Introduction to Phylogenetics

Workshop on Molecular Evolution 2018

Marine Biological Lab, Woods Hole, MA. USA

Mark T. Holder
University of Kansas

Outline

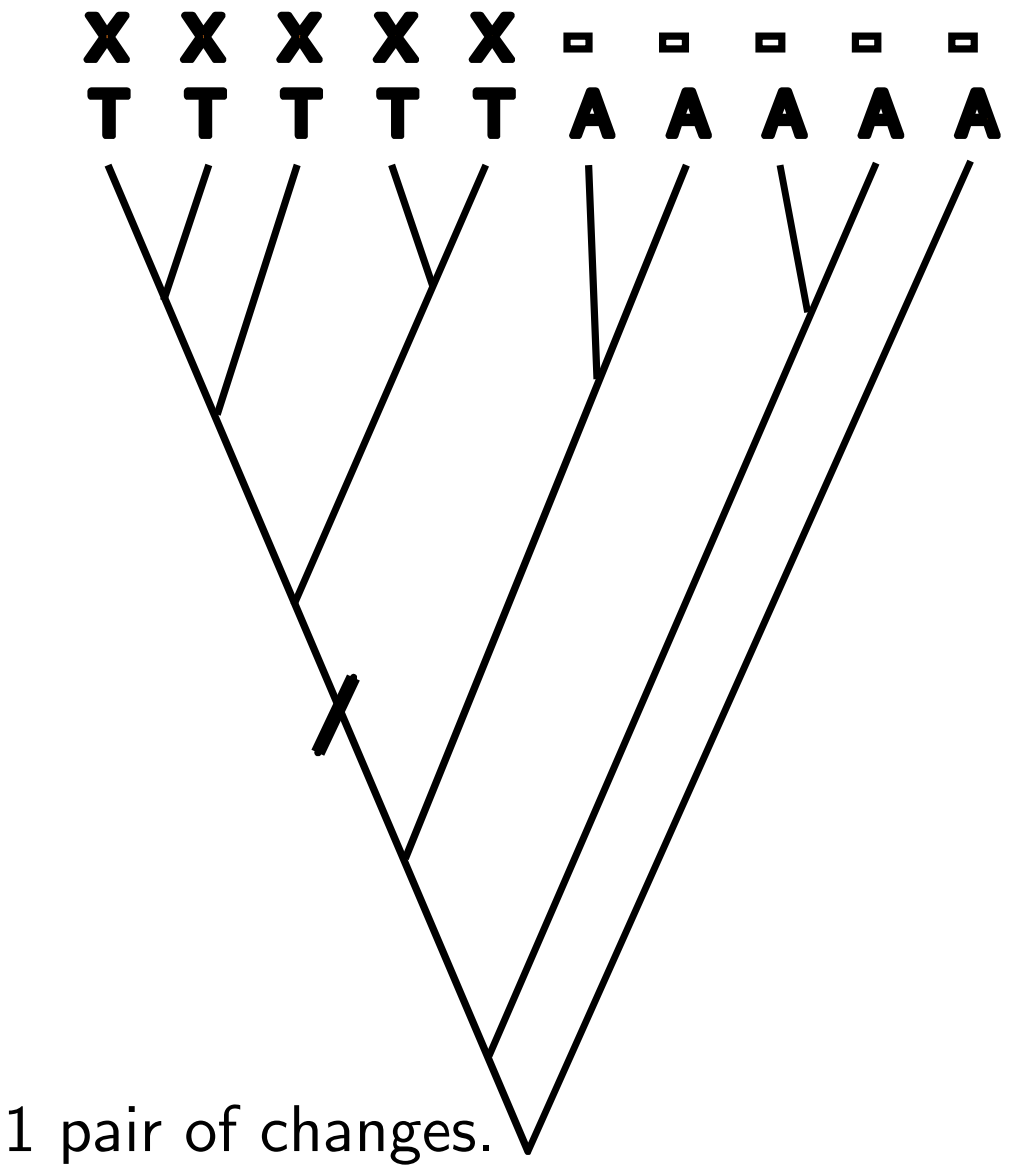
1. phylogenetics is crucial for comparative biology
2. tree terminology
3. why phylogenetics is difficult
4. parsimony
5. distance-based methods
6. theoretical basis of multiple sequence alignment

Part #1: phylogenetics is crucial for biology

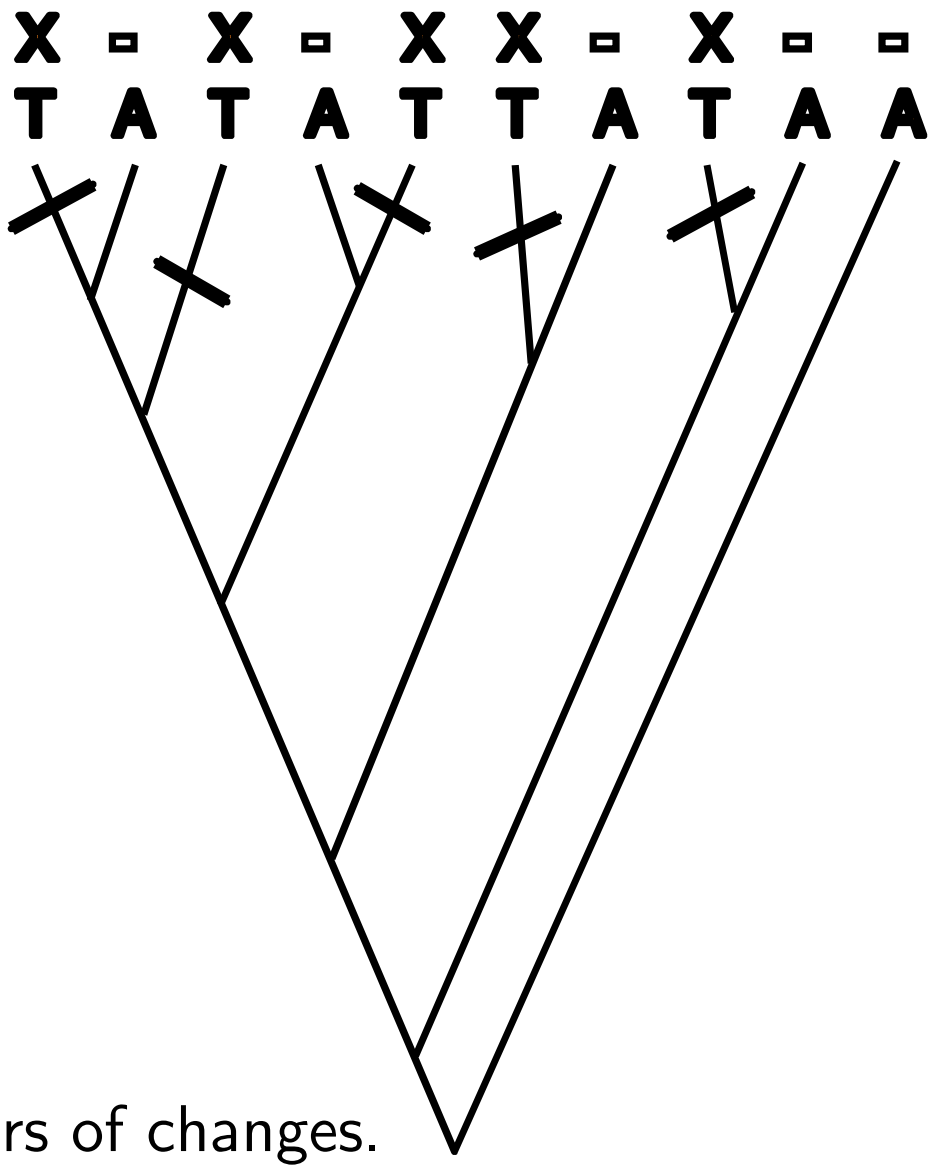
Species	Habitat	Photoprotection
1	terrestrial	xanthophyll
2	terrestrial	xanthophyll
3	terrestrial	xanthophyll
4	terrestrial	xanthophyll
5	terrestrial	xanthophyll
6	aquatic	none
7	aquatic	none
8	aquatic	none
9	aquatic	none
10	aquatic	none

slides by Paul Lewis

Phylogeny reveals the events that generate the pattern



1 pair of changes.
Coincidence?

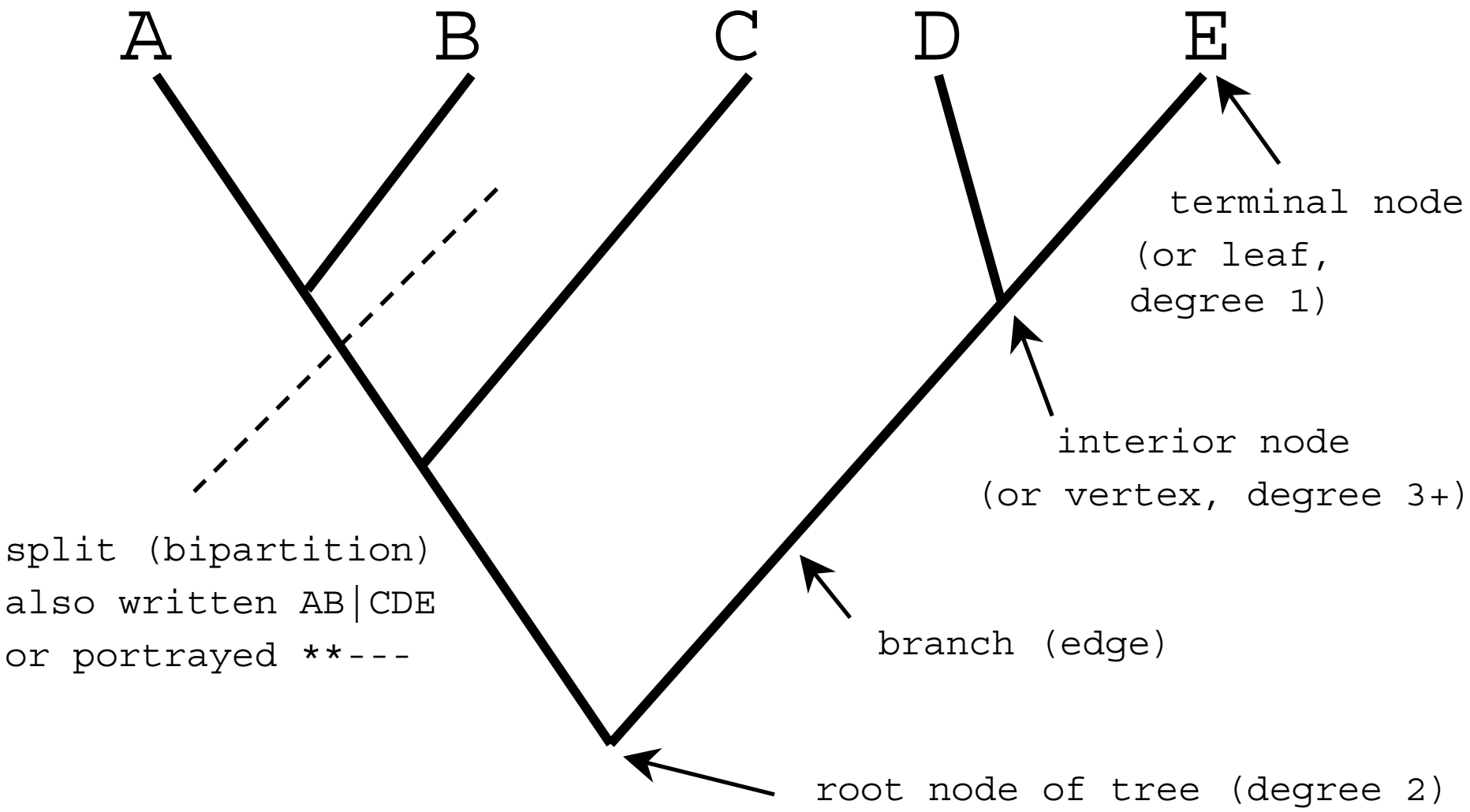


5 pairs of changes.
Much more convincing

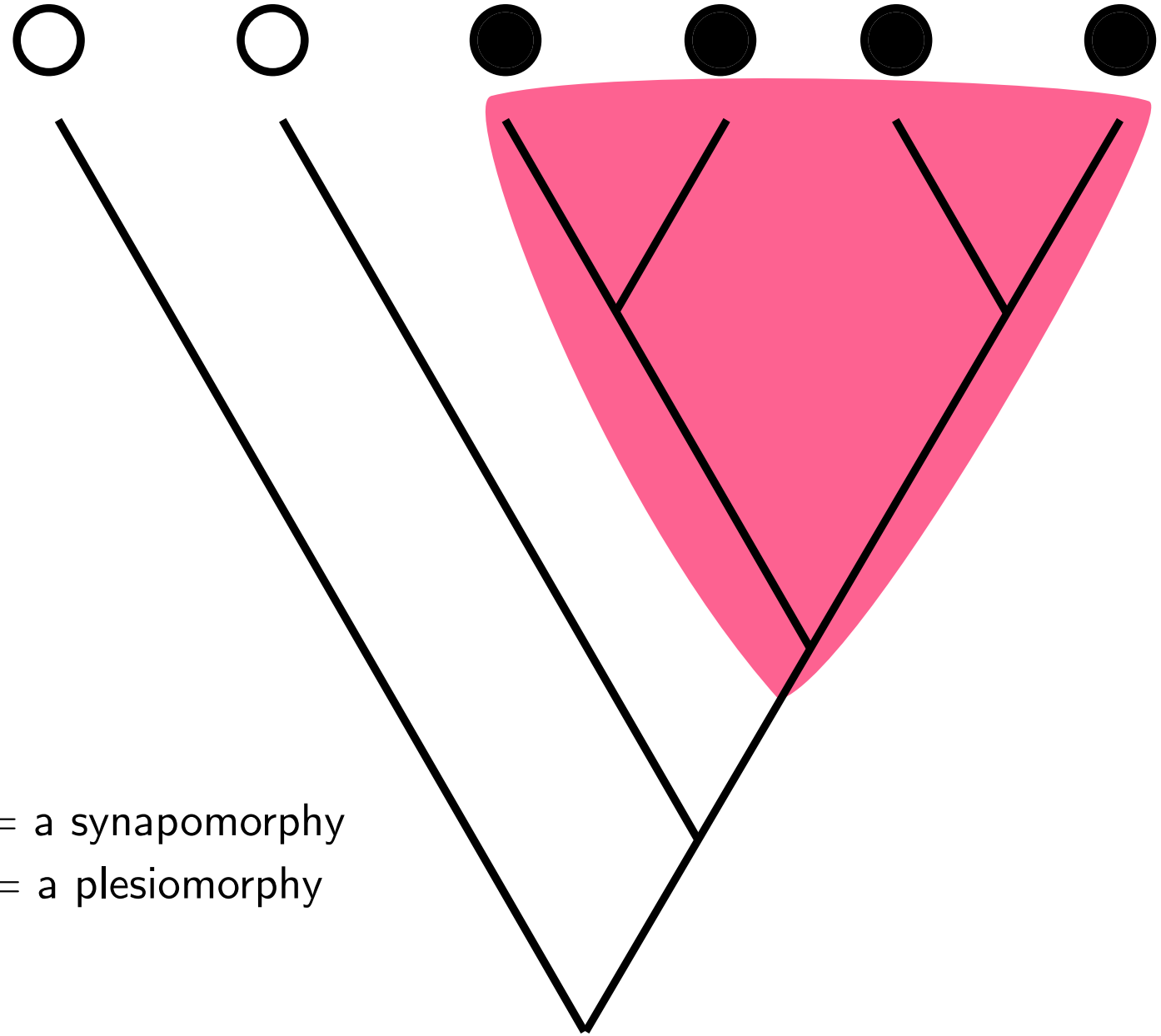
Many evolutionary questions require a phylogeny

- Determining whether a trait tends to be lost more often than gained, or vice versa
- Estimating divergence times (Tracy Heath Sunday + next Saturday)
- Distinguishing homology from analogy
- Inferring parts of a gene under strong positive selection (Joe Bielawski and Belinda Chang next Monday)

Part 2: Tree terminology

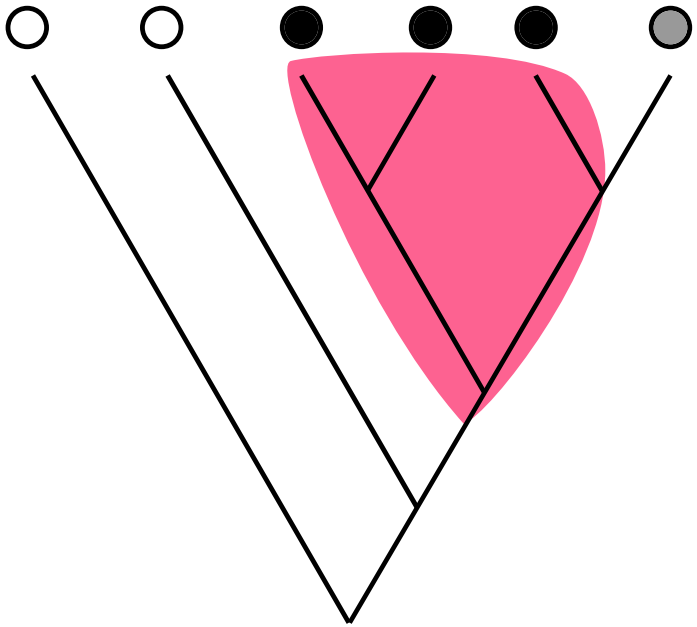


Monophyletic groups ("clades"): the basis of phylogenetic classification

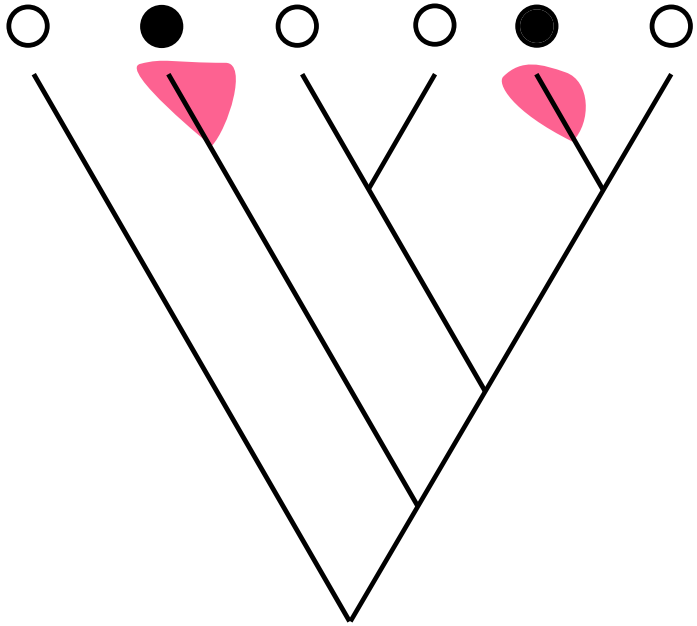


black state = a synapomorphy

white state = a plesiomorphy



Paraphyletic

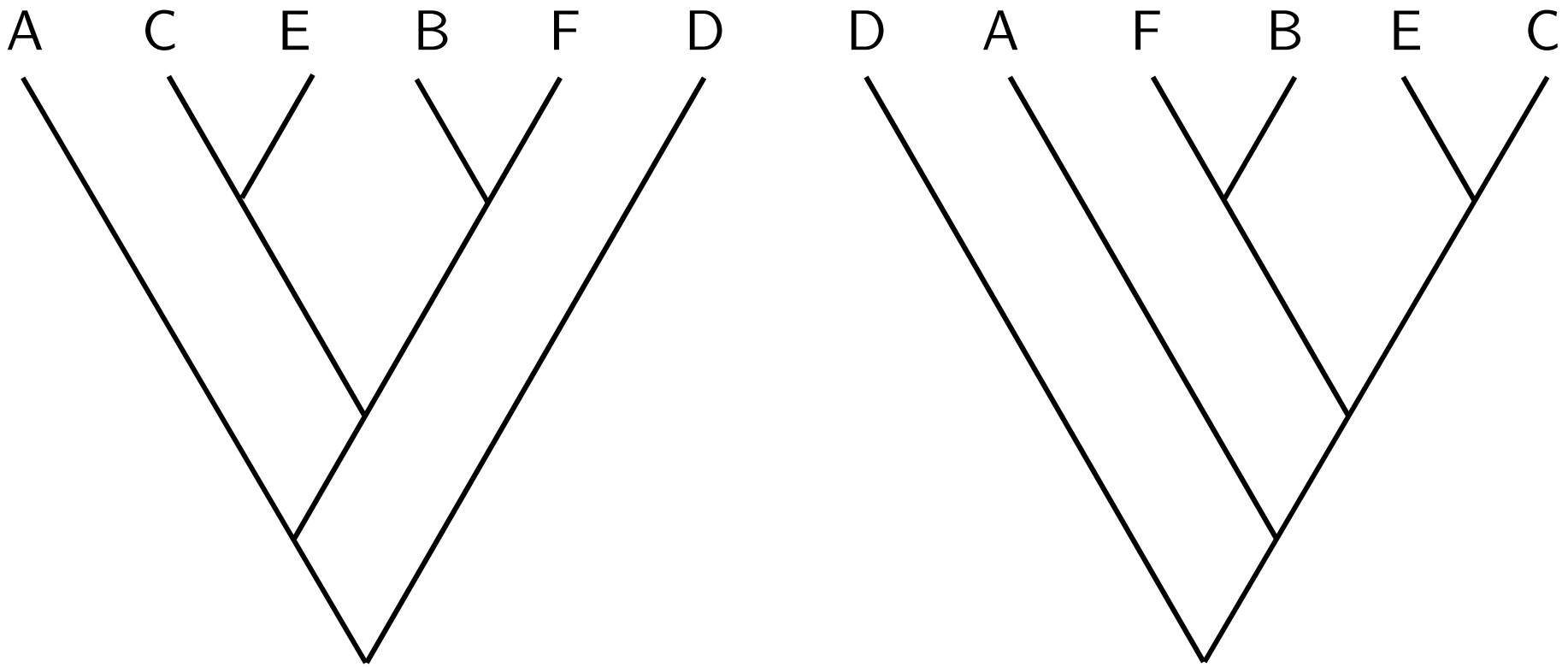


Polyphyletic

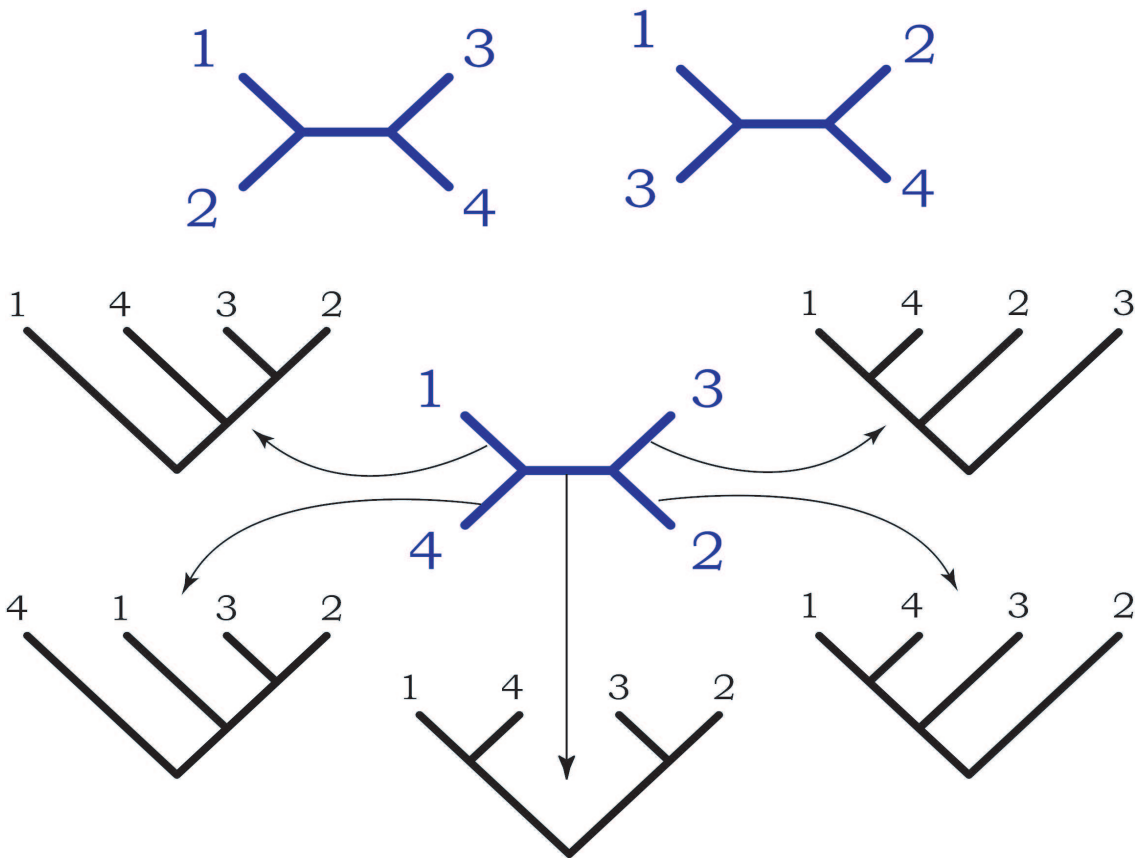
grey state is an autapomorphy

(images from Wikipedia)

Branch rotation does not matter



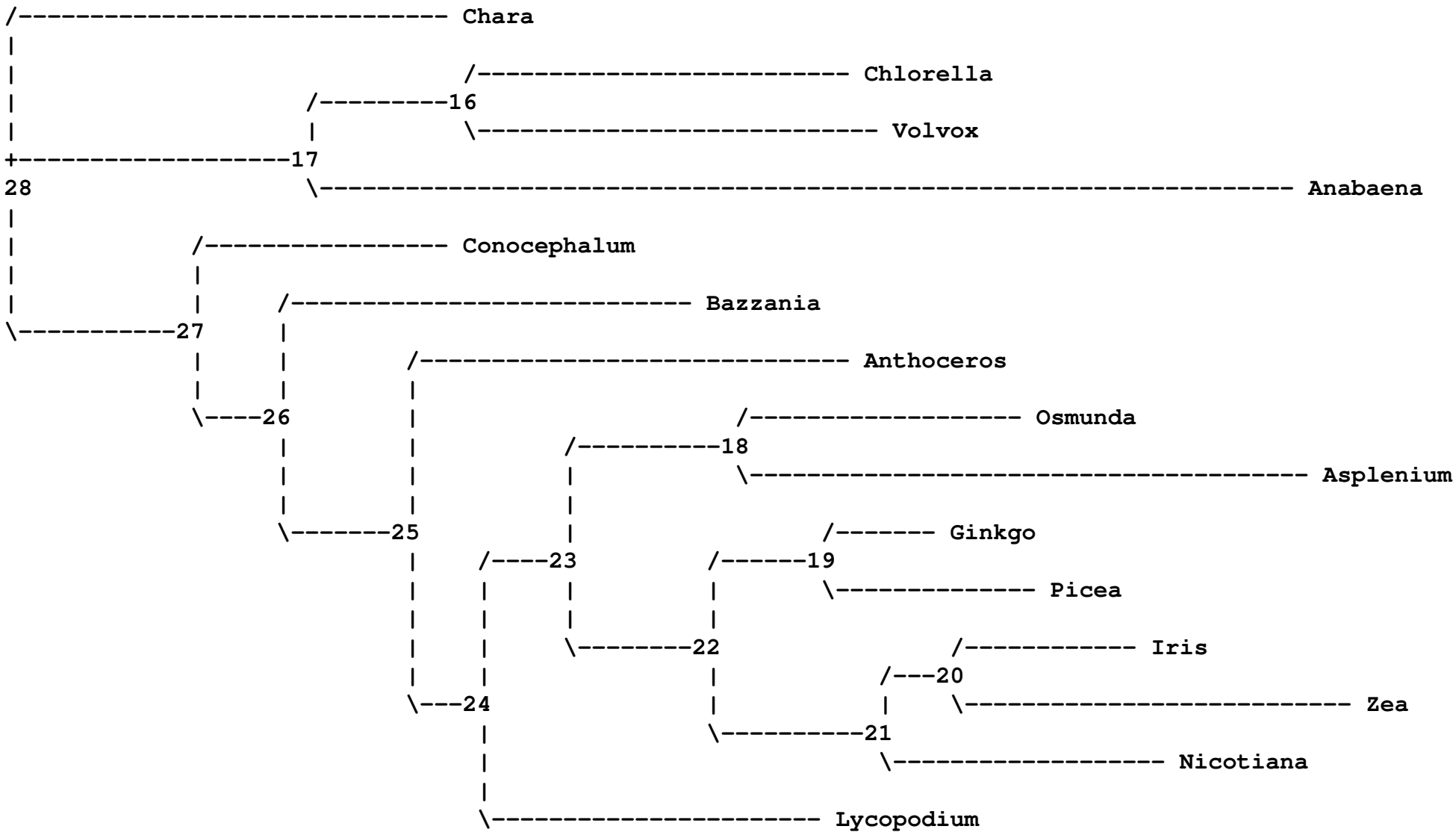
Rooted vs unrooted trees



ingroup: the focal taxa

outgroup: the taxa that are more distantly related. Assuming that the ingroup is monophyletic with respect to the outgroup can root a tree.

