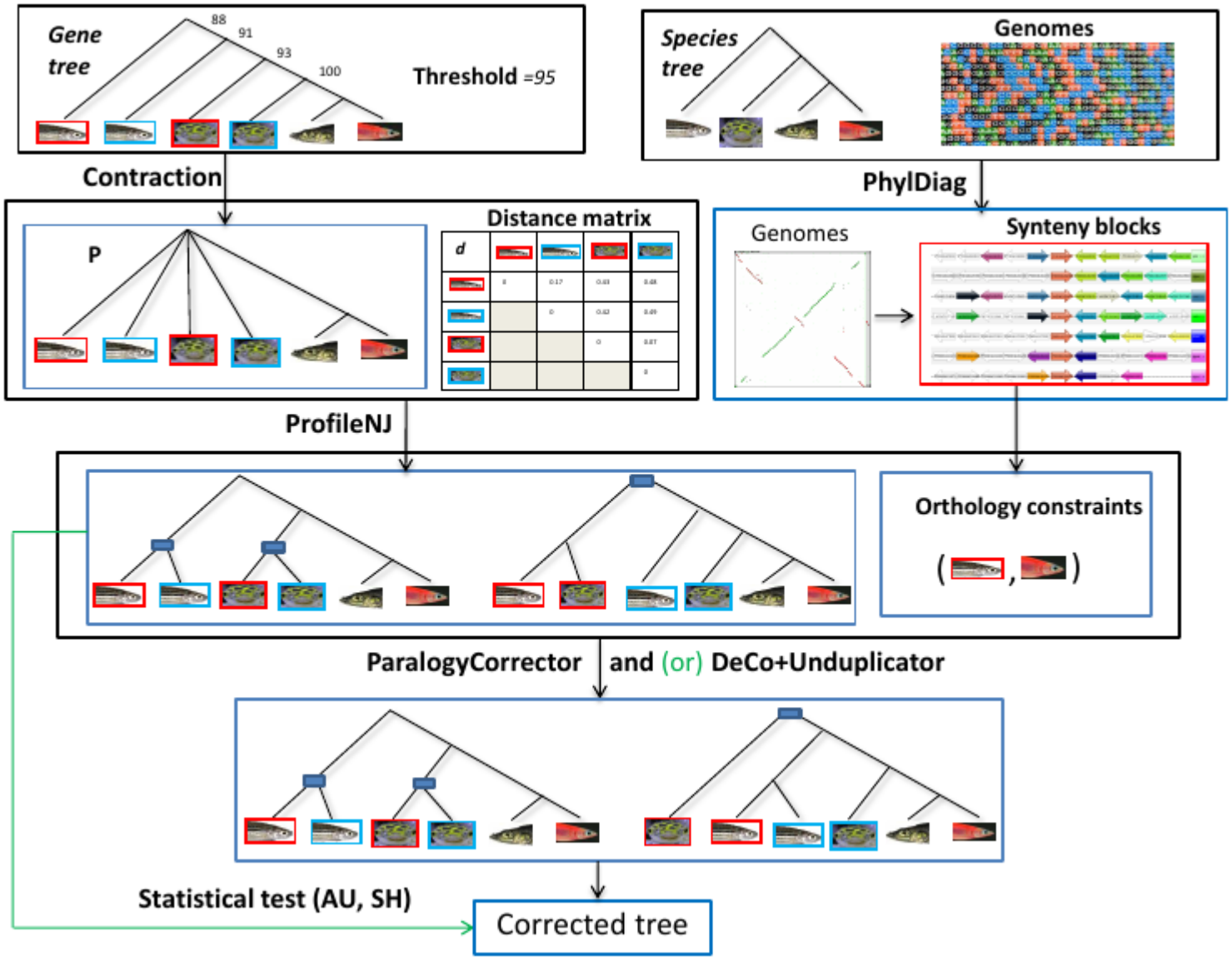


Figure 4 from Noutahi et al. (2016)

The main subject of this module: estimating a tree from sequence data

Tree construction:

- strictly algorithmic approaches - use a "recipe" to construct a tree
- optimality based approaches - choose a way to "score" a trees and then search for the tree that has the best score.

Expressing support for aspects of the tree:

- bootstrapping,
- testing competing trees against each other,
- posterior probabilities (in Bayesian approaches).

Optimality criteria

A rule for ranking trees (according to the data).
Each criterion produces a score.

Examples:

- Parsimony (Maximum Parsimony, MP)
- Maximum Likelihood (ML)
- Minimum Evolution (ME)
- Least Squares (LS)

Why doesn't simple clustering work?

Step 1: use sequences to estimate pairwise distances between taxa.

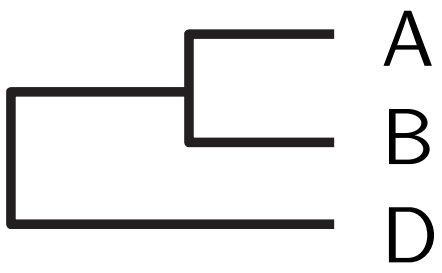
	A	B	C	D
A	-	0.2	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				-

Why doesn't simple clustering work?

	A	B	C	D
A	-	**0.2**	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				-

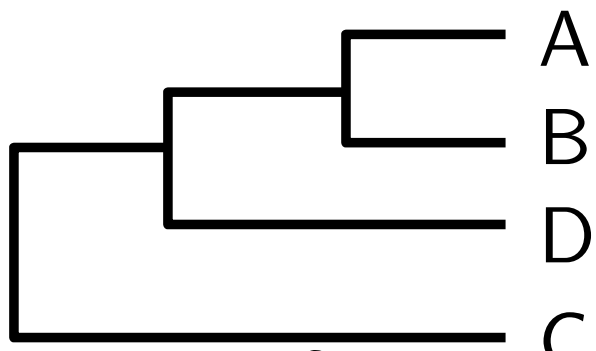
Why doesn't simple clustering work?

	A	B	C	D
A	-	0.2	0.5	**0.4**
B		-	0.46	**0.4**
C			-	0.7
D				-



Why doesn't simple clustering work?

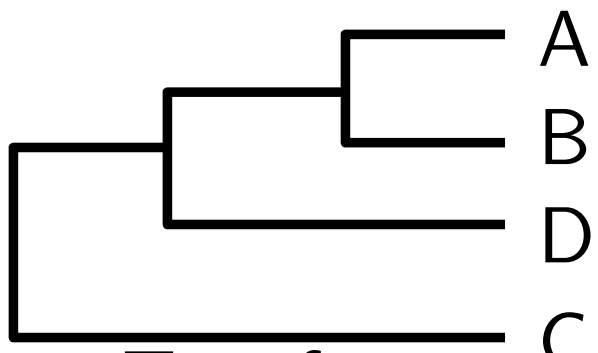
	A	B	C	D
A	-	0.2	0.5	**0.4**
B		-	0.46	**0.4**
C		-	-	**0.7**
D				0



Tree from clustering

Why doesn't simple clustering work?

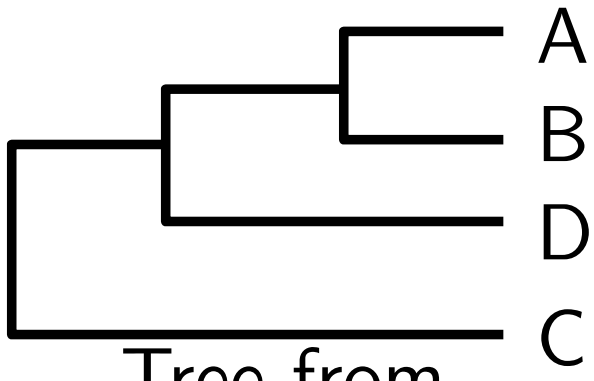
	A	B	C	D
A	0	0.2	**0.5**	0.4
B	0.2	0.2	**0.46**	0.4
C	**0.5**	**0.46**	0	**0.7**
D	0.4	0.4	**0.7**	0



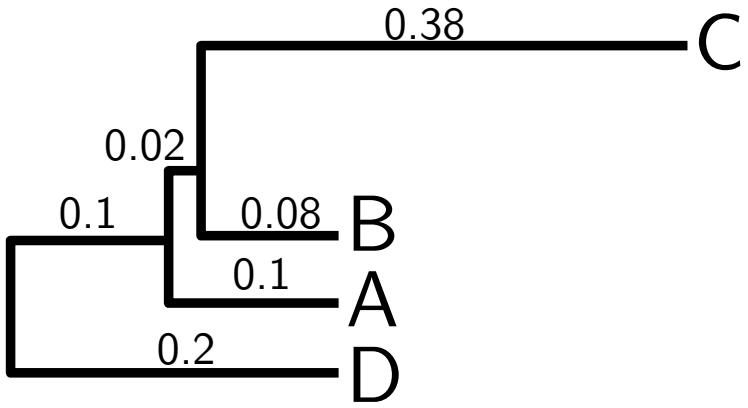
Tree from clustering

Why doesn't simple clustering work?

	A	B	C	D
A	0	0.2	0.5	0.4
B	0.2	0.	0.46	0.4
C	0.5	0.46	0	0.7
D	0.4	0.4	0.7	0



Tree from clustering



Tree with perfect fit

Why aren't the easy, obvious methods for generating trees good enough?

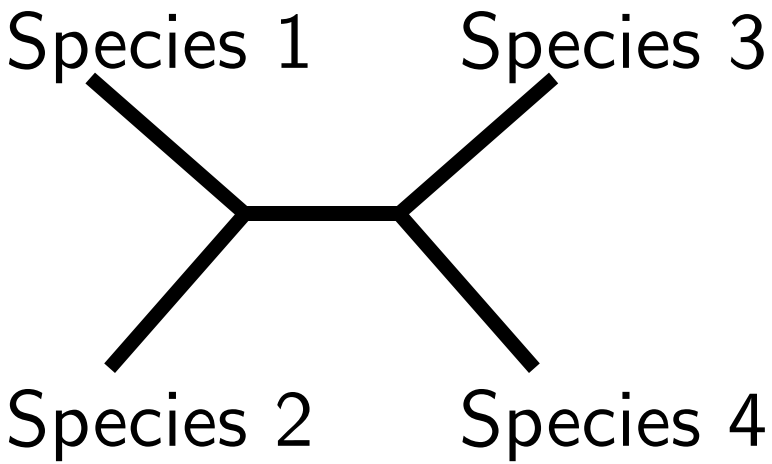
1. Simple clustering methods are sensitive to differences in the rate of sequence evolution (and this rate can be quite variable).

2. The "multiple hits" problem. When some sites in your data matrix are affected by more than 1 mutation, then the phylogenetic signal can be obscured. More on this later...

	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	.	.	.

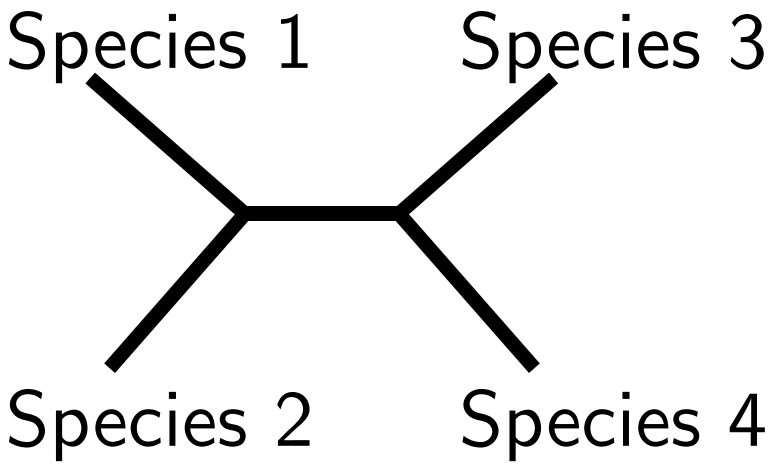
	1	2	3	4	5	6	7	8	9	.	.	.
Species 1	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 2	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 3	C	G	G	T	C	**C**	G	G	T	.	.	.
Species 4	C	G	G	C	C	**T**	G	G	T	.	.	.

One of the 3 possible trees:

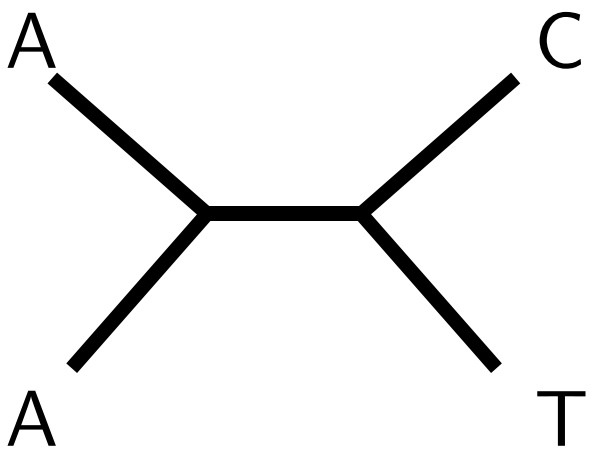


	1	2	3	4	5	6	7	8	9	.	.	.
Species 1	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 2	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 3	C	G	G	T	C	**C**	G	G	T	.	.	.
Species 4	C	G	G	C	C	**T**	G	G	T	.	.	.

