Phylogenetic inference from data

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Jeet Sukumaran and Mark Holder for slides)

How do we figure out what tree captures the relationships we're interested in?

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.
- Fundamental perspectives in all these approaches:

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.
- Fundamental perspectives in all these approaches:
  - Current patterns of biodiversity has been generated by processes of: (1) speciation, (2) extinction, and (3) character modification.

Phylogenetic Inference

- We cannot see, measure, collect, or otherwise obtain a phylogeny directly from nature.
- We have to *infer* it from data that we *can* collect from nature.
- A variety of different inferential approaches and data are used, with both increasing in sophistication and complexity over time.
- Fundamental perspectives in all these approaches:
  - Current patterns of biodiversity has been generated by processes of: (1) speciation, (2) extinction, and (3) character modification.
  - The phylogeny is an abstract representation ("model") of this diversification process.

Enormous numbers of topologies to consider

Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

Enormous numbers of topologies to consider

Taxa	Unrooted binary trees	Rooted binary trees
3	1	3
4	3	15
5	15	105
6	105	945
10	2,027,025	3×10^{7}
15	7×10^{12}	2×10^{14}
20	2×10^{20}	8×10^{21}
50	3×10^{74}	
100	2×10^{182}	
1,000	2×10^{2860}	
10,000	8×10^{38658}	
1,000,000	1×10^{5866723}	

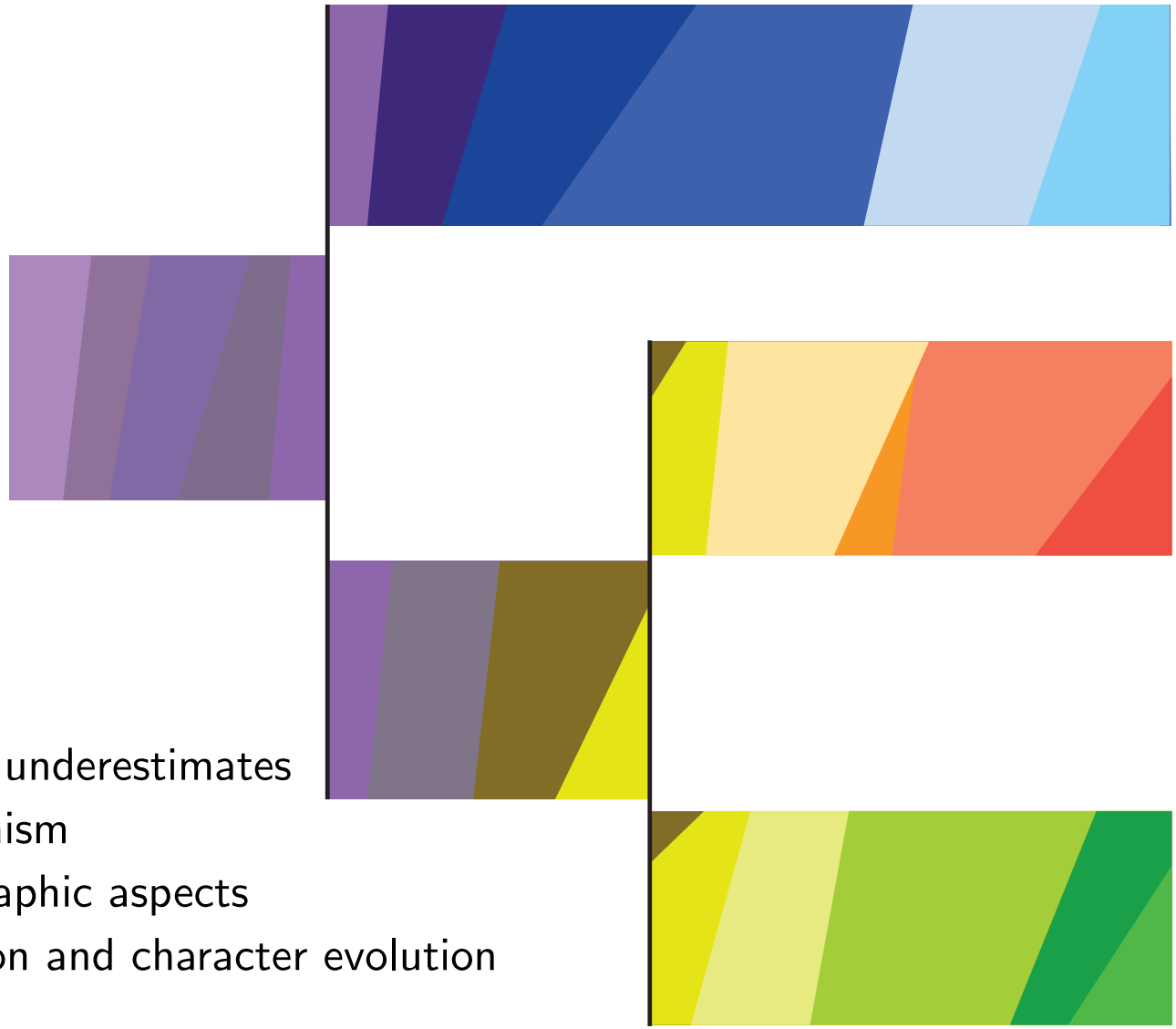
it is estimated that the there are between 10^{78} to 10^{82} atoms in the known, observable universe.

Estimating a tree from character 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.