Warning: software often displays unrooted trees like this:



Part 3: Phylogenetics is difficult

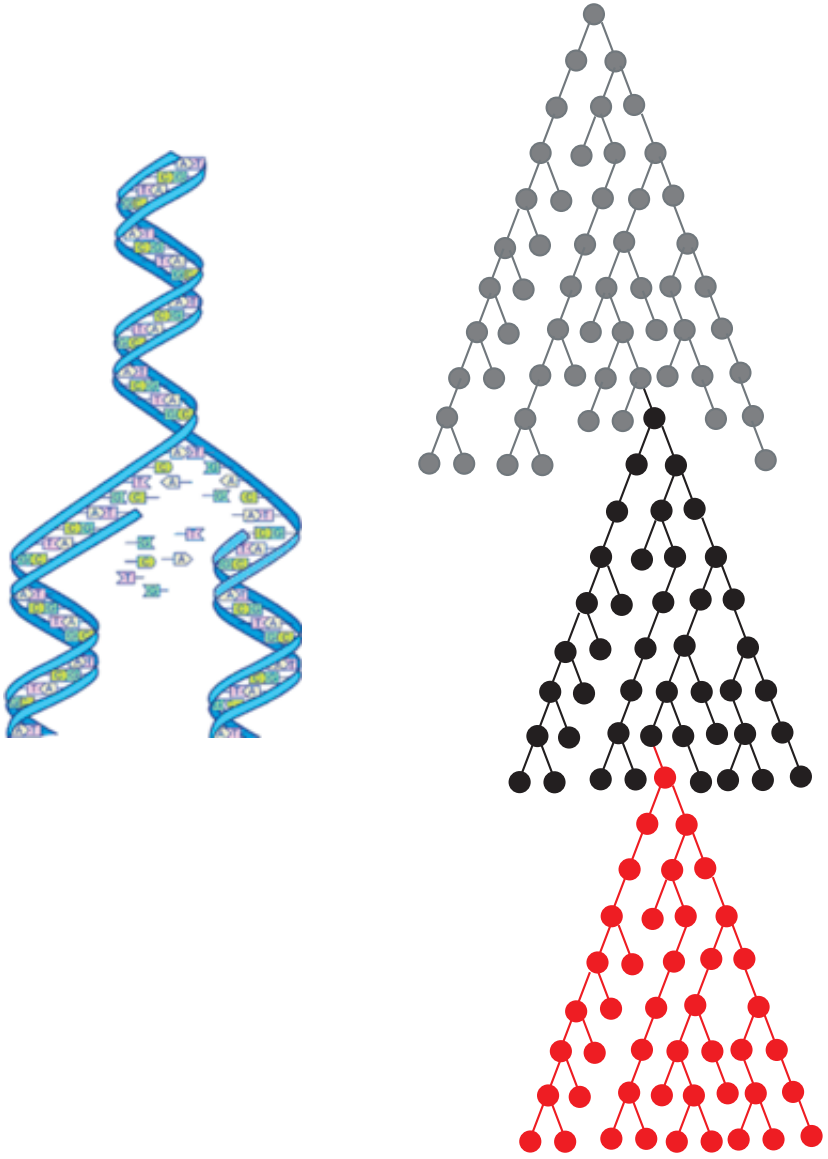
a. Many types of trees - species trees *vs* "gene trees" – coalescents or "gene family trees"

b. Many sources of error

c. No clean sampling theory that gives us clean hypothesis tests

d. Computational + statistical difficulties

(3a) Many types of trees: cellular genealogies



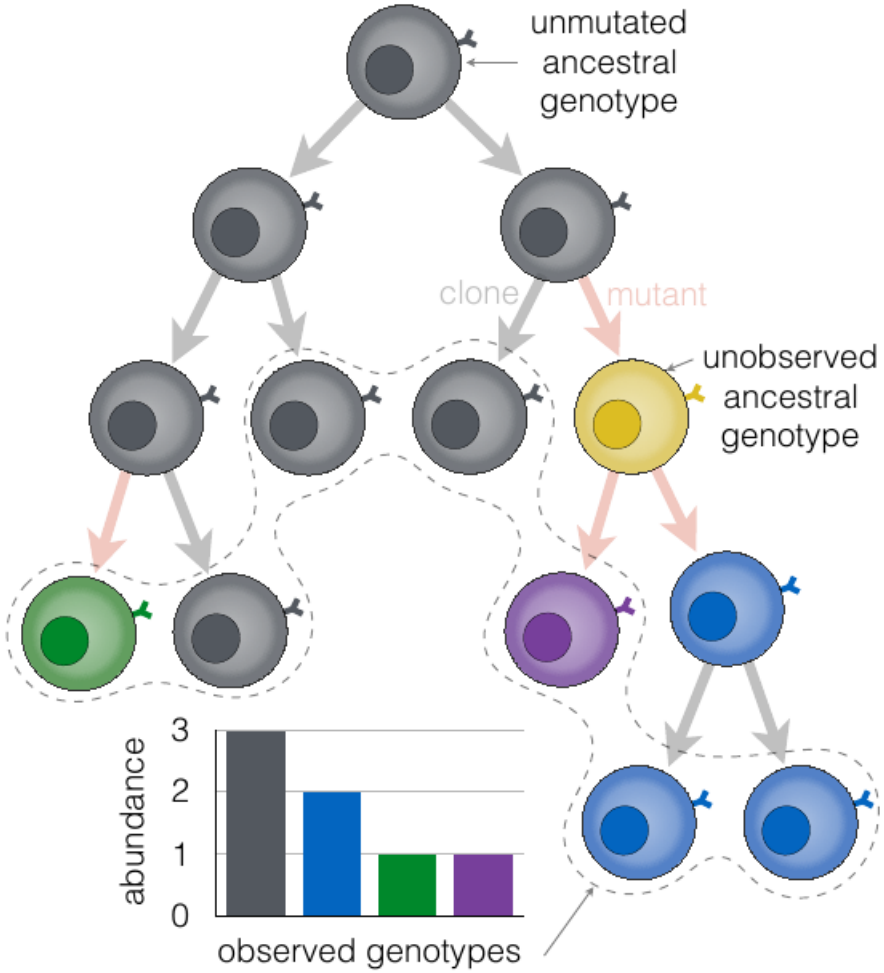
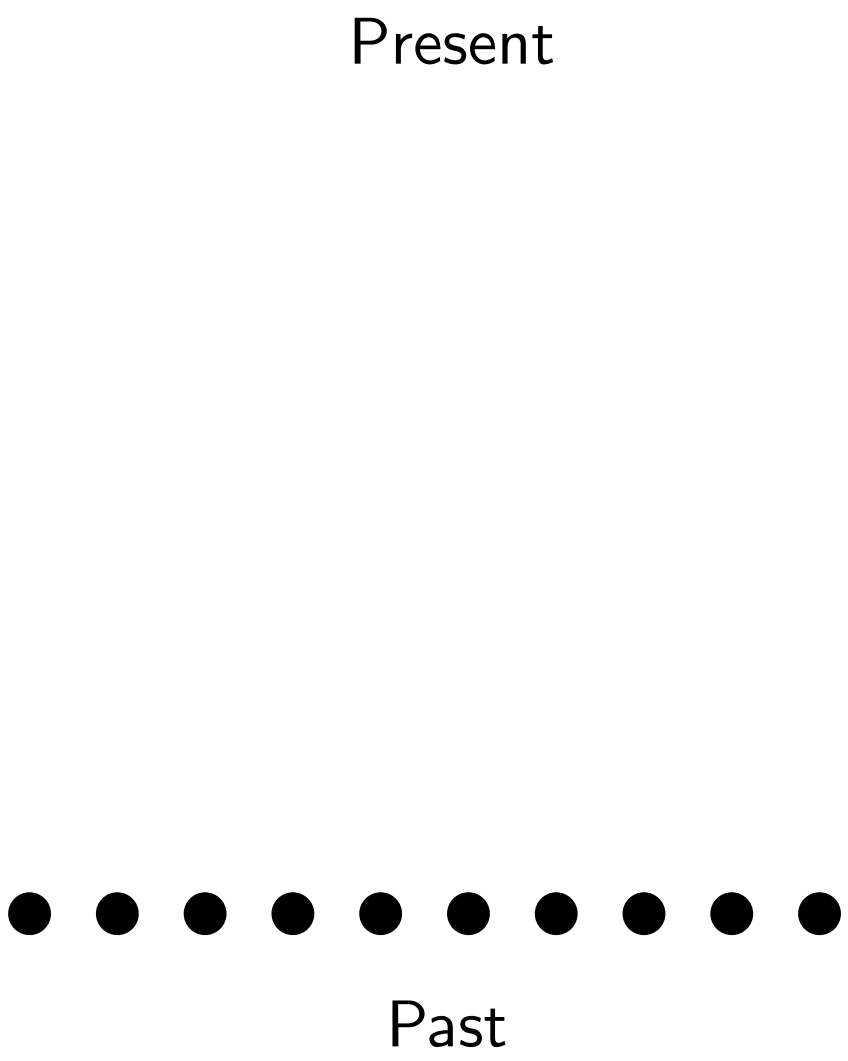


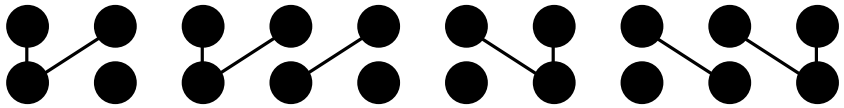
Figure 1 from DeWett *et al.* 2018

(3a) Many types of trees: genealogies in a population



(3a) Many types of trees: genealogies in a population

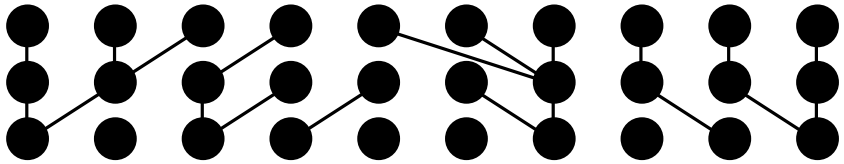
Present



Past

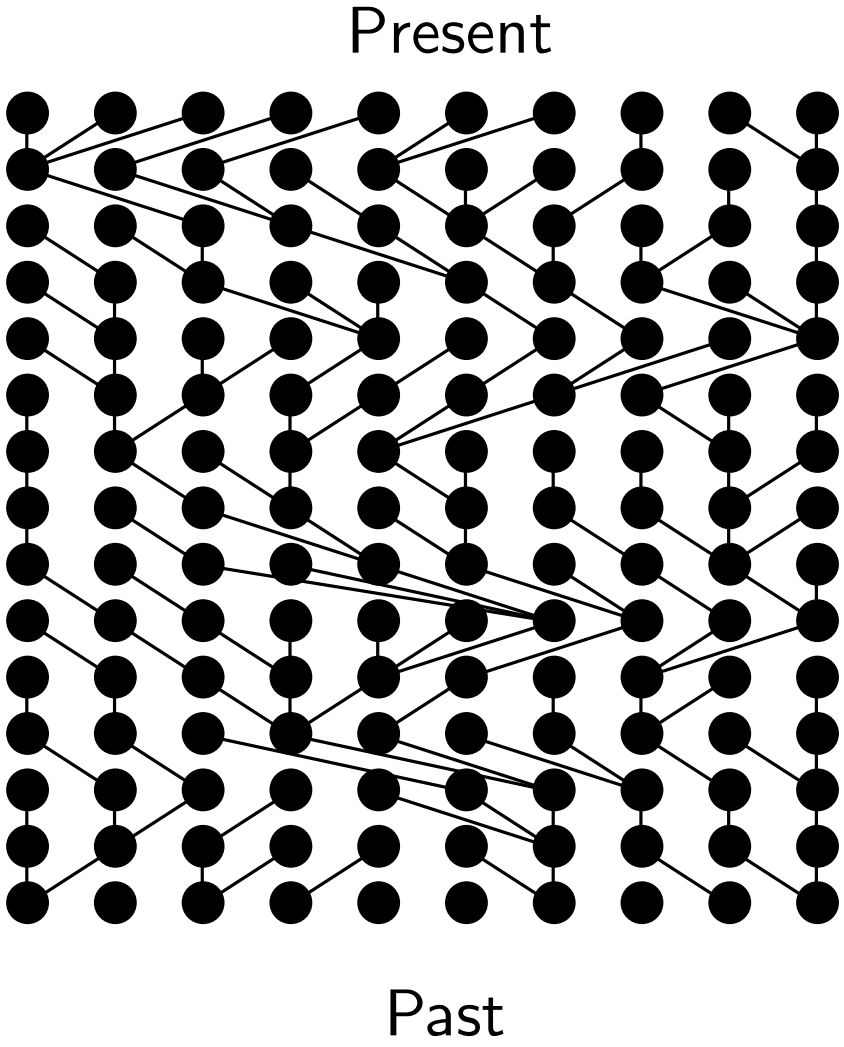
(3a) Many types of trees: genealogies in a population

Present

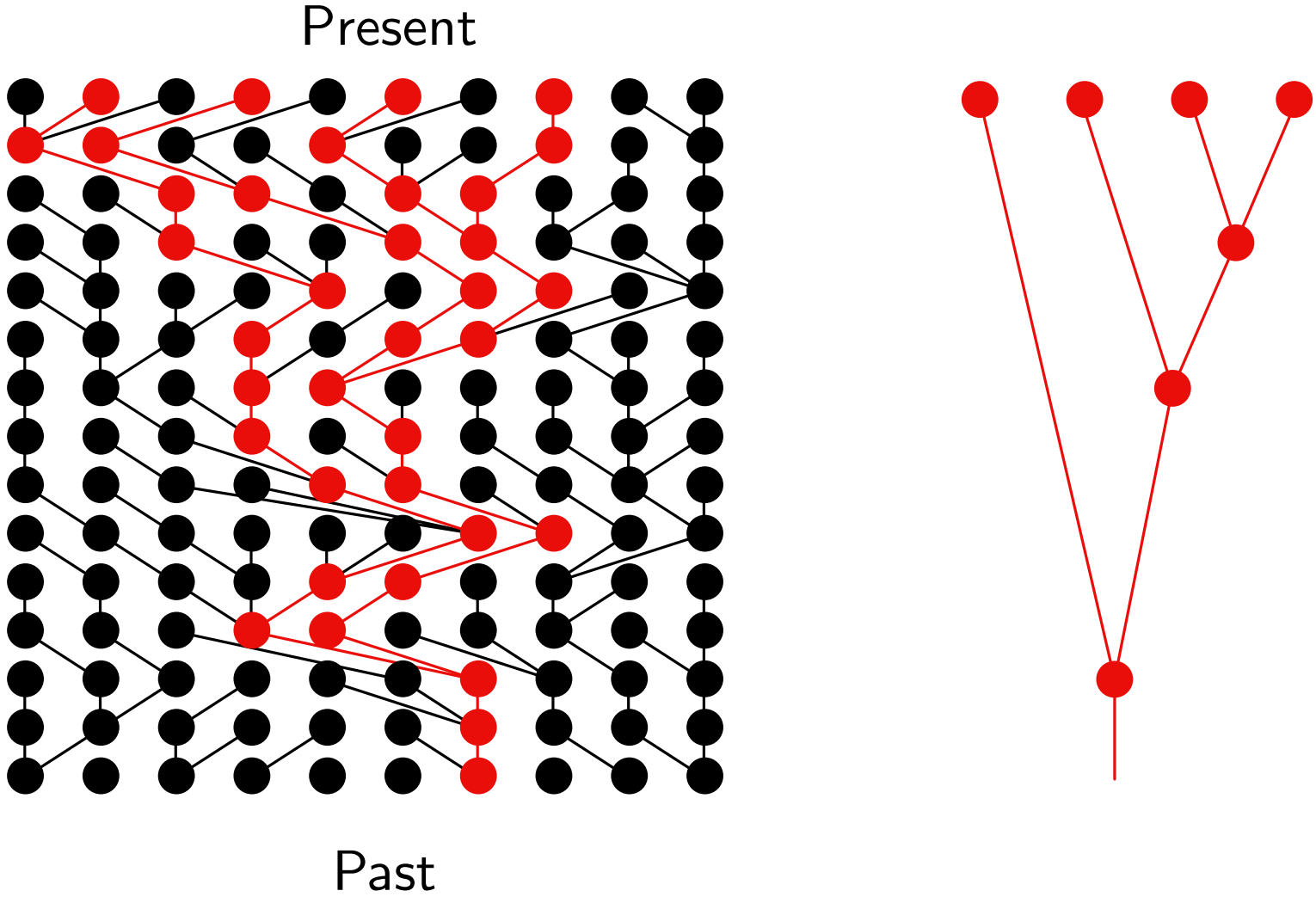


Past

(3a) Many types of trees: genealogies in a population



(3a) Many types of trees: genealogies in 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).

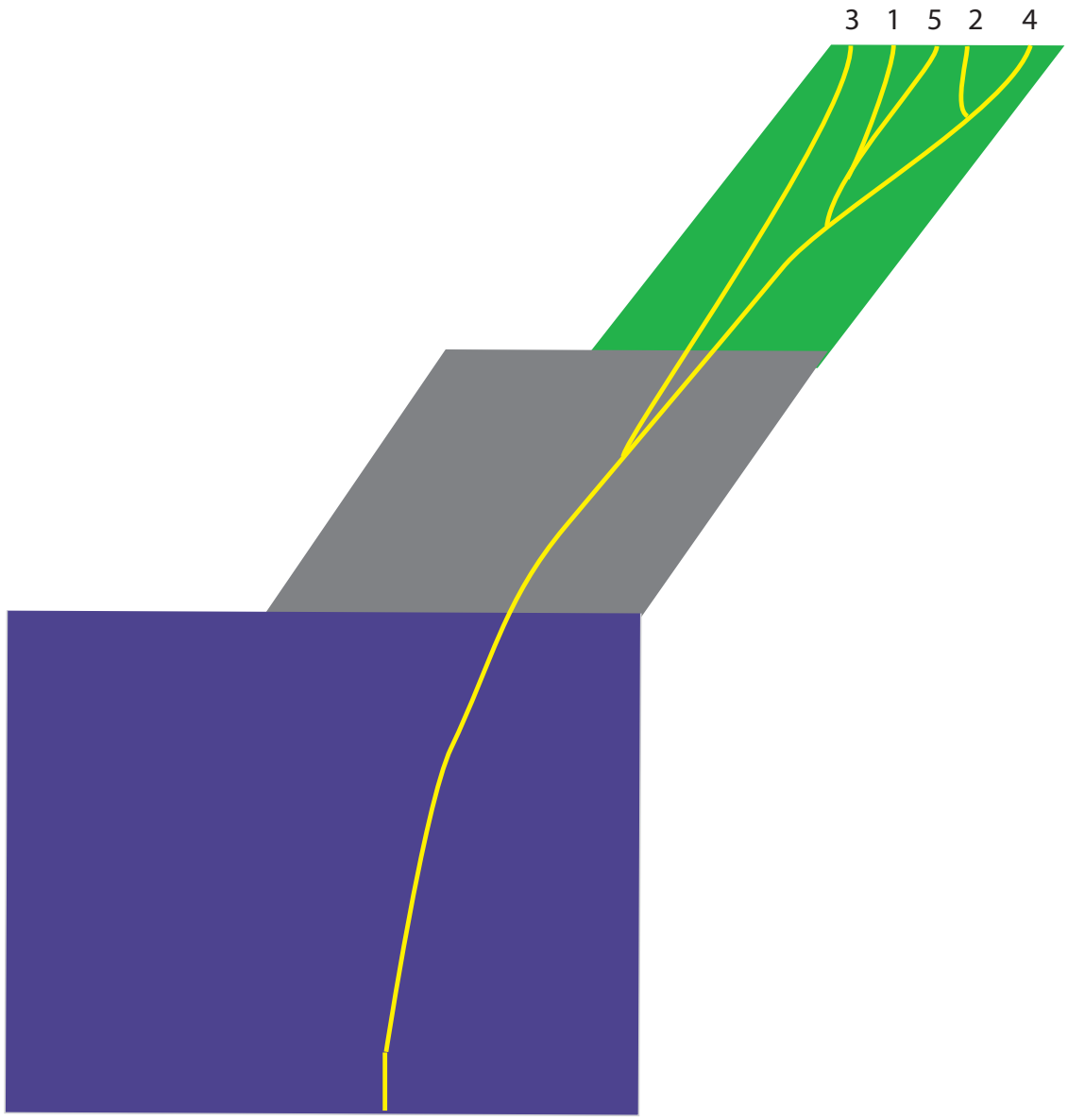
more terminology

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

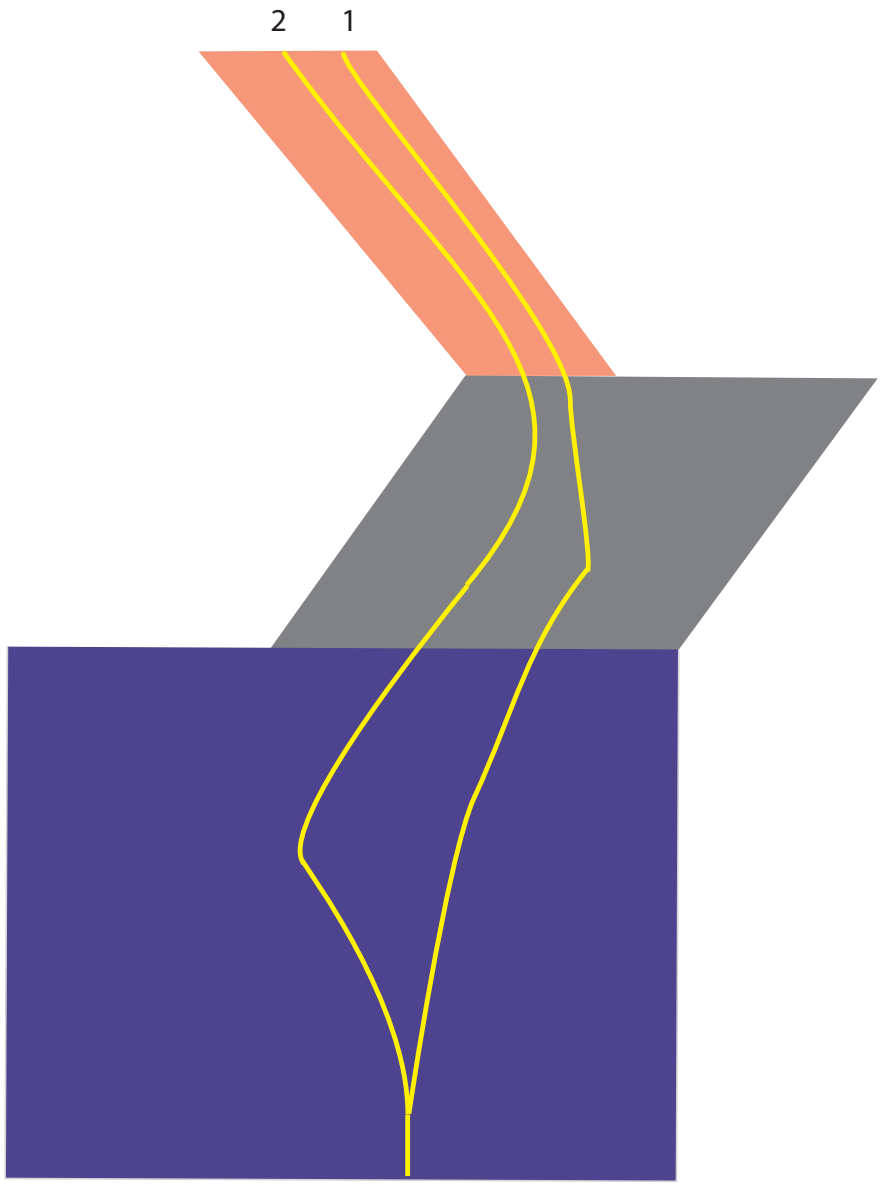
- allele – an alternative form a gene.
- haplotype – a linked set of alleles

But both of these terms require a differences in sequence.