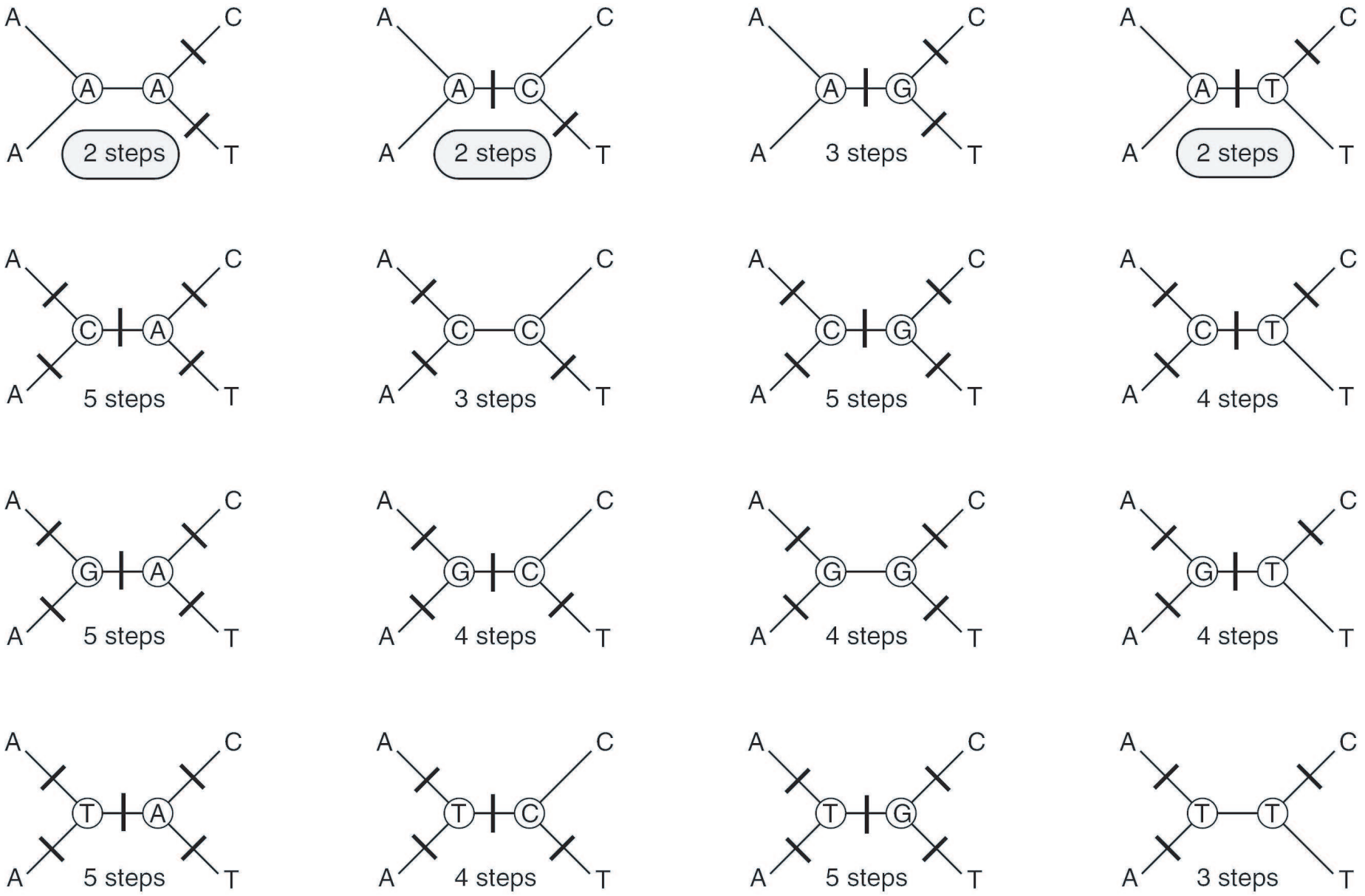
One of the 3 possible trees:



Same tree with states at character 6 instead of species names



Unordered Parsimony



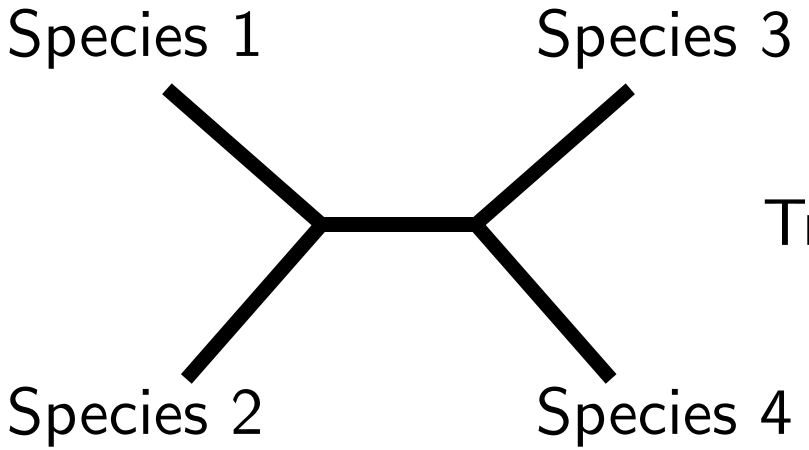
Things to note about the last slide

- 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 independent

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

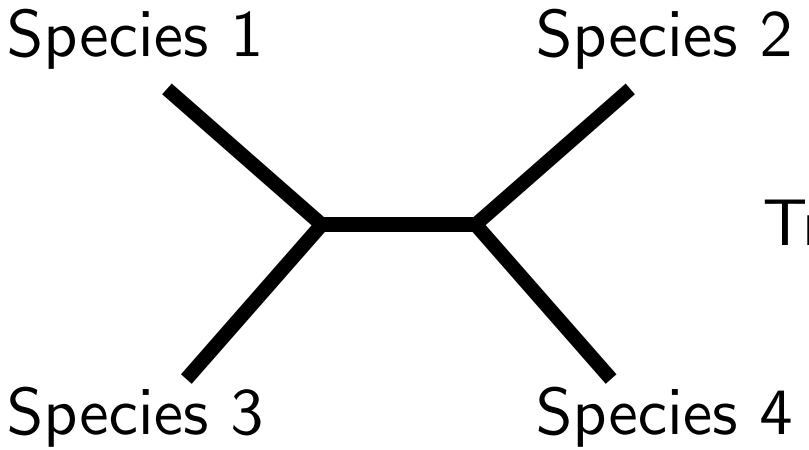


Tree 1 has a score of **4**

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

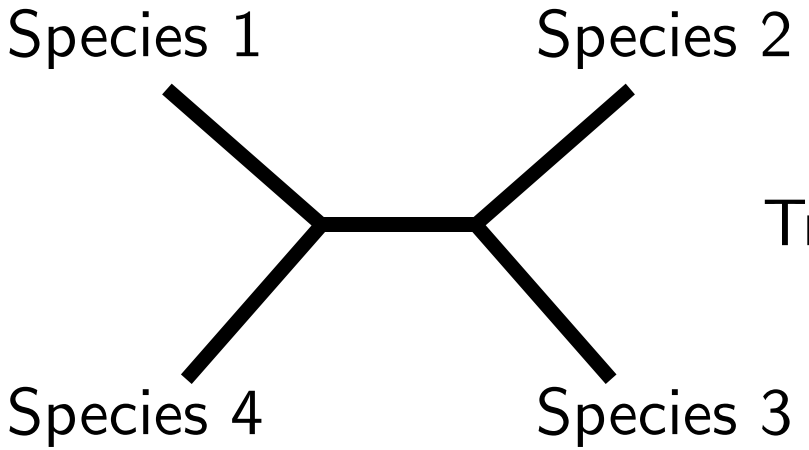


Tree 2 has a score of **5**

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.

Parsimony terms

- *homoplasy* multiple acquisitions of the same character state
  - parallelism, reversal, convergence
  - recognized by a tree requiring more than the minimum number of steps
  - minimum number of steps is the number of observed states minus 1

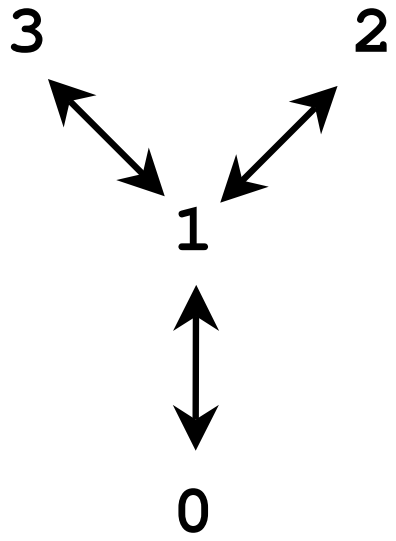
The parsimony criterion is equivalent to minimizing homoplasy.

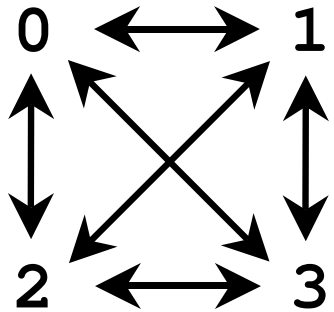
Homoplasy is one form of the multiple hits problem. In pop-gen terms, it is a violation of the infinite-alleles model.

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

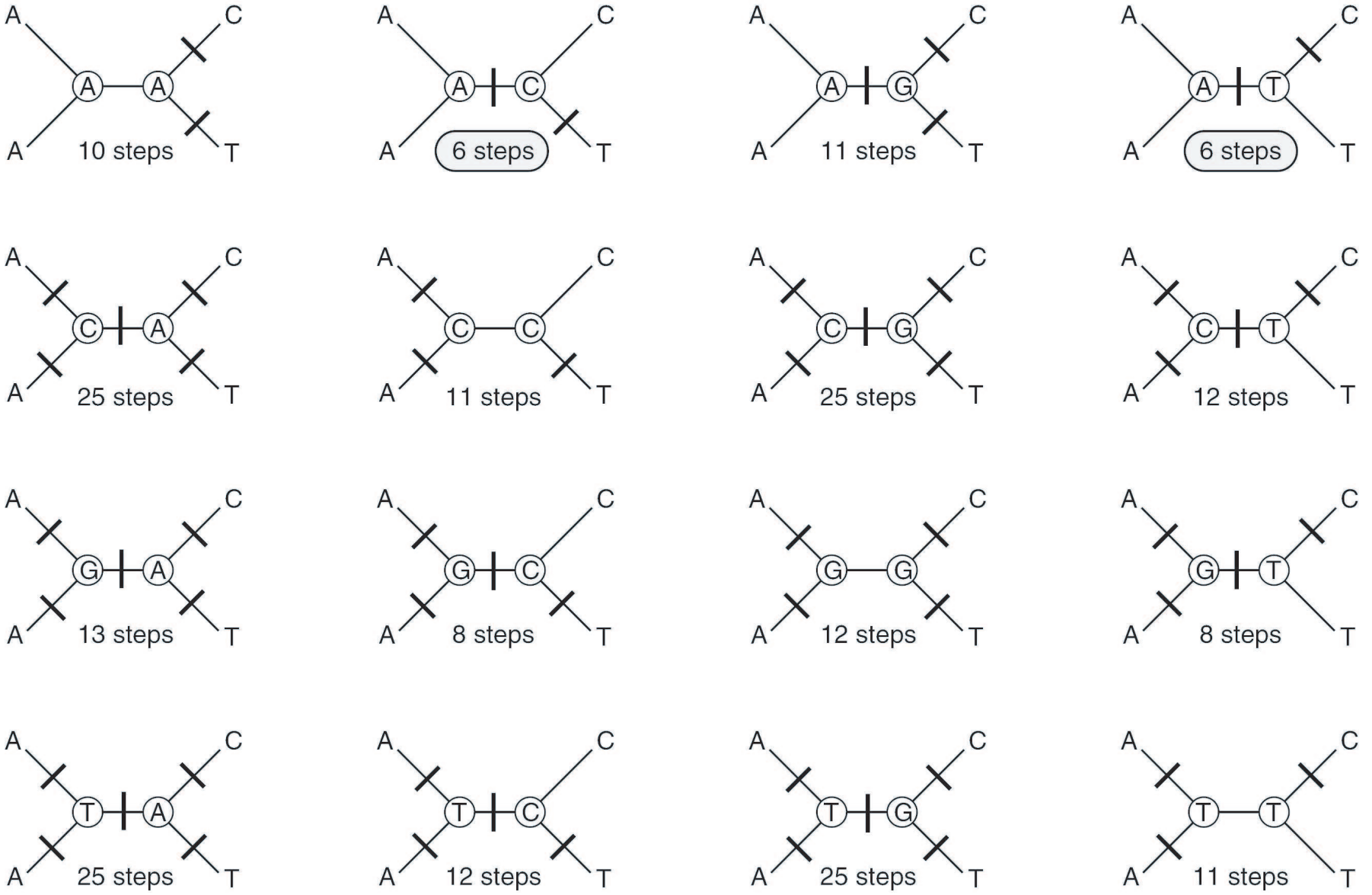
		To A	To C	To G	To T
From	A	0	1	1	1
	C	1	0	1	1
	G	1	1	0	1
	T	1	1	1	0

