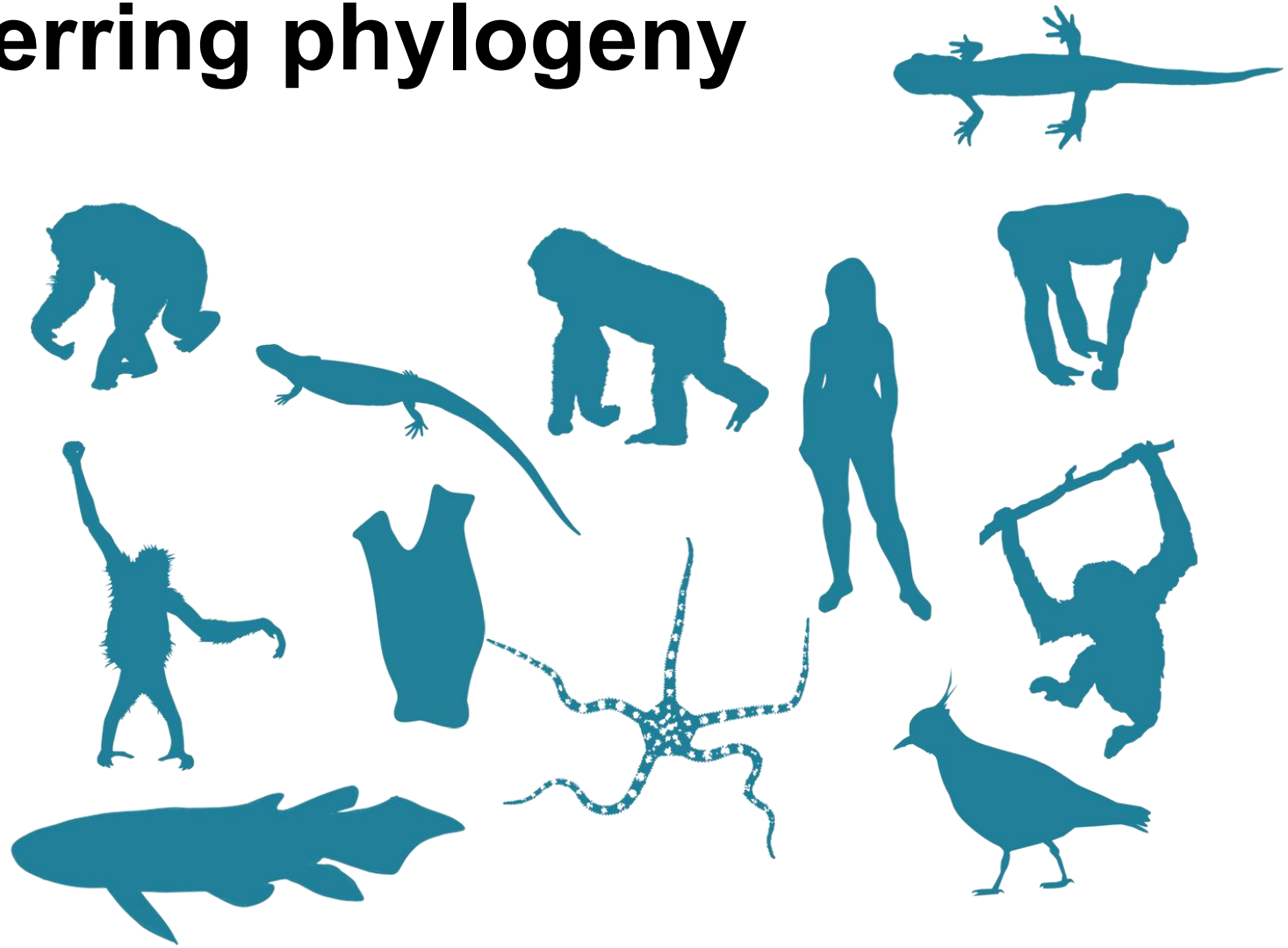
Going backwards- Inferring phylogeny



How many possible trees for 4 taxa?

Tree space is huge!!

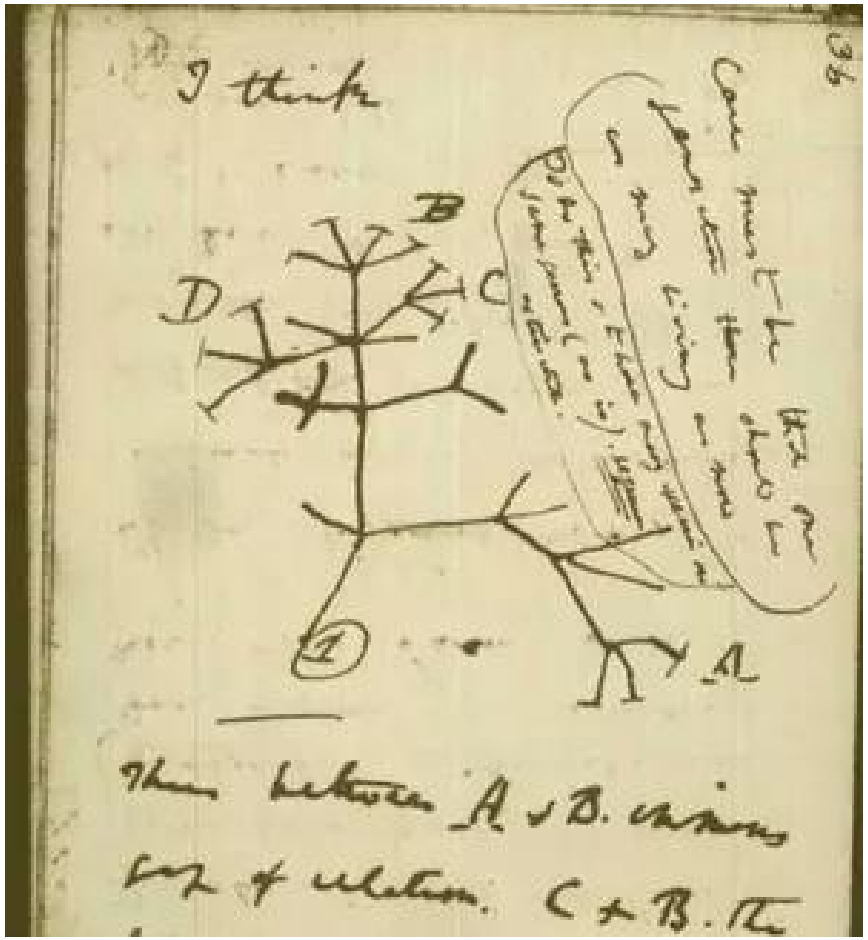
= Estimating phylogeny is a very hard problem

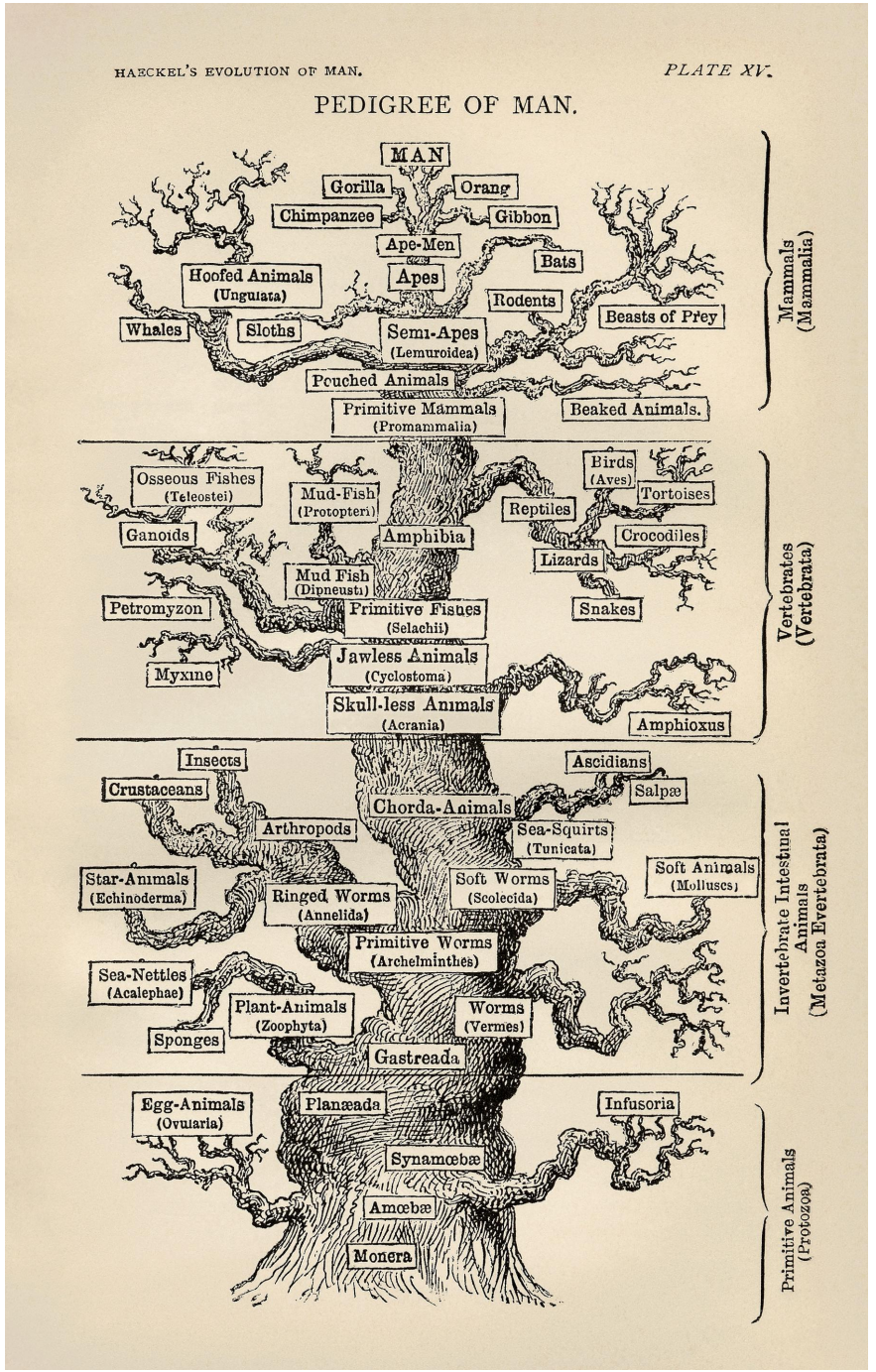
Table 3.1: The number of rooted, bifurcating, labeled trees for n species, for various values of n. The numbers for more than 20 species are approximate.