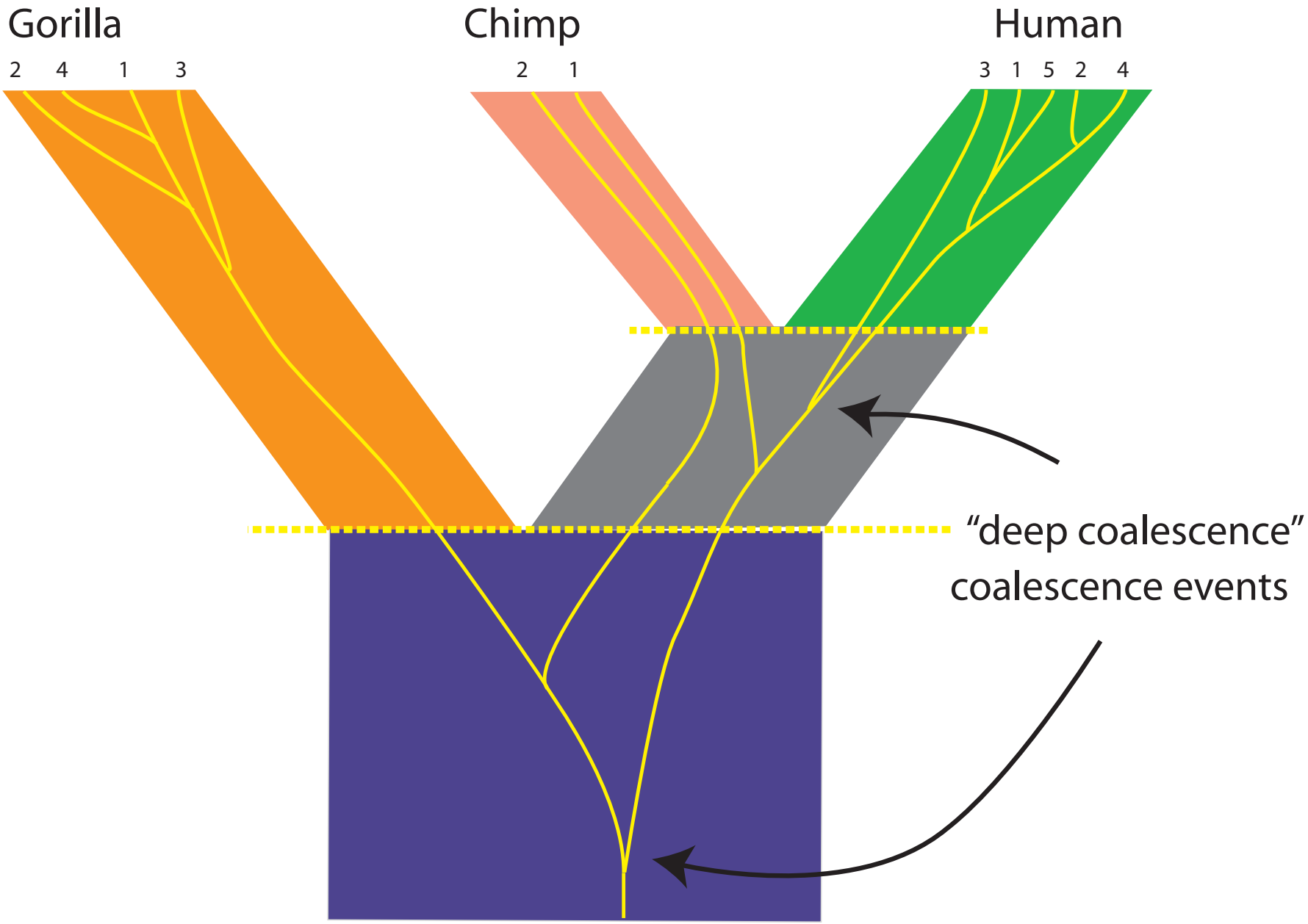
The gene trees that we draw depict genealogical relationships – regardless of whether or not nucleotide differences distinguish the "gene copies" at the tips of the tree.



(thanks to Peter Beerli for the images - next 3 slides)


2
1

(3a) 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

coalescent theory + estimating migration – Peter Beerli (next Thursday)

(3a) Inferring a species tree while accounting for the coalescent

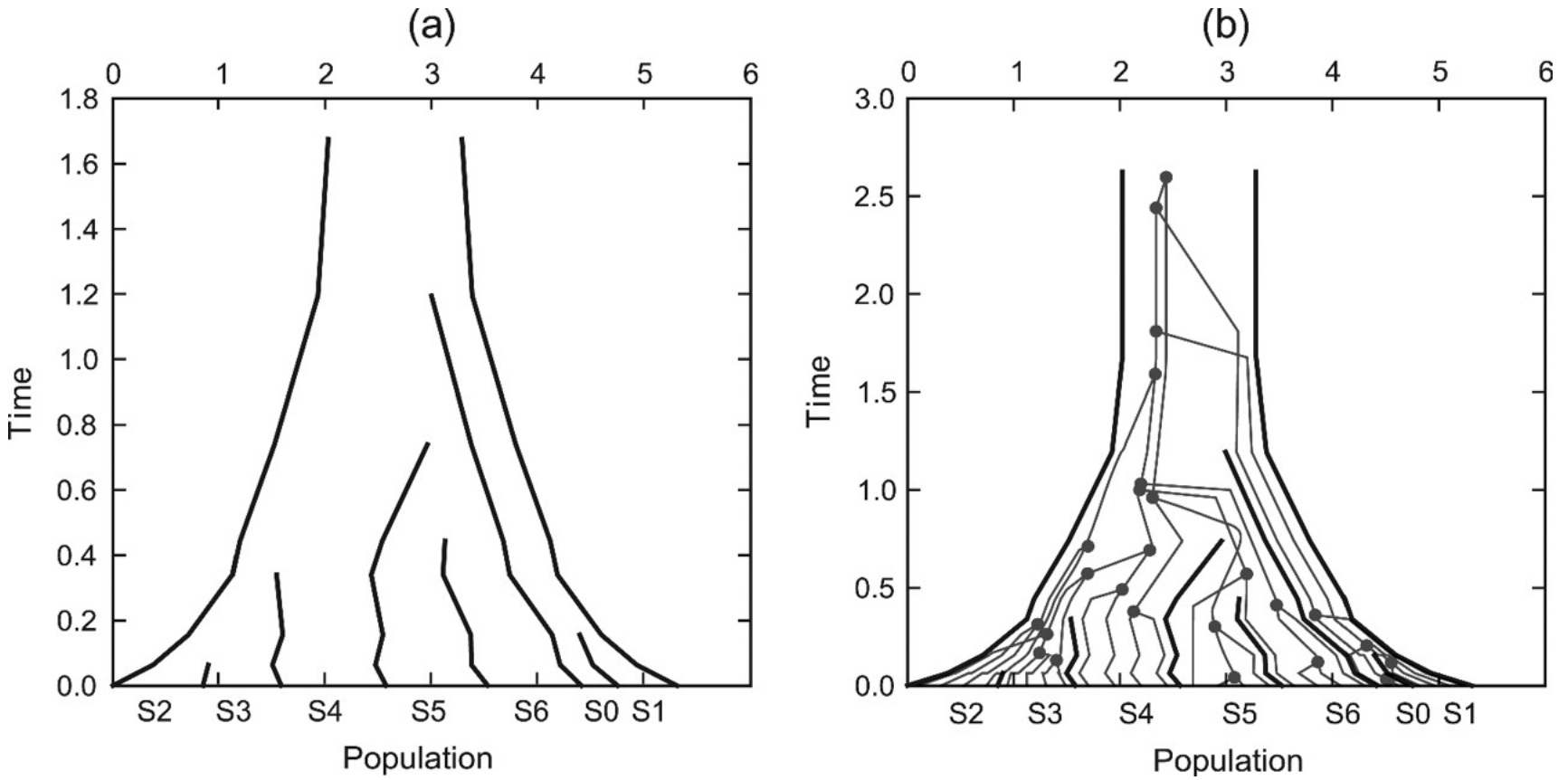


Figure 2 from Heled and Drummond (2010) *BEAST

See also the recent work by Huw Ogilvie and colleagues on StarBEAST2.

(3a) Considering coalescent effects without modeling gene trees

SVDQuartets
(Kubatko + Swofford
next Thursday)

PoMo model

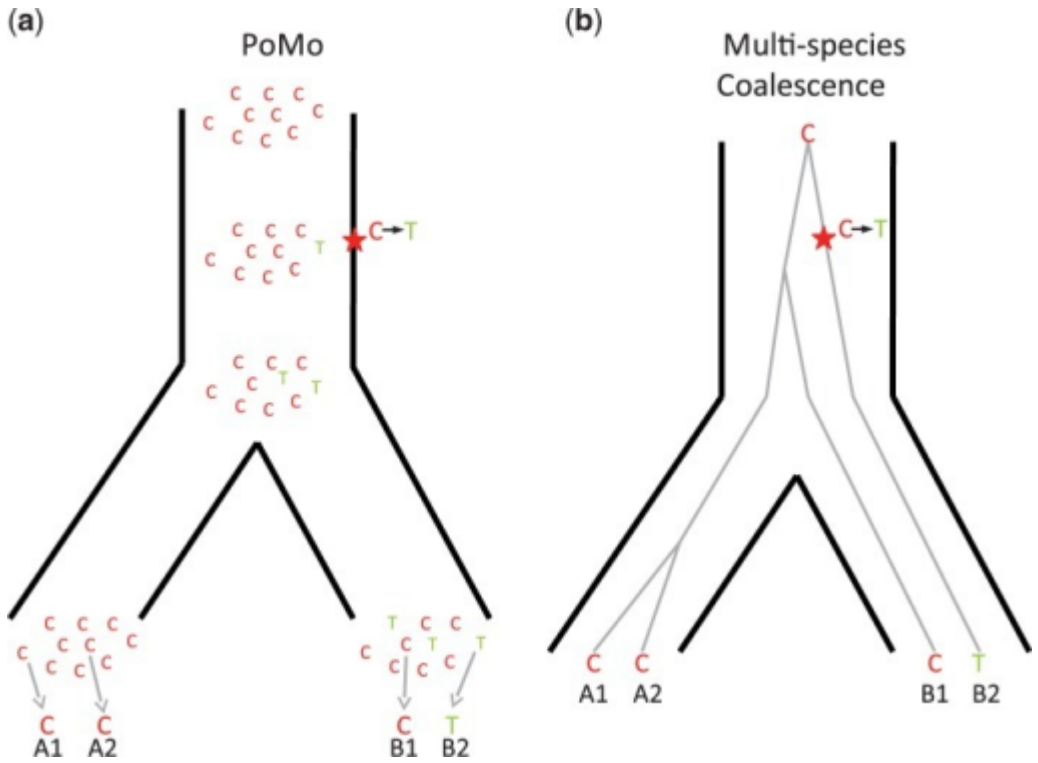
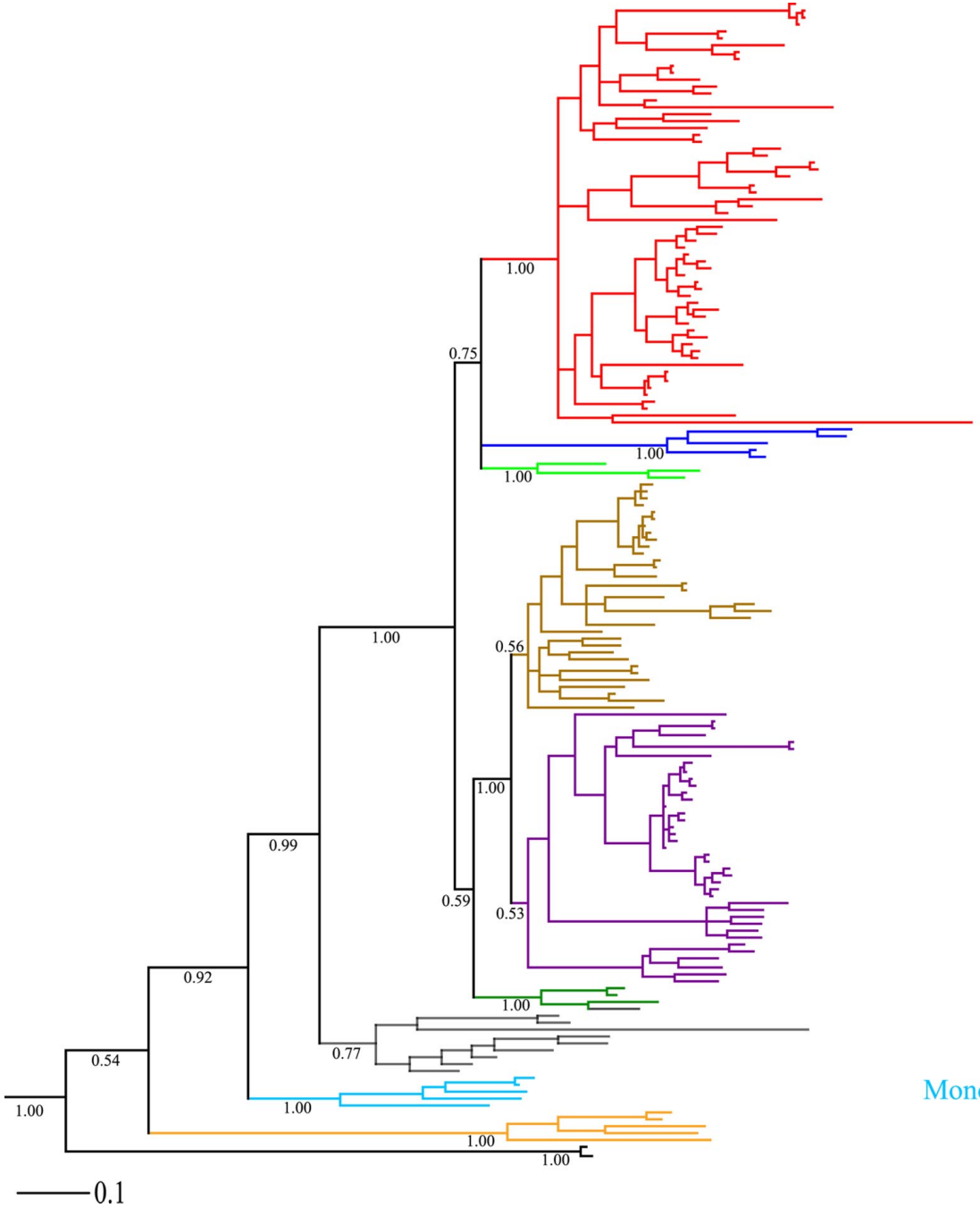


Figure 1 from De Maio et al. (2015)

(3a) Many types of tree: 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

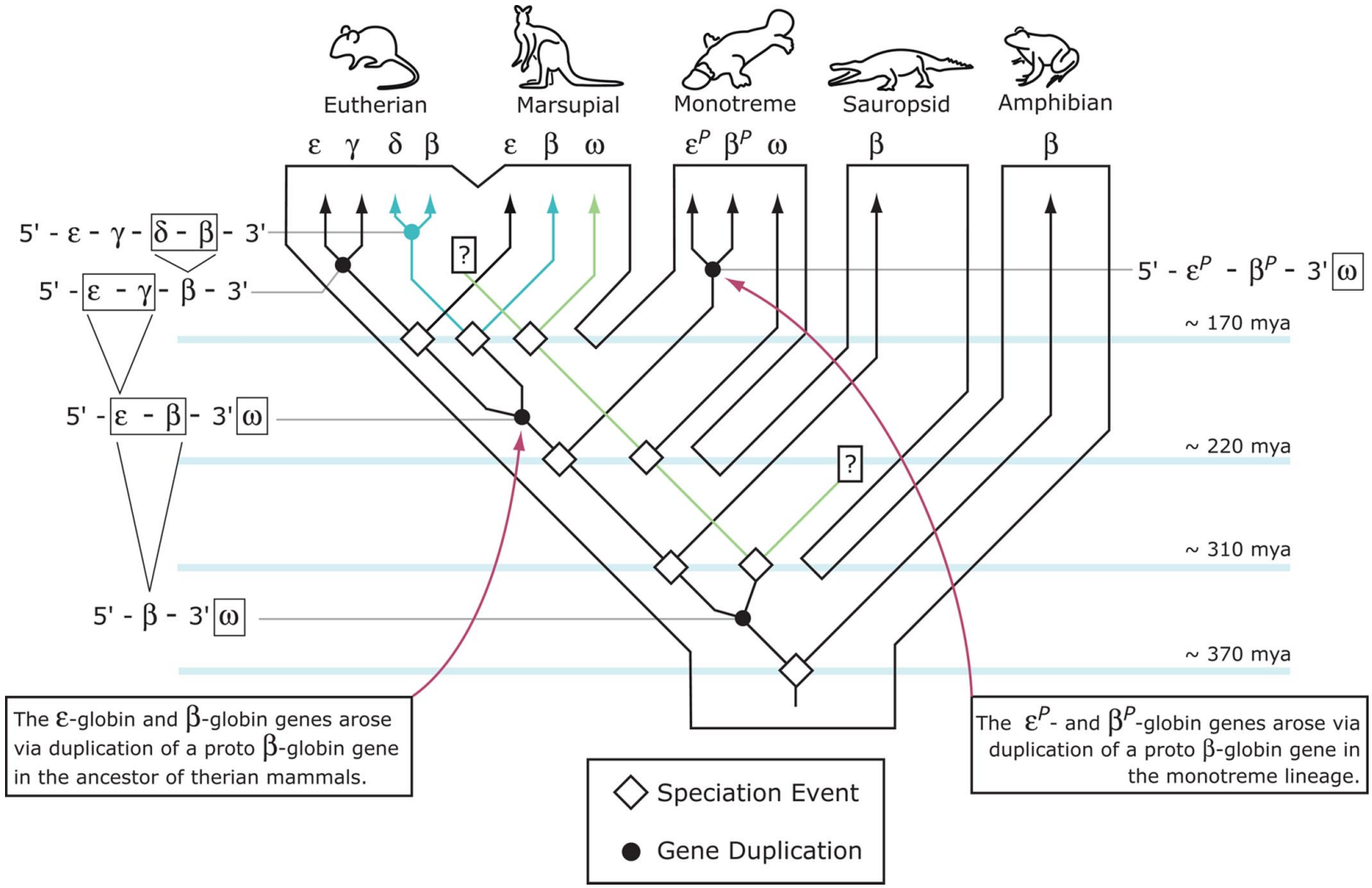
Sauropsida β-like globins

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.

Casey Dunn (today) and Laura Eme (next Tuesday)

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

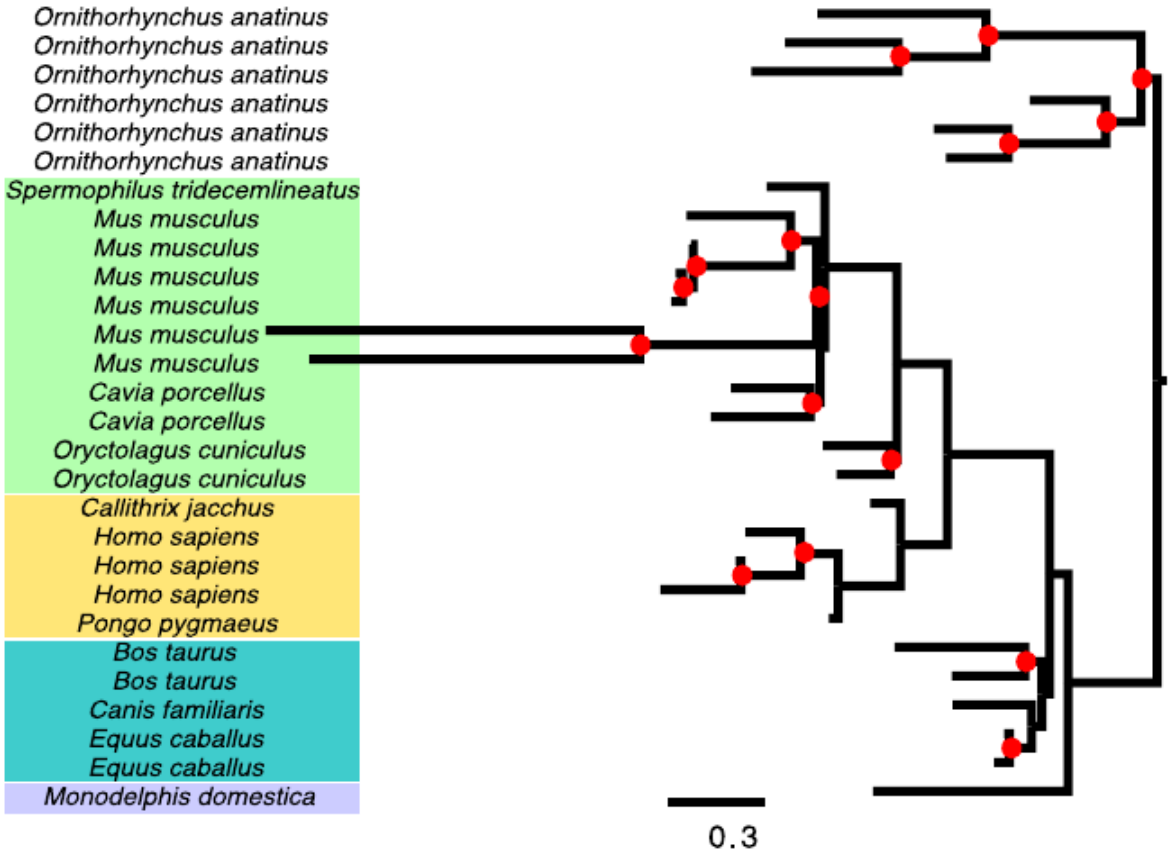


Figure 2A from Boussau et al. (2013)

(3a) Many types of trees:

	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

(3a) Joint estimation of gene duplication, loss, and coalescence with DLCoalRecon

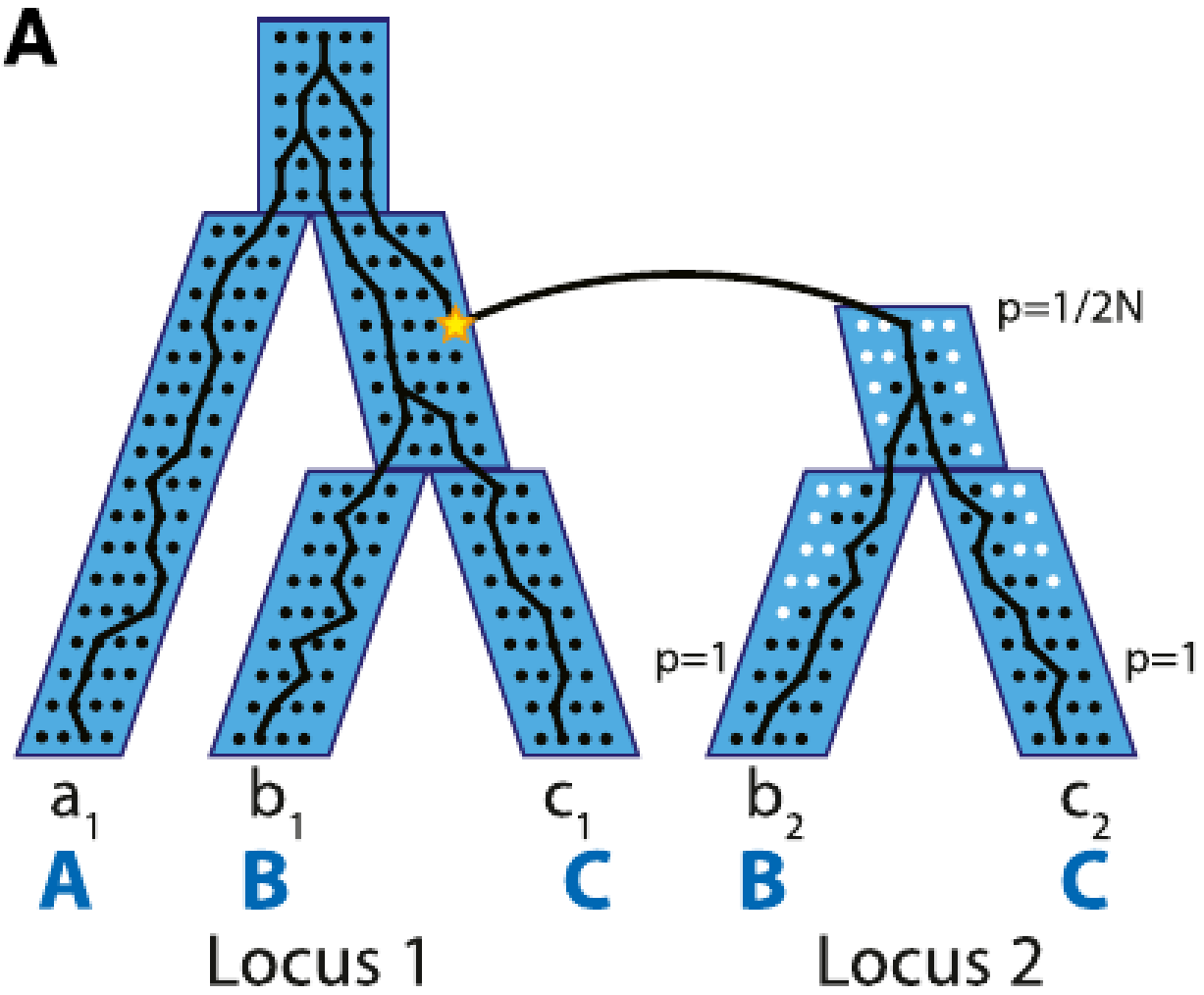


Figure 2A from Rasmussen and Kellis (2012)

(3a) DL models and coalescence

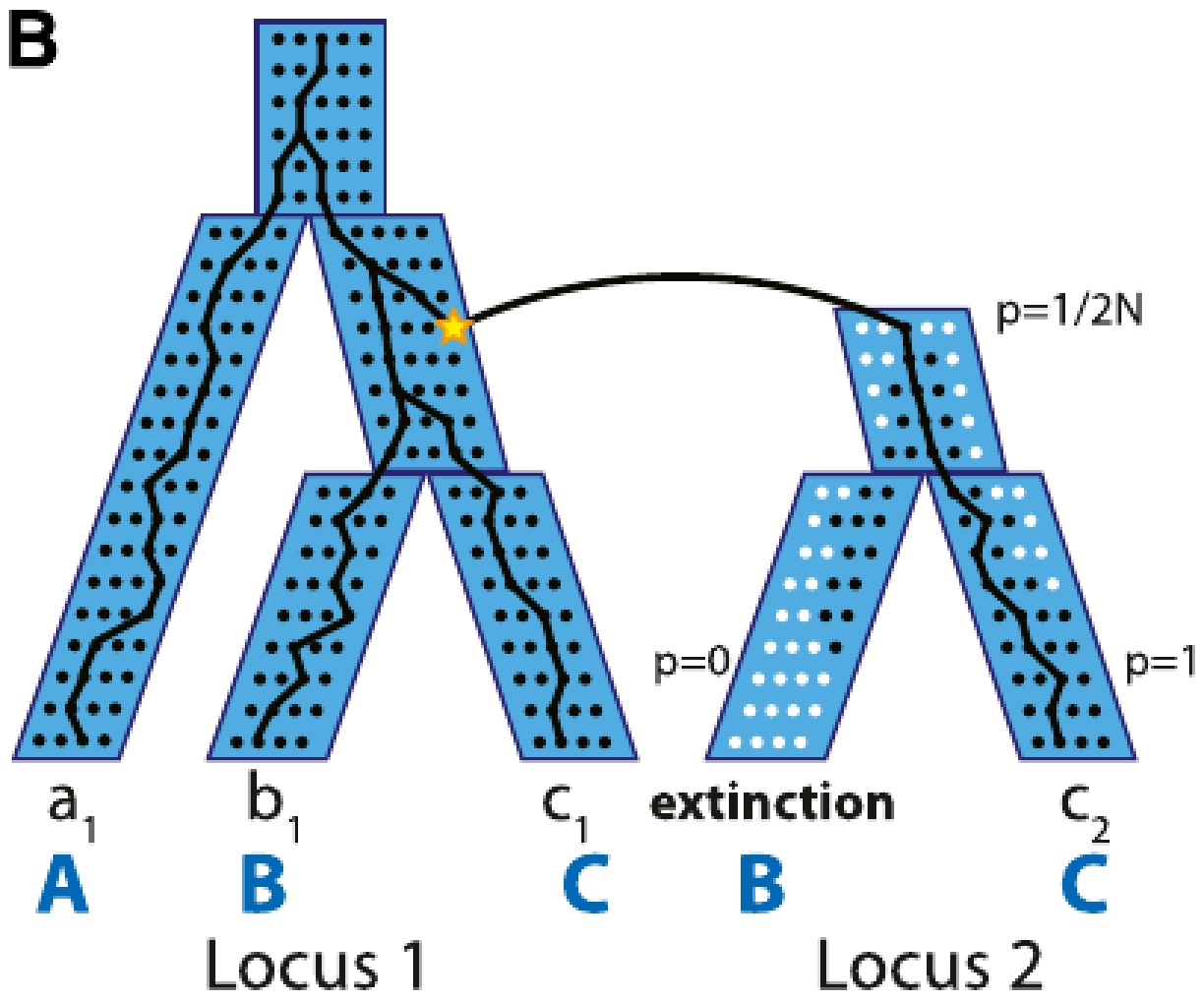


Figure 2B from Rasmussen and Kellis (2012)

(3a) Many types of trees: Lateral Gene Transfer

tree - a graph without cycles (loops)
network - general graph; cycles allowed

Cycles can represent

- lateral ("horizontal") gene transfer ,
- hybridization between species,
- introgression between populations.