Phylogeny with complete genome + "phenome" as colors:

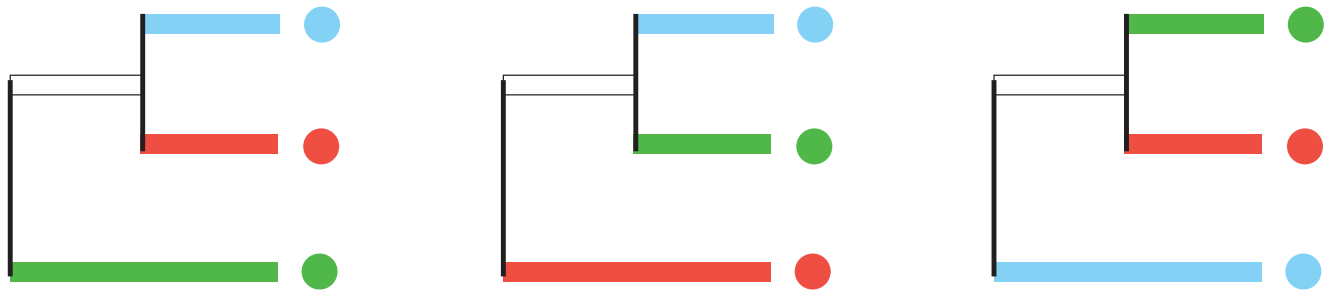


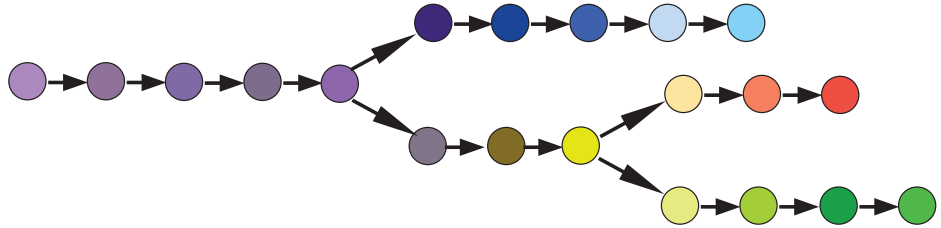
This figure:
dramatically underestimates
polymorphism
ignore geographic aspects
of speciation and character evolution

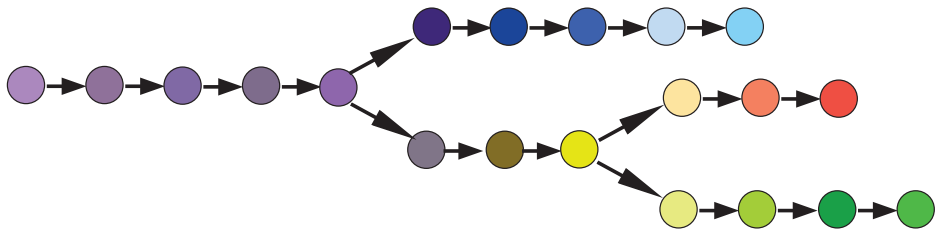
Extant species are just a thin slice of the phylogeny:

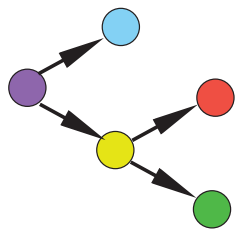
Our exemplar specimens are a subset of the current diversity:

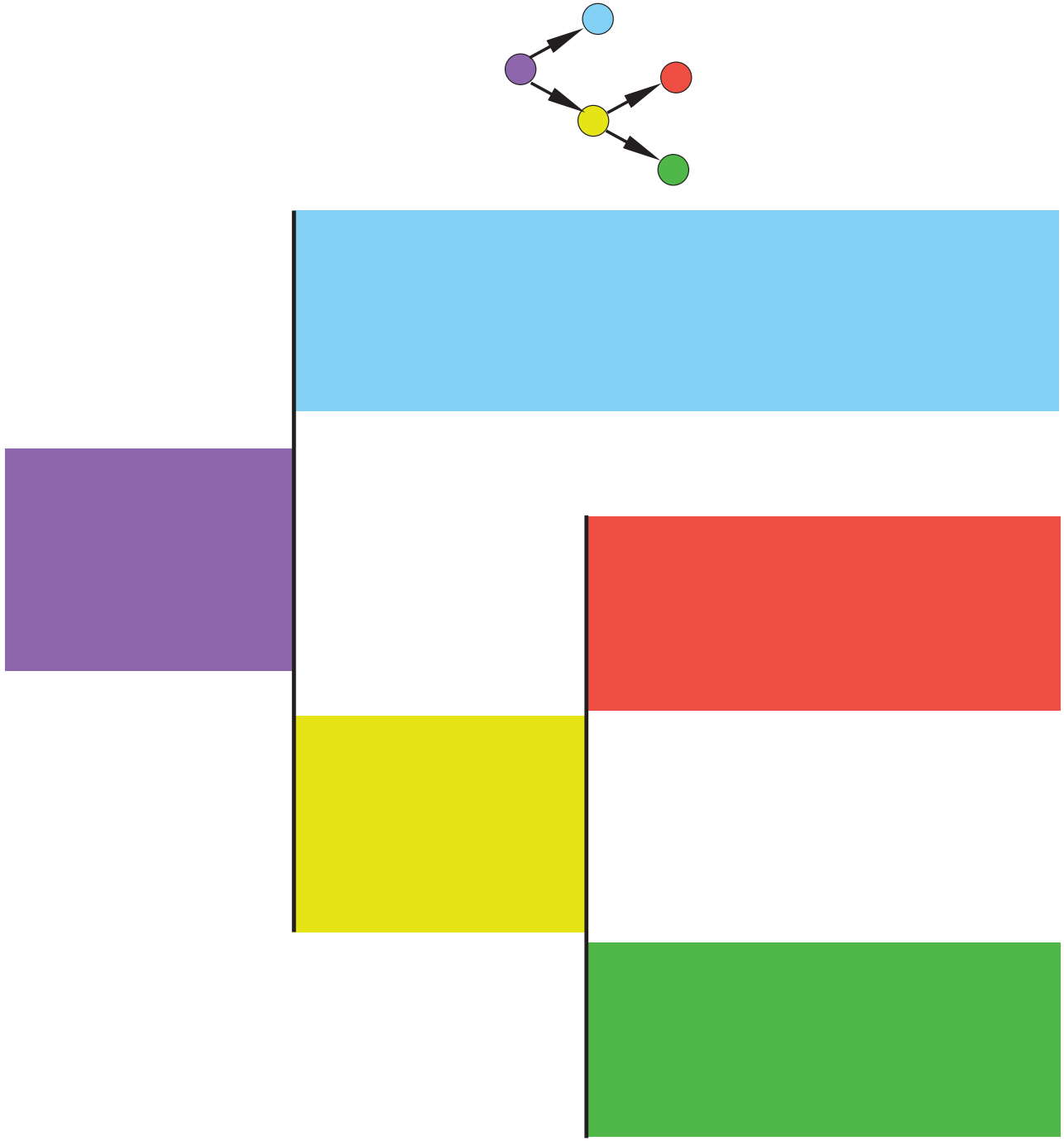
The phylogenetic inference problem:

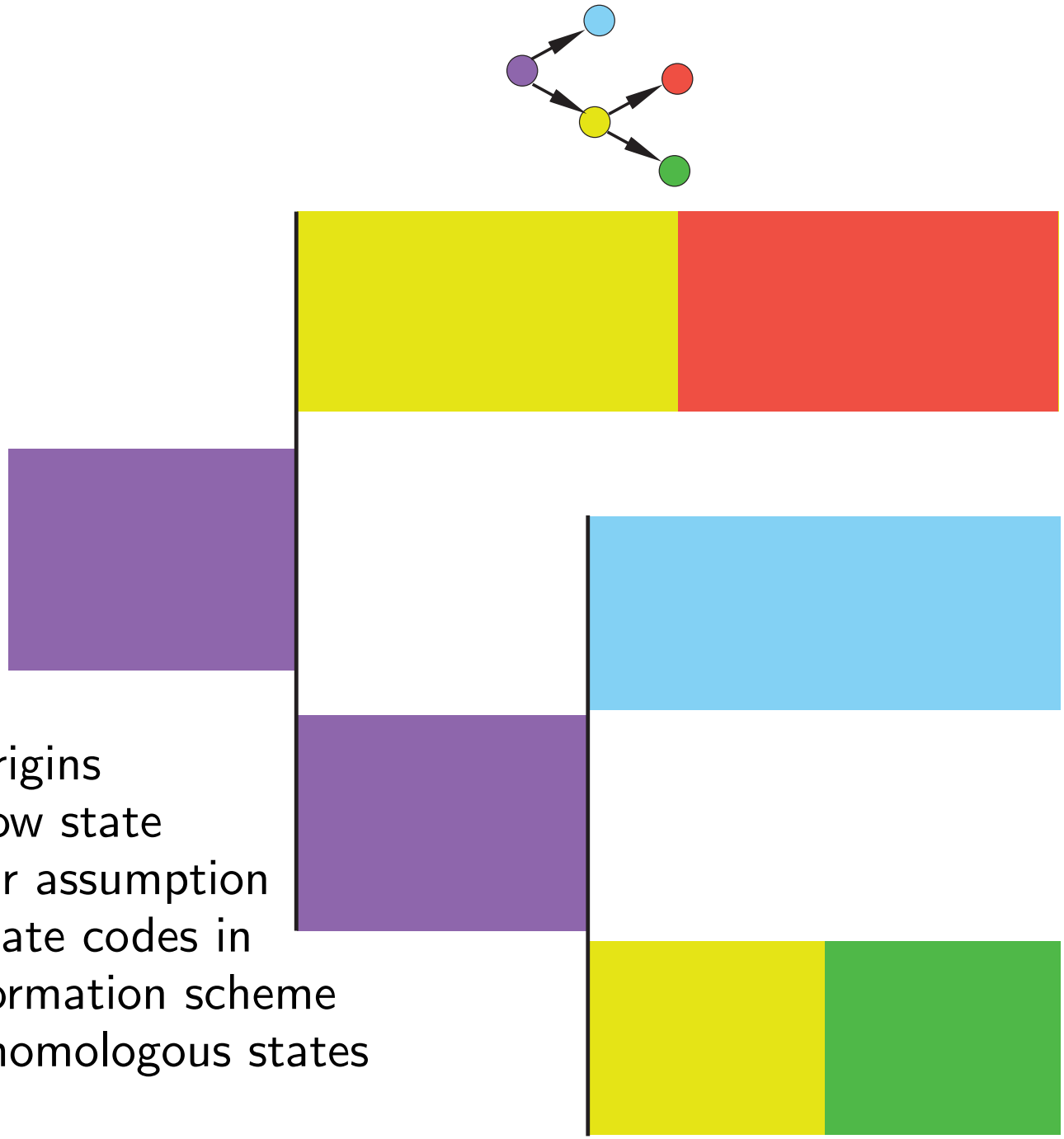




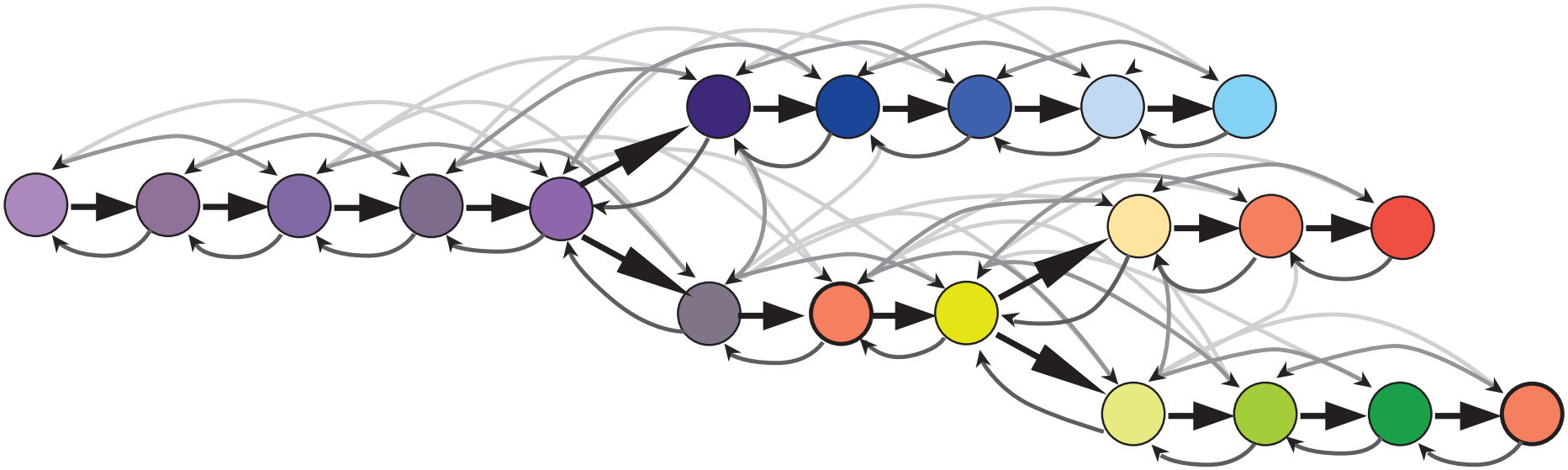








Multiple origins
of the yellow state
violates our assumption
that the state codes in
our transformation scheme
represent homologous states