Species	Number of trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	*34,459,425*
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

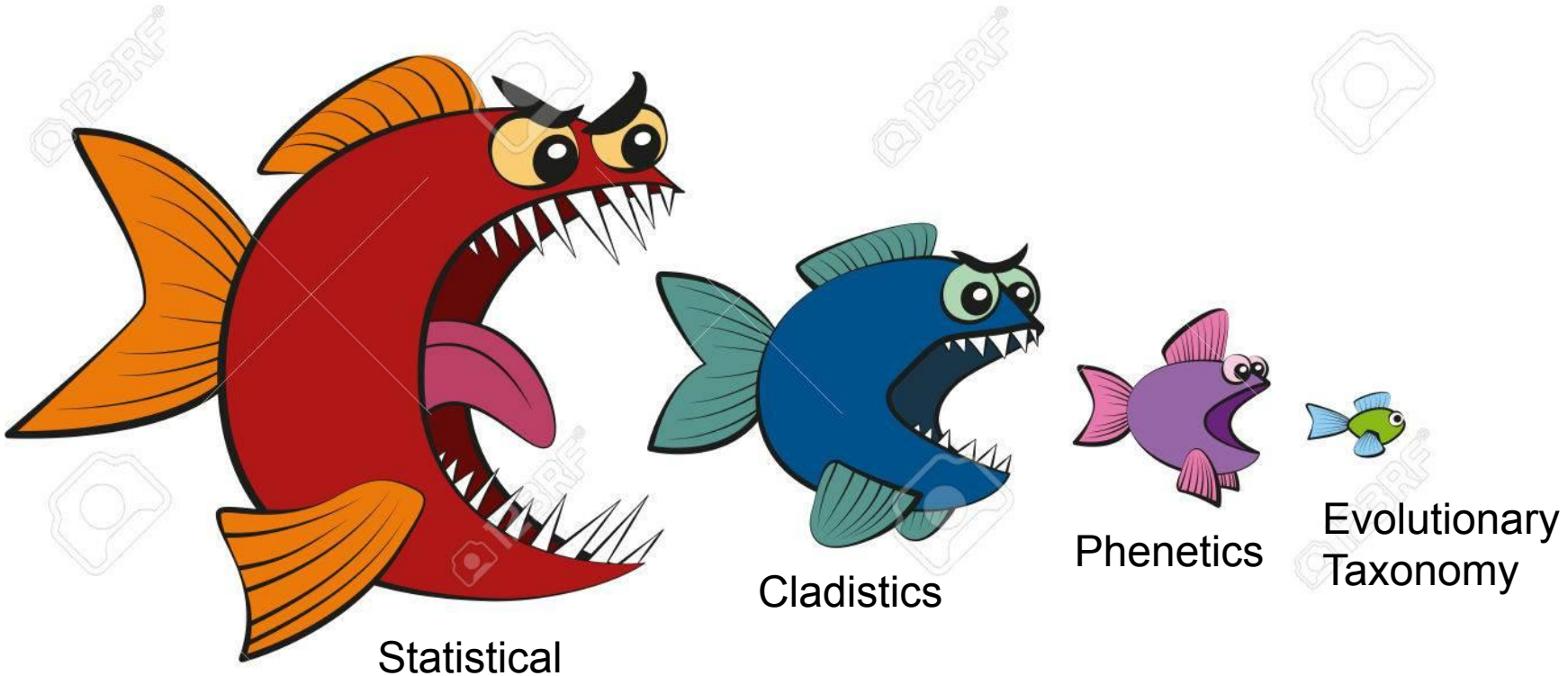
How can we find the “true” tree?
(Is there a true tree?!?)

History of phylogenetic inference





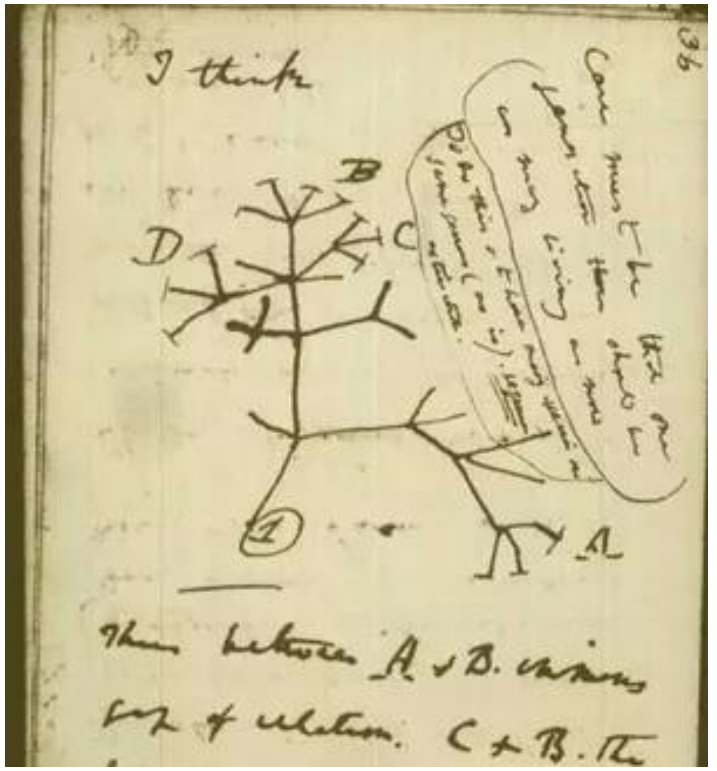
History of phylogenetic inference



History of phylogenetic inference: Evolutionary Taxonomy

No objective methodology

Relies on expert opinion & a lifetime of study



History of phylogenetic inference: Phenetics

Numerical Taxonomy & distance methods - measure overall similarity

How do you measure similarity?

What are the drawbacks of this approach?

Sokal, Sneath & Rohlf (1960's) argued phenetics provided much needed reproducibility

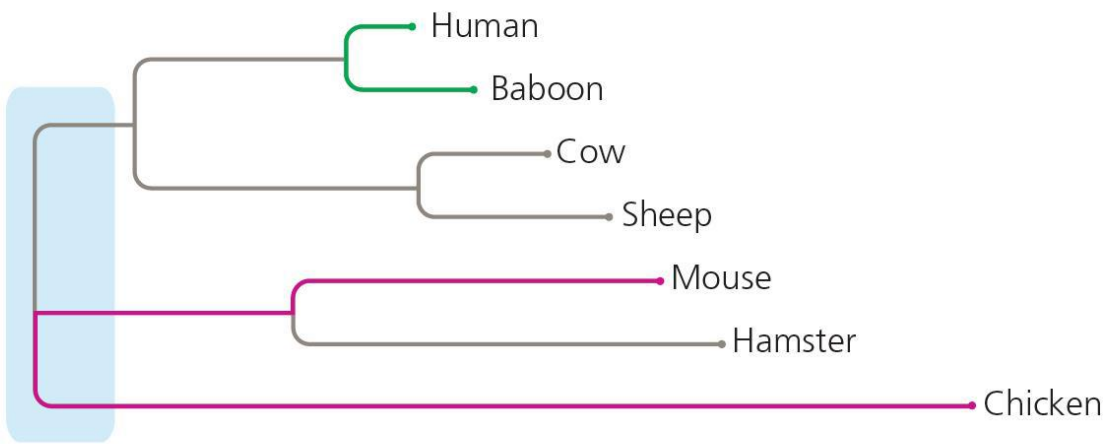
A

```
Human    M V H L T P E E K S A V T A L W G K V N V D E V G G E A L G R L L V V Y P W T Q R F F E S F G D L S
Baboon   • • • • • • • • • N • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • D • • • • •
Cow      • – – • • A • • • A • • • • F • • • • K • • • • • • • • • • • • • • • • • • • • • • • • • • • • • •
Sheep    • – – • • A • • • A • • • G F • • • • K • • • • • A • • • • • • • • • • • • • • • • • • H • • • • •
Mouse    • • • • • D A • • A • • S G • • • • • • A • • • • • • • • • • • • • • • • • • • • Y • D • • • • • •
Hamster  • • • • • D A • • A L • • G • • • • • • A • A • • A • • • • • • • • • • • • • • • • • • H • • • • •
Chicken  • • • W • A • • • Q L I • G • • • • • • • A • C • A • • • A • • • I • • • • • • • • • • A • • • N • •
```