Cécile Ané (next Friday)

(3a) Many types of trees: Lateral Gene Transfer

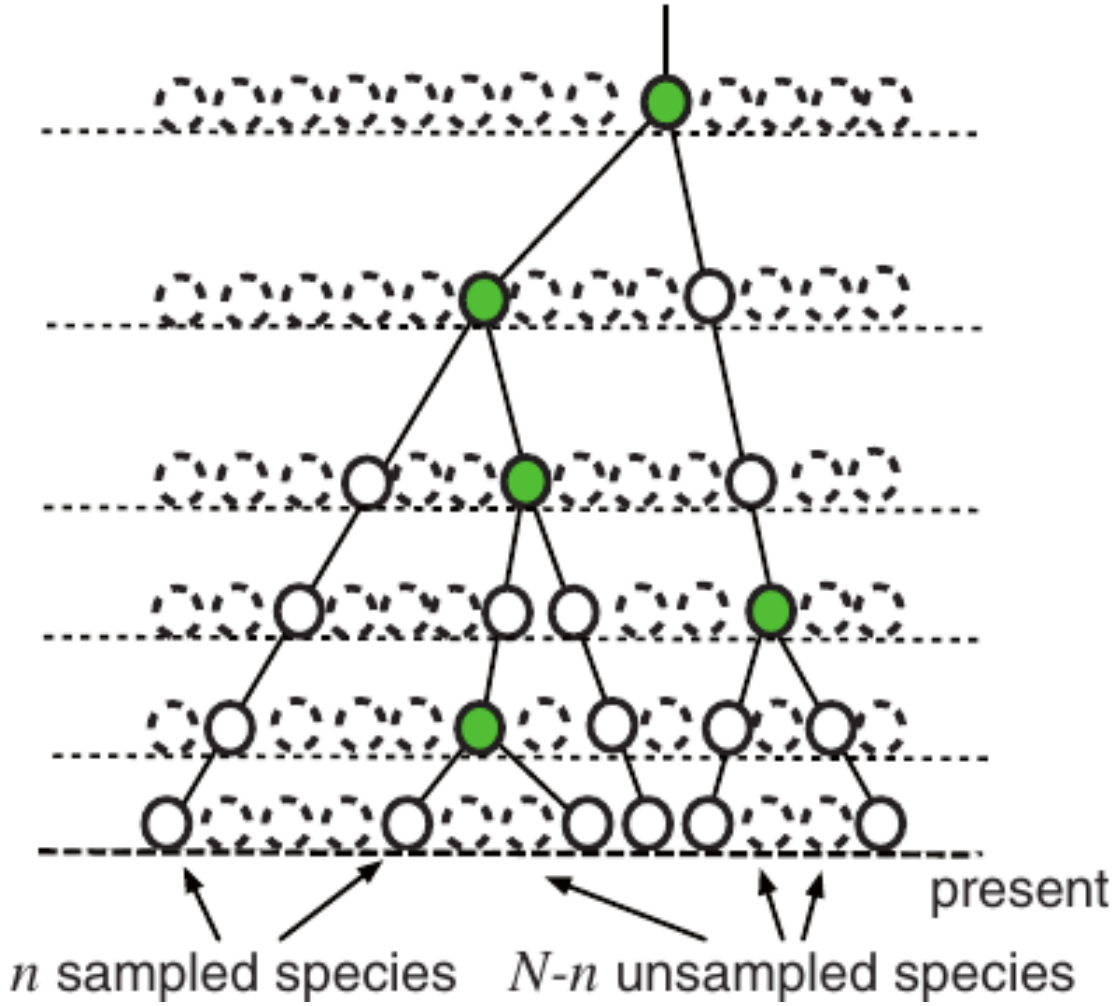


Figure 2c from Szöllősi et al. (2013)

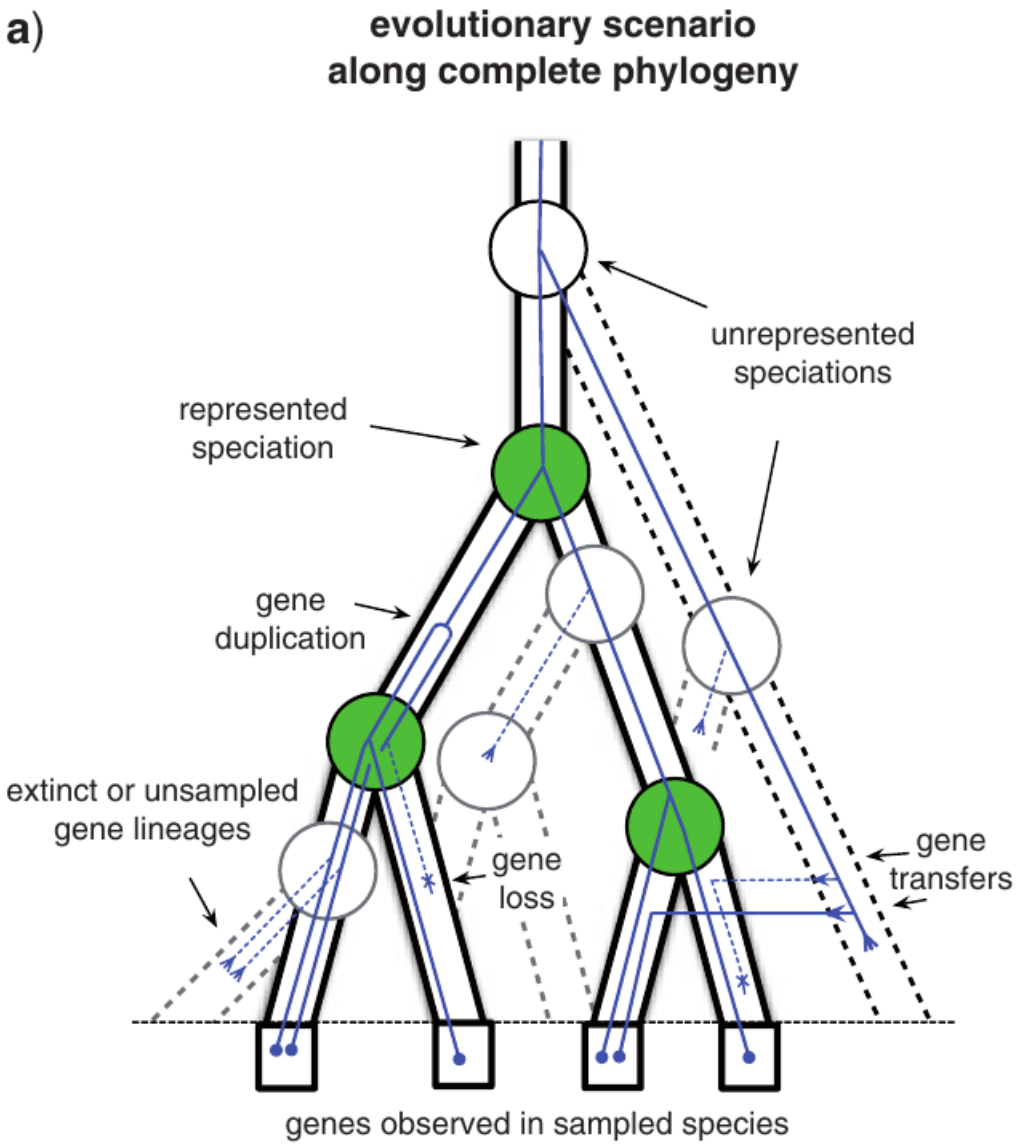


Figure 3 from Szöllősi et al. (2013)

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

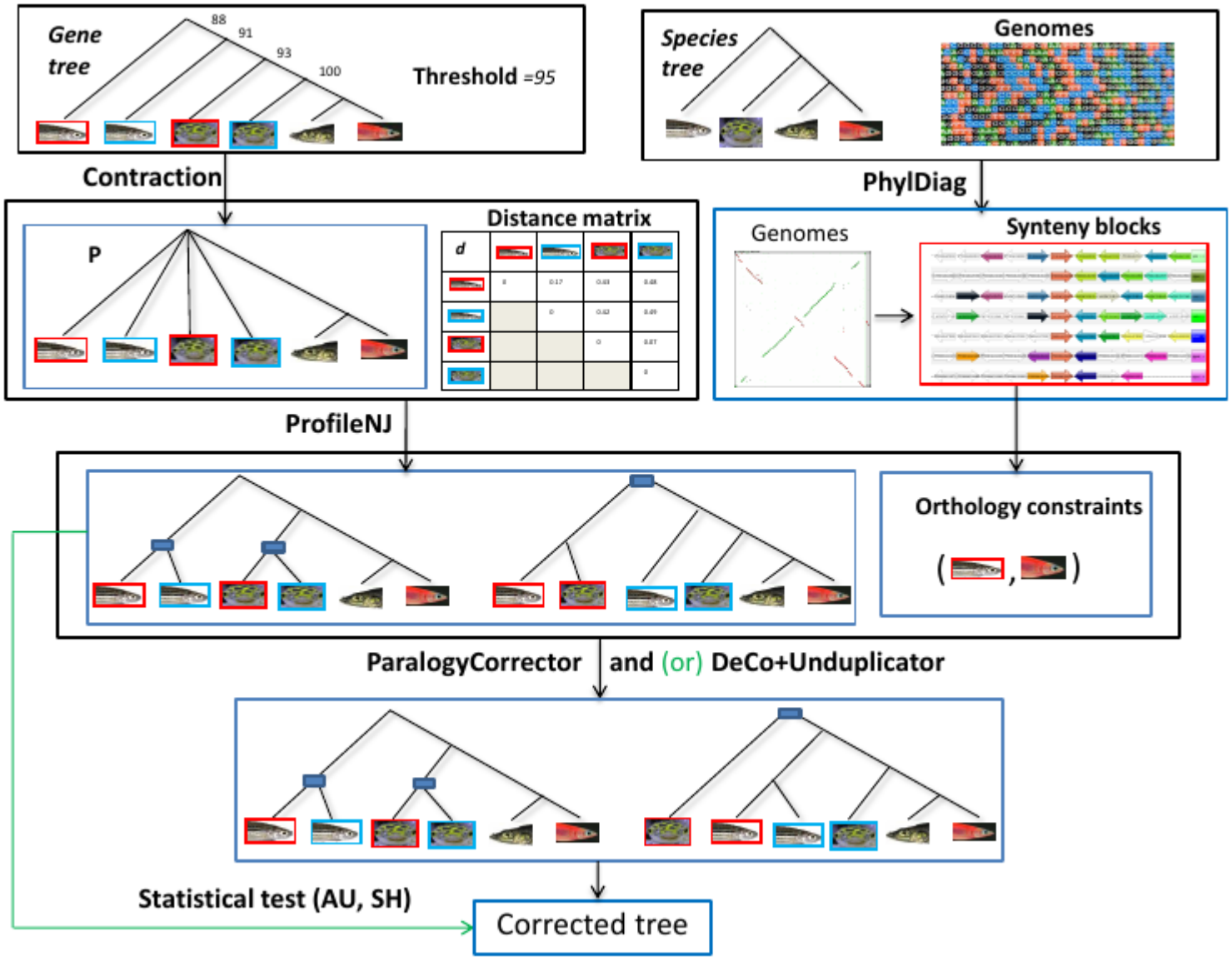


Figure 4 from Noutahi et al. (2016)

Part 3: Phylogenetics is difficult

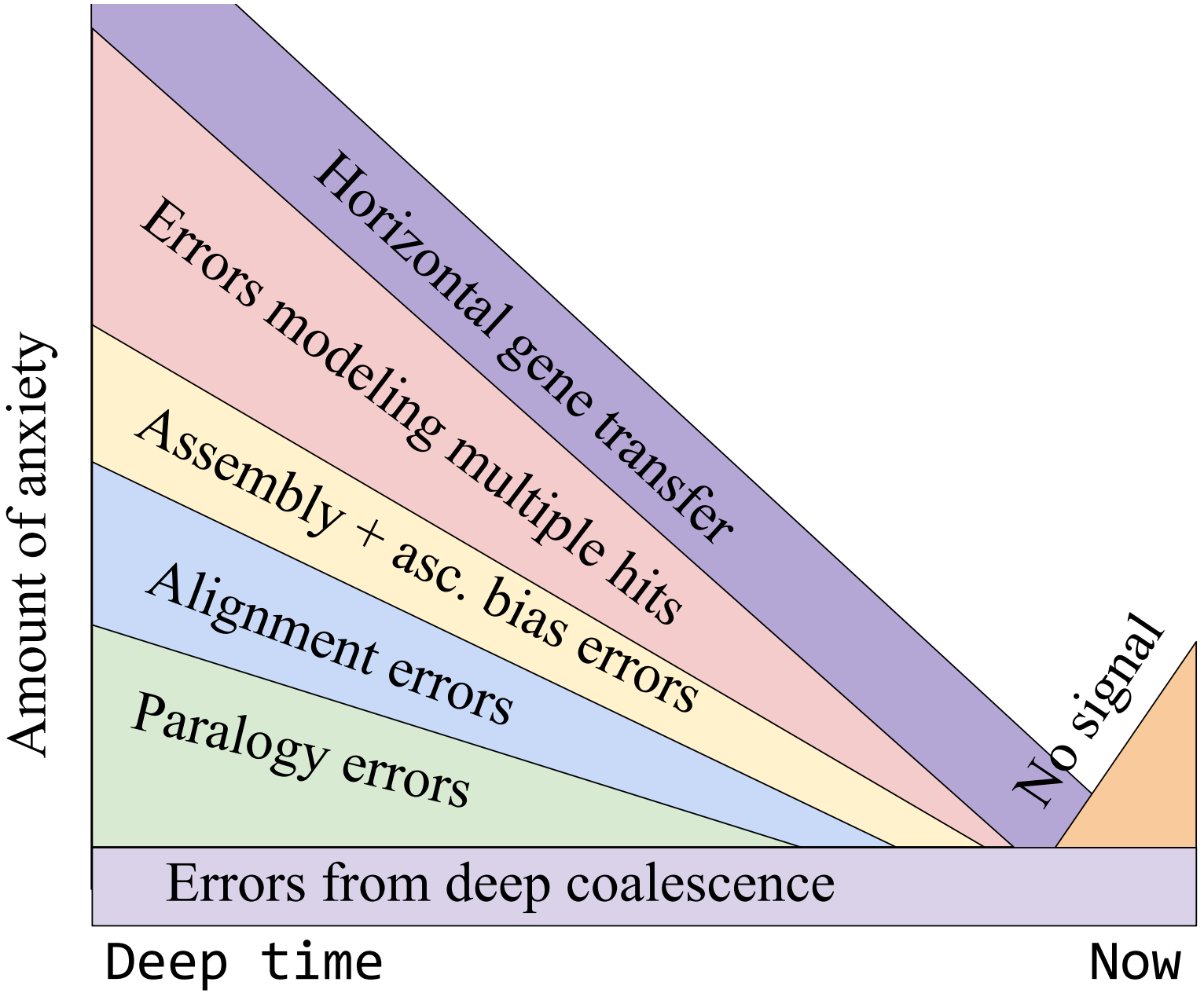
a. Many types of trees - species trees *vs* "gene trees" – coalescents or "gene family trees"

b. Many sources of error

c. No clean sampling theory that gives us clean hypothesis tests

d. Computational + statistical difficulties

(3b) sources of error cartoon



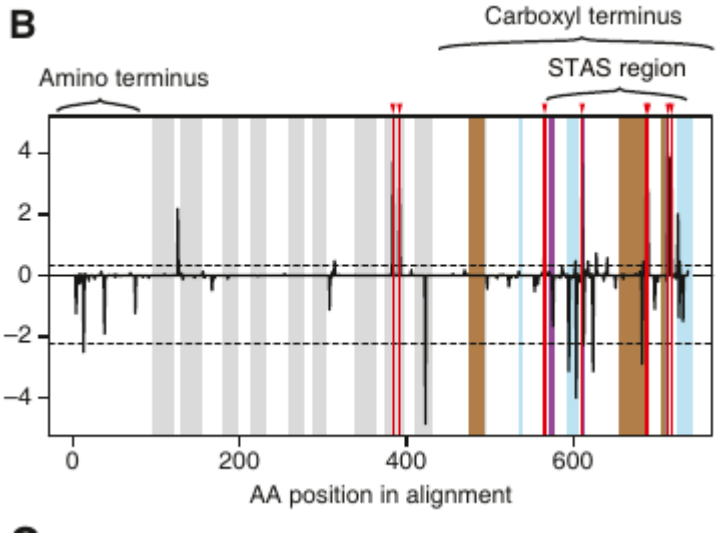


Figure 1 from Liu et al. (2010)

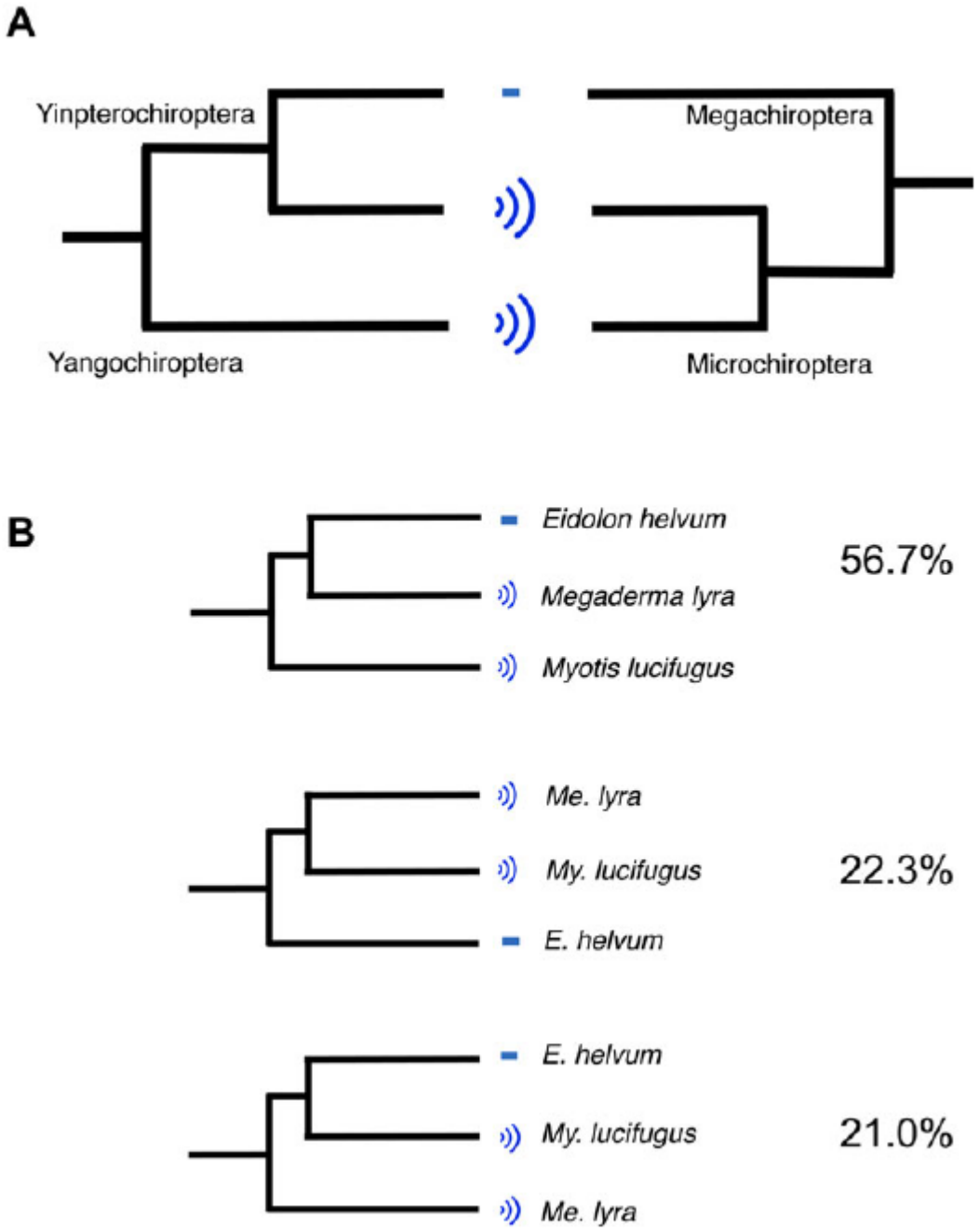
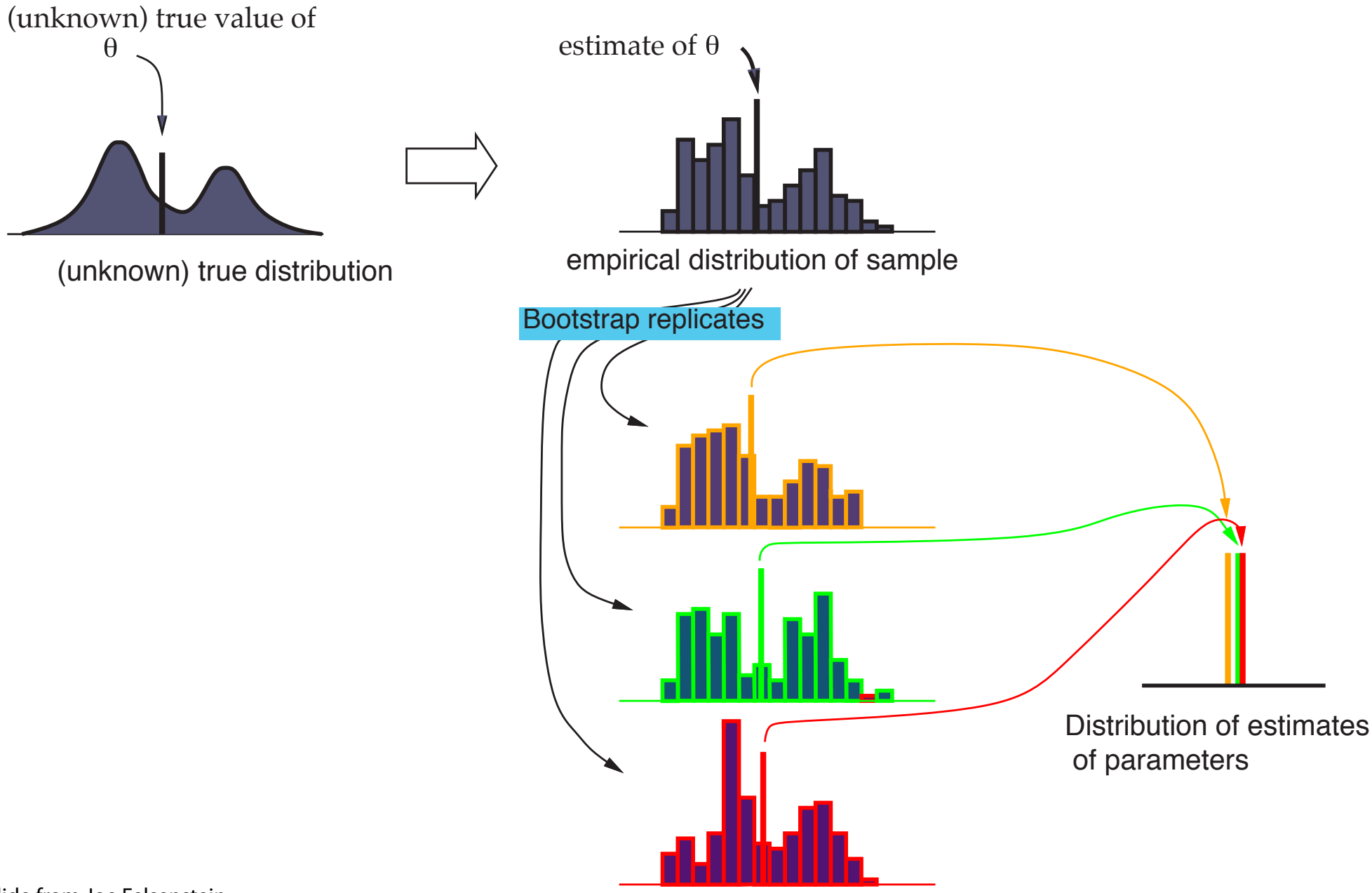


Figure 2 from Hahn and Nakhleh (2016)