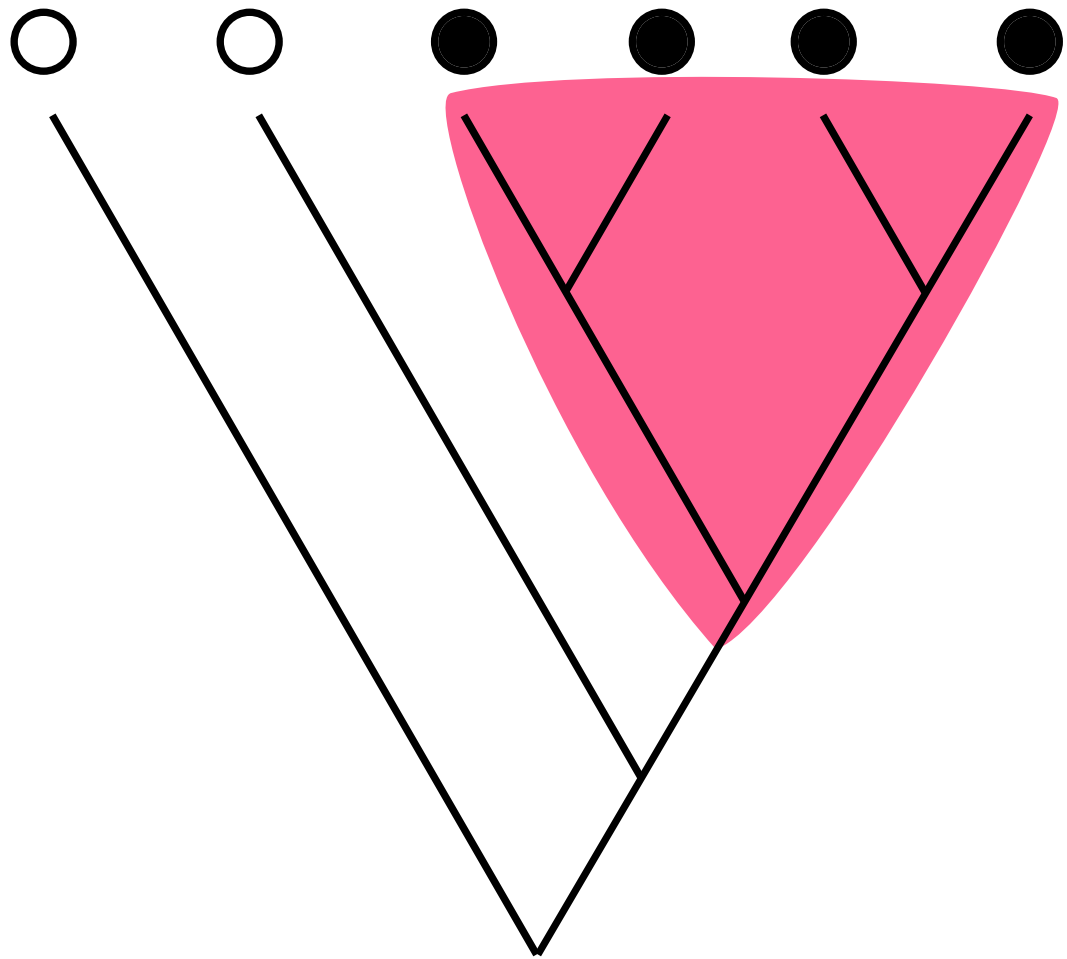


The meaning of homology (**very roughly**):

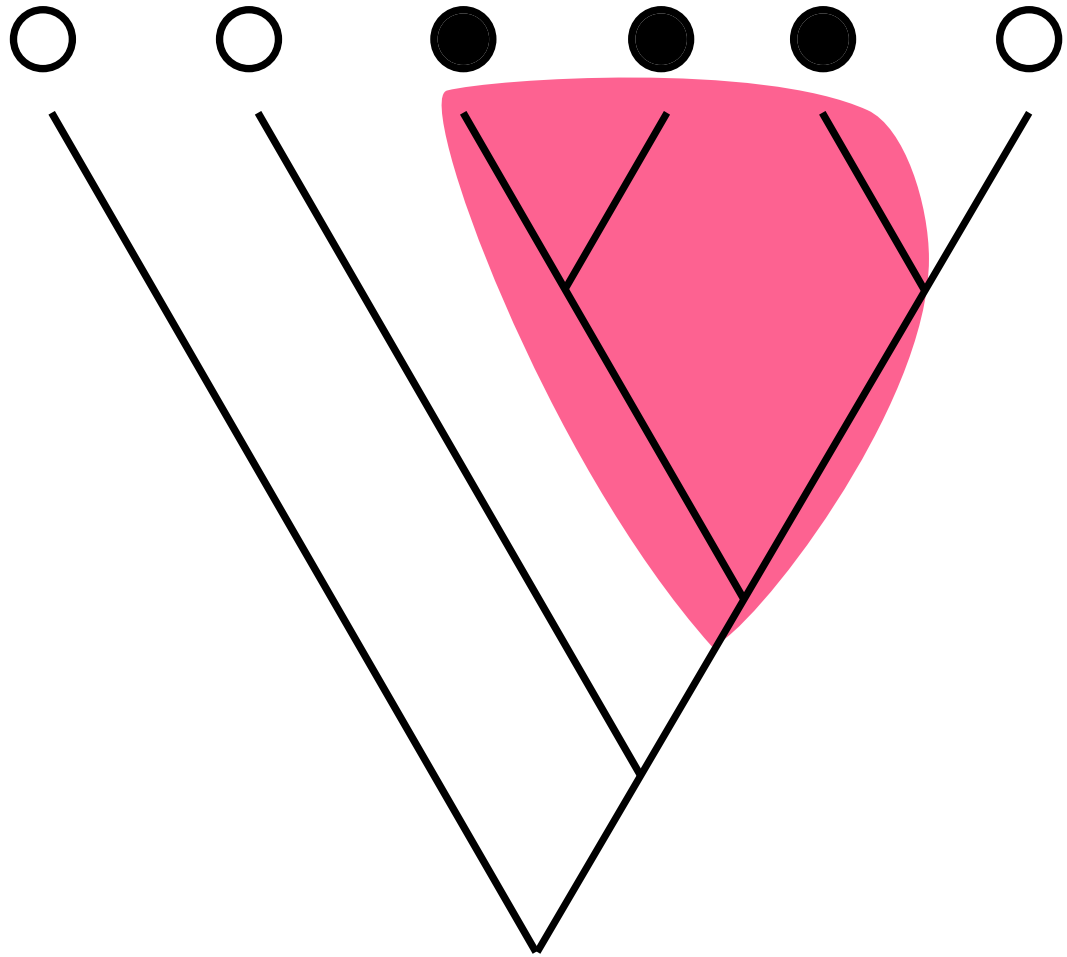
1. comparable (when applied to characters)
2. identical by descent (when applied to character states)