B

	Baboon	Cow	Sheep	Mouse	Hamster	Chicken
Human	2	6	9	8	9	13
Baboon		7	10	7	10	13
Cow			3	11	12	16
Sheep				12	9	15
Mouse					7	16
Hamster						14

C



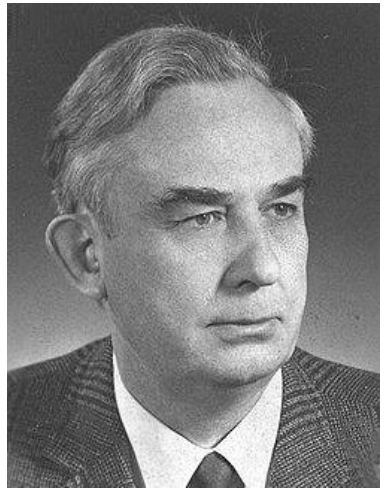
History of phylogenetic inference: Cladistics

"Phylogenetic systematics"

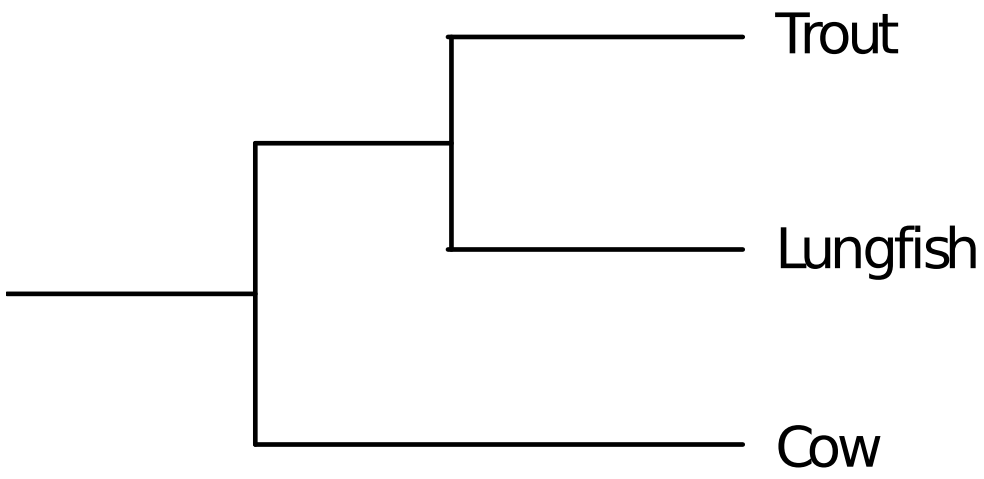
Argued that grouping should not be based on overall trait similarity, but only on shared derived traits (synapomorphies)

Taxonomy based on monophyly

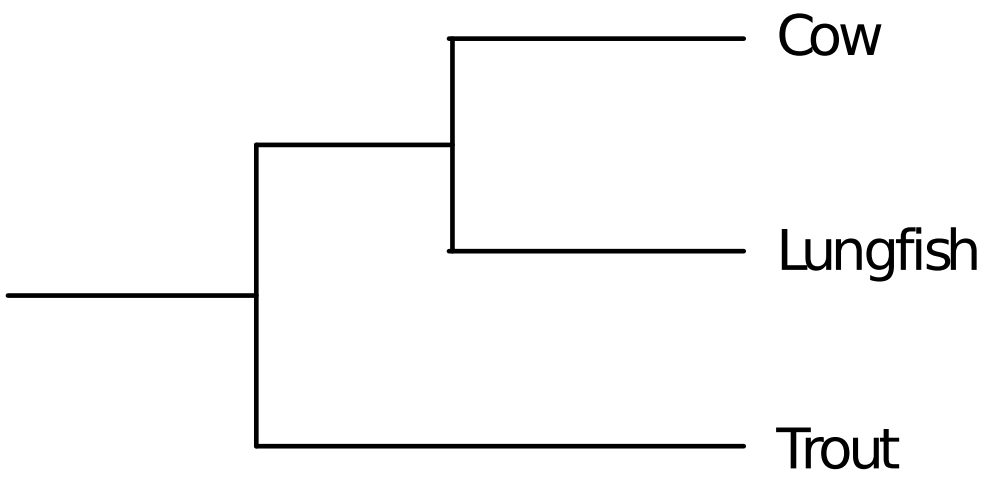
Parsimony & Occam's Razor- The simplest explanation is probably the correct one



Willi Hennig
(1966)