(3c) Hypothesis testing in phylogenetics is tricky

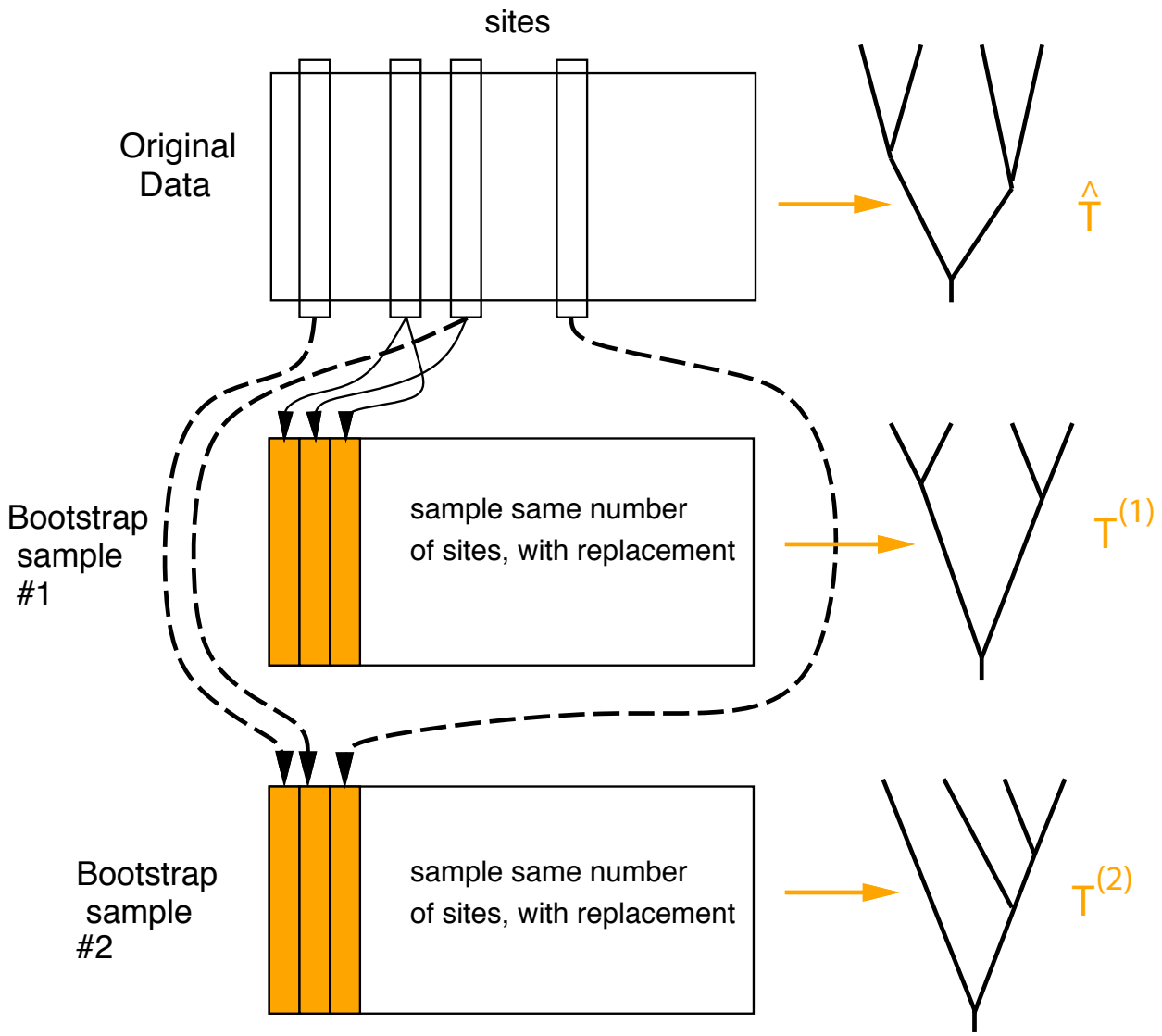
- complex literature on frequentists tests of topology (Holder last day)
- bootstrapping - examining effects of sampling error using resampling via computer
- Bayesian methods (Paul Lewis, John Huelsenbeck, Tracy Heath, and Michael Landis - this Sunday and the last Saturday)

The bootstrap



Slide from Joe Felsenstein

The bootstrap for phylogenies

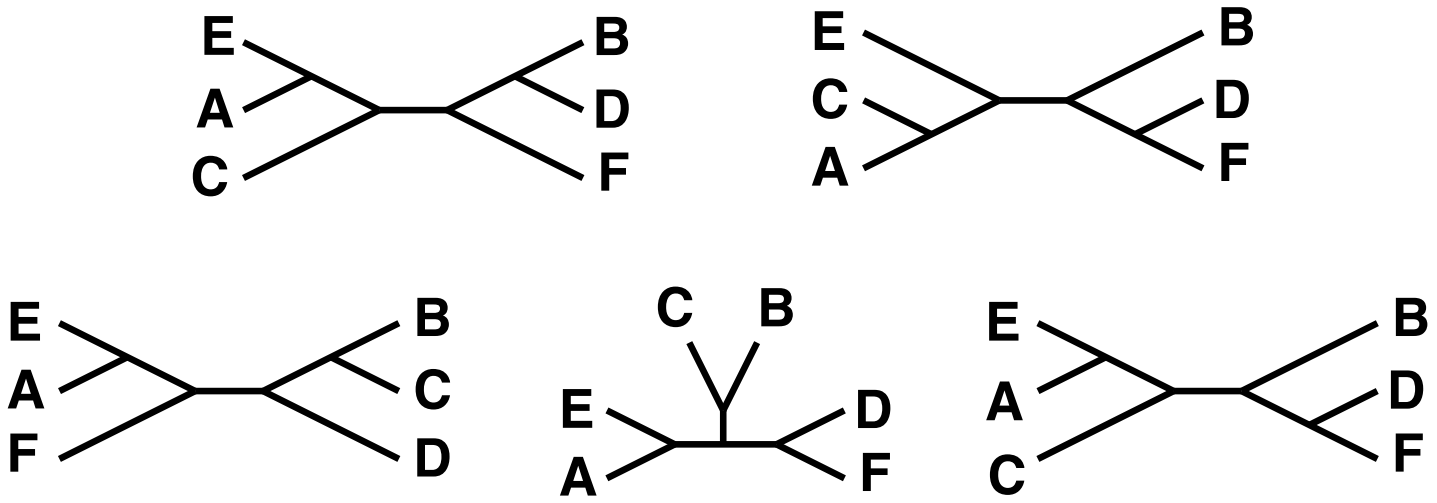


Slide from Joe Felsenstein

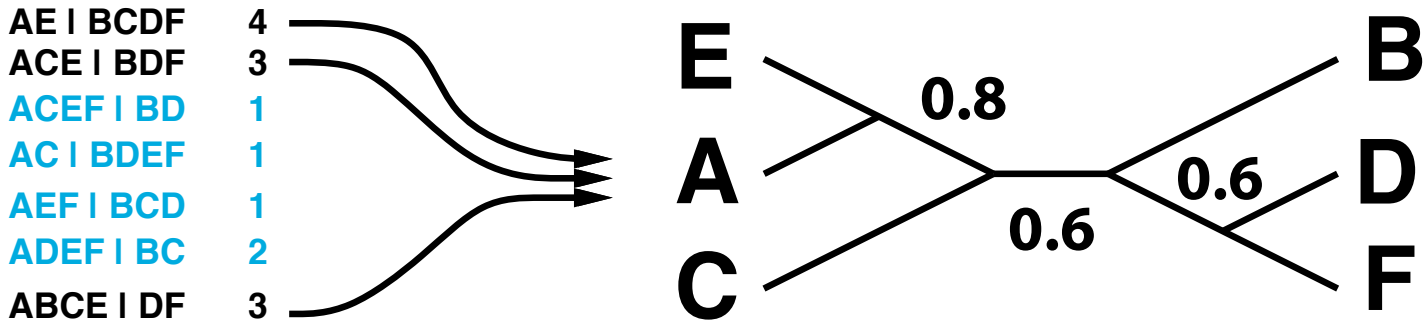
(and so on)

The majority-rule consensus tree

Trees:



How many times each partition of species is found:



Slide from Joe Felsenstein

(3c) bootstrapping

- http://phylo.bio.ku.edu/mephytis/boot-sample.html
- http://phylo.bio.ku.edu/mephytis/bootstrap.html

(3d) Phylogenetics is computationally difficult

Problems:

- Huge number of trees
- Strange geometry of tree space
- Large number of numerical parameters that need to be considered.

Some strategies:

- Pragmatic computational heuristics for tree searching – Emily Jane McTavish (tomorrow) and Bui Quang Minh (Tuesday)
- Markov chain Monte Carlo (Paul Lewis, John Huelsenbeck, Tracy Heath, and Michael Landis - this Sunday and the last Saturday)

Optimality criteria

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

Examples:

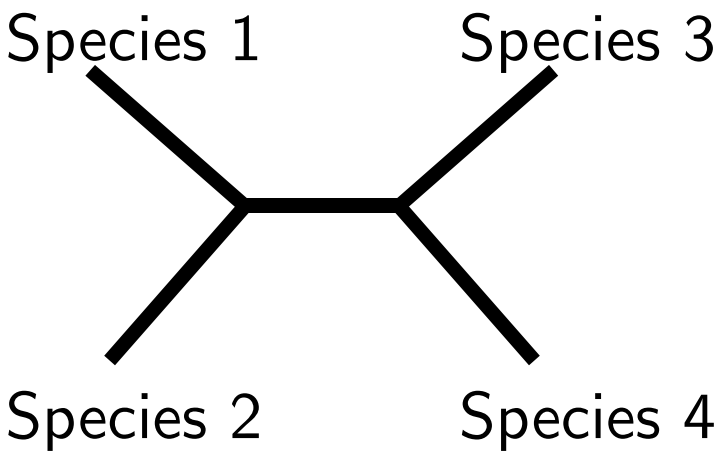
- Parsimony (Maximum Parsimony, MP)
- Maximum Likelihood (ML)
- Minimum Evolution (ME)
- Least Squares (LS)

	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	.	.	.

next few slides from Paul Lewis

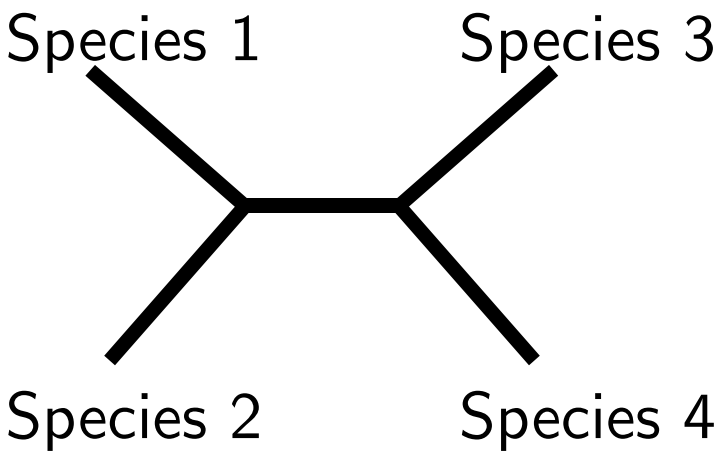
	1	2	3	4	5	6	7	8	9	.	.	.
Species 1	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 2	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 3	C	G	G	T	C	**C**	G	G	T	.	.	.
Species 4	C	G	G	C	C	**T**	G	G	T	.	.	.

One of the 3 possible trees:

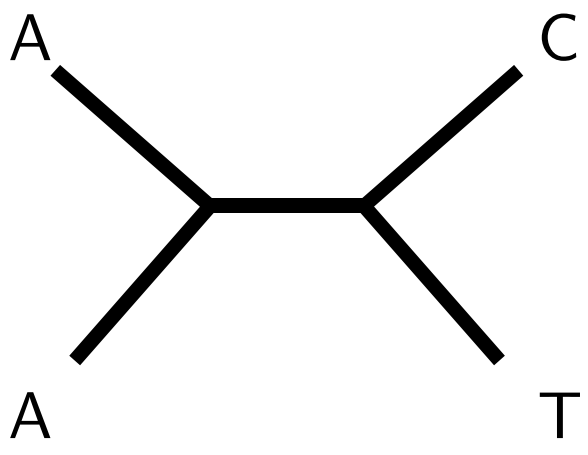


	1	2	3	4	5	6	7	8	9	.	.	.
Species 1	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 2	C	G	A	C	C	**A**	G	G	T	.	.	.
Species 3	C	G	G	T	C	**C**	G	G	T	.	.	.
Species 4	C	G	G	C	C	**T**	G	G	T	.	.	.

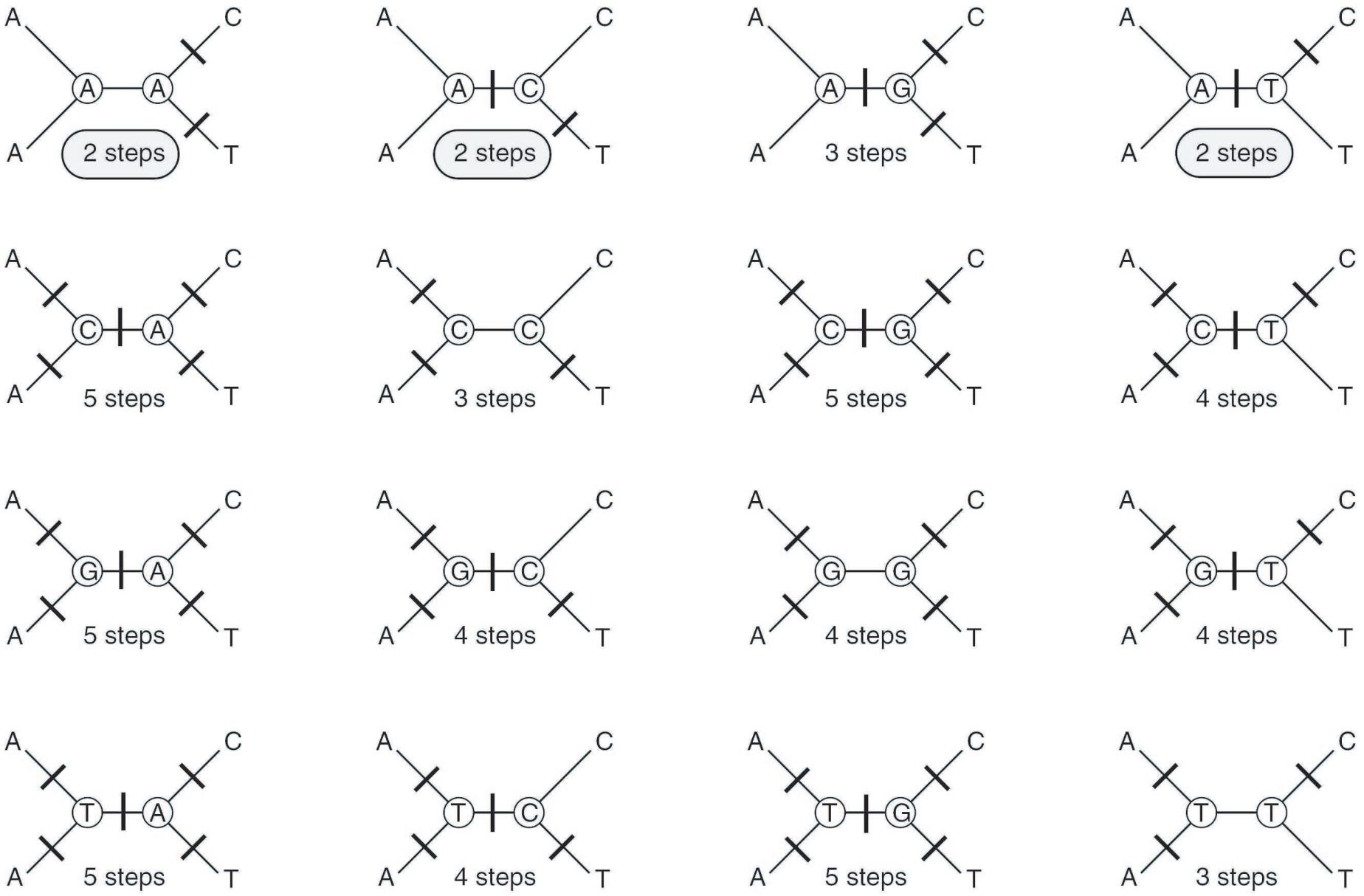
One of the 3 possible trees:



Same tree with states at character 6 instead of species names



Unordered Parsimony



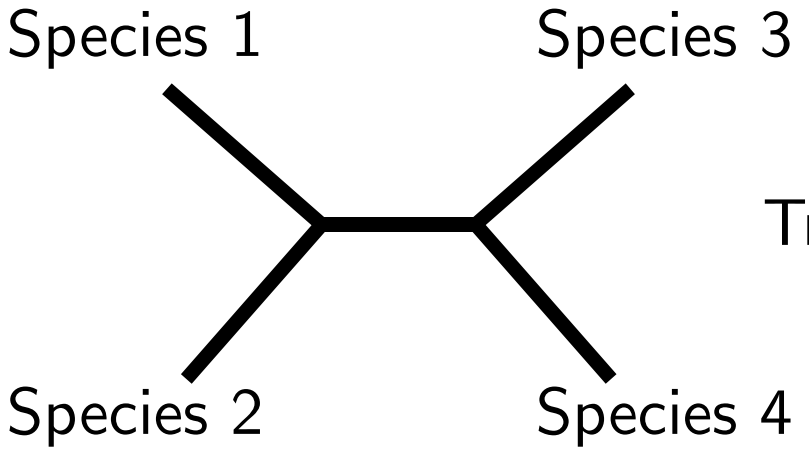
Things to note about the last slide

- 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

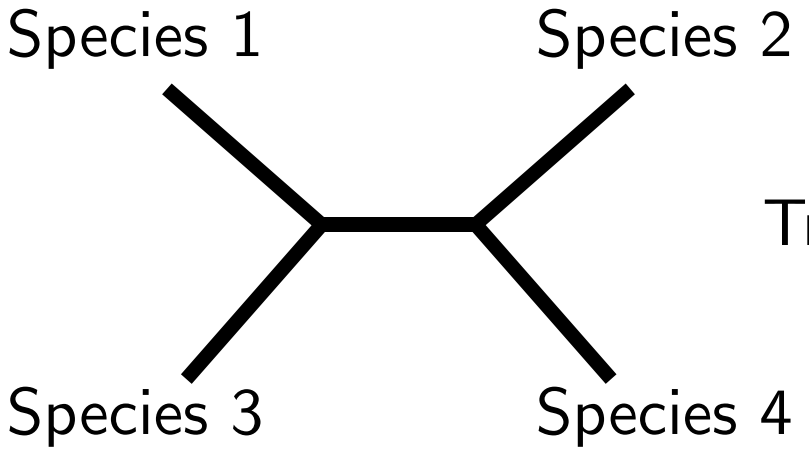


Tree 1 has a score of **4**

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

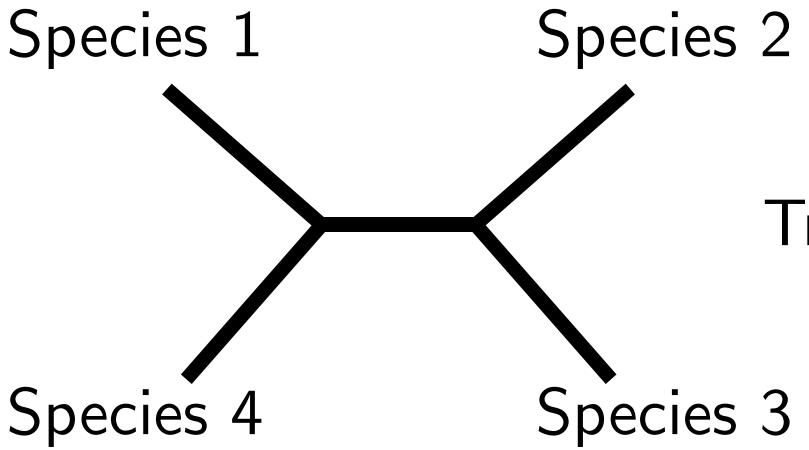


Tree 2 has a score of **5**

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.

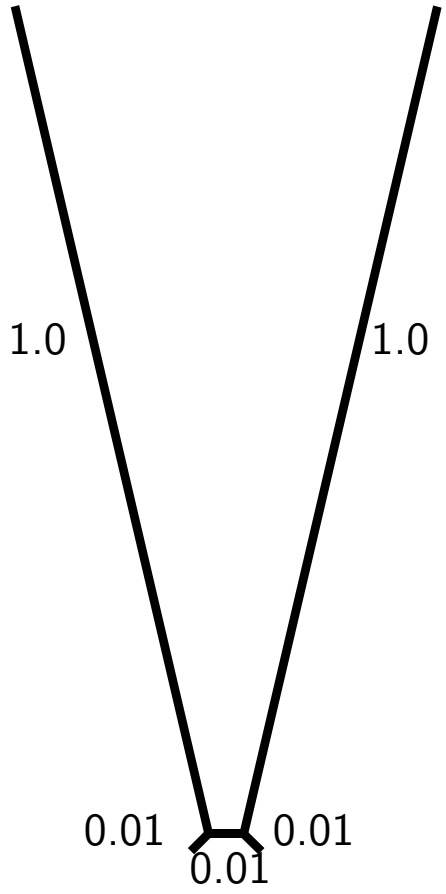
In the example matrix at the beginning of these slides, only character 3 is parsimony informative.

	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

Qualitative description of parsimony

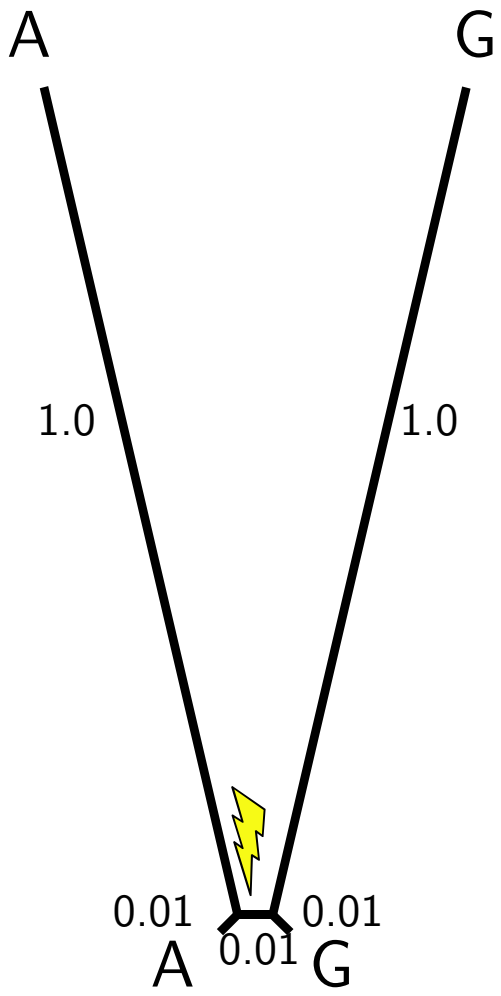
- Enables estimation of ancestral sequences.
- Even though parsimony always seeks to minimizes the number of changes, it can perform well even when changes are not rare.
- Does not "prefer" to put changes on one branch over another
- Hard to characterize statistically
  - the set of conditions in which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

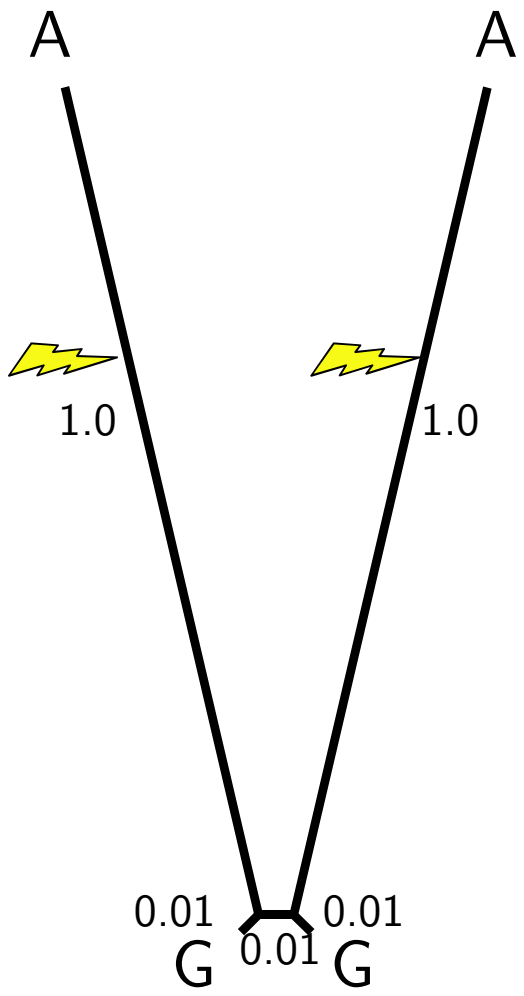
Long branch attraction



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



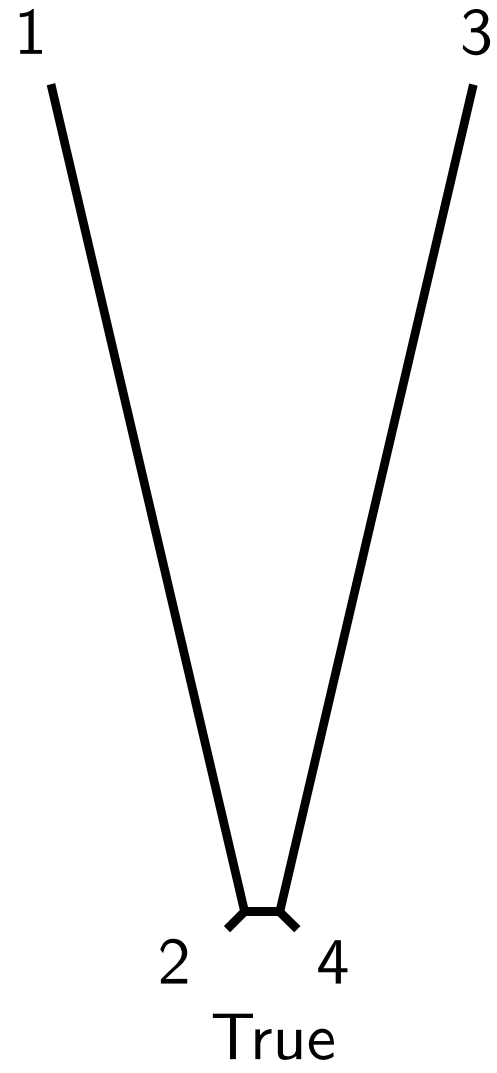
Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

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

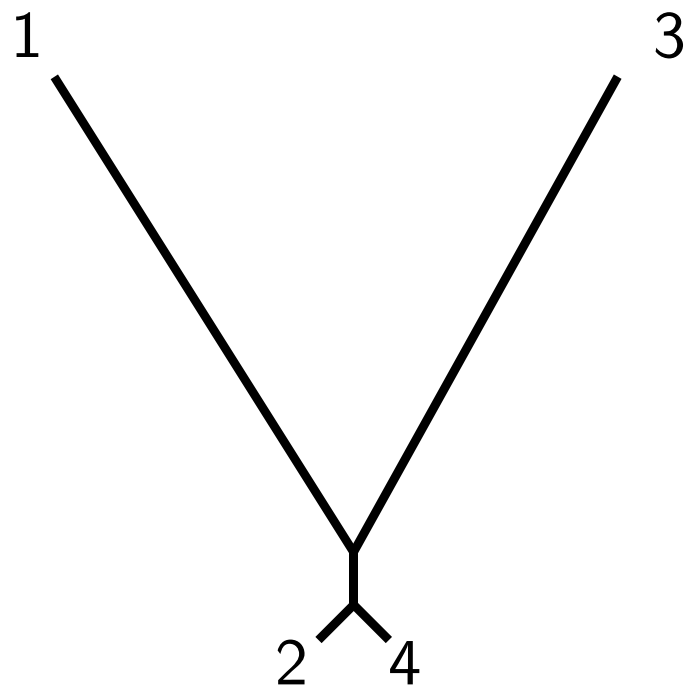
The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).