Ideally, each possible character state would arise once in the entire history of life on earth.

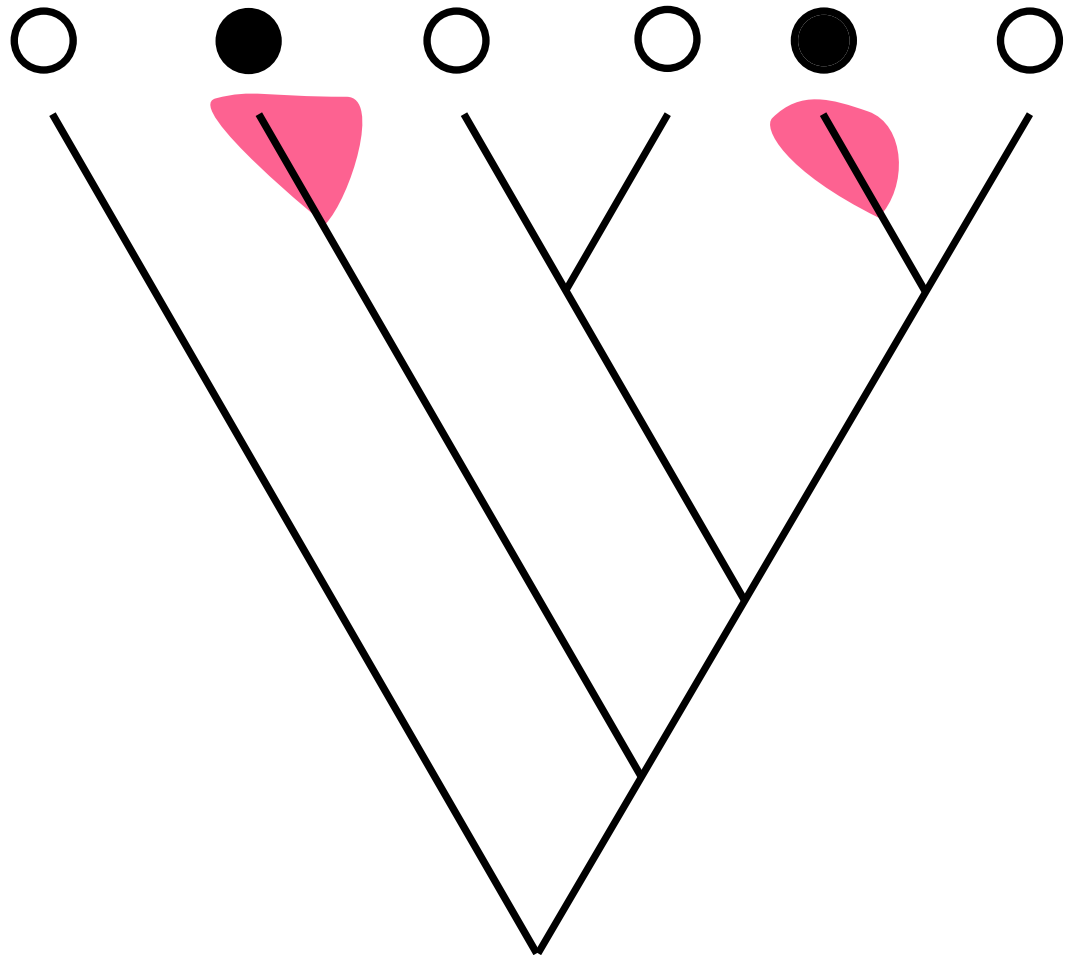
Instances of the filled character state are homologous
Instances of the hollow character state are homologous



Instances of the filled character state are homologous
Instances of the hollow character state are NOT homologous



Instances of the filled character state are NOT homologous
Instances of the hollow character state are homologous



Rule: Two taxa that share a character state must be more closely related to each other than either is to a taxon that displays a different state.(method suggested by Hennig)
Is this a valid rule?

Hennigian logical analysis

The German entomologist Willi Hennig (in addition to providing strong arguments for phylogenetic classifications) clarified the logic of phylogenetic inference.

Hennig's correction to our rule: Two taxa that share a **derived** character state must be more closely related to each other than either is to a taxon that displays the **primitive** state.

Hennig's logic is valid

Here we will use 0 for the primitive state, and 1 for the derived state.

	placenta	vertebra
Homo sapiens	1	1
Rana catesbiana	0	1
Drosophila melanogaster	0	0

Now the character "placenta" does not provide a grouping, but "vertebra" groups human and frog as sister taxa.