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

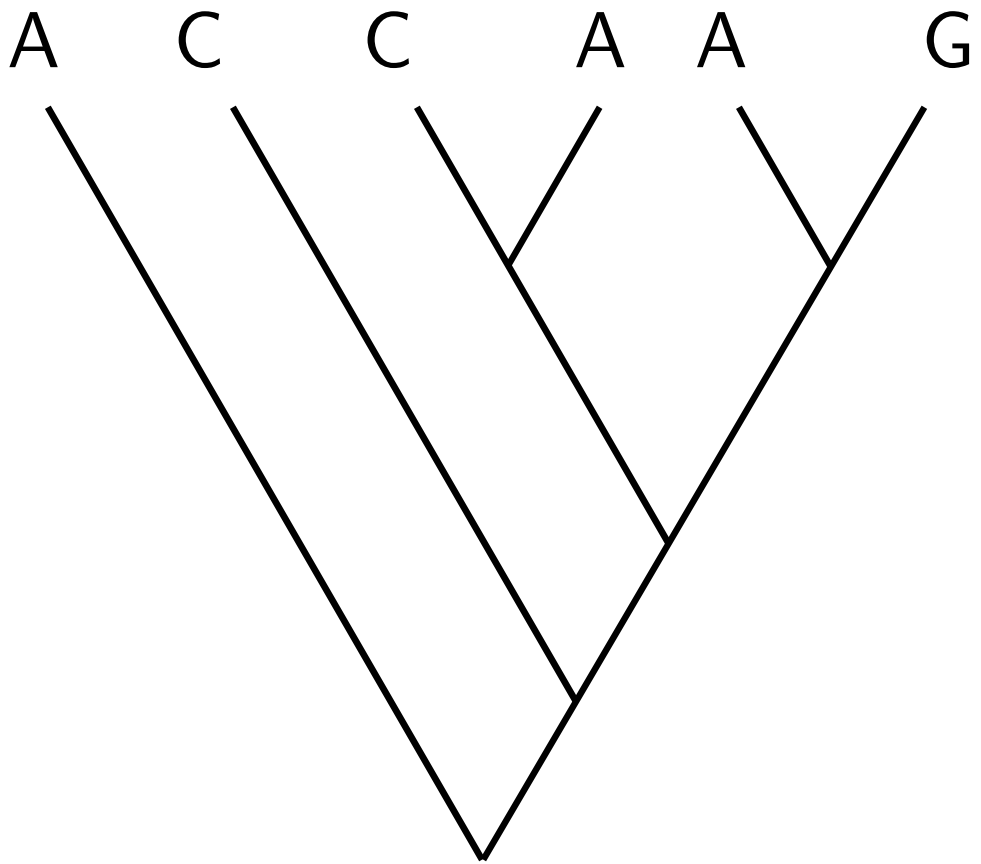
- Parsimony scores from different stepmatrices cannot be meaningfully compared (31 under Fitch is not "better" than 45 under a transversion:transition stepmatrix)

- Parsimony cannot be used to infer the stepmatrix weights

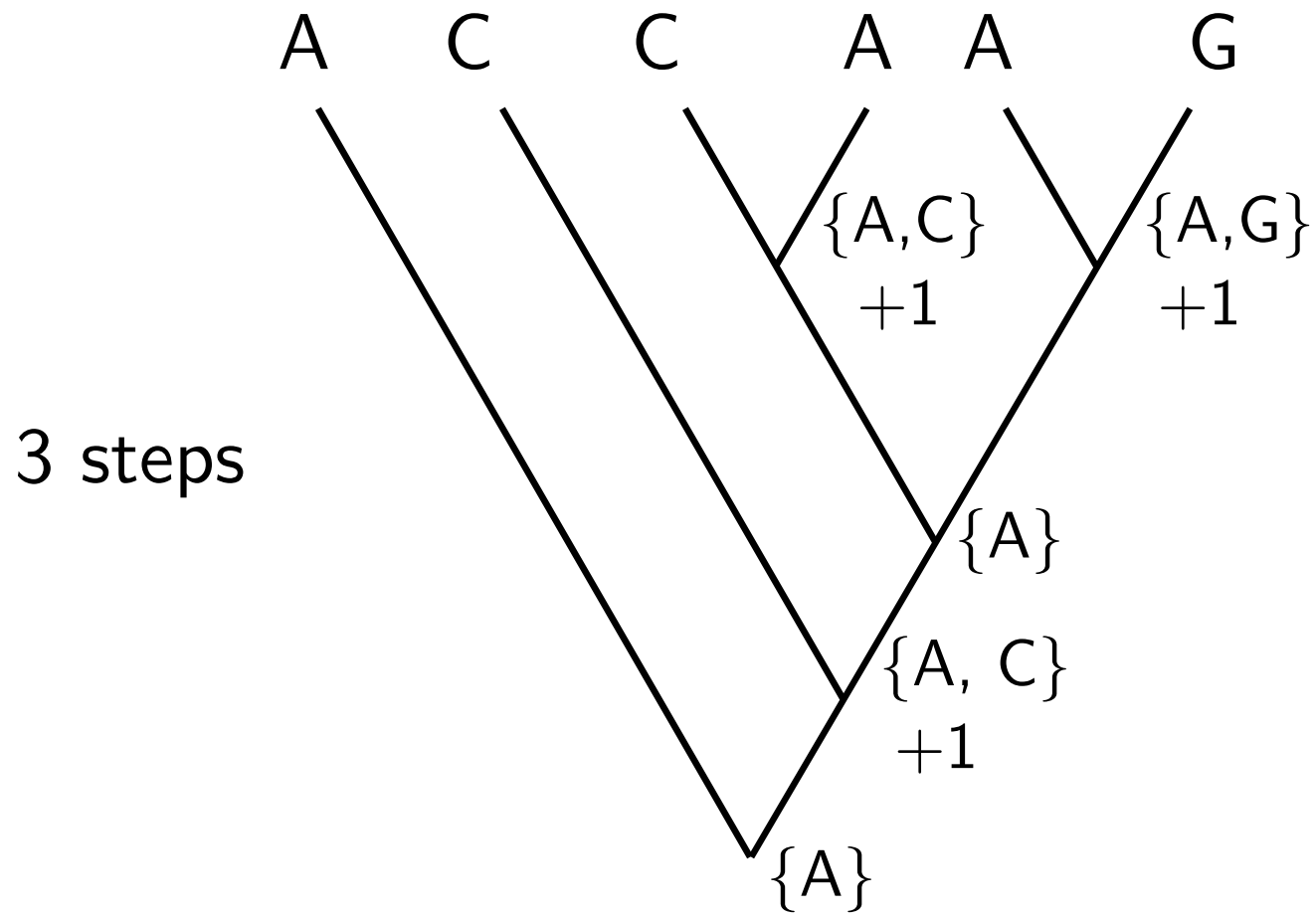
Other Parsimony variants

- *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 Goloboff (1993)

Scoring trees under parsimony is fast



Scoring trees under parsimony is fast – Fitch algorithm



Scoring trees under parsimony is fast

The "down-pass state sets" calculated in the Fitch algorithm can be stored at an internal node.

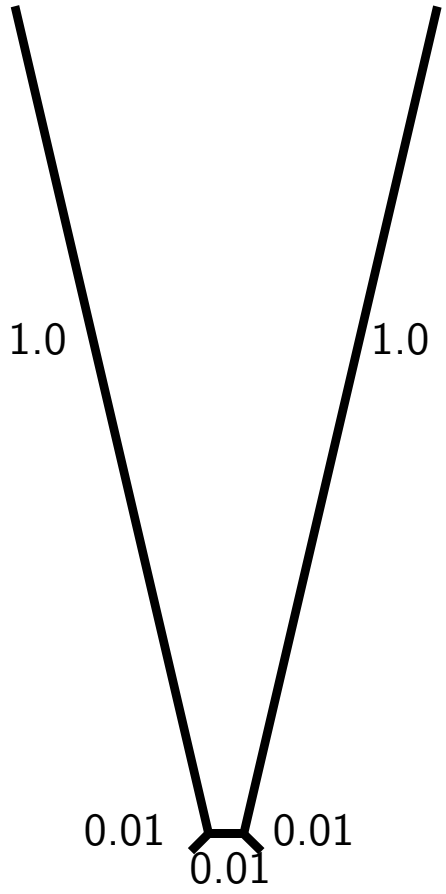
This lets you treat those internal nodes as pseudo-tips:

- avoid rescoring the entire tree if you make a small change, and
- break up the tree into smaller subtrees (Goloboff's sectorial searching).

Qualitative description of parsimony

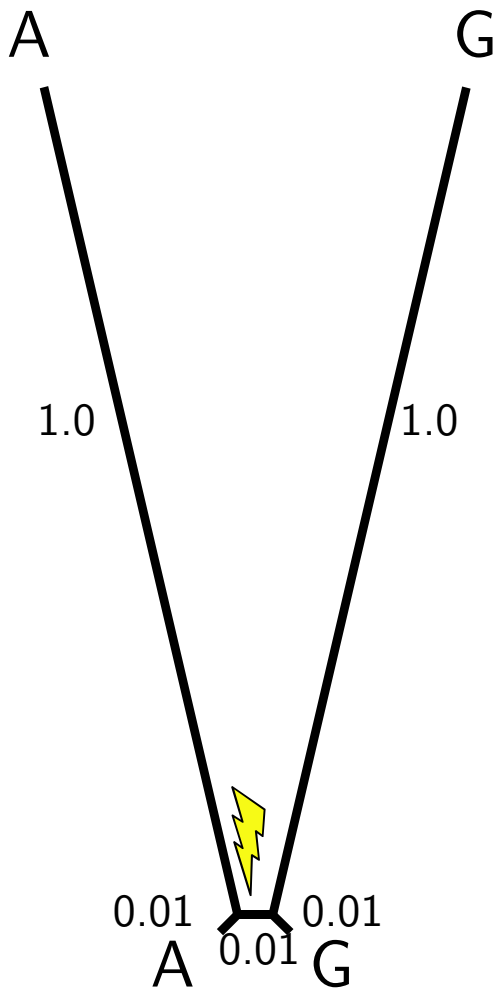
- Enables estimation of ancestral sequences.
- Even though parsimony always seeks to minimizes the number of changes, it can perform well even when changes are not rare.
- Does not "prefer" to put changes on one branch over another
- Hard to characterize statistically
  - the set of conditions in which 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.

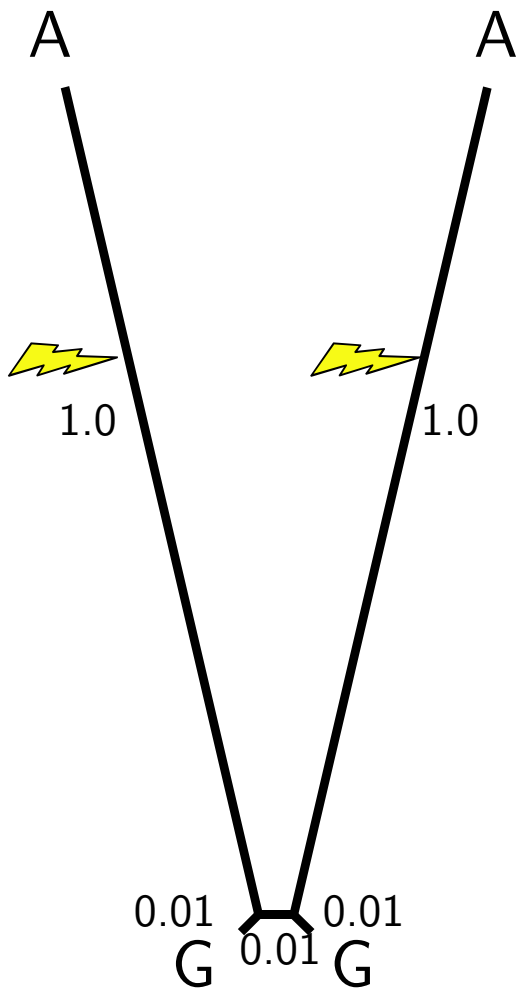
Long branch attraction



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



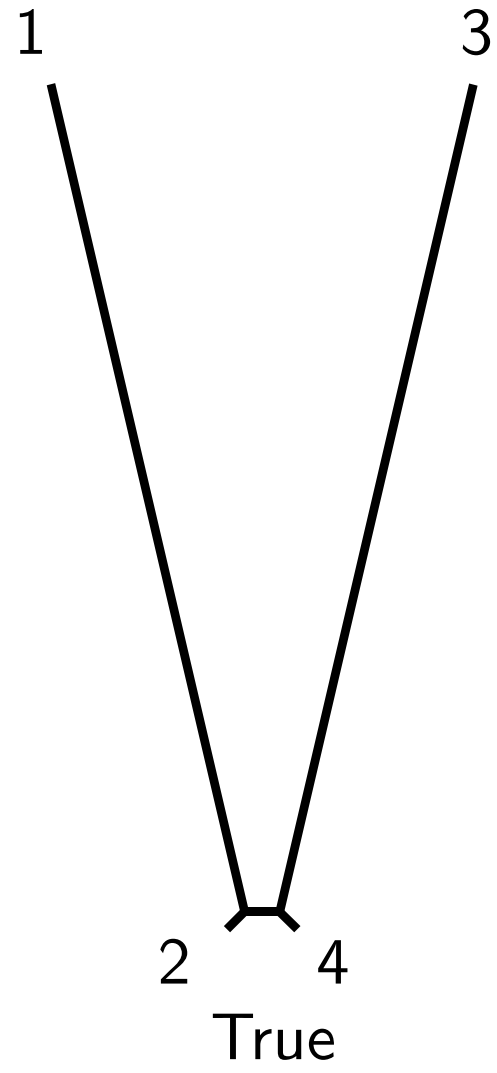
Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