Long branch attraction

Parsimony is almost guaranteed to get this tree wrong.



True



Inferred

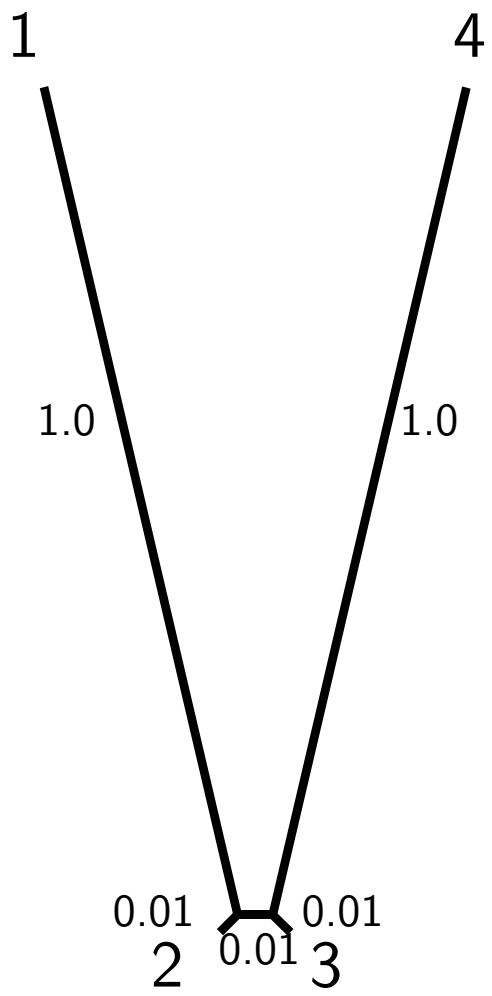
Inconsistency

- 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.

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:

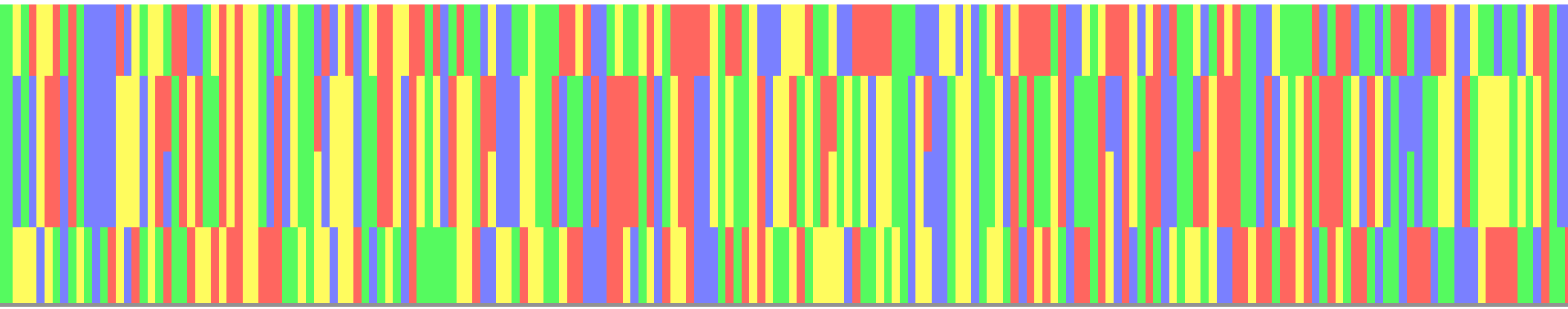
- start at a node,
- if the next symbol is '(' then add a child to the current node and move to this child,
- if the next symbol is a label, then label the node that you are at,
- if the next symbol is a comma, then move back to the current node's parent and add another child,
- if the next symbol is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

If the data is generated such that:

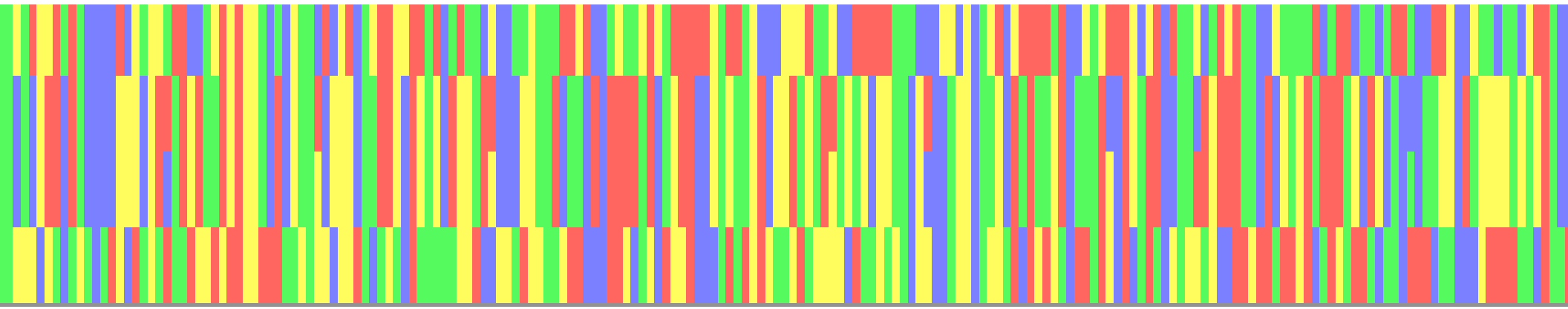
$$\Pr\begin{pmatrix} A \\ A \\ G \\ G \end{pmatrix} \approx 0.0003 \quad \text{and} \quad \Pr\begin{pmatrix} A \\ G \\ G \\ A \end{pmatrix} \approx 0.008$$

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...

Why doesn't simple clustering work?

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

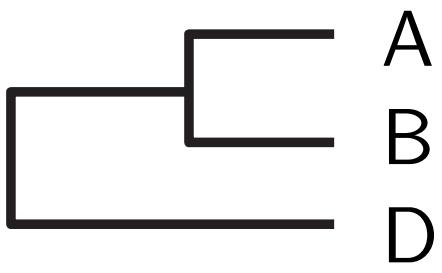
	A	B	C	D
A	-	0.2	0.5	0.4
B		-	0.46	0.4
C			-	0.7
D				-

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				-

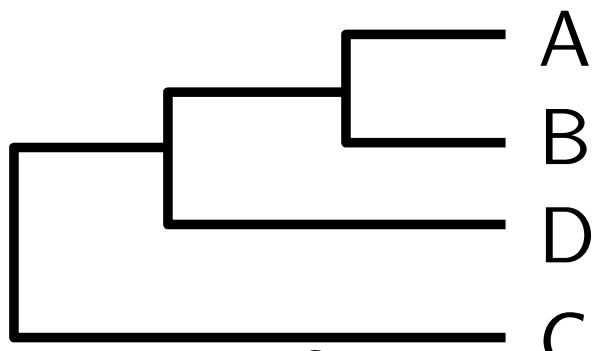
Why doesn't simple clustering work?

	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?

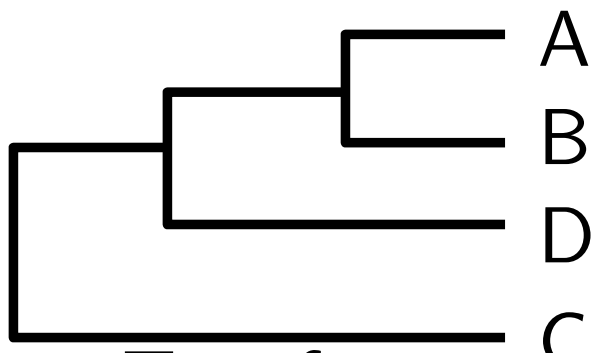
	A	B	C	D
A	-	0.2	0.5	**0.4**
B		-	0.46	**0.4**
C		-	-	**0.7**
D				0



Tree from clustering

Why doesn't simple clustering work?

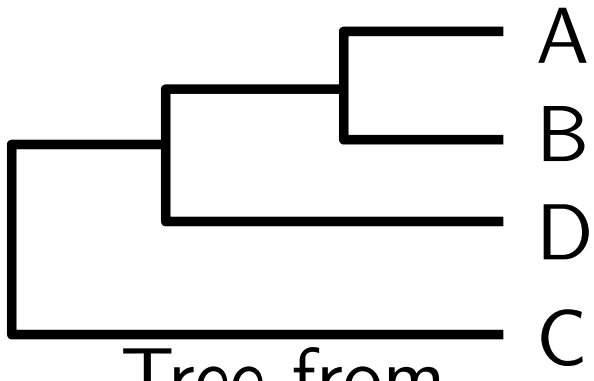
	A	B	C	D
A	0	0.2	**0.5**	0.4
B	0.2	0.2	**0.46**	0.4
C	**0.5**	**0.46**	0	**0.7**
D	0.4	0.4	**0.7**	0



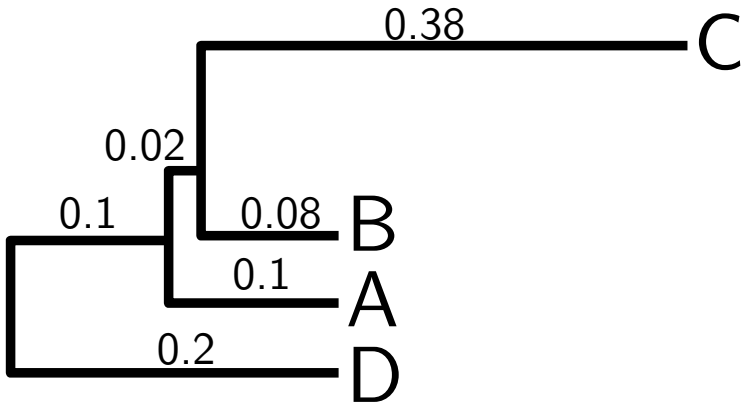
Tree from clustering

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

Simple clustering methods are sensitive to...

1. differences in the rate of sequence evolution.

2. The "multiple hits" problem. – some sites are affected by more than 1 mutation

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

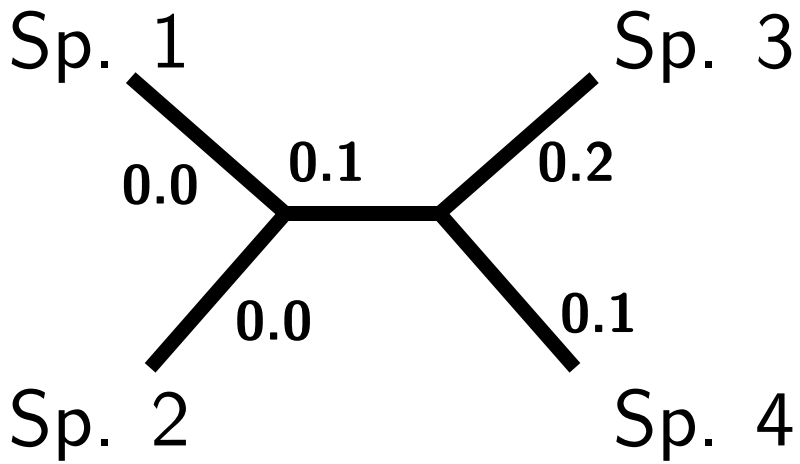
. . . so we can just use the lower triangle.

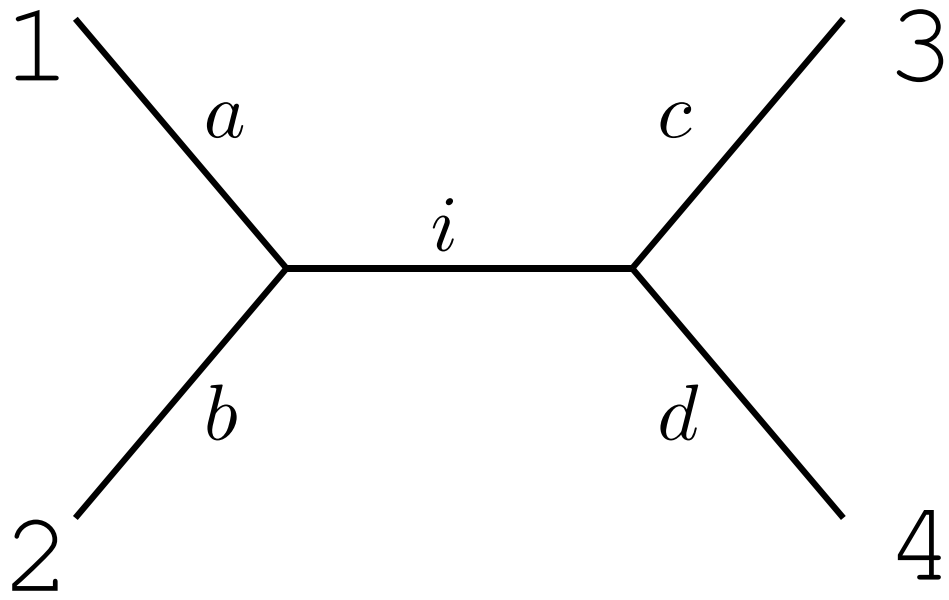
	Species 1	Species 2	Species 3
Species 2	0		
Species 3	0.3	0.3	
Species 4	0.2	0.2	0.3

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

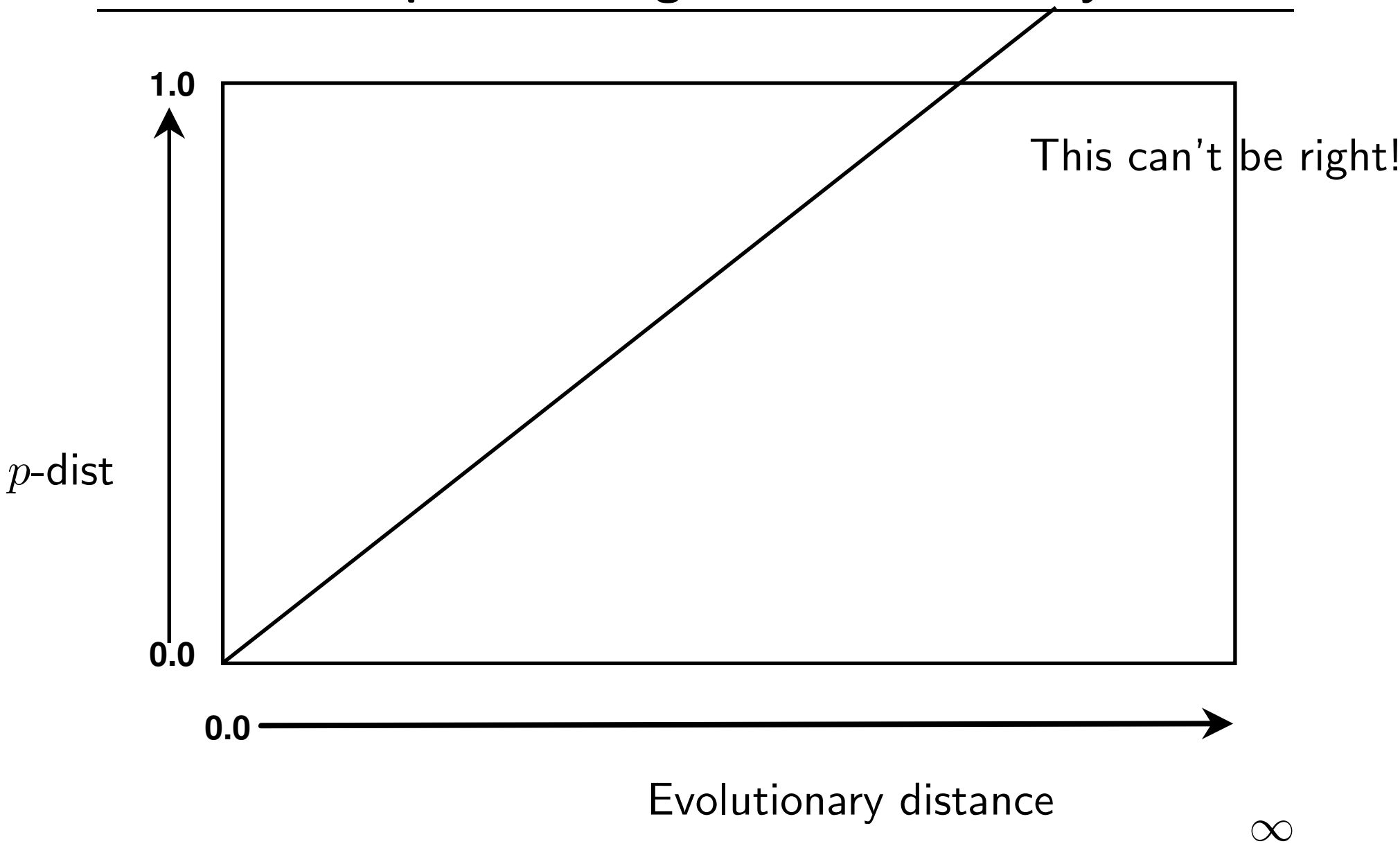
	1	2	3
2	d_{12}		
3	d_{13}	d_{23}	
4	d_{14}	d_{24}	d_{34}

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.

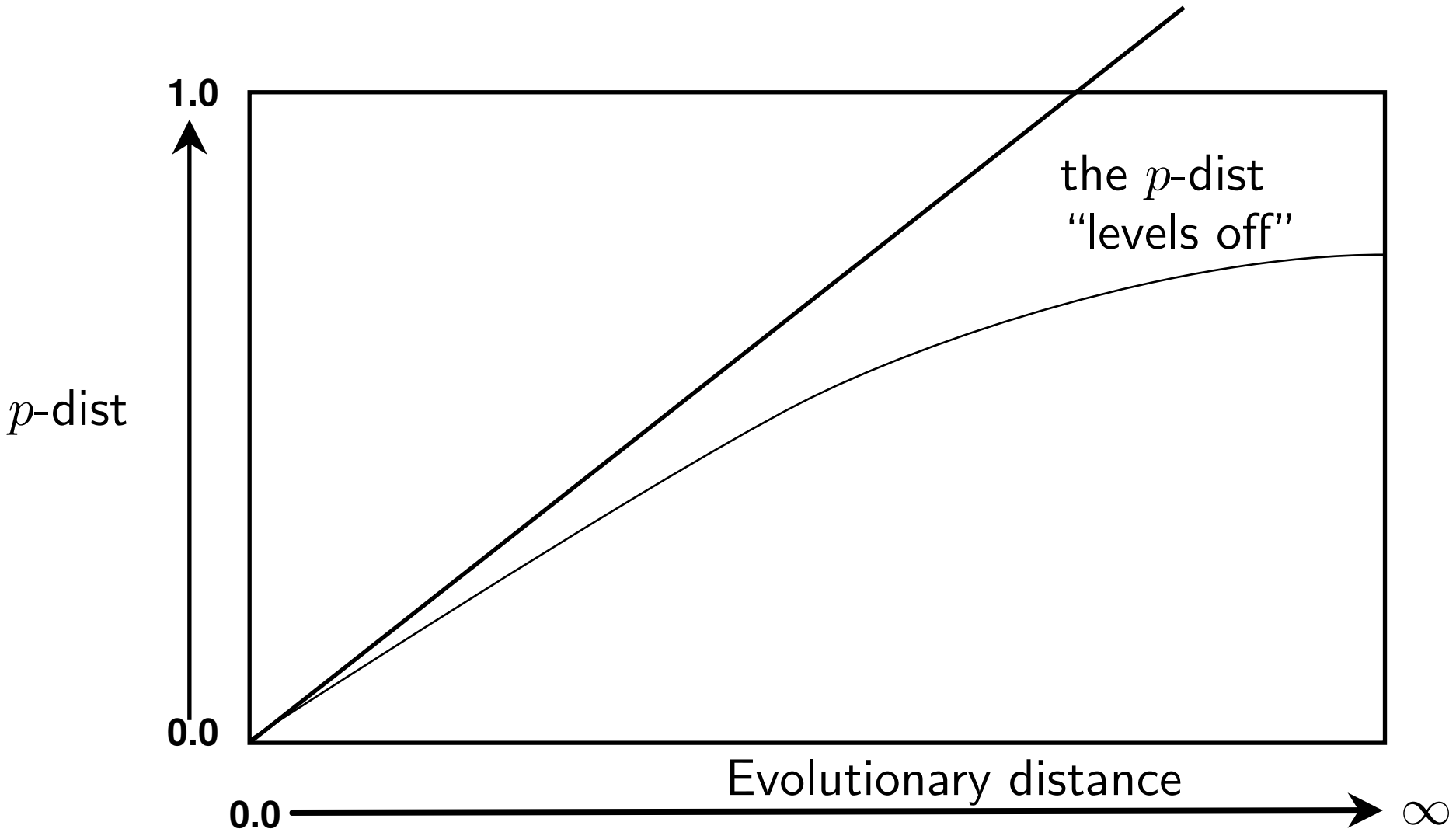
We hope the distances that we measure can produce good estimates of the evolutionary distance, but we know that they cannot be perfect.

Intuition of sequence divergence vs evolutionary distance



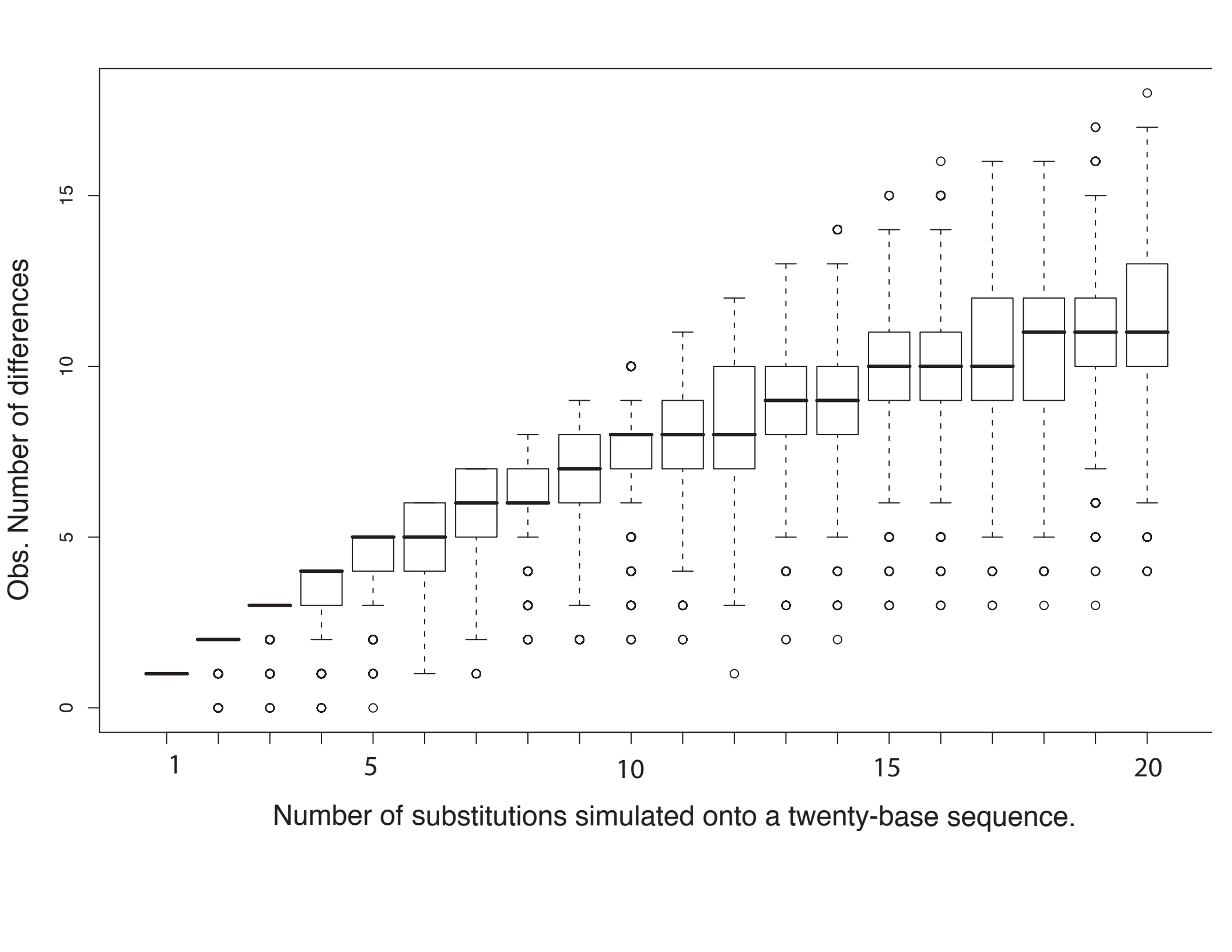
slide from Paul Lewis

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 Paul Lewis (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