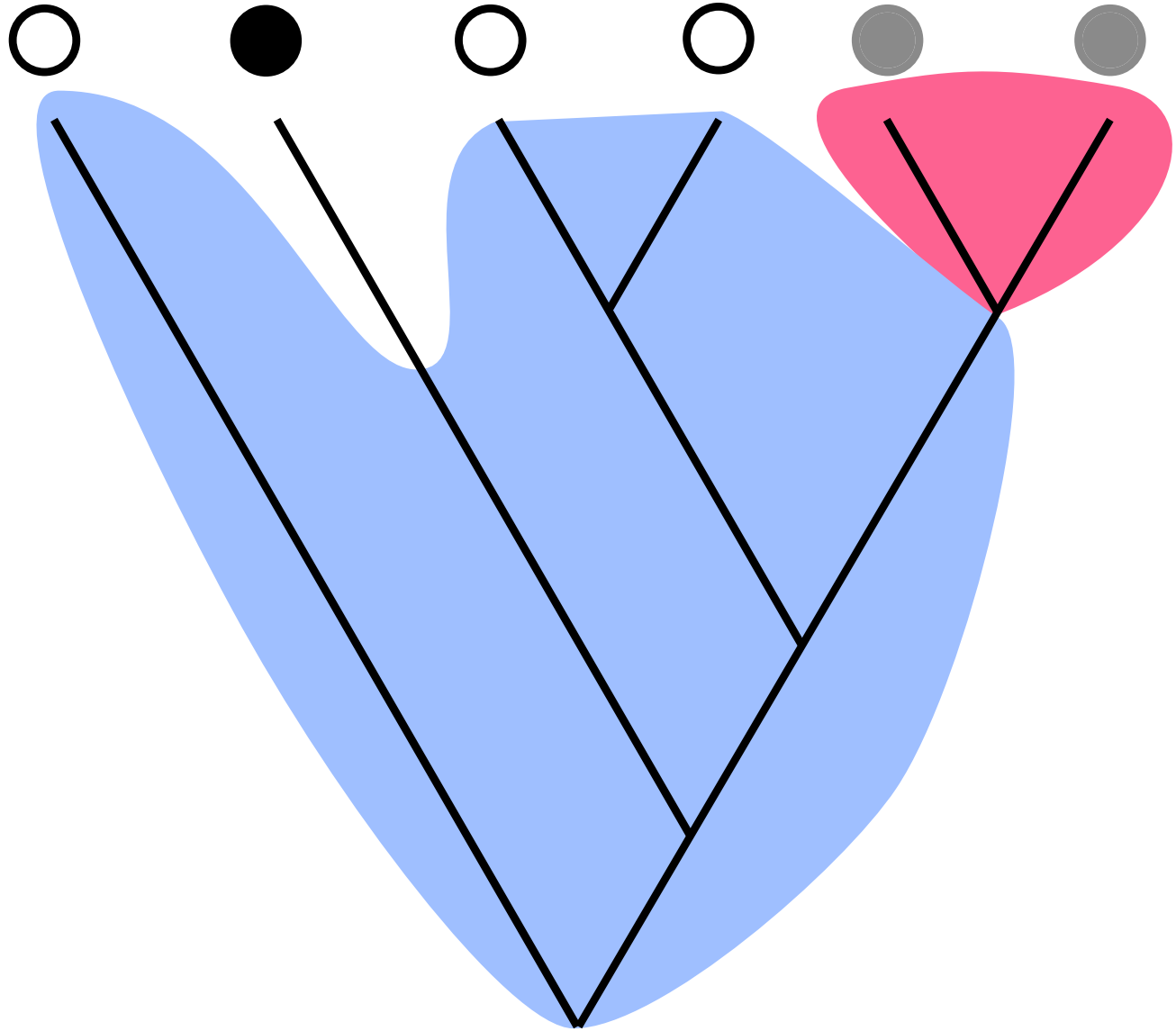
Hennigian terminology

prefixes:

- “apo” - refers to the new or derived state
- “plesio” - refers to the primitive state
- “syn” or “sym” - used to indicate shared between taxa
- “aut” - used to indicate a state being unique to one taxon

Hennigian rules

- synapomorphy - shared, derived states. Used to diagnose monophyletic groups.
- symplesiomorphy - shared, primitive states. Diagnose icky, unwanted paraphyletic groups.
- autapomorphy – a unique derived state. **No** evidence of phylogenetic relationships.
- constant characters – columns in a matrix with no variability between taxa. **No** evidence of phylogenetic relationships.



Hennigian inference

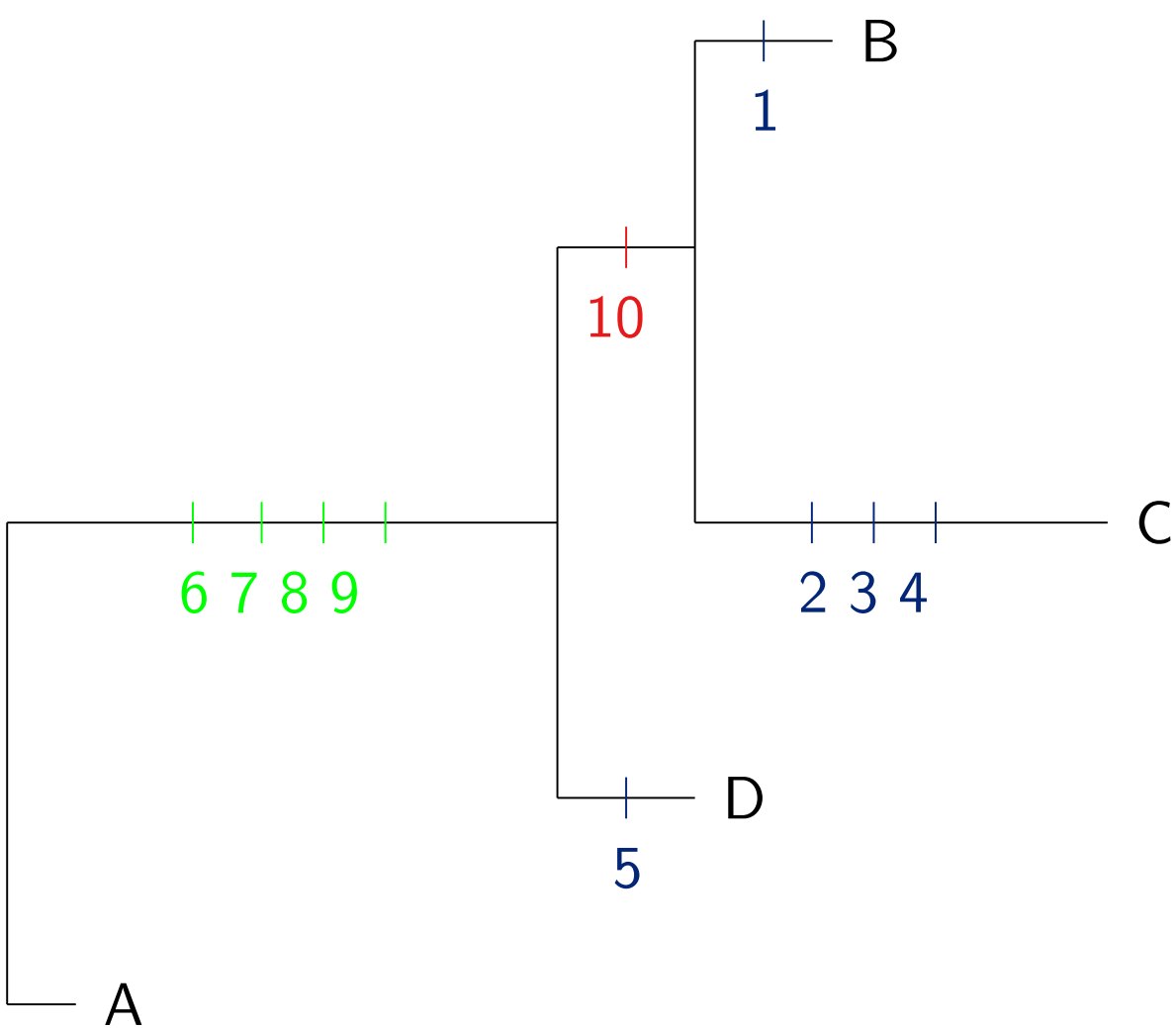
When we create a character matrix for Hennig's system, it is crucial that:

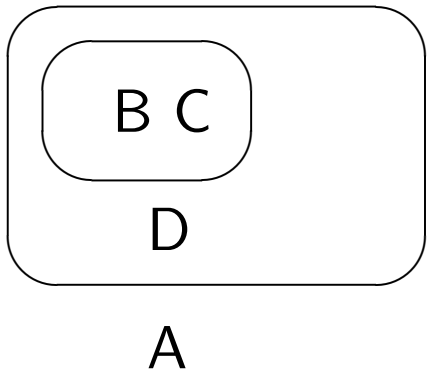
- traits assigned the same state represent homologous states (trace back to the MRCA)
- we correctly identify the directionality of the transformations (which state is plesiomorphic and which is apomorphic). The process of identifying the direction of change is called polarization.

Polarization could be done based on developmental considerations, paleontological evidence, or biogeographic considerations, but the most common technique is outgroup polarization.

Taxon	Character #									
	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1
D	0	0	0	0	1	1	1	1	1	0

Breakout room! What are the relationships between these taxa?


B
1
10
C
6 7 8 9
2 3 4
D
5
A


B C
D
A

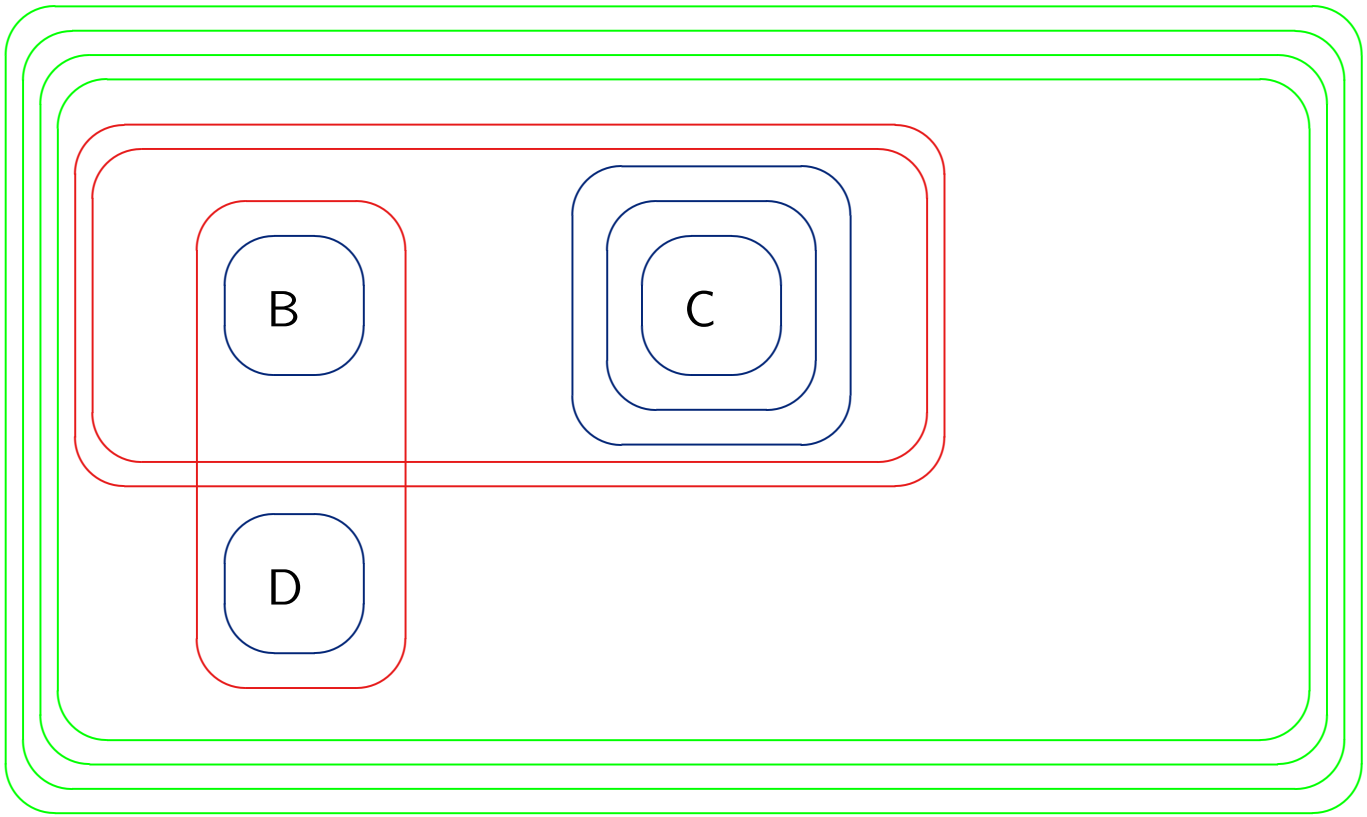
If characters are not polarized (ancestral and descendent states known) this method can infer unrooted trees.