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

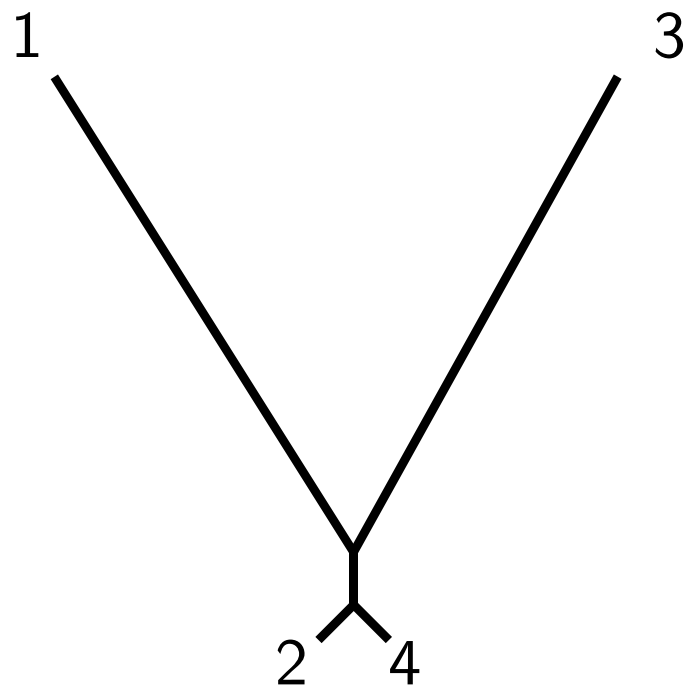
The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).

Long branch attraction

Parsimony is almost guaranteed to get this tree wrong.



True



Inferred

Inconsistency

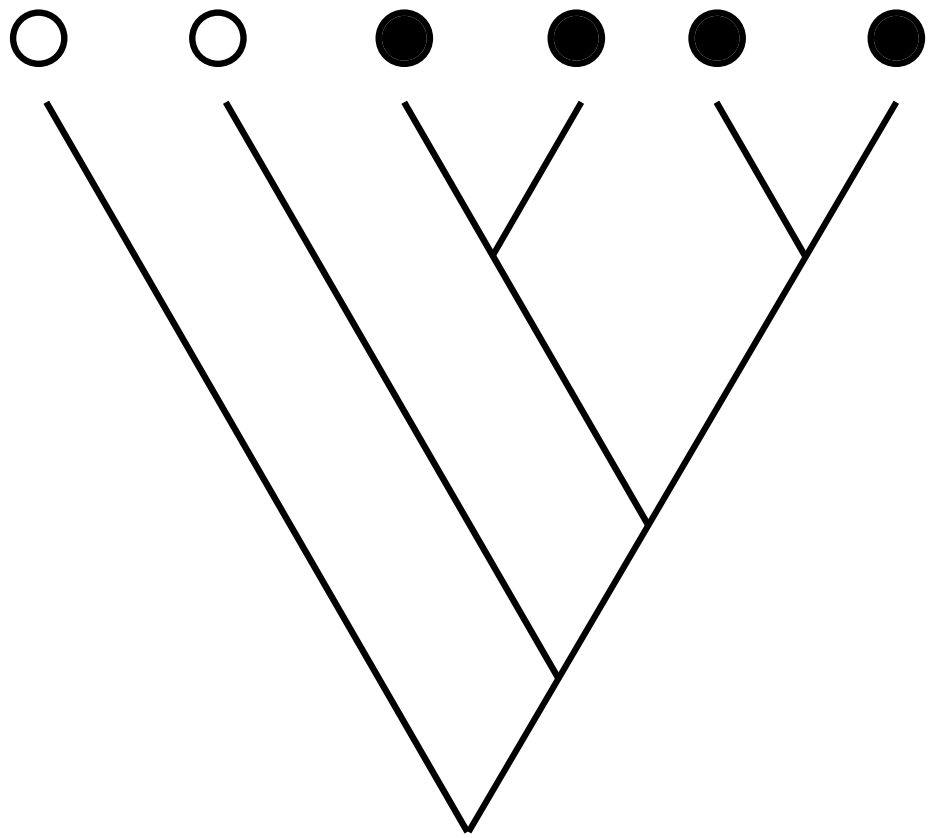
- Statistical Consistency (roughly speaking) is converging to the true answer as the amount of data goes to ∞.

- 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.

Parsimony terms

- *synapomorphy* – a shared derived (newly acquired) character state. Evidence of monophletic groups.



Parsimony terms

- *parsimony informative* – a character with parsimony score variation across trees
  - *min* score $\neq$ *max* score
  - must be variable.
  - must have more than one *shared* state

Consistency Index (CI)

- minimum number of changes divided by the number required on the tree.
- CI=1 if there is no homoplasy
- negatively correlated with the number of species sampled

Retention Index (RI)

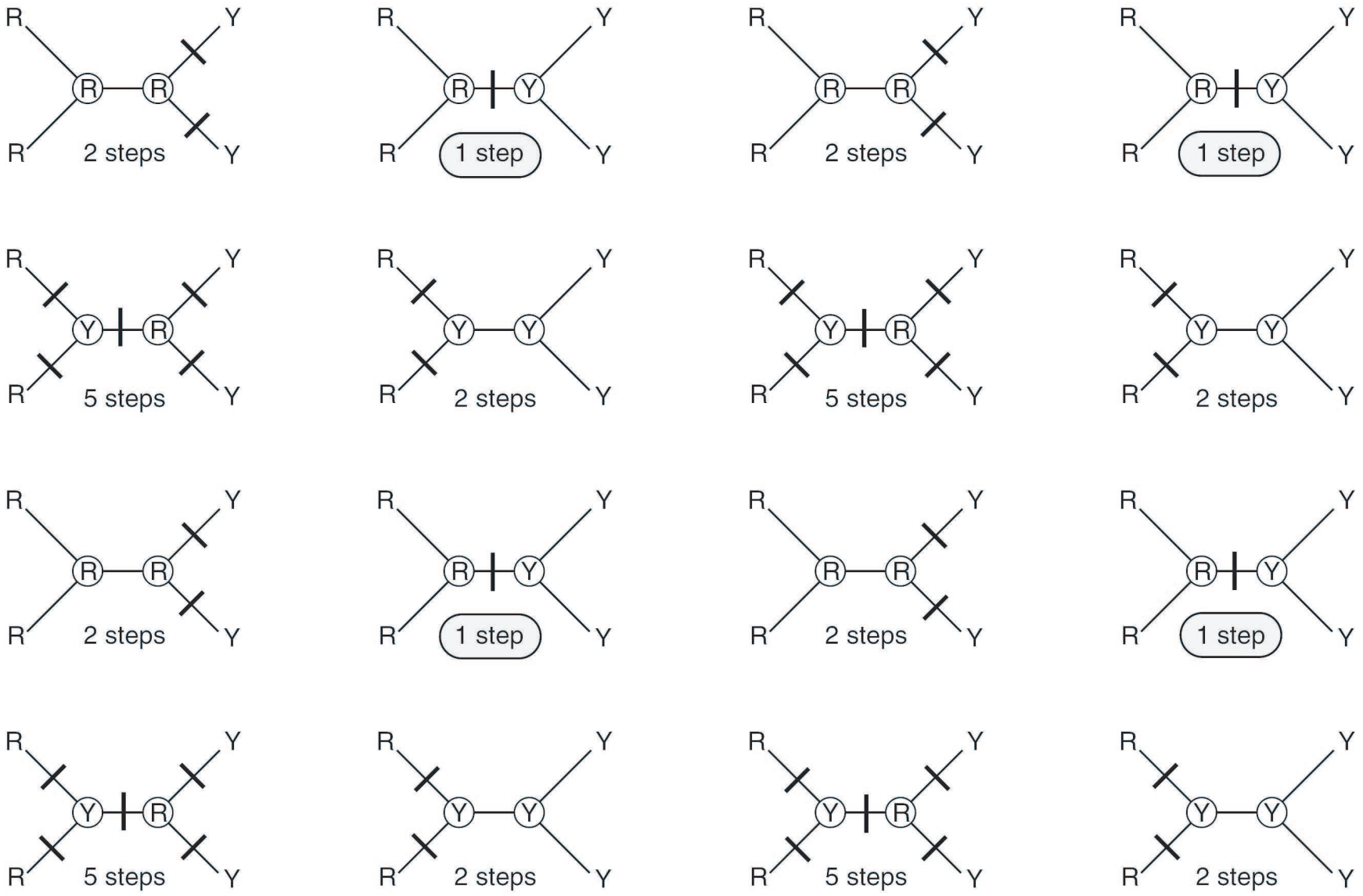
$$\mathrm{RI} = \frac{\mathrm{MaxSteps} - \mathrm{ObsSteps}}{\mathrm{MaxSteps} - \mathrm{MinSteps}}$$

- defined to be 0 for parsimony uninformative characters
- RI=1 if the character fits perfectly
- RI=0 if the tree fits the character as poorly as possible

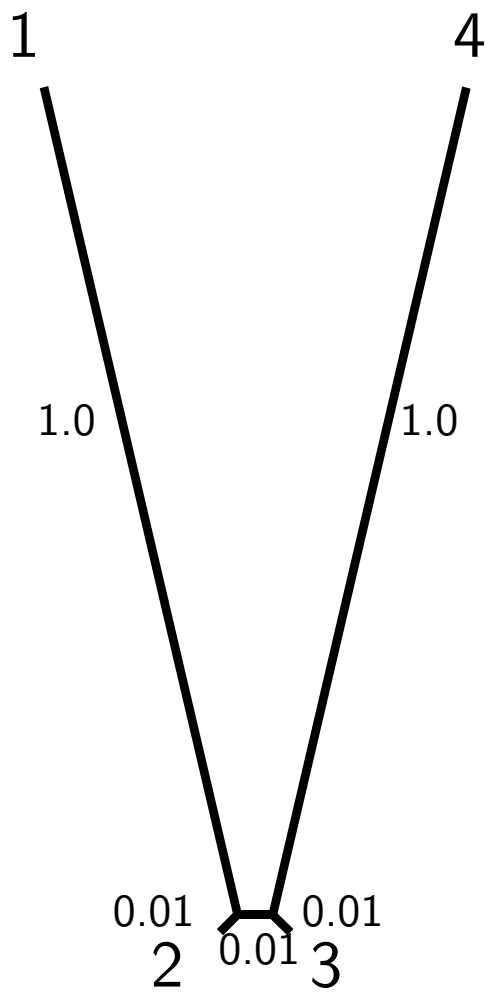
Transversion parsimony

- Transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) occur more frequently than transversions (purine $\leftrightarrow$ pyrimidine)
- So, *homoplasy* involving transitions is much more common than transversions (*e.g.* $A \rightarrow G \rightarrow A$)
- Transversion parsimony (also called RY-coding) ignores all transitions

Transversion parsimony



Long branch attraction tree again



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).

If the data is generated such that:

$$\Pr\begin{pmatrix} A \\ A \\ G \\ G \end{pmatrix} \approx 0.0003 \quad \text{and} \quad \Pr\begin{pmatrix} A \\ G \\ G \\ A \end{pmatrix} \approx 0.008$$

then how can we hope to infer the tree ((1,2),3,4) ?

Note: ((1,2),3,4) is referred to as Newick or New Hampshire notation for the tree.

You can read it by following the rules:

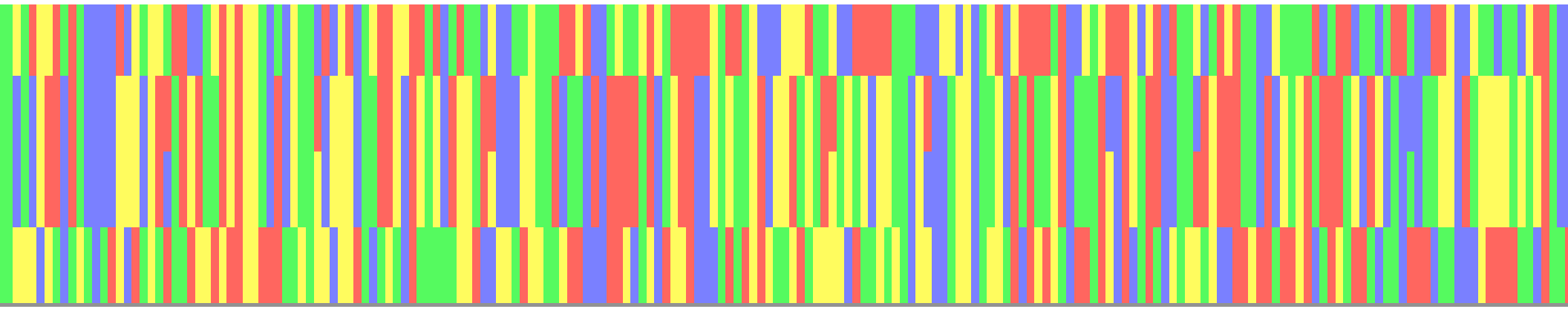
- start at a node,
- if the next symbol is '(' then add a child to the current node and move to this child,
- if the next symbol is a label, then label the node that you are at,
- if the next symbol is a comma, then move back to the current node's parent and add another child,
- if the next symbol is a ')', then move back to the current node's parent.