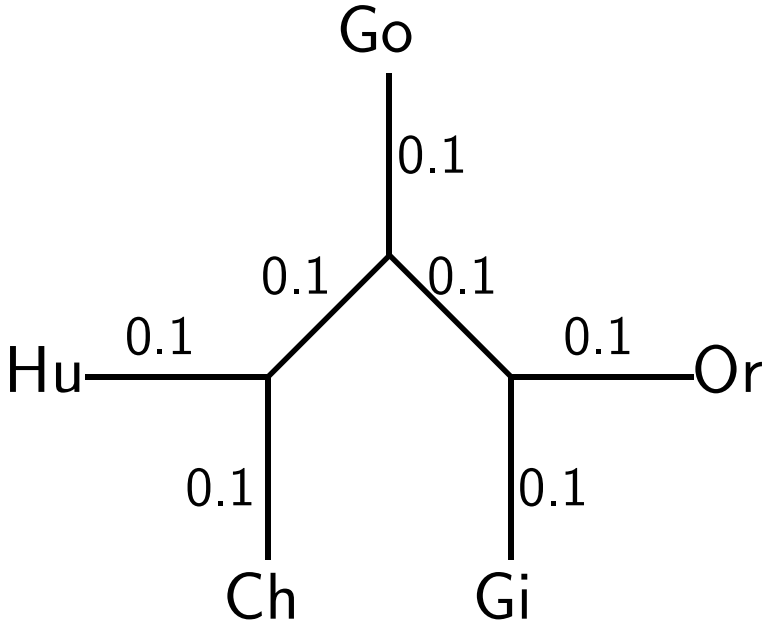
Least Squares Branch Lengths

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

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

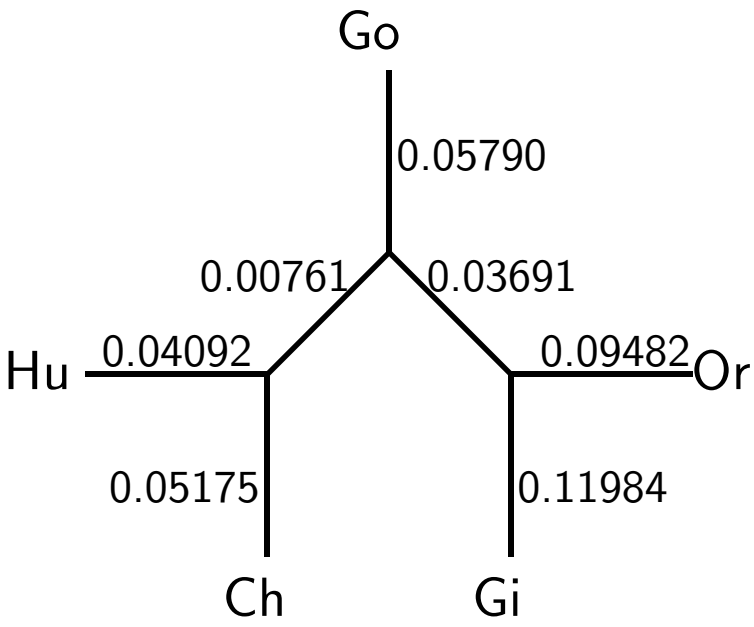
Poor fit using arbitrary branch lengths

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**

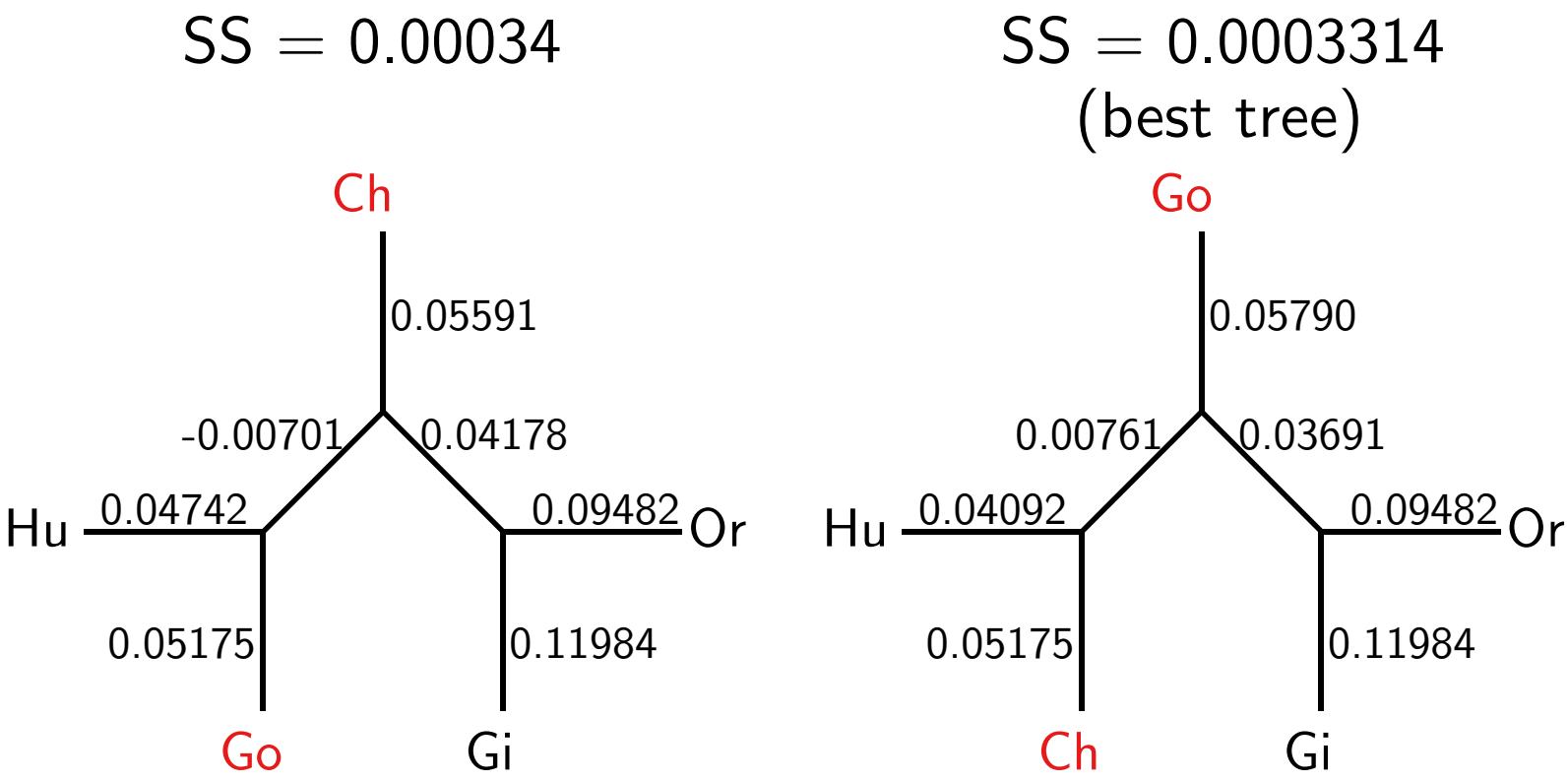


Optimizing branch lengths yields the least-squares score

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**

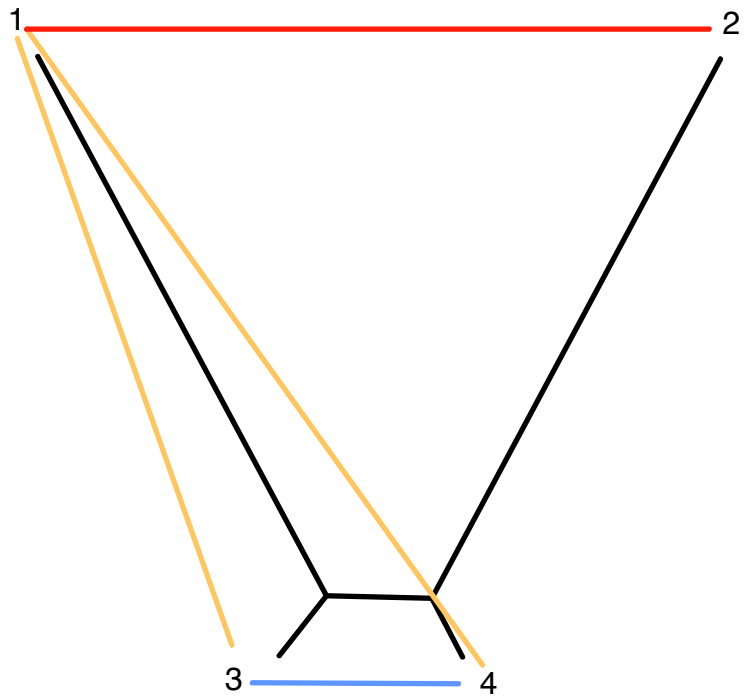


Least squares as an optimality criterion



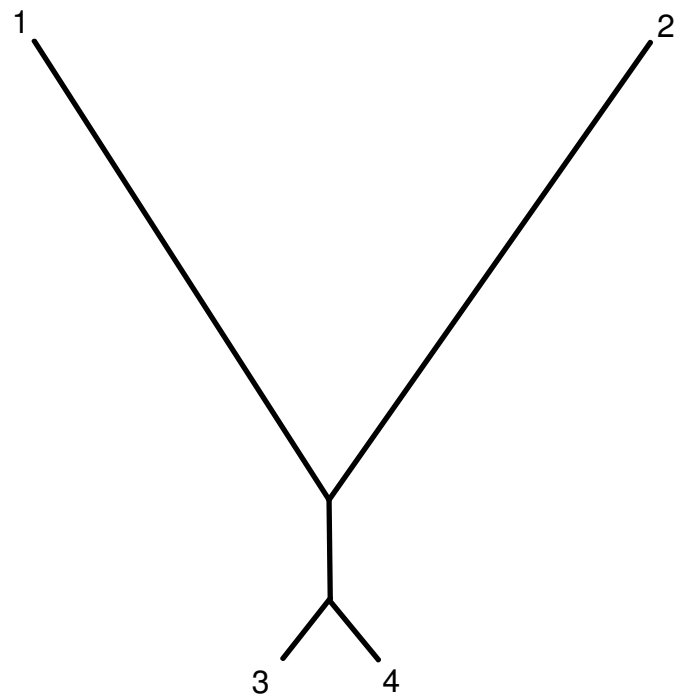
Failure to correct distance sufficiently leads to poor performance

"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.

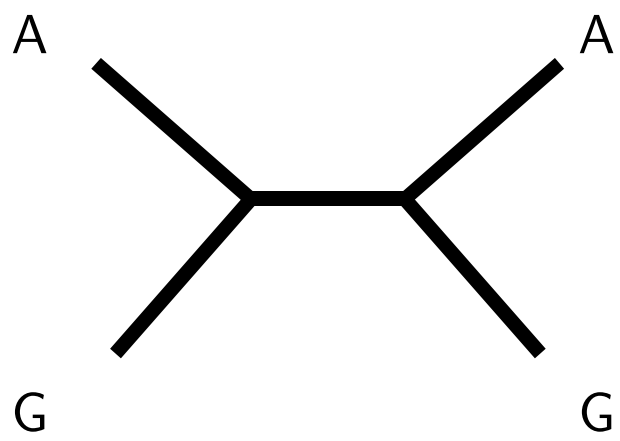


Distance methods: pros

- Fast – the 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

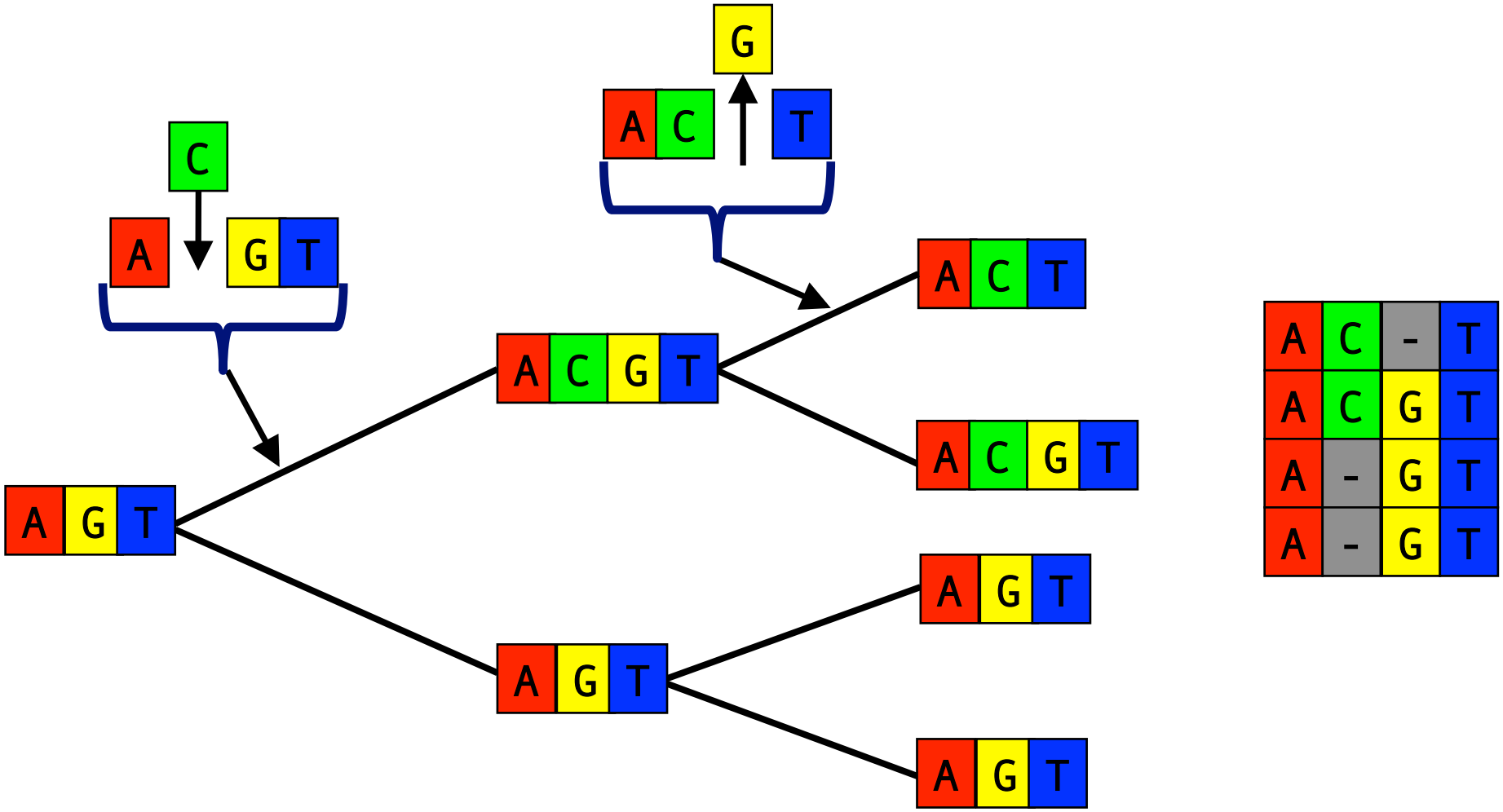


Outline

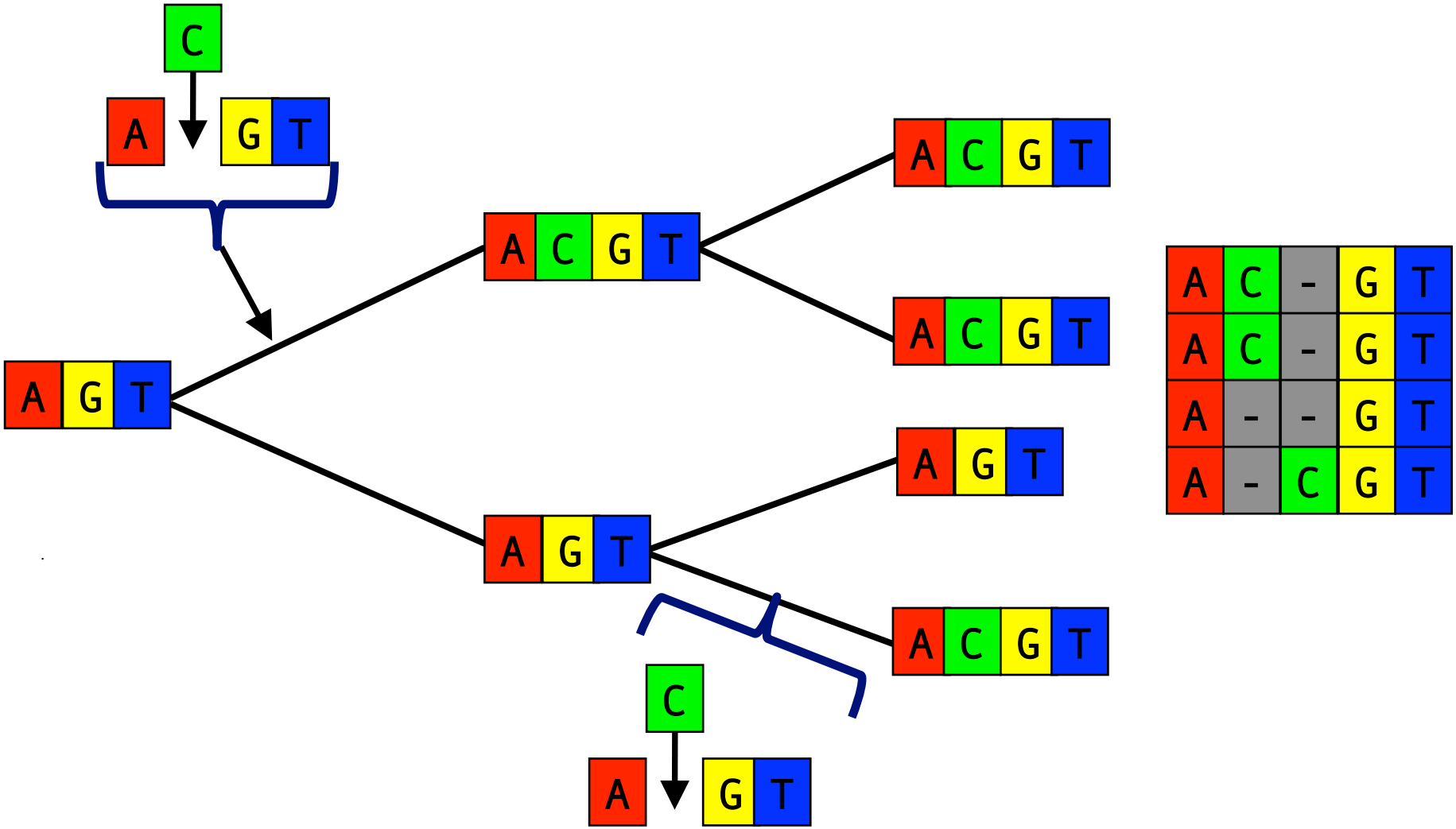
1. phylogenetics is crucial for comparative biology
2. tree terminology
3. why phylogenetics is difficult
4. parsimony
5. distance-based methods
6. theoretical basis of multiple sequence alignment

Multiple Sequence Alignment (MSA) - main points

- The goal of MSA is to introduce gaps such that residues in the same column are homologous (all residues in the column descended from a residue in their common ancestor).



slide by Derrick Zwickl



slide by Derrick Zwickl

Expressing homology detection as a bioinformatics challenge

- The problem is recast as:
  - reward matches (+ scores)
  - penalize rare substitutions (– scores),
  - penalize gaps (– scores),
  - try to find an alignment that maximizes the total score

- Pairwise alignment is tractable

- Most MSA programs use progressive alignment:
  - this reduces MSA to a series of pairwise operations.
  - these algorithms are heuristic. They are not guaranteed to return the optimal solution.
  - the criteria used are not ideal from an evolutionary standpoint (and this has implications for tree inference).

- Simultaneous inference of MSA and tree is the most appropriate choice (see Hossain et al., 2015), but is computationally demanding. See: Poisson Indel Process (Bouchard-Côté and Jordan, 2013), Bali-Phy, Handel, AliFritz, and POY software
- Many people filter the automatically generated alignments: GUIDANCE2 (and similar tools) cull ambiguously aligned regions to lower the chance that misalignment leads to errors in downstream analyses.

BLOSUM 62 Substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	−1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Scoring an alignment with the BLOSUM 62 matrix

Pongo	V	D	E	V	G	G	E	L	G	R	L	F	V	V	P	T	Q
Gorilla	V	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	2	-2	0	6	-6	-3	-4	-2	-2	4	0	4	-1	7	4	1

The score for the alignment is

$$D_{ij} = \sum_k d_{ij}^{(k)}$$

If i indicates *Pongo* and j indicates *Gorilla*. (k) is just an index for the column.

$$D_{ij} = 12$$

Scoring an alignment with gaps

If we were to use a gap penalty of -8:

Pongo	V	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	V	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

By introducing gaps we have improved the score:

$$D_{ij} = 40$$

Gap Penalties

Penalizing gaps more heavily than substitutions avoids alignments like this:

```
Pongo     VDEVGGE-LGRLFVVPTQ
Gorilla   VDEVGG-DLGRLFVVPTQ
```

Affine gap penalties are often used to accommodate multi-site indels:

$$\mathsf{GP} = \mathsf{GO} + (l)\mathsf{GE}$$

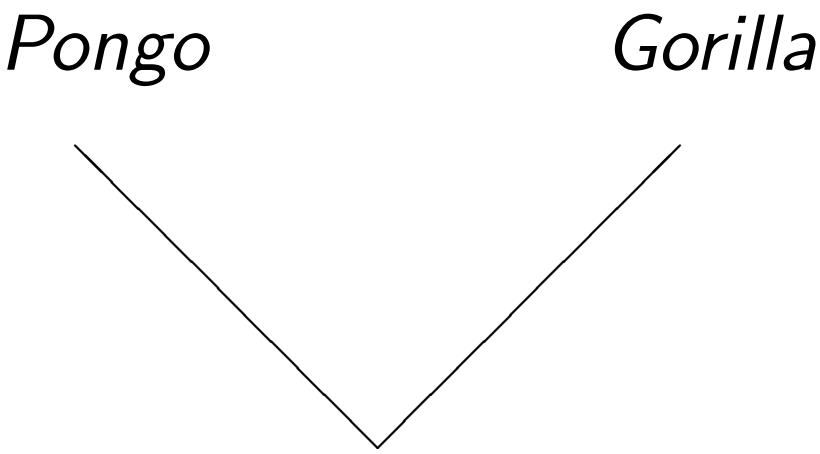
where:

- GP is the gap penalty.
- GO is the "gap-opening penalty"
- GE is the "gap-extension penalty"
- l is the length of the gap

Pairwise alignment costs

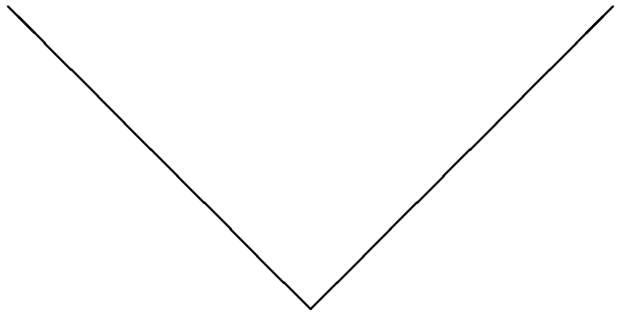
- Paul Lewis will explain likelihood tomorrow,
- Additive costs can be justified as approximations to the log of likelihoods if:
  - we can identify the events that must have occurred in generate the data, and
  - we can assign (relative) probabilities based on whether these events are rare or common.

Pongo	V	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	V	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1



Pongo	**V**	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	**V**	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

Pongo *Gorilla*

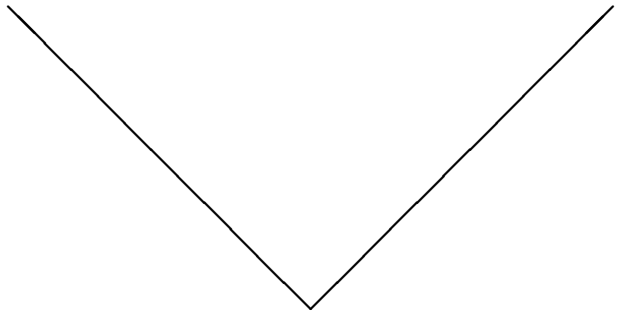


V $\leftrightarrow$ V

$$\mathbb{P}(\text{pos. } 1) = \mathbb{P}(V \leftrightarrow V)$$

Pongo	**V**	**D**	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	**V**	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

Pongo *Gorilla*

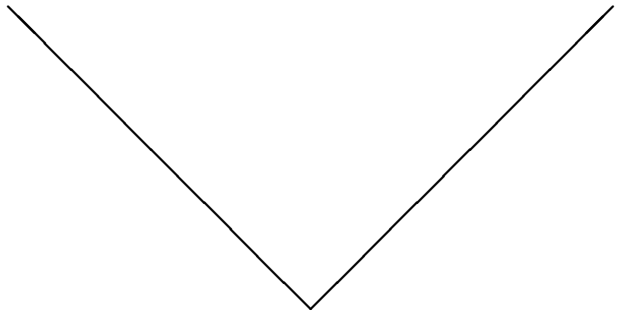


V $\leftrightarrow$ V
D $\leftrightarrow$ –

$$\mathbb{P}(\text{pos. } 1-2) = \mathbb{P}(V \leftrightarrow V) \times \mathbb{P}(D \leftrightarrow -)$$

Pongo	**V**	**D**	**E**	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	**V**	-	**E**	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

Pongo *Gorilla*

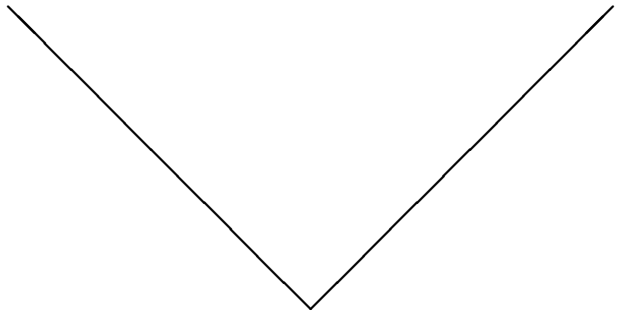


V ↔ V
D ↔ –
E ↔ E

$$\mathbb{P}(\text{pos. } 1-3) = \mathbb{P}(V \leftrightarrow V) \times \mathbb{P}(D \leftrightarrow -) \times \mathbb{P}(E \leftrightarrow E)$$

Pongo	**V**	**D**	**E**	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	**V**	-	**E**	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