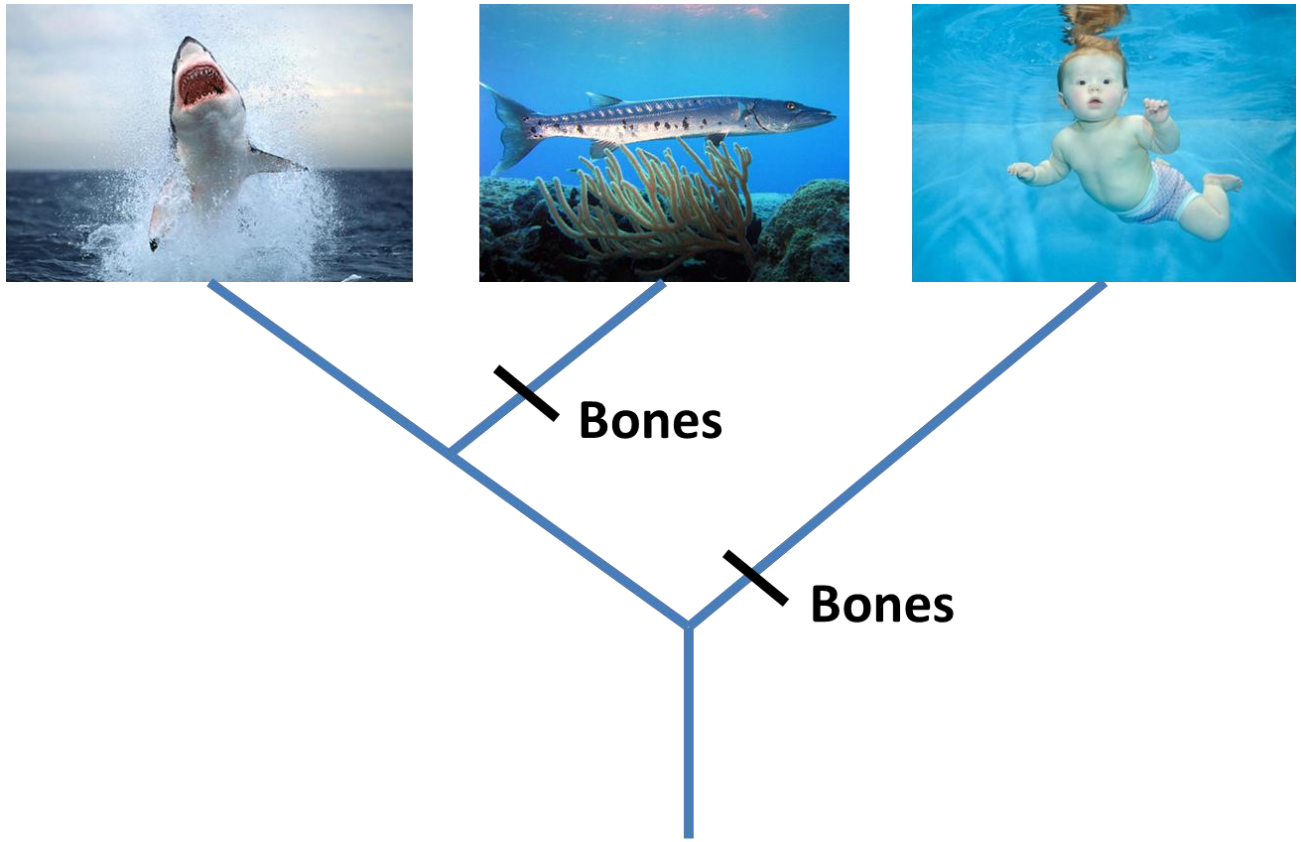


Phenogram

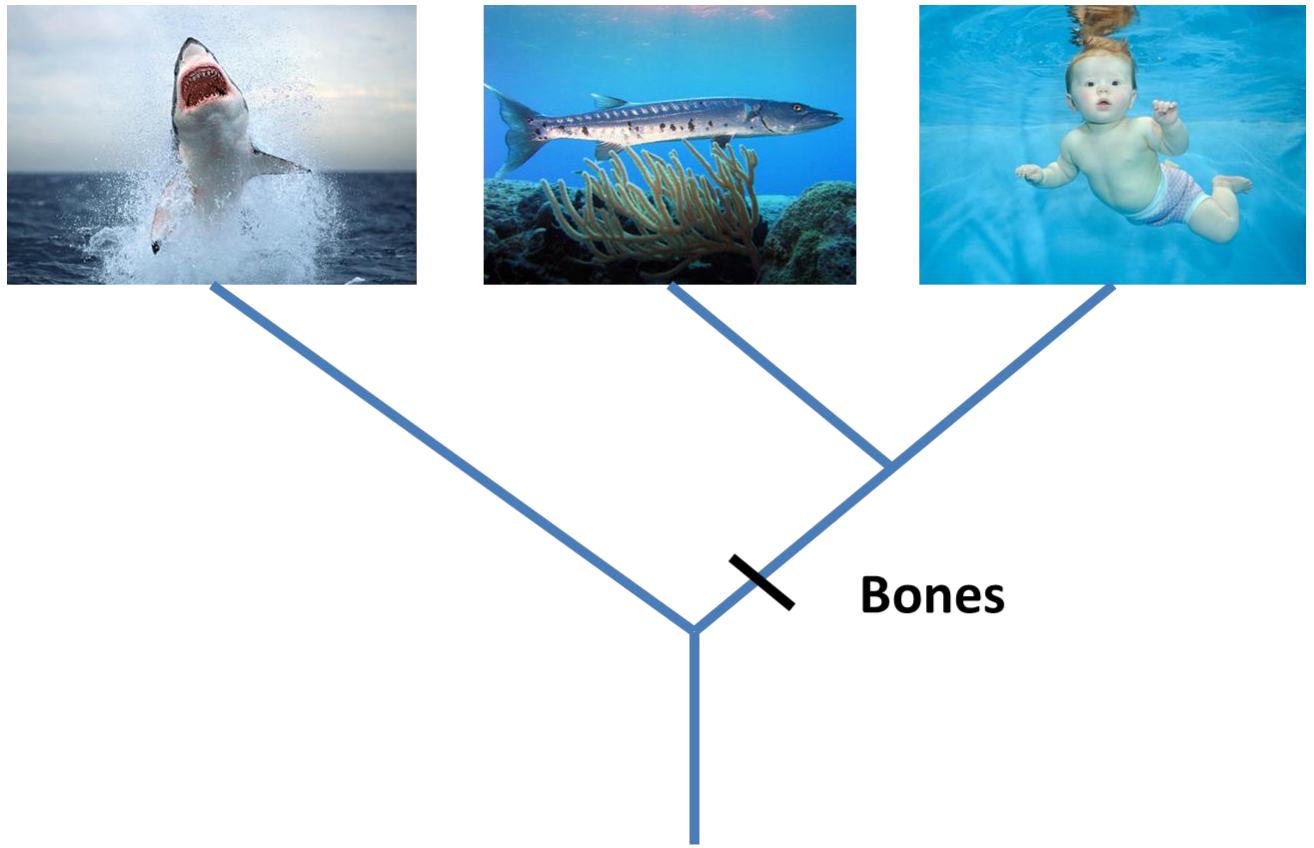


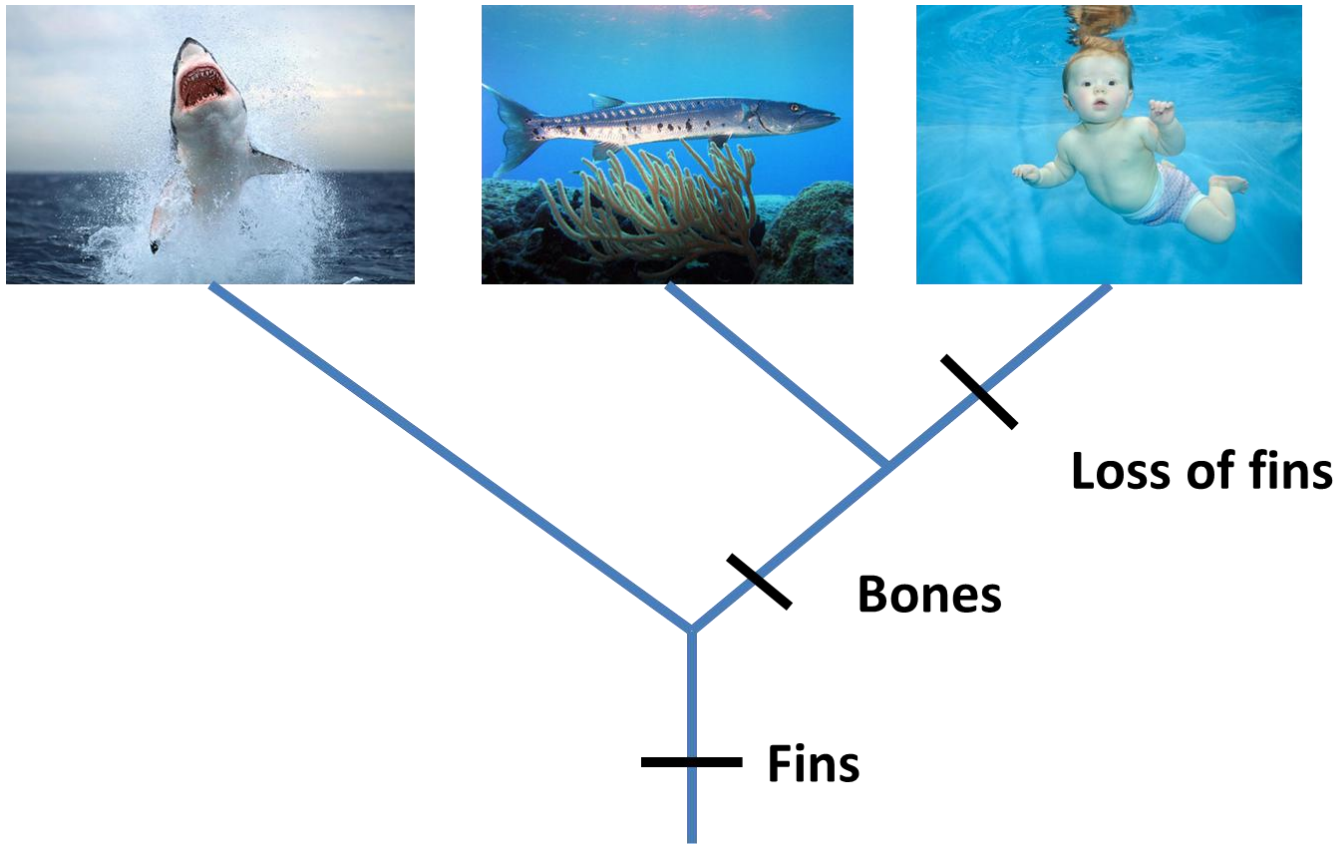
Cladogram

What’s the phylogenetic relationship?