If the data is generated such that:

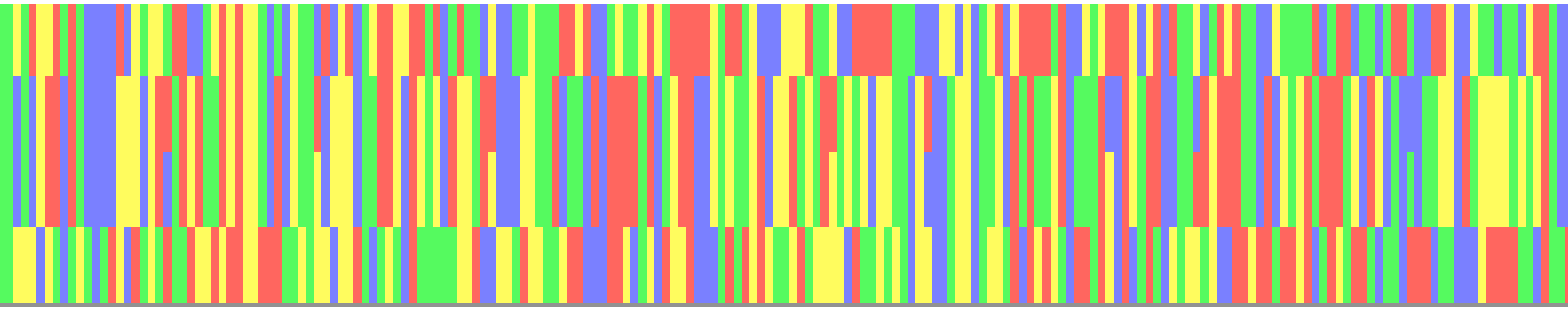
$$\Pr\begin{pmatrix} A \\ A \\ G \\ G \end{pmatrix} \approx 0.0003 \quad \text{and} \quad \Pr\begin{pmatrix} A \\ G \\ G \\ A \end{pmatrix} \approx 0.008$$

then how can we hope to infer the tree ((1,2),3,4) ?

Looking at the data in “bird’s eye” view (using Mesquite):



Looking at the data in "bird's eye" view (using Mesquite):



We see that sequences 1 and 4 are clearly very different.

Perhaps we can estimate the tree if we use the branch length information from the sequences...

Distance-based approaches to inferring trees

- Convert the raw data (sequences) to a pairwise distances
- Try to find a tree that explains these distances.
- *Not* simply clustering the most similar sequences.

	1	2	3	4	5	6	7	8	9	10
Species 1	C	G	A	C	C	A	G	G	T	A
Species 2	C	G	A	C	C	A	G	G	T	A
Species 3	C	G	G	T	C	C	G	G	T	A
Species 4	C	G	G	C	C	A	T	G	T	A

Can be converted to a distance matrix:

	Species 1	Species 2	Species 3	Species 4
Species 1	0	0	0.3	0.2
Species 2	0	0	0.3	0.2
Species 3	0.3	0.3	0	0.3
Species 4	0.2	0.2	0.3	0

Note that the distance matrix is symmetric.

	Species 1	Species 2	Species 3	Species 4
Species 1	0	0	0.3	0.2
Species 2	0	0	0.3	0.2
Species 3	0.3	0.3	0	0.3
Species 4	0.2	0.2	0.3	0

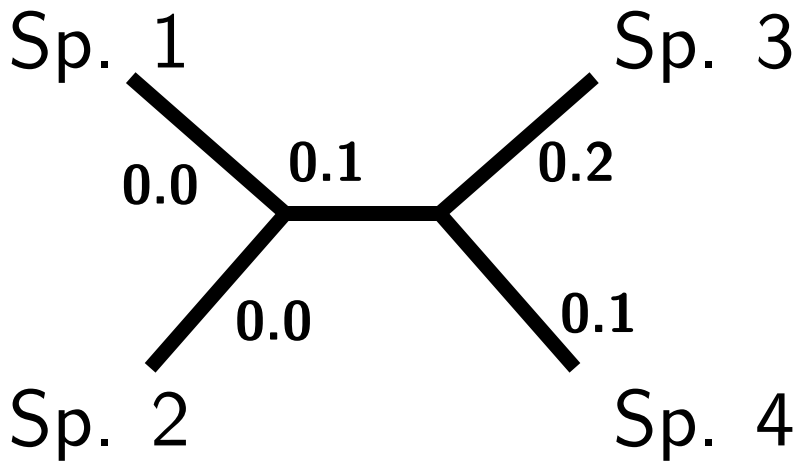
. . . so we can just use the lower triangle.

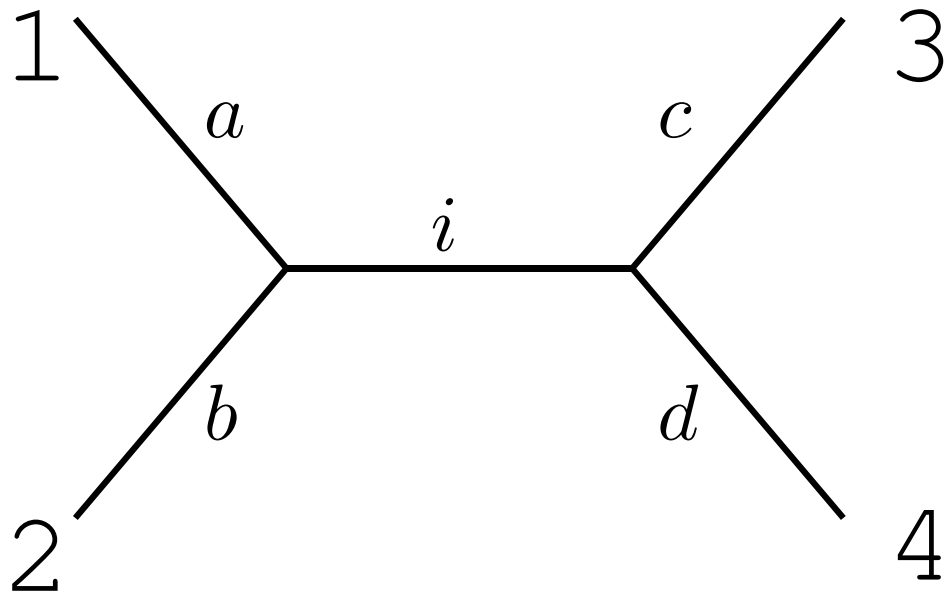
	Species 1	Species 2	Species 3
Species 2	0		
Species 3	0.3	0.3	
Species 4	0.2	0.2	0.3

Can we find a tree that would predict these observed character divergences?

	Species 1	Species 2	Species 3
Species 2	0		
Species 3	0.3	0.3	
Species 4	0.2	0.2	0.3

Can we find a tree that would predict these observed character divergences?





parameters

$$p_{12} = a + b$$
$$p_{13} = a + i + c$$
$$p_{14} = a + i + d$$
$$p_{23} = b + i + c$$
$$p_{23} = b + i + d$$
$$p_{34} = c + d$$

data

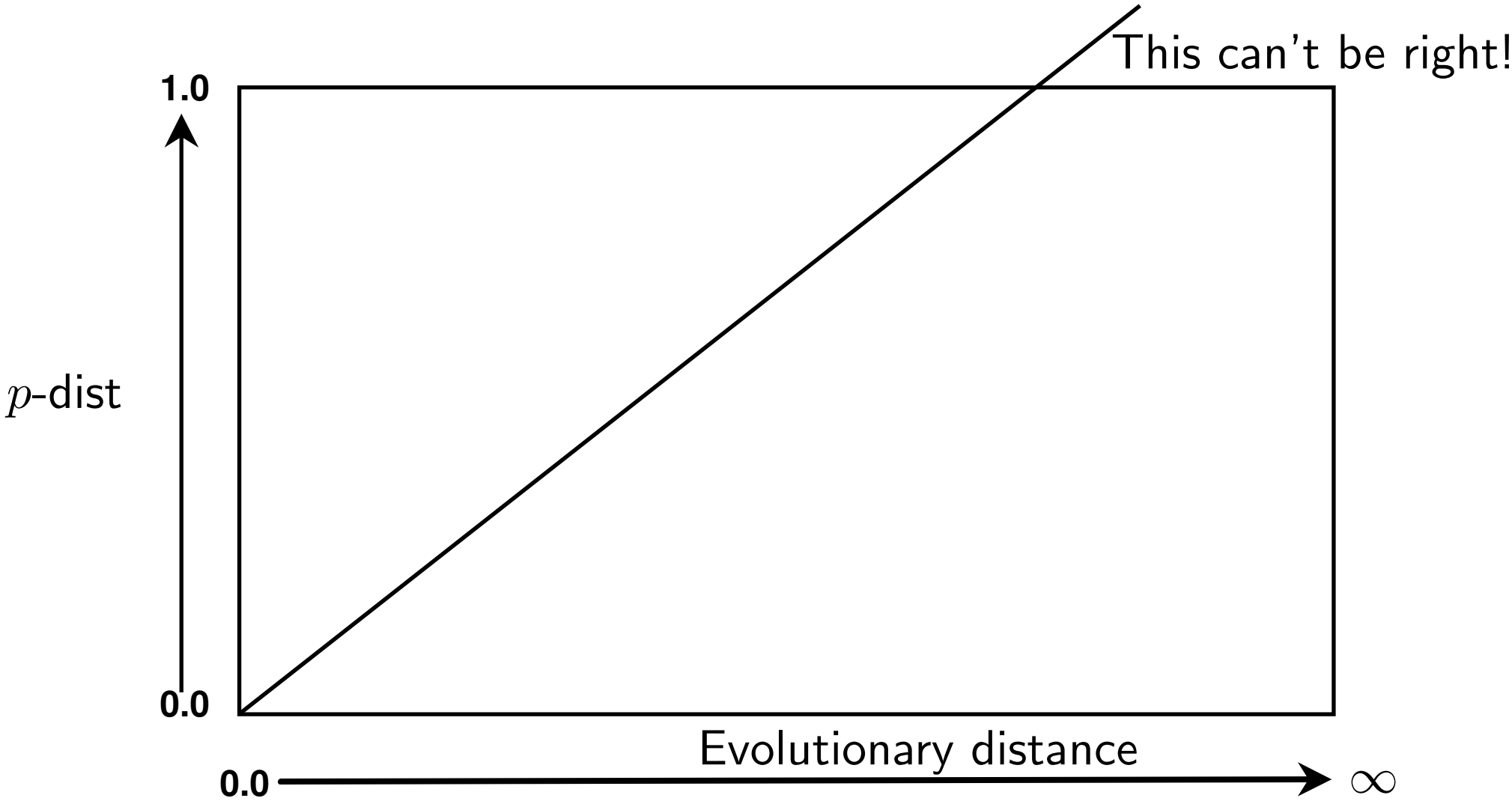
	1	2	3
2	d_{12}		
3	d_{13}	d_{23}	
4	d_{14}	d_{24}	d_{34}

If our pairwise distance measurements were error-free estimates of the *evolutionary distance* between the sequences, then we could always infer the tree from the distances.