Character conflict

Homo sapiens	AGTTCAAGT
Rana catesbiana	AATTCAAGT
Drosophila melanogaster	AGTTCAAGC
C. elegans	AATTCAAGC

The red character implies that either (*Homo* + *Drosophila*) is a group (if G is derived) and/or (*Rana* + *C. elegans*) is a group.
The green character implies that either (*Homo* + *Rana*) is a group (if T is derived) and/or (*Drosophila* + *C. elegans*) is a group.
The green and red character cannot both be correct.

Taxon	Character #											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1	1	0
D	0	0	0	0	1	1	1	1	1	0	0	1



A

Character conflict

Two characters are compatible if they can both be mapped on the same tree so that all of the character states displayed could be homologous.

Character conflict

Two characters are compatible if they can both be mapped on the same tree so that all of the character states displayed could be homologous.

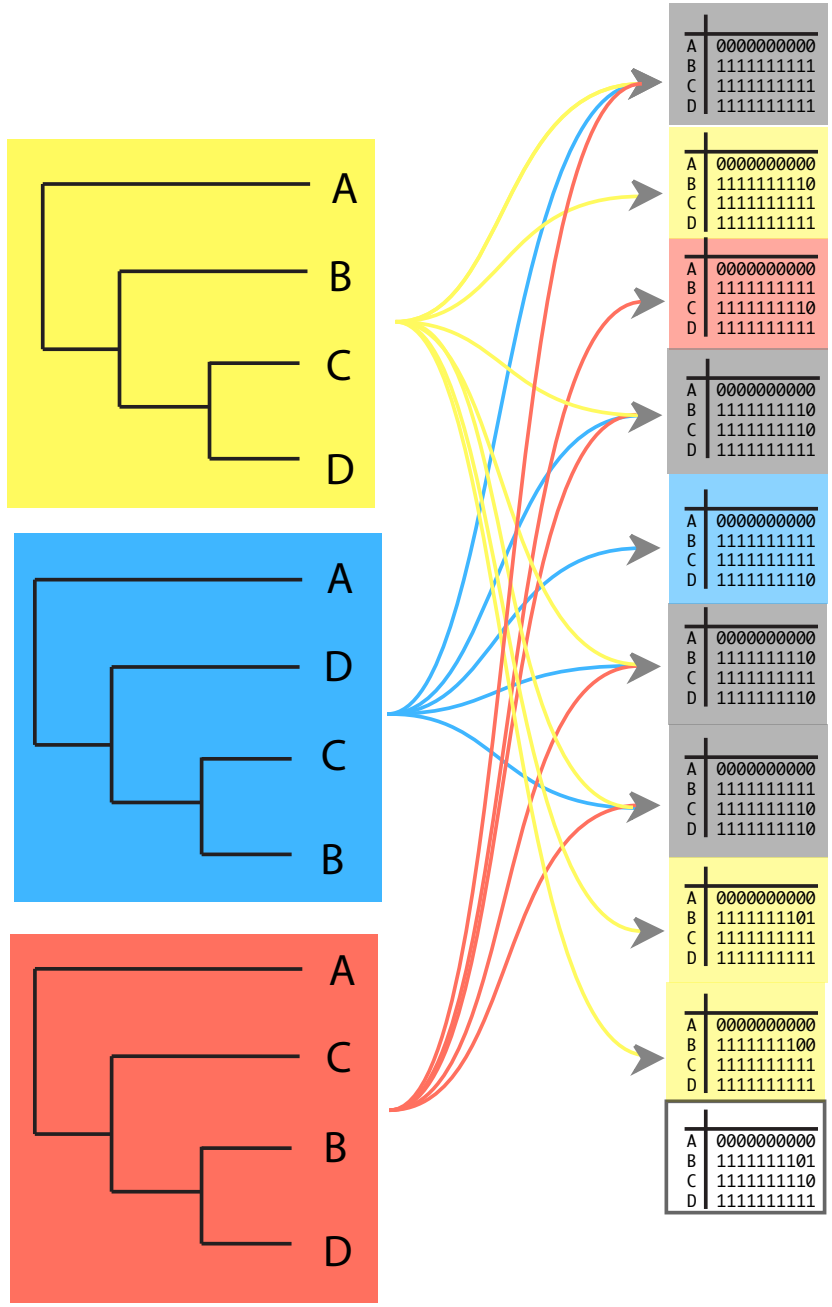
Incompatible characters are evidence of homoplasy in the data

Character conflict

Two characters are compatible if they can both be mapped on the same tree so that all of the character states displayed could be homologous.

Incompatible characters are evidence of homoplasy in the data Homoplasy

literally means the "same change" has occurred more than once in the evolutionary history of the group. The presence of homoplasy undermines Hennigian and Parsimony analyses.


A
B
C
D
A
D
C
B
A
C
B
D
A 0000000000
B 1111111111
C 1111111111
D 1111111111
A 0000000000
B 1111111110
C 1111111111
D 1111111111
A 0000000000
B 1111111111
C 1111111110
D 1111111111
A 0000000000
B 1111111110
C 1111111110
D 1111111111
A 0000000000
B 1111111111
C 1111111111
D 1111111110
A 0000000000
B 1111111110
C 1111111111
D 1111111110
A 0000000000
B 1111111111
C 1111111110
D 1111111110
A 0000000000
B 1111111101
C 1111111111
D 1111111111
A 0000000000
B 1111111100
C 1111111111
D 1111111111
A 0000000000
B 1111111101
C 1111111110
D 1111111111

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- can we detect character conflict?
- is there a logic-based solution to the problem of character conflict?

Detecting character conflict in binary characters

Consider the four possible combinations of states in a two-character matrix.

The characters are incompatible *iff* (when you look across all taxa) you see all four state combinations.

		Char 1	
		0	1
Char 2	0	×	×
	1	×	×

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- Can we detect character conflict? Yes
- Is there a logic-based solution to the problem of character conflict? No, nothing purely based on logic (and the suggestions for culling data to make matrices suitable for logical inference can lead to unsatisfyingly subjecive analyses).
- What can we do?

We must have an "error model"

In this class we will focus on Maximum Likelihood and Bayesian statistical estimates for evolutionary models.