Pongo *Gorilla*

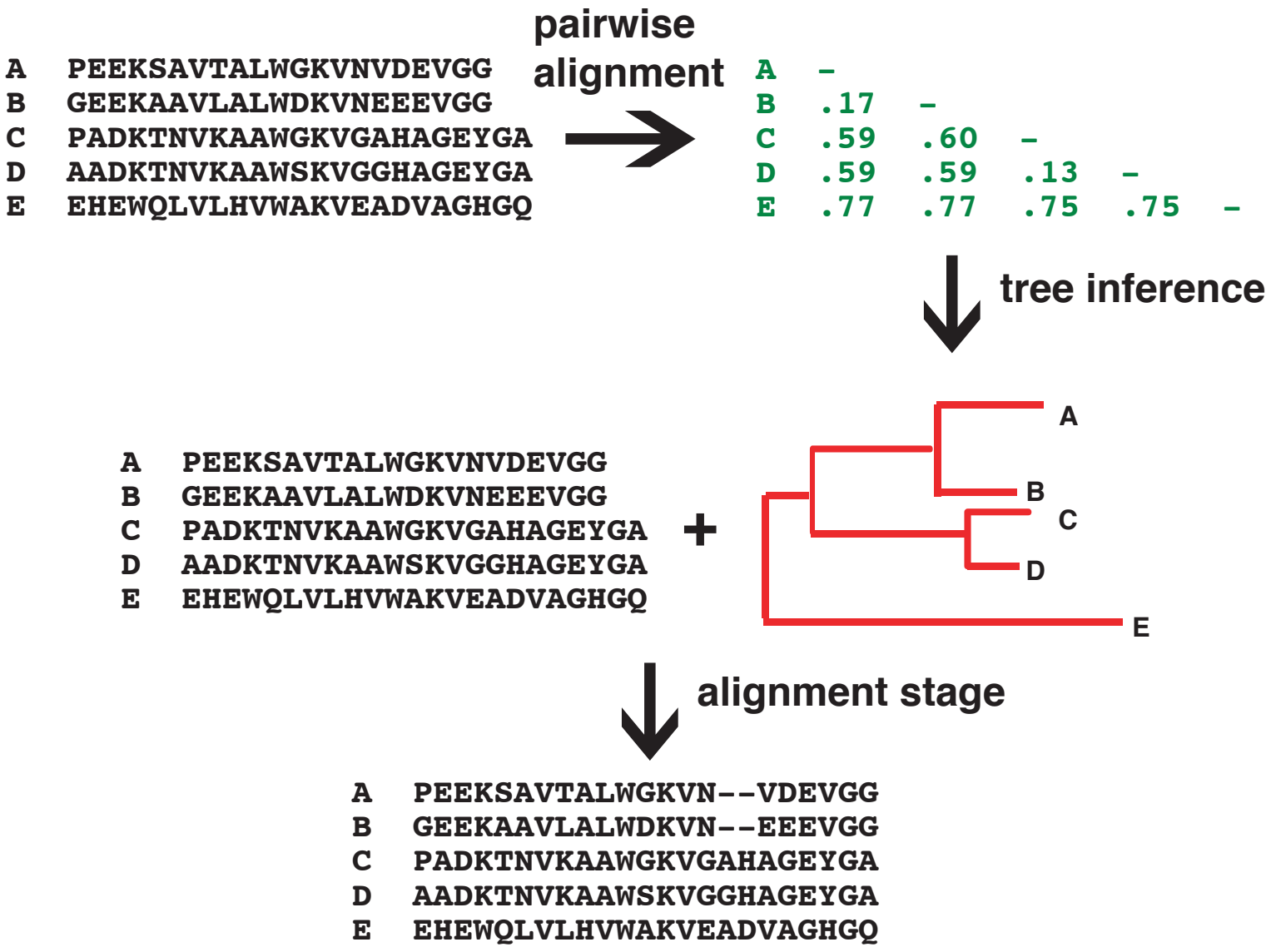


V ↔ V
D ↔ –
E ↔ E

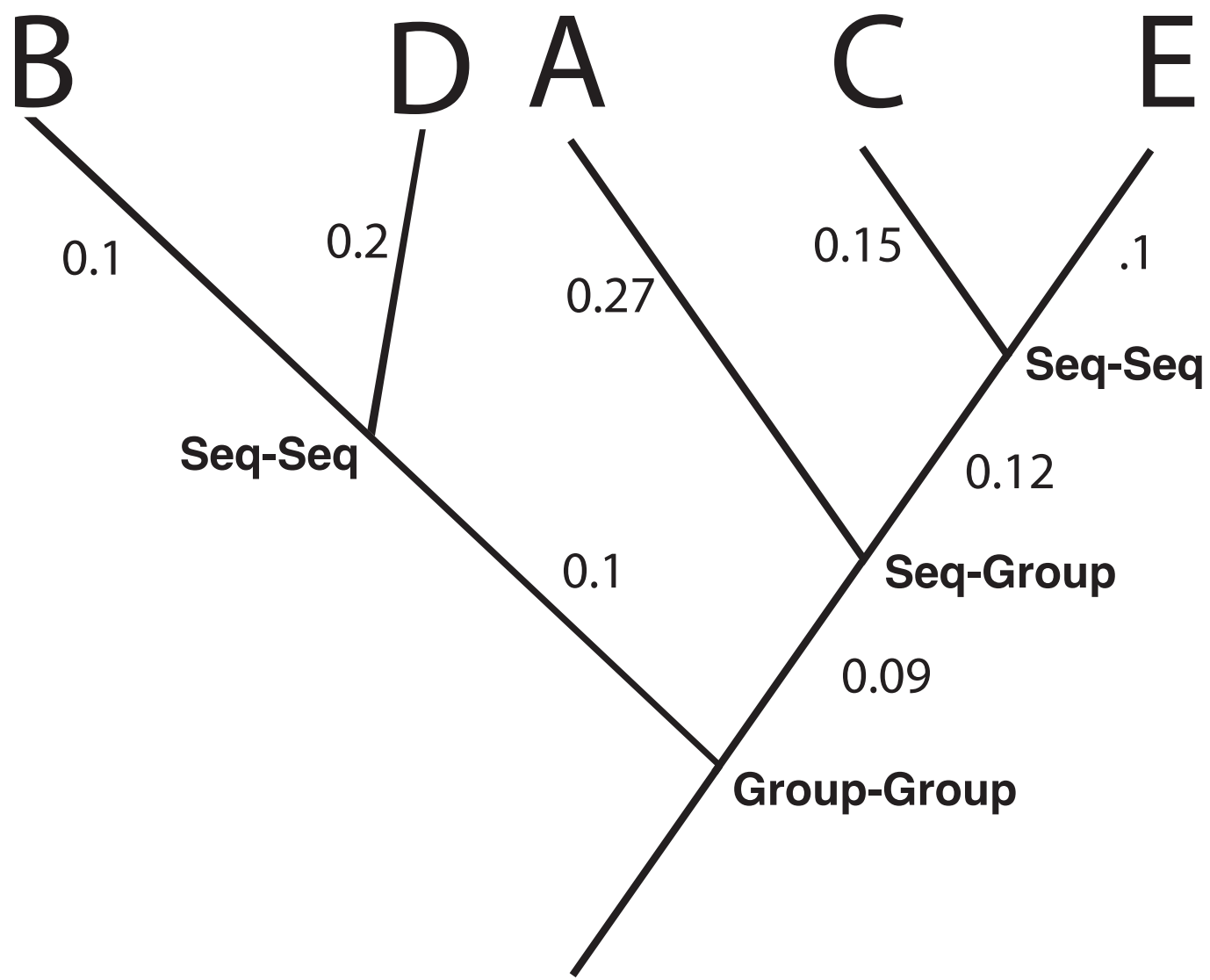
$$\ln \mathbb{P}(\text{pos. } 1-3) = \ln \mathbb{P}(V \leftrightarrow V) + \ln \mathbb{P}(D \leftrightarrow -) + \ln \mathbb{P}(E \leftrightarrow E)$$

Multiple sequence alignment is an ugly topic in bioinformatics

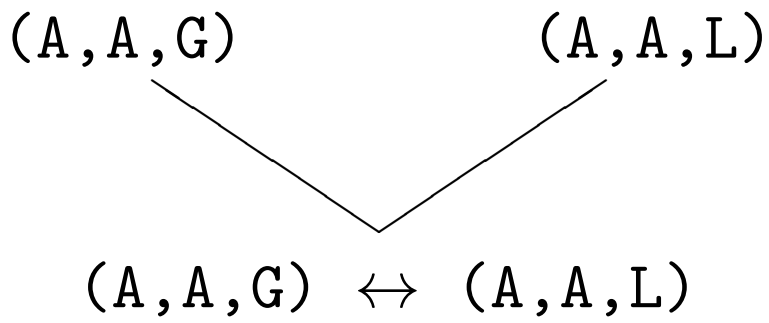
- Clever programming tricks help, but we still have to rely on *heuristics* – approaches that provide good solutions, but are not guaranteed to find the best solution.
- The additive scoring system suffers from the fact that we do not observe ancestral sequences.


pairwise
alignment
A PEEKSAVTALWGKVNVDEVGG
B GEEKAAVLALWDKVNEEEVGG
C PADKTNVKAAWGKVGAHAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA
E EHEWQLVLHVWAKVEADVAGHGQ
A -
B .17 -
C .59 .60 -
D .59 .59 .13 -
E .77 .77 .75 .75 -
tree inference
A PEEKSAVTALWGKVNVDEVGG
B GEEKAAVLALWDKVNEEEVGG
C PADKTNVKAAWGKVGAHAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA
E EHEWQLVLHVWAKVEADVAGHGQ
A
B
C
D
E
alignment stage
A PEEKSAVTALWGKVN--VDEVGG
B GEEKAAVLALWDKVN--EEEVGG
C PADKTNVKAAWGKVGAHAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA
E EHEWQLVLHVWAKVEADVAGHGQ

Aligning multiple sequences



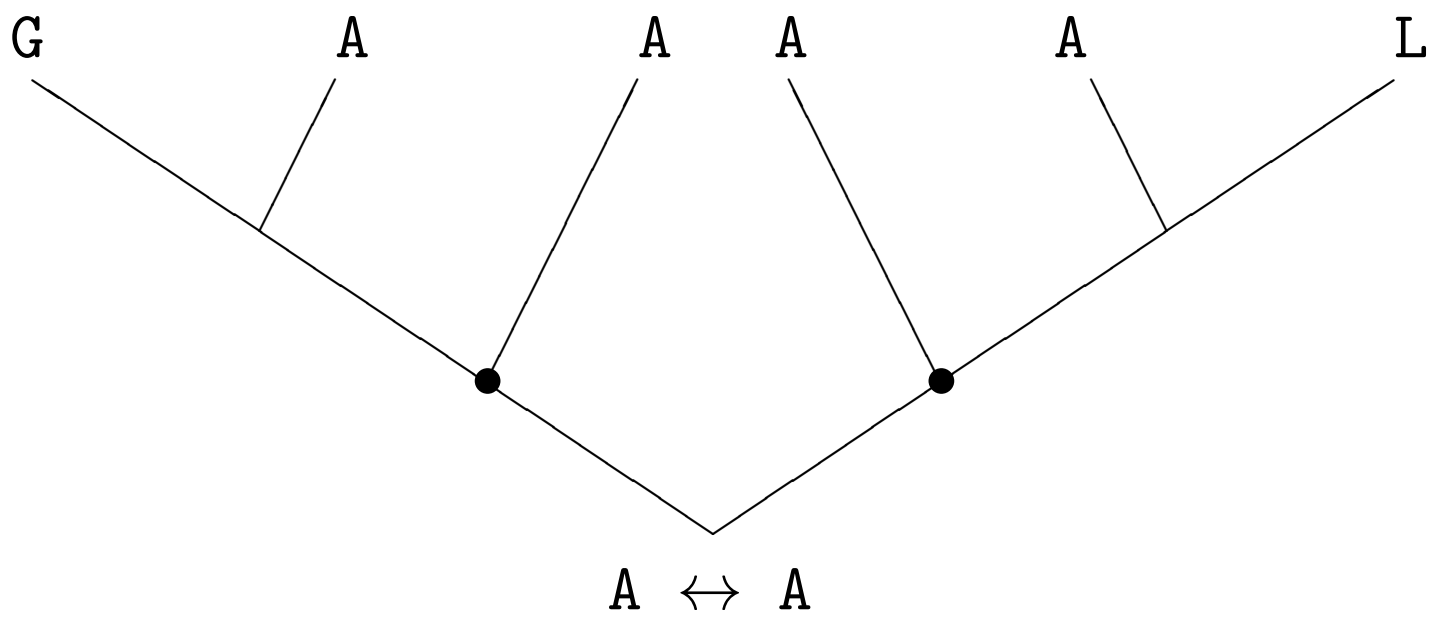
Imperfect scoring system. Consider one position in a group-to-group alignment:



The sum-of-pairs score for aligning would be:

$$\frac{4}{9}(A \leftrightarrow A) + \frac{2}{9}(A \leftrightarrow L) + \frac{2}{9}(G \leftrightarrow A) + \frac{1}{9}(G \leftrightarrow L)$$

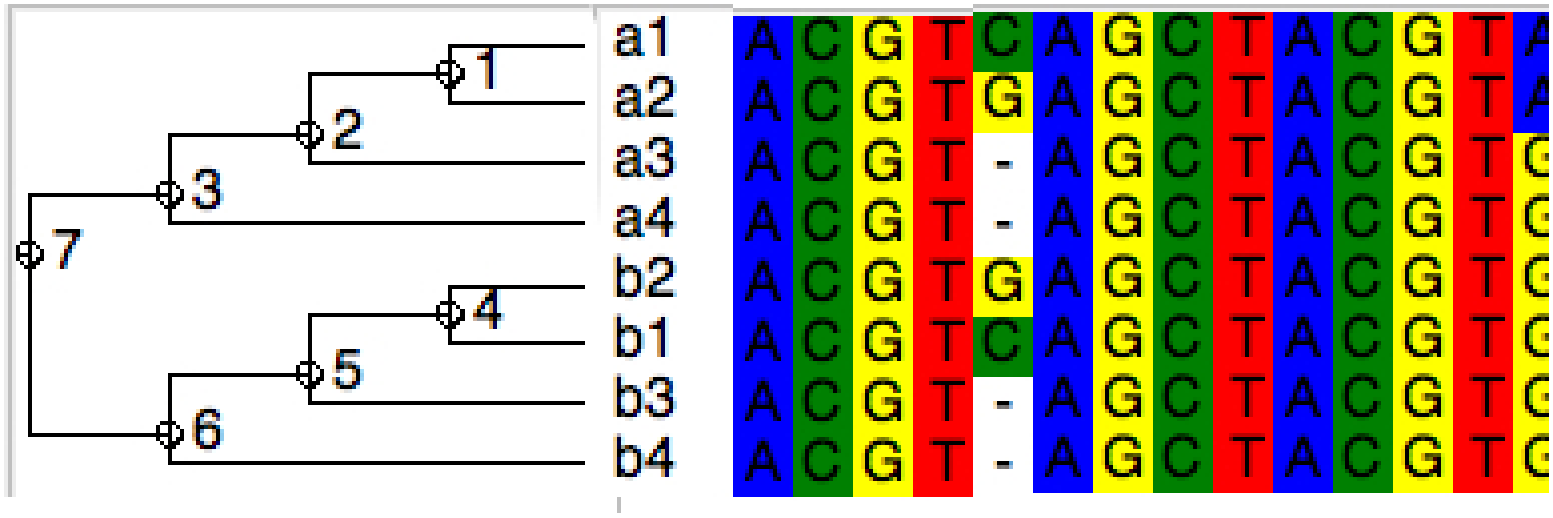
But in the context of the tree we might be pretty certain of an A↔A event



Note: weighted sum-of-pairs would help reflect the effect of ancestry better (but still not perfectly; sum-of-pairs techniques are simply not very sophisticated forms of ancestral sequence reconstruction).

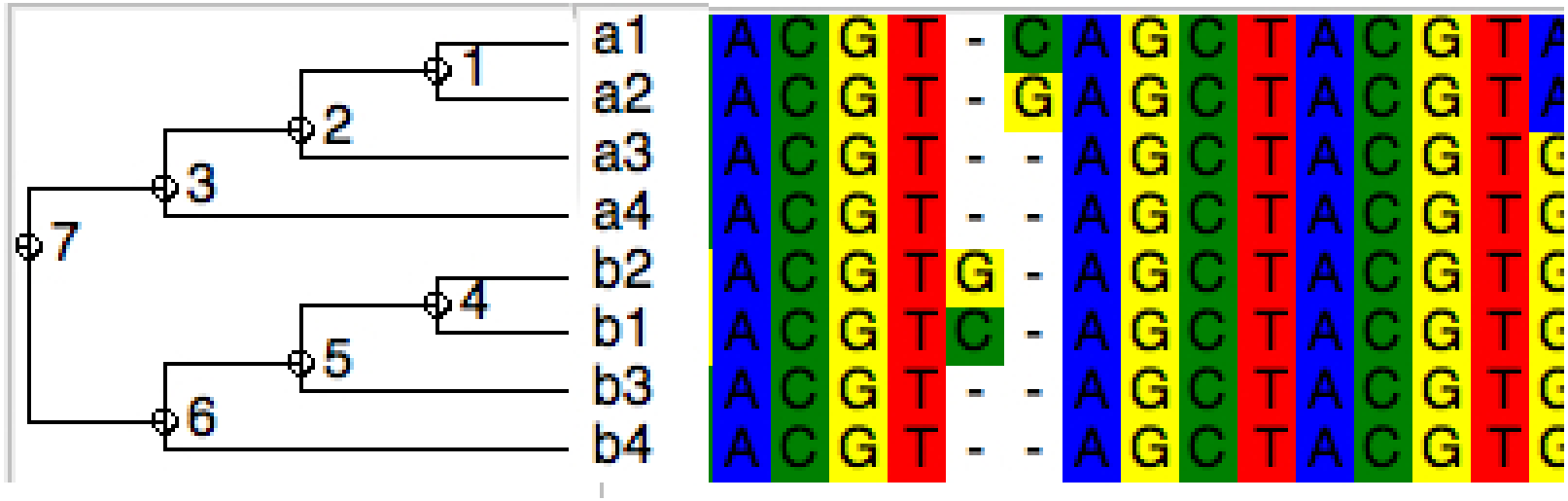
PRANK

Löytynoja and Goldman (2005) showed most progressive alignment techniques were particularly prone to compression because of poor ancestral reconstruction:



PRANK

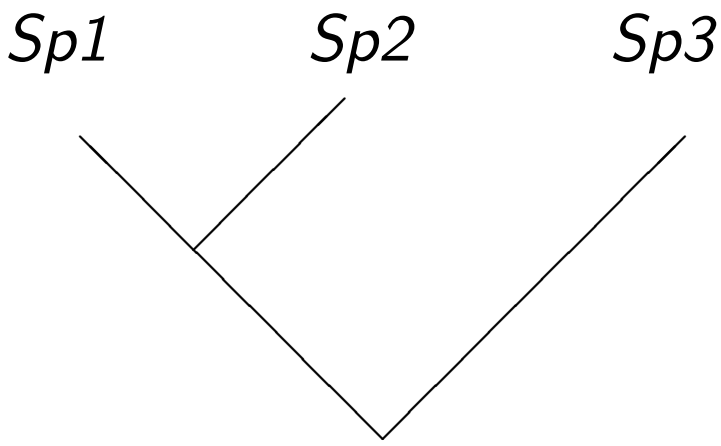
Flagging inserted residues allows PRANK to effectively skip over these positions in the ancestor, producing more phylogenetically-sensible alignments:



Greedy choices leading to failure to find the best alignment

Consider the scoring scheme:
match = 0 mismatch = -3 gap = -7

Guide Tree:



Sequences:

Sp1	GACCGTG
Sp2	GCCGTAG
Sp3	GACCGTAG

Greedy choices leading to failure to find the best alignment

match = 0 mismatch = -3 gap = -7

ungapped1*vs*2

Sp1	G	A	C	C	G	T	G		
Sp2	G	C	C	G	T	A	G		
Score	0	-3	0	-3	-3	-3	0	**Total=**	-12

would be preferred over gapped1*vs*2:

Sp1	G	A	C	C	G	T	-	G	
Sp2	G	-	C	C	G	T	A	G	
Score	0	-7	0	0	0	0	-7	0	**Total=** -14

Adding a *Sp3* to ungapped1*vs*2:

Sp1	G	-	A	C	C	G	T	G
Sp2	G	-	C	C	G	T	A	G
Sp3	G	A	C	C	G	T	A	G

This implies 1 indel, and 4 substitutions. Score = -19 *

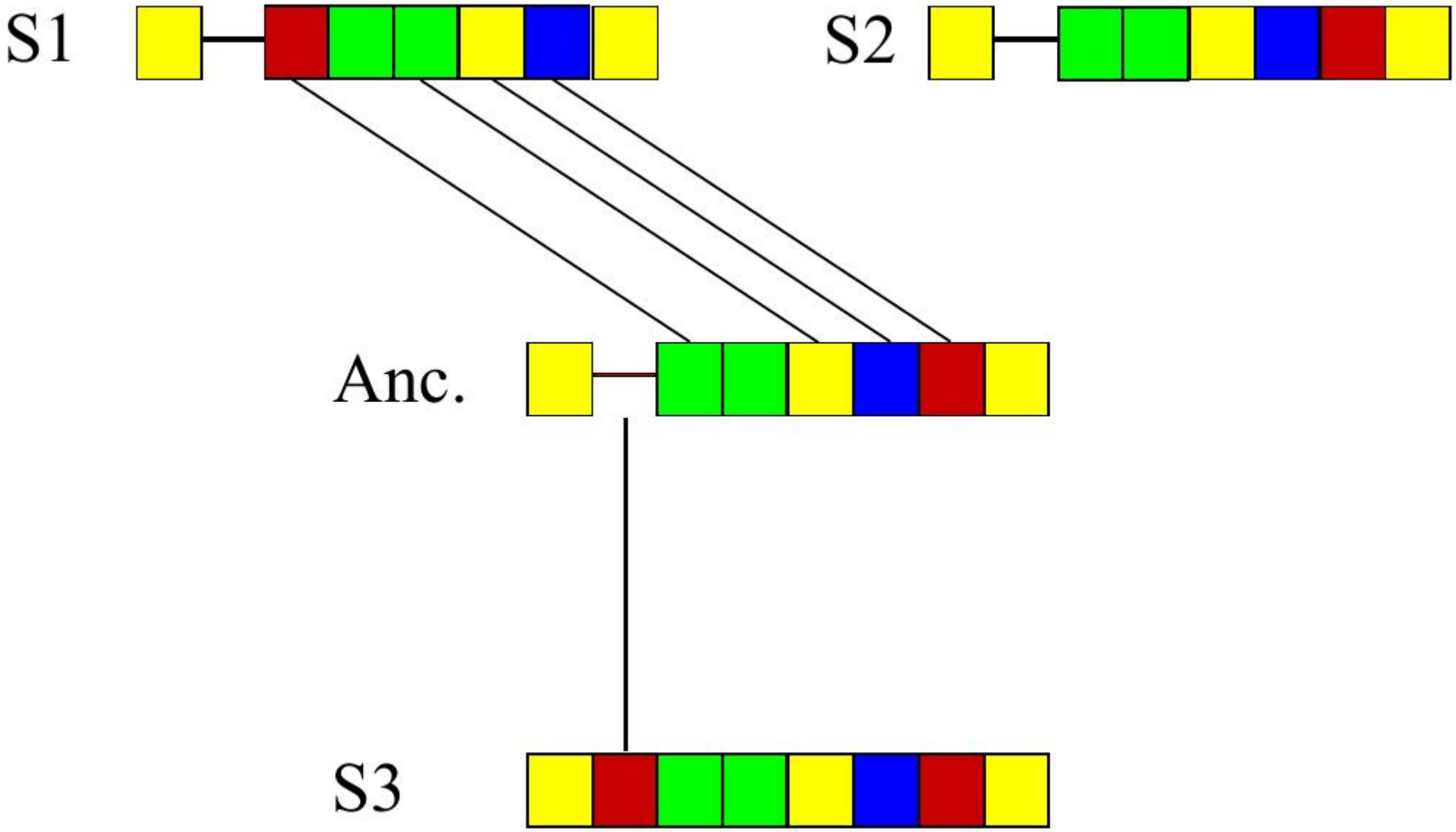
If we had been able to use gapped1*vs*2 then we could have:

Sp1	G	A	C	C	G	T	-	G
Sp2	G	-	C	C	G	T	A	G
Sp3	G	A	C	C	G	T	A	G

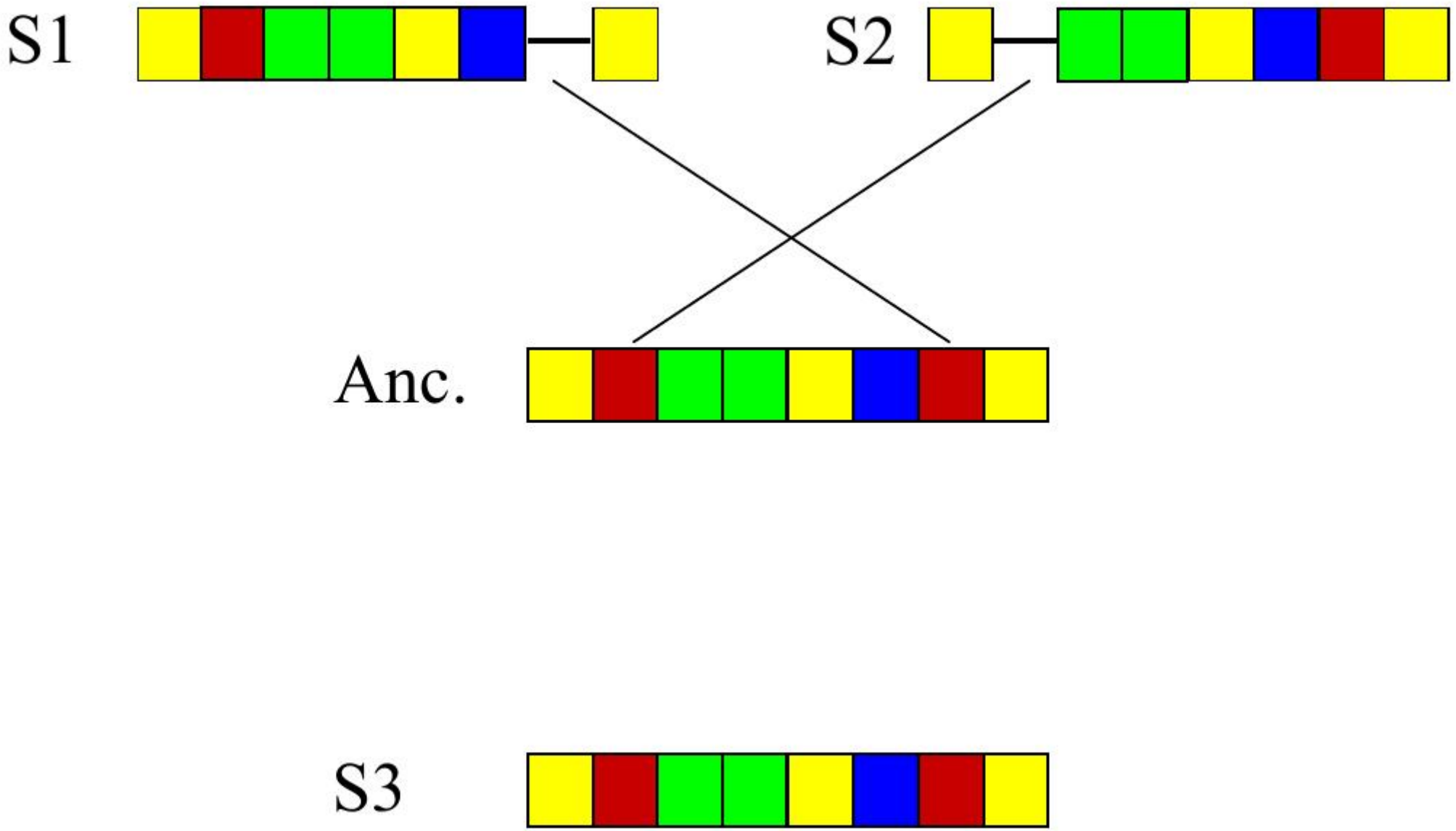
score = -14 *

* = "sort of..."

Score = -19 if we count events, but sum of pairs score would differ



Score = -14 if we count events, but sum of pairs score would differ



Polishing (aka "iterative alignment" can correct some errors caused by greedy heuristics)

1. break the alignment into 2 groups of sequences (often by breaking an edge in the merge tree).

2. realign those 2 groups to each other

3. keep the realignment if it improves the score

Opal also uses random 3-group polishing.

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.

De Maio, N., Schrempf, D., and Kosiol, C. (2015). Pomo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6):1018–1031.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.

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

Liu, Y., Cotton, J. A., Shen, B., Han, X., Rossiter, S. J., and Zhang, S. (2010). Convergent sequence evolution between echolocating bats and dolphins. *Current Biology*, 20(2):R53 – R54.

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. *PLoS One*.

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.

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