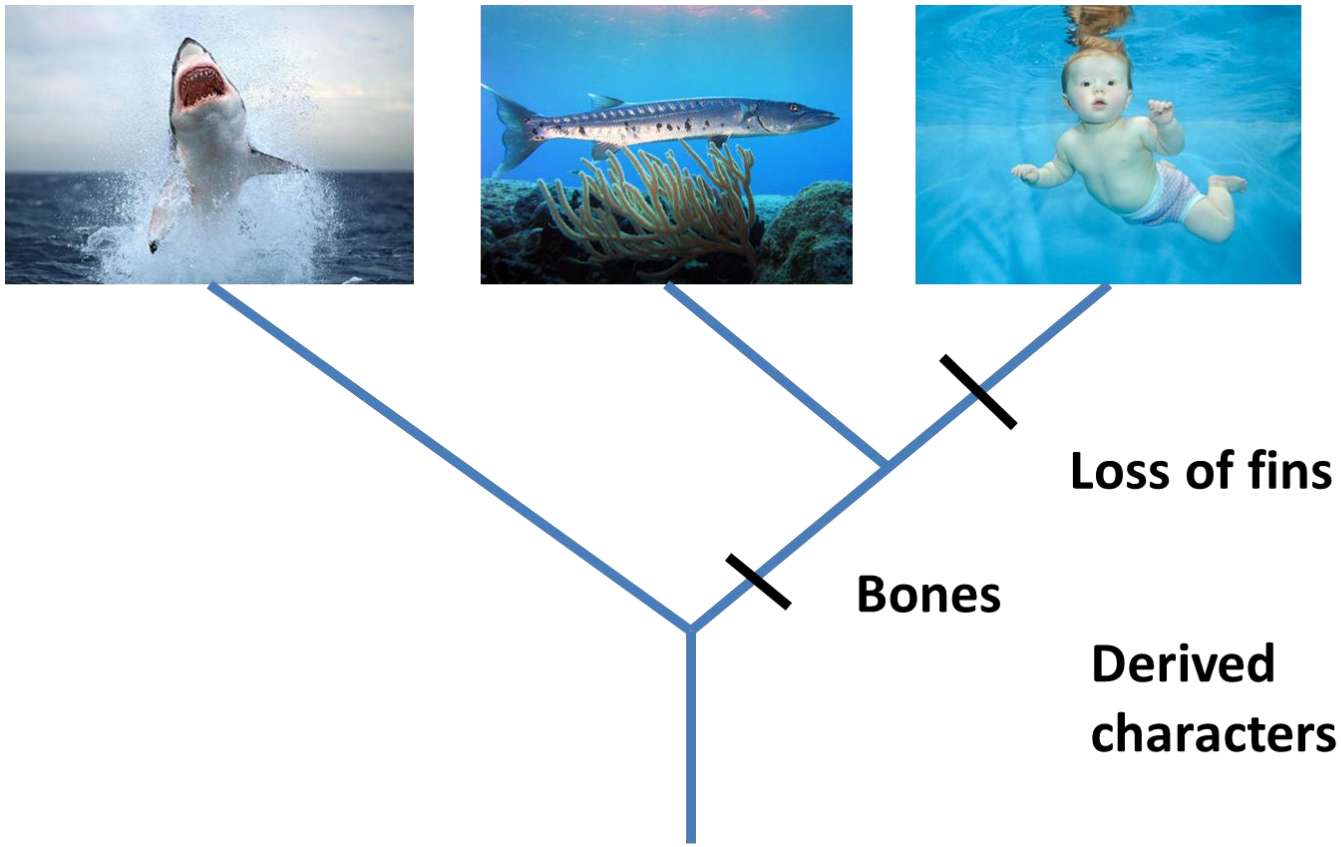

Bones
Bones

Parsimony

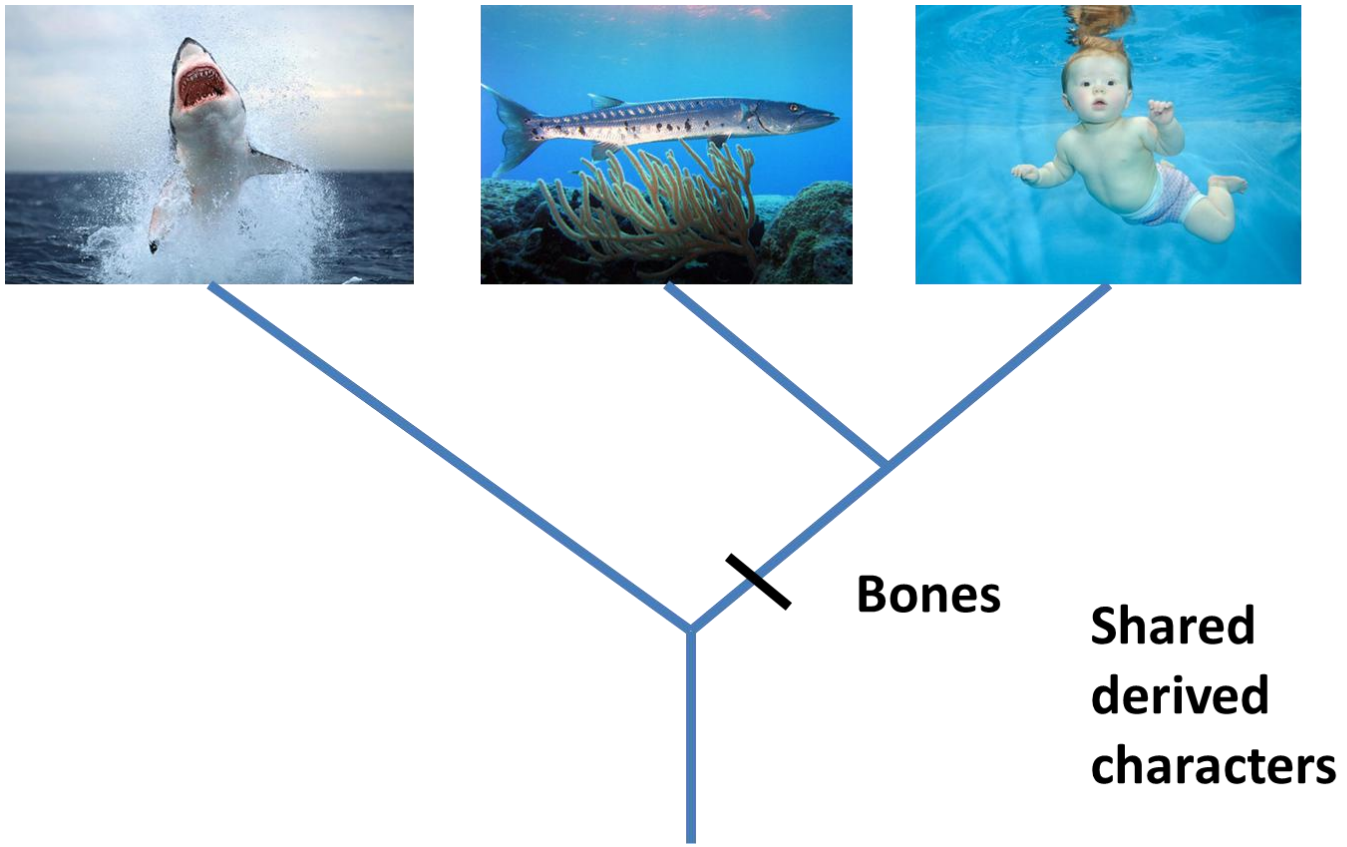



Loss of fins
Bones
Fins

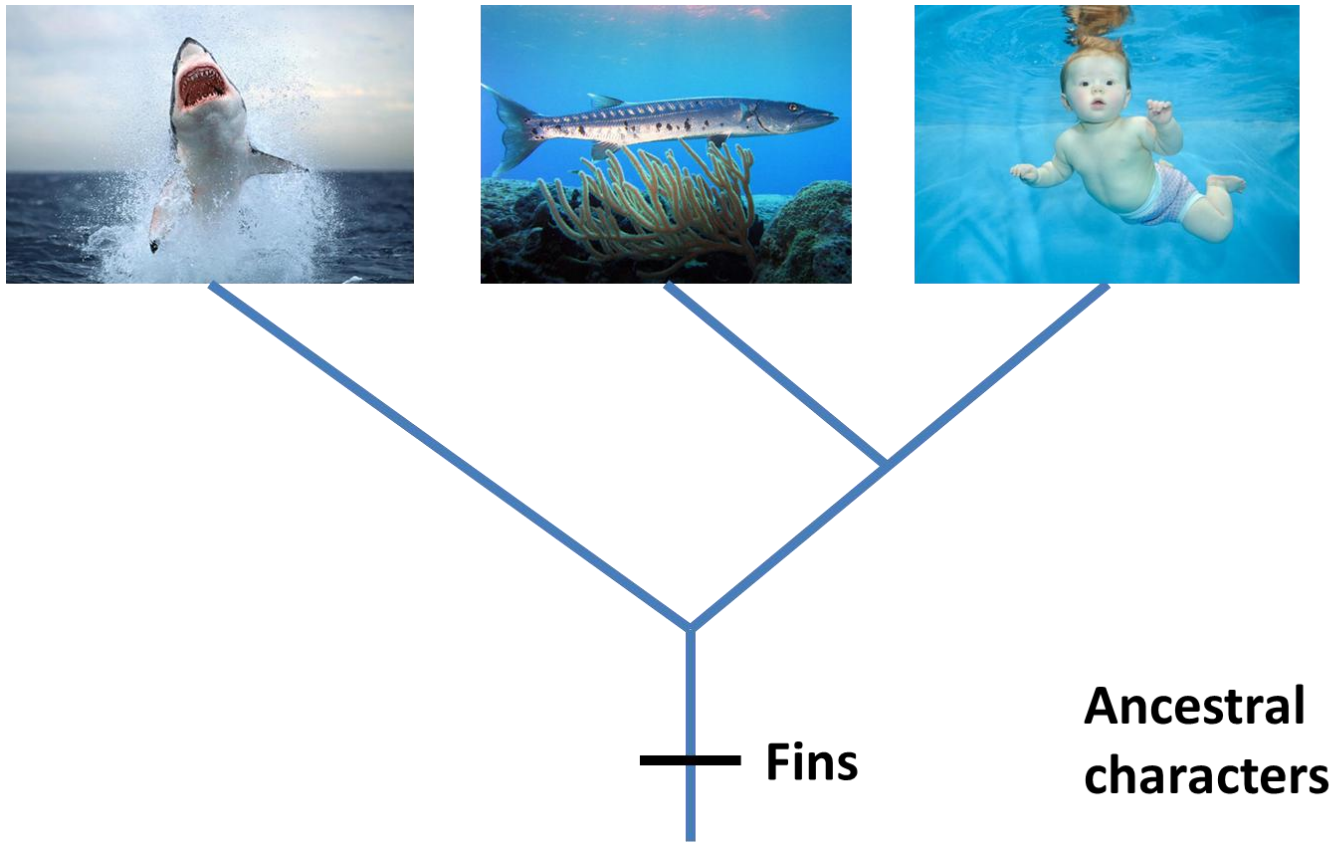
Derived Characters = Apomorphies



Shared derived characters = Synapomorphies



Shared ancestral characters = Symplesiomorphies



Overall similarity takes all characters equally

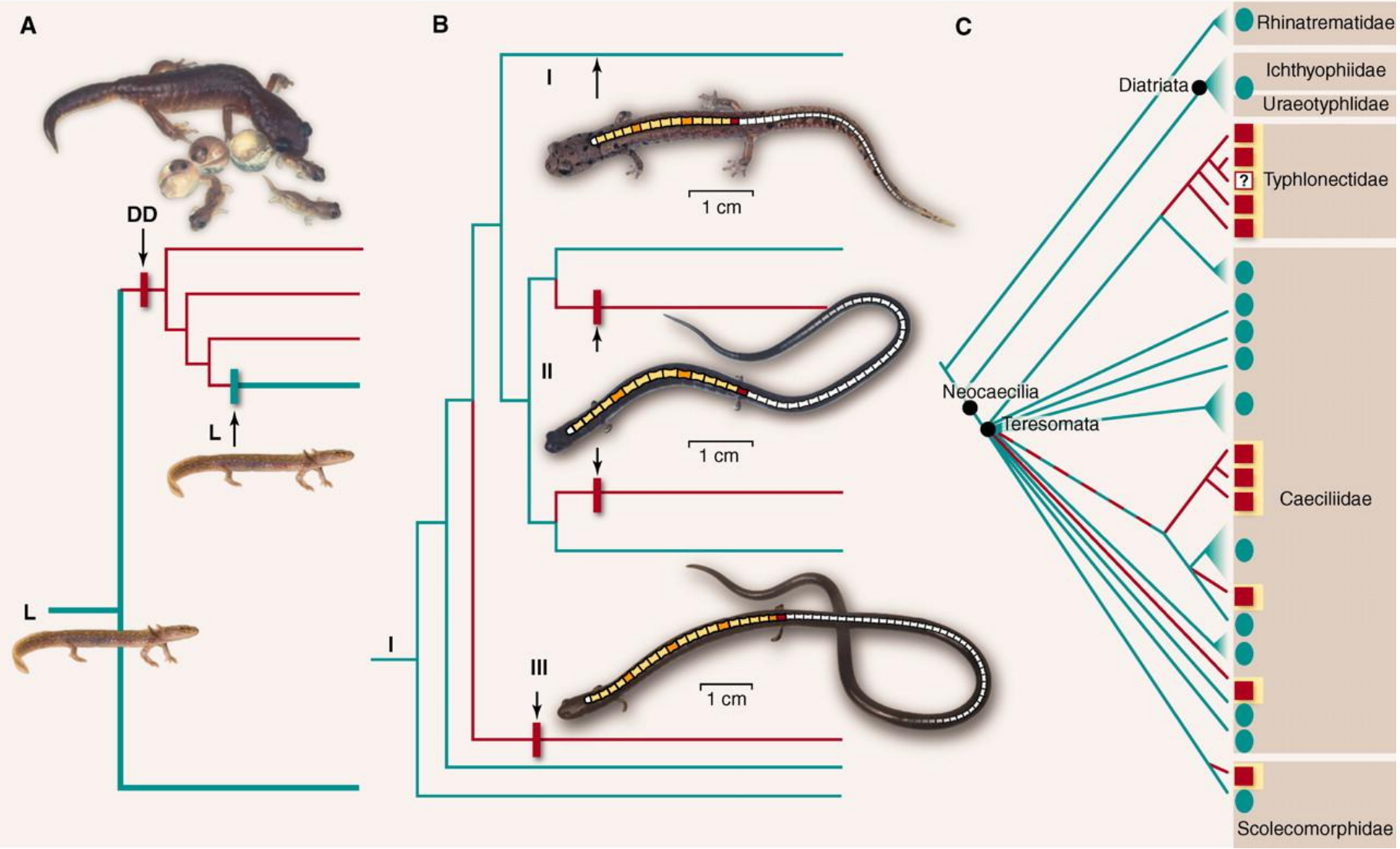
Cladistics only considers *synapomorphies*, and excludes *autapomorphies & symplesiomorphies*

Best supported hypothesis: Minimizing # of steps

Critical role of outgroup to distinguish derived vs. ancestral characters

Also note that characters are defined as derived/ancestral relative to the taxa being considered (the ingroup)

Conflict = Homoplasy
Evolutionary reversals and/or convergence



Parsimony informative vs. uninformative

Must have at least 2 taxa with shared traits in 2 different states.

TABLE 7.7 Examples of informative and uninformative characters, with ? used to represent uncertain or missing character states.

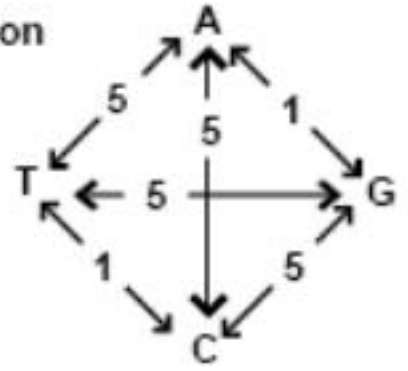
Taxon	Characters													
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	1	0	?	1	0	1	0	0	0	1	0
3	0	1	1	2	1	0	2	0	1	1	0	0	2	0
4	1	1	1	1	?	1	2	0	1	0	1	1	2	1
5	1	1	2	3	1	1	0	0	1	2	?	2	2	2
6	1	0	2	2	1	?	1	0	1	0	?	3	2	?

Generalized Parsimony

Step matrices

	To	Ordered				Unordered				Dollo				Camin-Sokal			