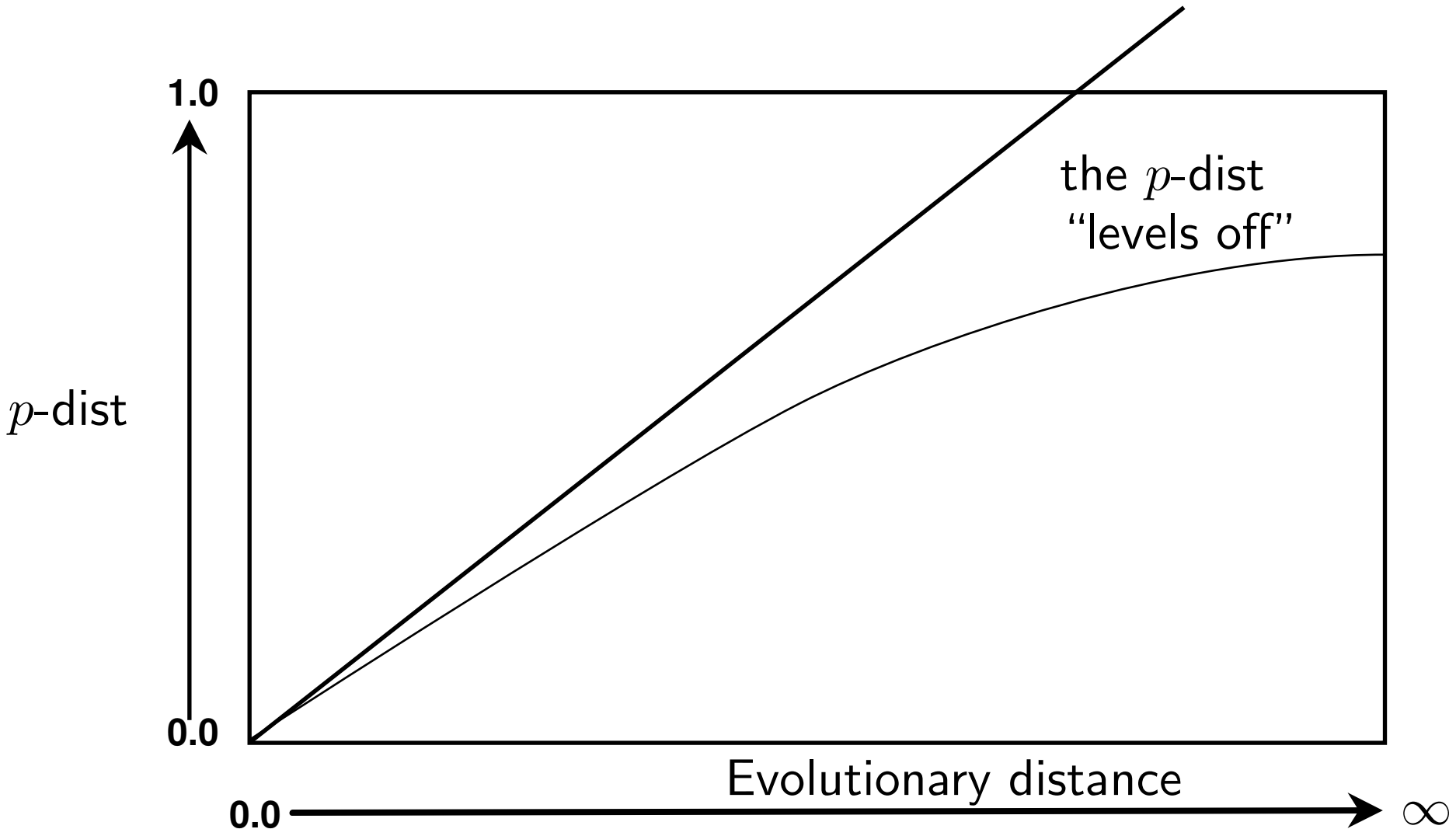
The evolutionary distance is the number of mutations that have occurred along the path that connects two tips.

We hope the distances that we measure can produce good estimates of the evolutionary distance, but we know that they cannot be perfect.

Intuition of sequence divergence vs evolutionary distance

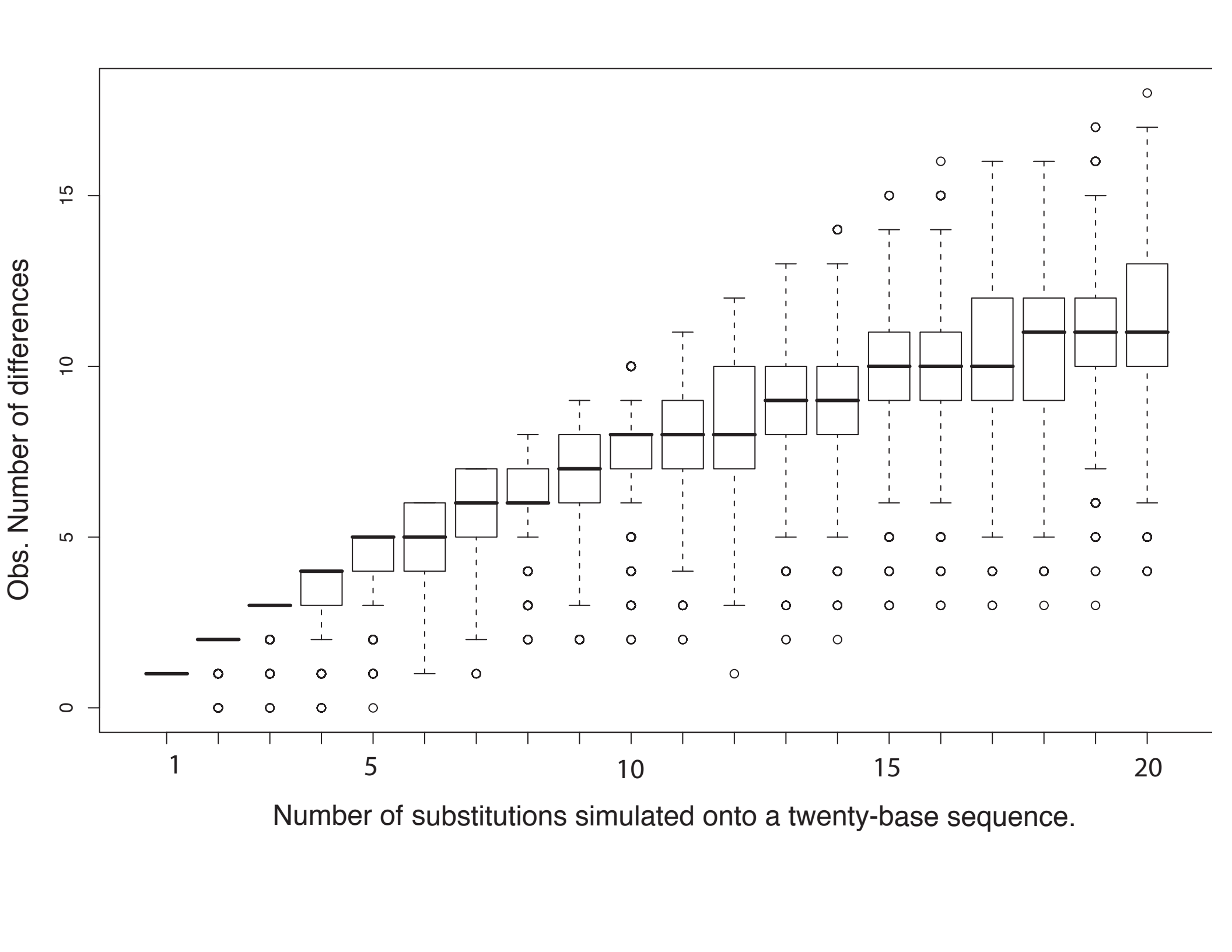


Sequence divergence vs evolutionary distance



"Multiple hits" problem (also known as saturation)

- Levelling off of sequence divergence vs time plot is caused by multiple substitutions affecting the same site in the DNA.
- At large distances the "raw" sequence divergence (also known as the p-distance or Hamming distance) is a poor estimate of the true evolutionary distance.
- Statistical models must be used to correct for unobservable substitutions (much more on these models tomorrow!)
- Large p-distances respond more to model-based correction – and there is a larger error associated with the correction.


Obs. Number of differences
0
5
10
15
1
5
10
15
20
Number of substitutions simulated onto a twenty-base sequence.

Distance corrections

- applied to distances before tree estimation,
- converts raw distances to an estimate of the evolutionary distance

$$d = -\frac{3}{4} \ln\left(1 - \frac{4c}{3}\right)$$

"raw" p-distances

	1	2	3
2	c_{12}		
3	c_{13}	c_{23}	
4	c_{14}	c_{24}	c_{34}

corrected distances

	1	2	3
2	d_{12}		
3	d_{13}	d_{23}	
4	d_{14}	d_{24}	d_{34}

$$d = -\frac{3}{4} \ln\left(1 - \frac{4c}{3}\right)$$

"raw" p-distances

	1	2	3
2	0.0		
3	0.3	0.3	
4	0.2	0.2	0.3

corrected distances

	1	2	3
2	0		
3	0.383	0.383	
4	0.233	0.233	0.383

Least Squares Branch Lengths

$$\text{Sum of Squares} = \sum_i \sum_j \frac{(p_{ij} - d_{ij})^2}{\sigma_{ij}^k}$$

- minimize discrepancy between path lengths and observed distances
- σ_{ij}^k is used to "downweight" distance estimates with high variance

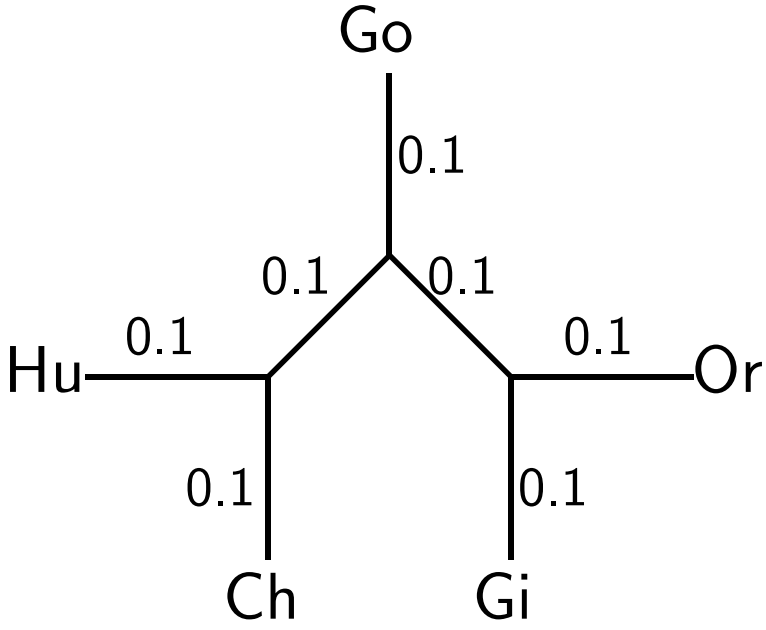
Least Squares Branch Lengths

$$\text{Sum of Squares} = \sum_i \sum_j \frac{(p_{ij} - d_{ij})^2}{\sigma_{ij}^k}$$

- in unweighted least-squares (Cavalli-Sforza & Edwards, 1967): $k = 0$
- in the method Fitch-Margoliash (1967): $k = 2$ and $\sigma_{ij} = d_{ij}$

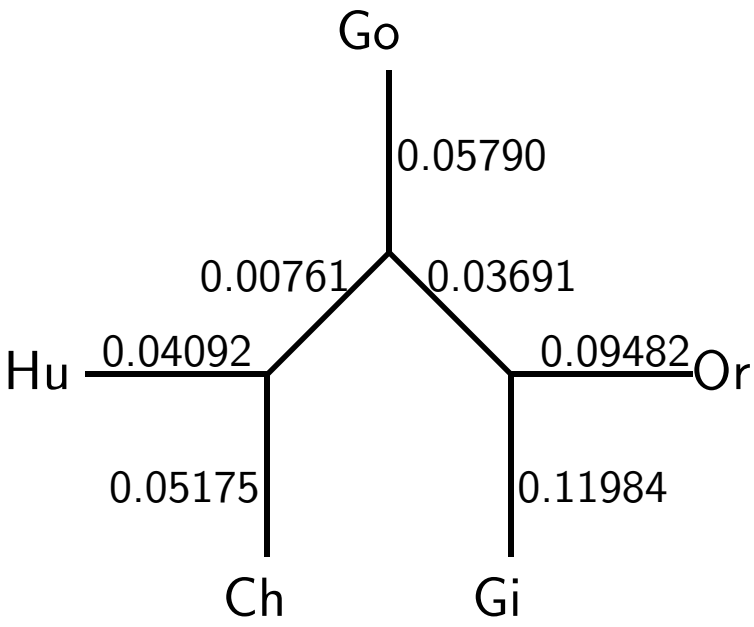
Poor fit using arbitrary branch lengths

Species	d_{ij}	p_{ij}	$(p-d)^2$
Hu-Ch	0.09267	0.2	0.01152
Hu-Go	0.10928	0.3	0.03637
Hu-Or	0.17848	0.4	0.04907
Hu-Gi	0.20420	0.4	0.03834
Ch-Go	0.11440	0.3	0.03445
Ch-Or	0.19413	0.4	0.04238
Ch-Gi	0.21591	0.4	0.03389
Go-Or	0.18836	0.3	0.01246
Go-Gi	0.21592	0.3	0.00707
Or-Gi	0.21466	0.2	0.00021
		S.S.	**0.26577**