		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
F	0	-	1	2	3	-	1	1	1	-	M	2M	3M	-	1	2	3
r	1	1	-	1	2	1	-	1	1	1	-	M	2M	i	-	1	2
o	2	2	1	-	1	1	1	-	1	2	1	-	M	i	i	-	1
m	3	3	2	1	-	1	1	1	-	3	2	1	-	i	i	i	-

Transition/Transversion Parsimony



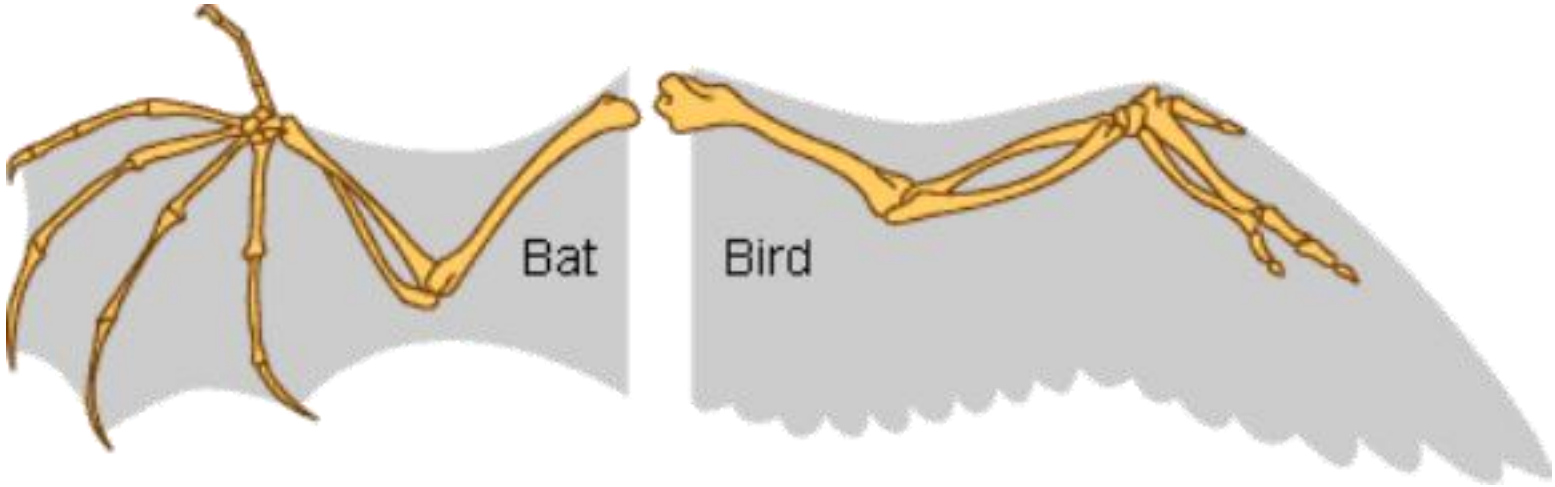
	A	C	G	T
A	-	5	1	5
C	5	-	5	1
G	1	5	-	5
T	5	1	5	-

By Peter Forey
https://www.palass.org/publications/newsletter/cladistics-palaeontologists/cladistics-palaeontologists-part-4-optimisation

Homology

Primary homology - Initial hypothesis of homology due to similarity

Secondary homology - Implied homology single origin on phylogeny by parsimony analysis



Underlines importance of character construction - often surprisingly subjective
"Reciprocal illumination" Hennig 1966

History of phylogenetic inference: Statistical phylogenetics

Luca Cavalli-Sforza, Anthony Edwards and later, Joseph Felsenstein

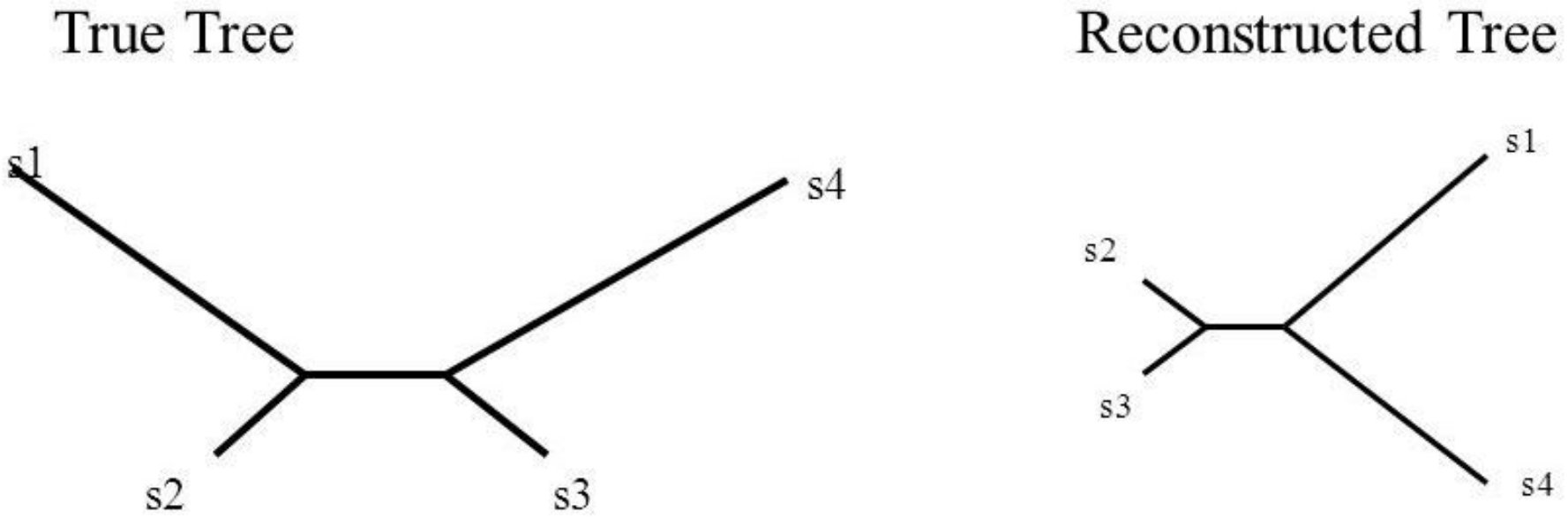
Probabilistic models of trait evolution

Showed statistical issues with parsimony, e.g. "the Felsenstein zone"

Maximum Likelihood & Bayesian inference

Joe Felsenstein

The Felsenstein Zone & "Long-branch attraction"



More generally, parsimony fell out of favor for molecular data due to more flexibility in incorporating substition models in statistical phylogenetics

My oversimplification of the problem...

Under the cladistics philosophy (model), homoplasy is a “mistake” in character construction

Coding independently evolved traits the same, look closer and you'll see they're different!! (e.g. wings of birds & wings of bats)

But for DNA, complete convergence is real (a “T” is a “T” is a “T”, no matter how it became that)

As similarity by convergence becomes more common than similarity by shared ancestry, parsimony fails, and no “looking closer” at the character will save it.

Parsimony main assumption “Occam's Razor” has unclear implications for what this means for the assumed evolutionary process

Ultimately -> rare changes
or "no common mechanism" (Tuffley and Steele 1997)

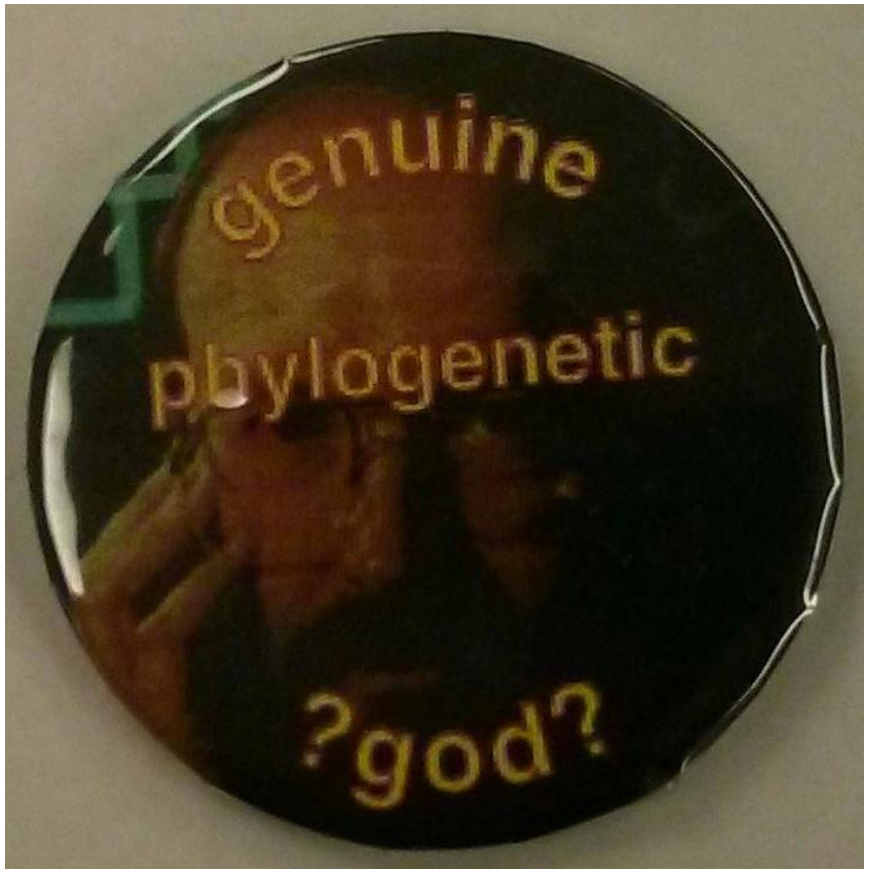
Inconsistent estimator - An estimator that gets more and more certain of the wrong answer as more data is added

Conflict in science

MATT SIMON SCIENCE 02.03.16 07:00 AM

TWITTER NERD-FIGHT REVEALS A LONG, BIZARRE SCIENTIFIC FEUD

#ParsimonyGate



The epistemological paradigm of this journal is parsimony. There are strong philosophical arguments in support of parsimony versus other methods of phylogenetic inference (e.g. Farris, 1983).

The high citation index of *Cladistics* shows that the journal is publishing some of the most ground-breaking empirical and theoretical research on the history of life, and we remain committed to the publication of outstanding systematics research. As a community of scientists, the Willi Hennig Society is always open to new methods and ideas, and to well-reasoned criticisms of old ones. However, we do not hold in special esteem any method solely because it is novel or purportedly sophisticated.

Phylogenetic data sets submitted to this journal should be analysed using parsimony. If alternative methods are also used and there is no difference among the results, the author should defer to the principles of the Society and present the tree obtained by parsimony. Unless there is a pertinent reason to include multiple trees from alternative methods, a tree based on parsimony is sufficient as an intelligible, informative and repeatable hypothesis of relationships, and articles should not be cluttered with multiple, often redundant, trees produced from other methods. If alternative methods give different results and the author prefers an unparsimonious topology, he or she is welcome to present that result, but should be prepared to defend it on philosophical grounds.

In keeping with numerous theoretical and empirical discussions of methodology published in this journal, we do not consider the hypothetical problem of statistical inconsistency to constitute a philosophical argument for the rejection of parsimony. All phylogenetic methods, including parsimony, may produce inconsistent or otherwise inaccurate results for a given data set. The absence of certain truth represents a philosophical limit of empirical science.

Cladistics will publish research based on methods that are repeatable, clearly articulated and philosophically sound. We believe these guidelines implement the vision of Willi Hennig (1965, p. 97), who said, "(i)nvestigation of the phylogenetic relationship between all existing species and the expression of the results of this research in a form which cannot be misunderstood, is the task of phylogenetic systematics."

Parsimony still commonly used for:

Morphological analyses (though probabilistic models exist here too)

Starting trees

Understanding history of trait change

Practical considerations

Heuristic searches

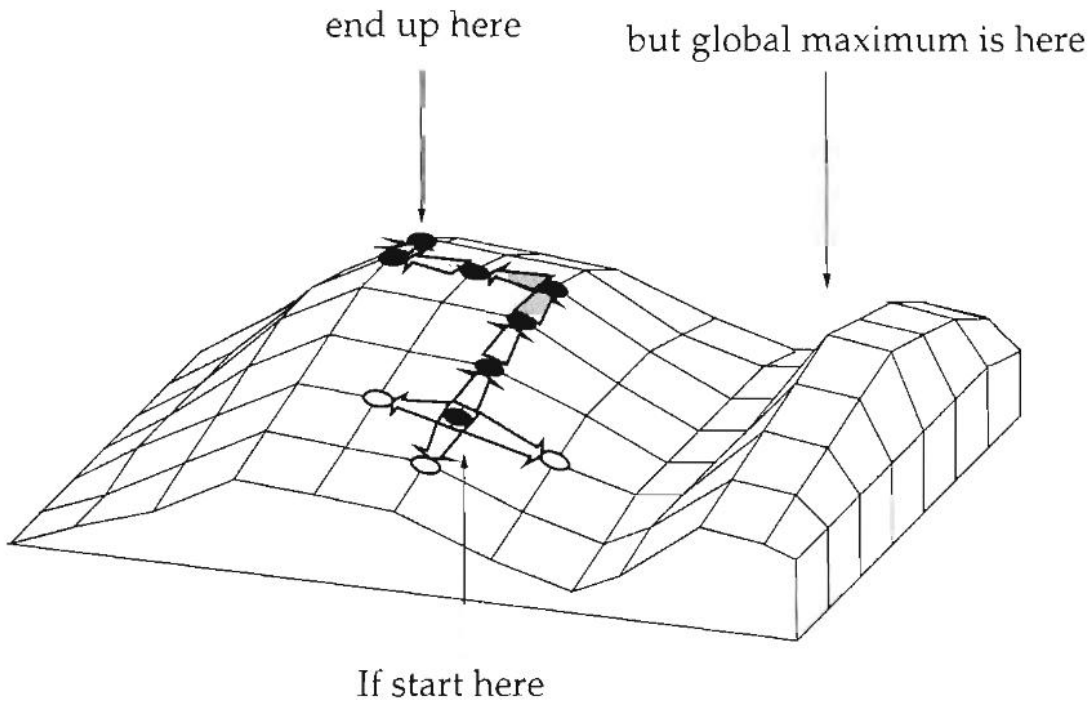


Figure 4.1: A surface rising above a two-dimensional plain (or plane). The process of climbing uphill on the surface is illustrated, as well as the failure to find a higher peak by this "greedy" method.