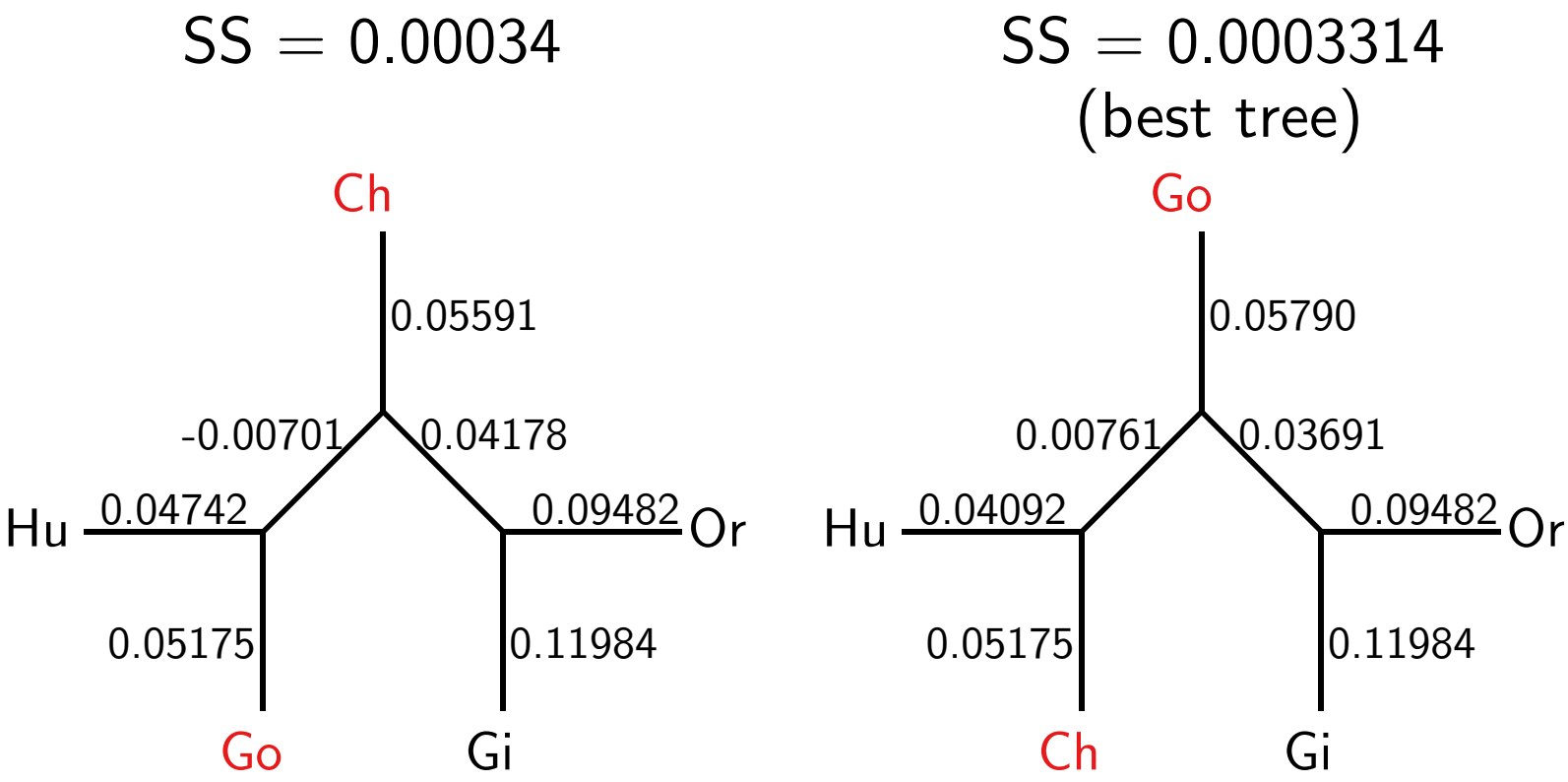


Optimizing branch lengths yields the least-squares score

Species	d_{ij}	p_{ij}	$(p-d)^2$
Hu-Ch	0.09267	0.09267	0.000000000
Hu-Go	0.10928	0.10643	0.000008123
Hu-Or	0.17848	0.18026	0.000003168
Hu-Gi	0.20420	0.20528	0.000001166
Ch-Go	0.11440	0.11726	0.000008180
Ch-Or	0.19413	0.19109	0.000009242
Ch-Gi	0.21591	0.21611	0.000000040
Go-Or	0.18836	0.18963	0.000001613
Go-Gi	0.21592	0.21465	0.000001613
Or-Gi	0.21466	0.21466	0.000000000
		S.S.	**0.000033144**

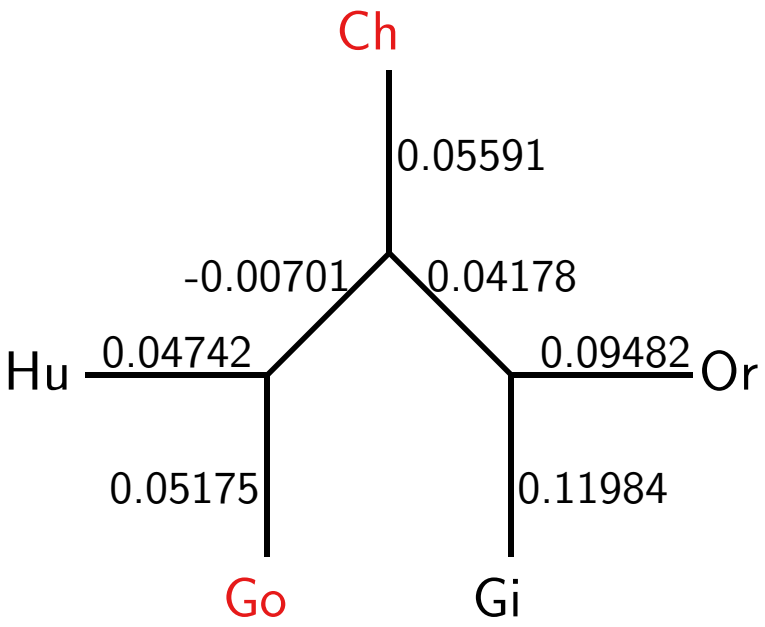


Least squares as an optimality criterion

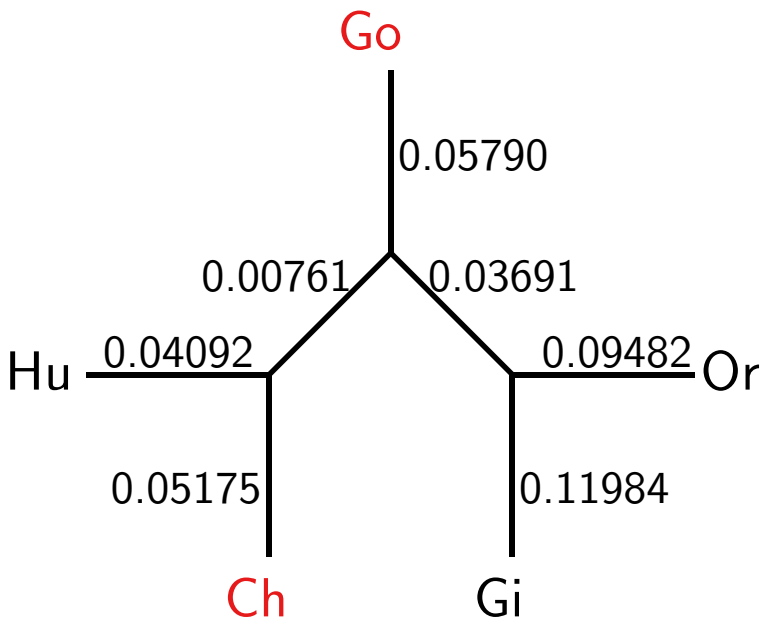


Minimum evolution optimality criterion

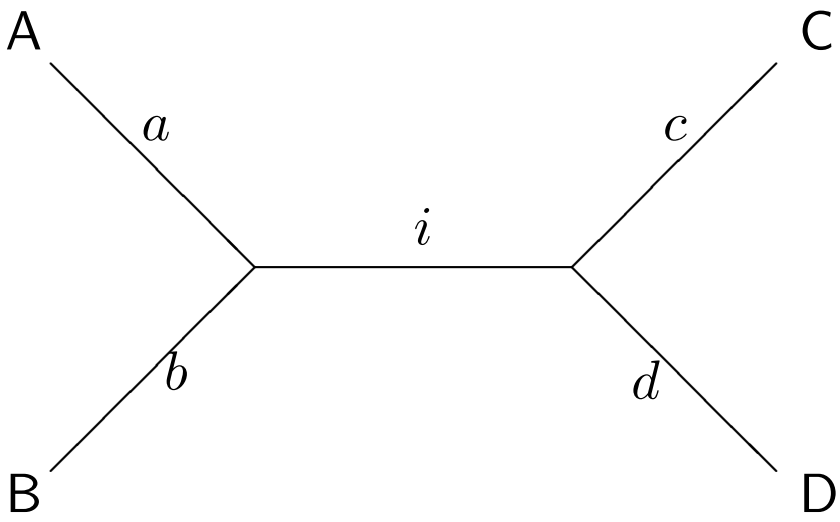
Sum of branch lengths
=0.41152



Sum of branch lengths
=0.40975
(best tree)



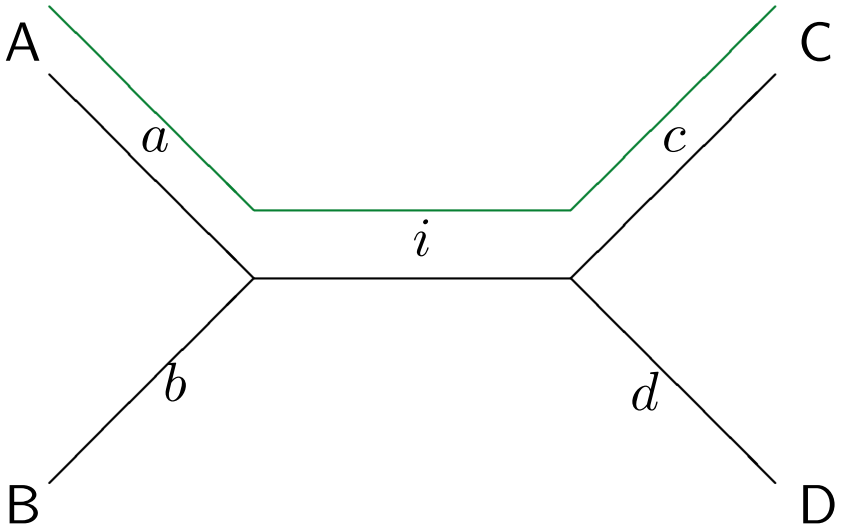
We still use least squares branch lengths when we use Minimum Evolution



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

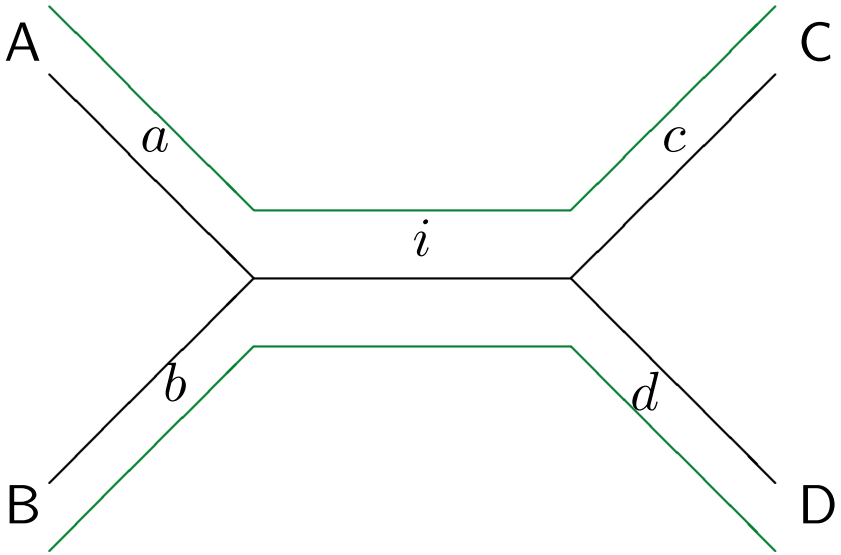
If the tree above is correct then:

$$
\begin{aligned}
p_{AB} &= a + b \\
p_{AC} &= a + i + c \\
p_{AD} &= a + i + d \\
p_{BC} &= b + i + c \\
p_{BD} &= b + i + d \\
p_{CD} &= c + d
\end{aligned}
$$



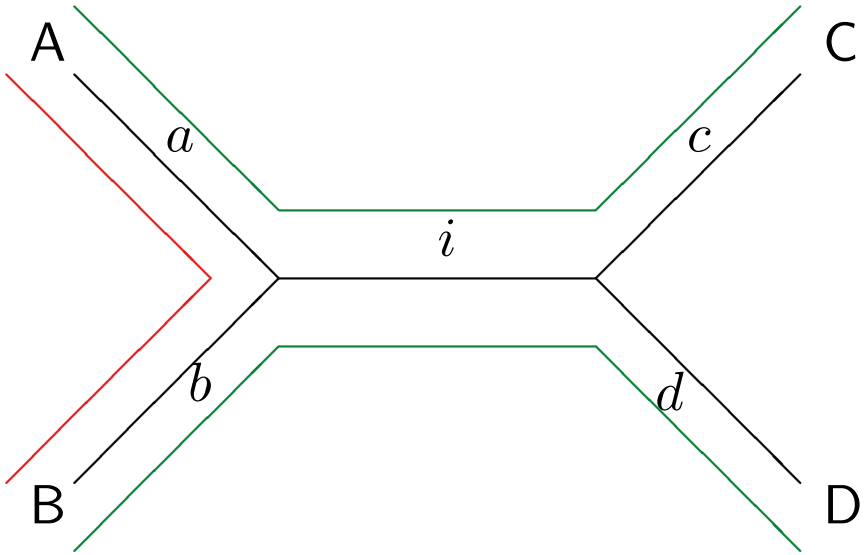
	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

d_{AC}



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

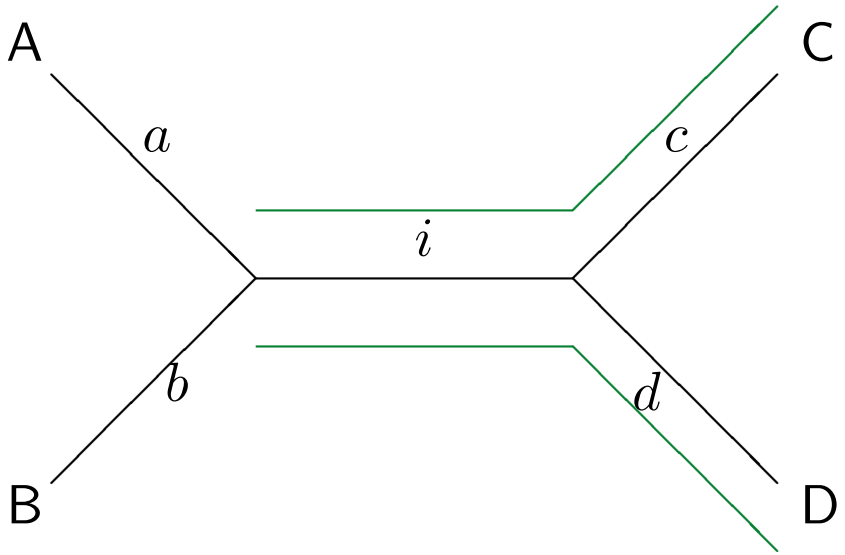
$$d_{AC} + d_{BD}$$



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

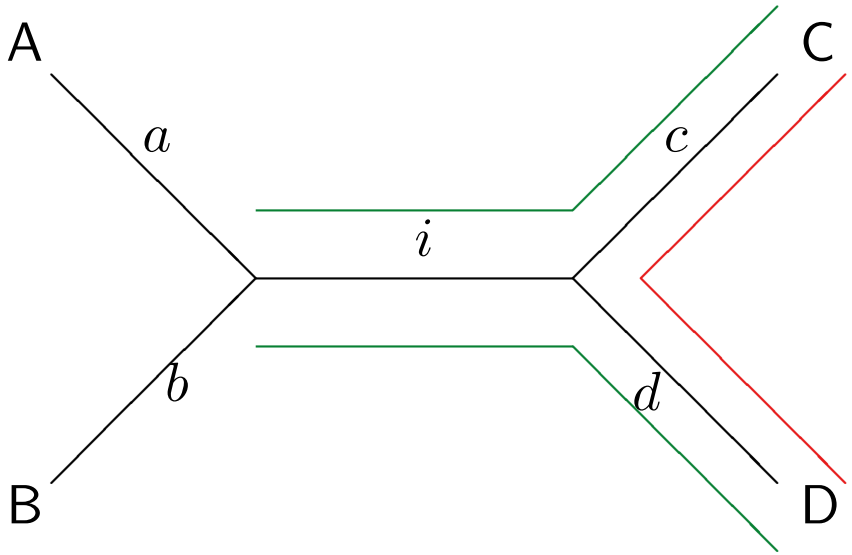
$$d_{AC} + d_{BD}$$

$$d_{AB}$$



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

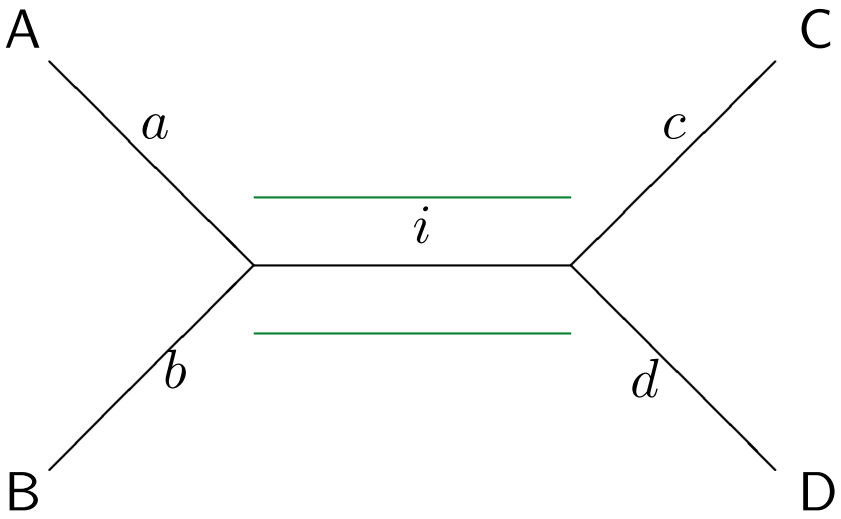
$$d_{AC} + d_{BD} - d_{AB}$$



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

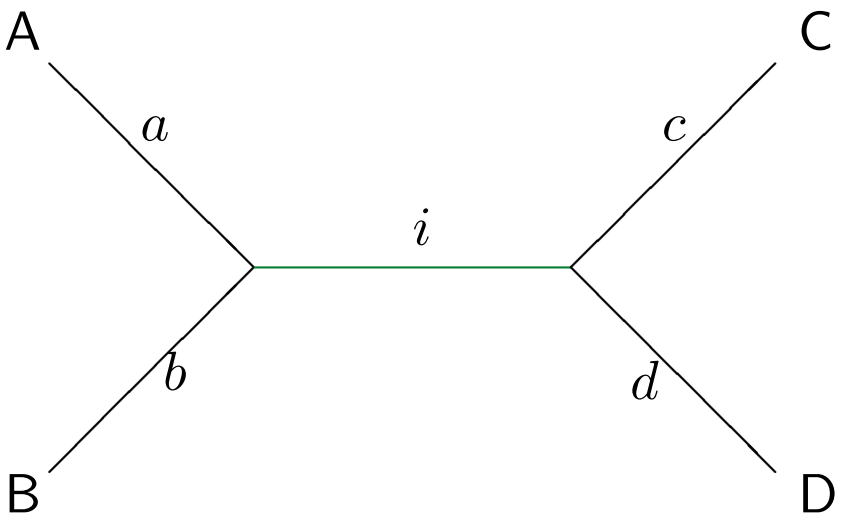
$$d_{AC} + d_{BD} - d_{AB}$$

$$d_{CD}$$



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

$$d_{AC} + d_{BD} - d_{AB} - d_{CD}$$



	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

$$i^{\dagger} = \frac{d_{AC}+d_{BD}-d_{AB}-d_{CD}}{2}$$

Note that our estimate

$$i^{\dagger} = \frac{d_{AC} + d_{BD} - d_{AB} - d_{CD}}{2}$$

does not use all of our data. d_{BC} and d_{AD} are ignored!

We could have used $d_{BC} + d_{AD}$ instead of $d_{AC} + d_{BD}$ (you can see this by going through the previous slides after rotating the internal branch).

$$i^{*} = \frac{d_{BC} + d_{AD} - d_{AB} - d_{CD}}{2}$$

A better estimate than either i or i^* would be the average of both of them:

$$i' = \frac{d_{BC} + d_{AD} + d_{AC} + d_{BD}}{2} - d_{AB} - d_{CD}$$

This logic has been extend to trees of more than 4 taxa by Pauplin (2000) and Semple and Steel (2004).

Balanced minimum evolution

Desper and Gascuel (2002, 2004) refer to fitting the branch lengths using the estimators of Pauplin (2000) and preferring the tree with the smallest tree length "Balanced Minimum Evolution."

They that it is equivalent to a form of weighted least squares in which distances are down-weighted by an exponential function of the topological distances between the leaves.

Desper and Gascuel (2005) showed that neighbor-joining is star decomposition (more on this later) under BME. See Gascuel and Steel (2006)

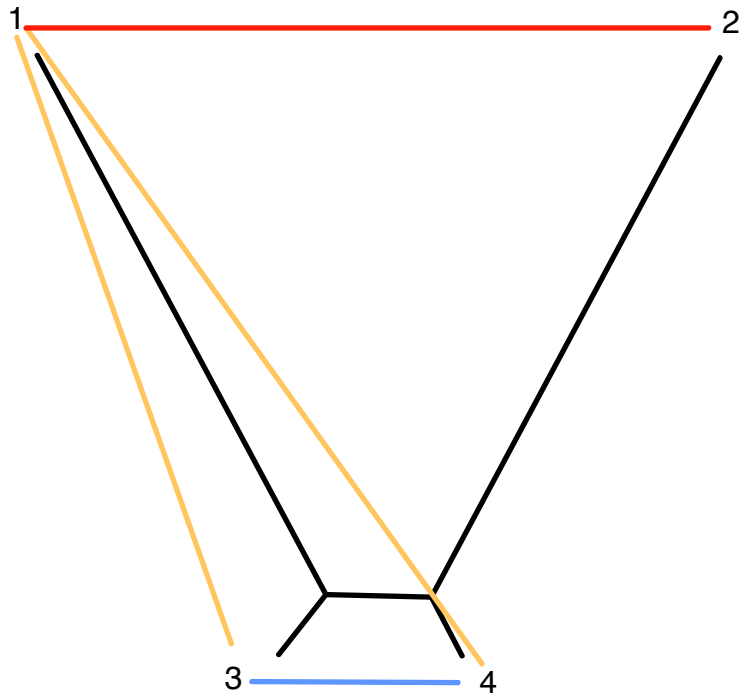
FastME

Software by Desper and Gascuel (2004) which implements searching under the balanced minimum evolution criterion.

It is extremely fast and is more accurate than neighbor-joining (based on simulation studies).

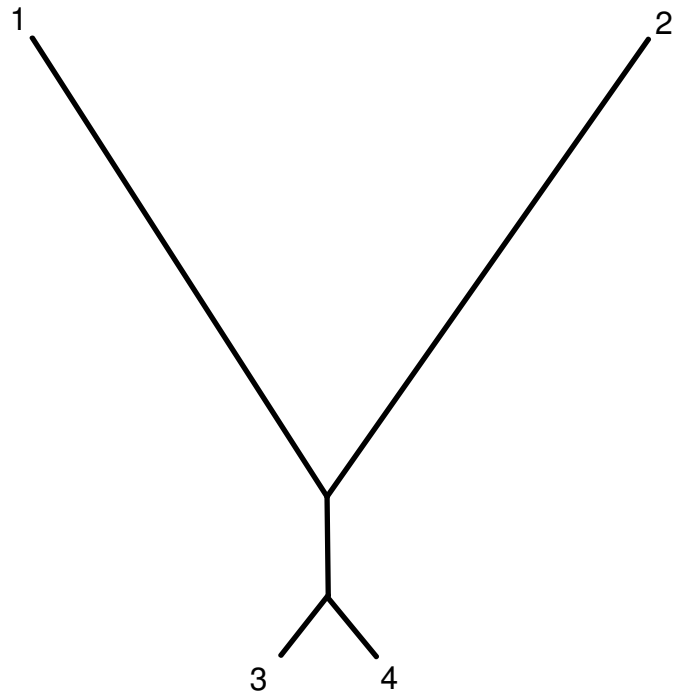
Failure to correct distance sufficiently leads to poor performance

"Under-correcting" will underestimate long evolutionary distances more than short distances



Failure to correct distance sufficiently leads to poor performance

The result is the classic "long-branch attraction" phenomenon.

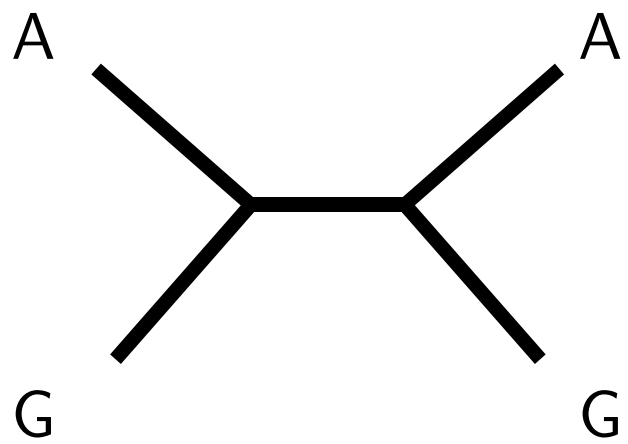


Distance methods: pros

- Fast – the new FastTree method Price et al. (2009) can calculate a tree in less time than it takes to calculate a full distance matrix!
- Can use models to correct for unobserved differences
- Works well for closely related sequences
- Works well for clock-like sequences

Distance methods: cons

- Do not use all of the information in sequences
- Do not reconstruct character histories, so they not enforce all logical constraints



References

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330.

Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9(5):687–705.

Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*.

Desper, R. and Gascuel, O. (2005). The minimum evolution distance-based approach to phylogenetic inference. In Gascuel, O., editor, *Mathematics of Evolution and Phylogeny*, pages 1–32. Oxford University Press.

Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000.

Goloboff, P. (1993). Estimating character weights during tree search. *Cladistics*, 9(1):83–91.

Heled, J. and Drummond, A. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.

Noutahi, E., Semeria, M., Lafond, M., Seguin, J., Boussau, B., Guéguen, L., El-Mabrouk, N., and Tannier, E. (2016). Efficient gene tree correction guided by species and synteny evolution.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 2000(51):41–47.

Price, M. N., Dehal, P., and Arkin, A. P. (2009). FastTree: Computing large minimum-evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7):1641–1650.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.

Semple, C. and Steel, M. (2004). Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, 32(4):669–680.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*.