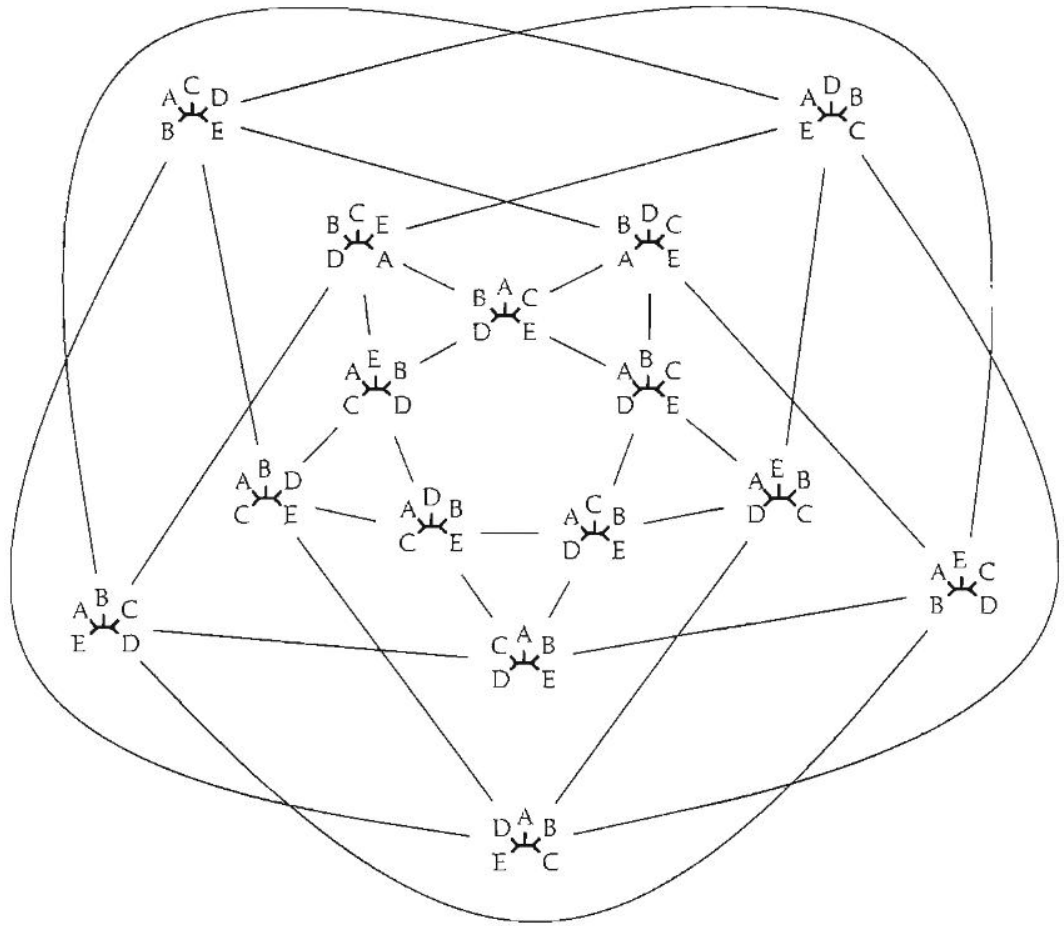


Figure 4.3: The space of all 15 possible unrooted trees with 5 tips. Neighbors are connected by lines when a nearest-neighbor interchange can convert one into the other. The labels A–E correspond to the species names Alpha through Epsilon in that data set. This symmetric arrangement of nodes was discovered by Ben Rudd Schoenberg (personal communication), and we thus denote this graph the Schoenberg graph.

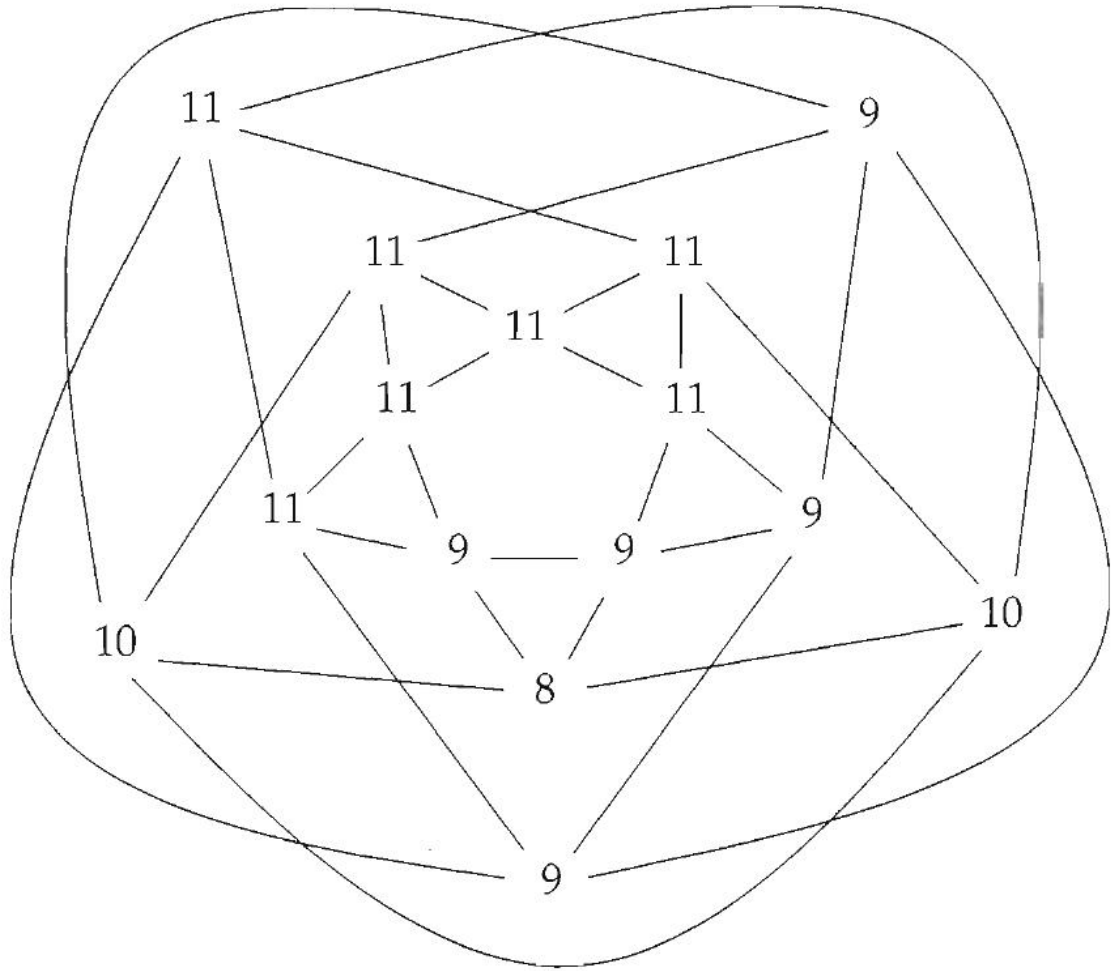
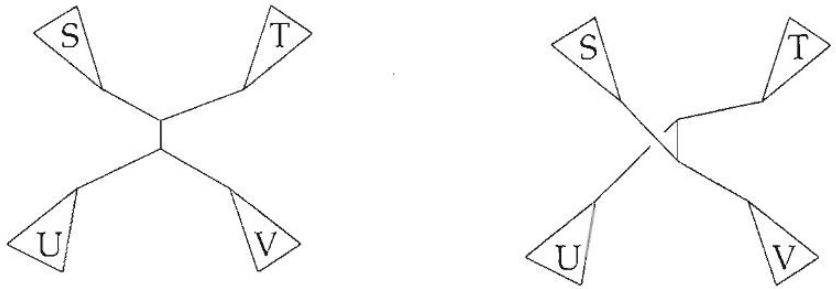
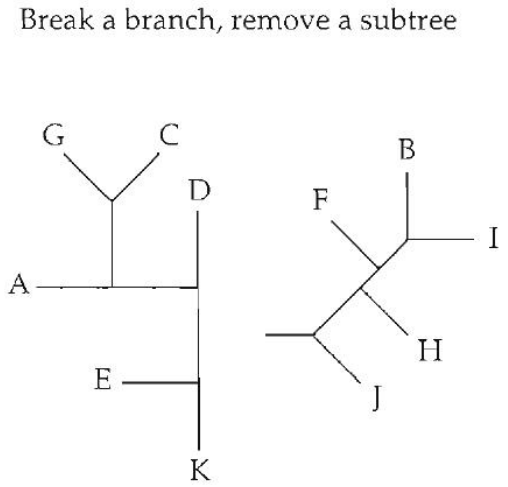


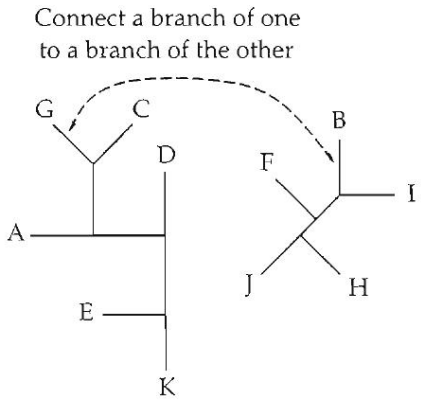
Figure 4.4: The space of all 15 possible trees, as in Figure 4.3, where the number of changes of state on the data set of Table 1.1 is shown. Nearest-neighbor interchanges search for the most parsimonious tree by moving in this graph.

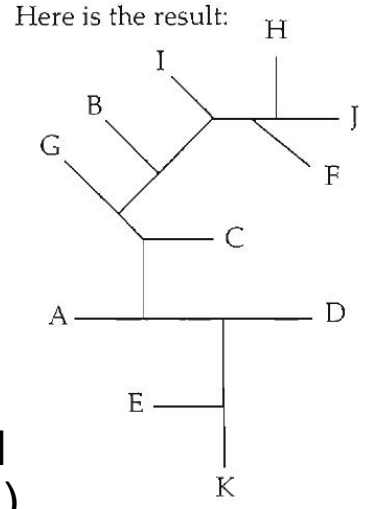
Tree space "moves"



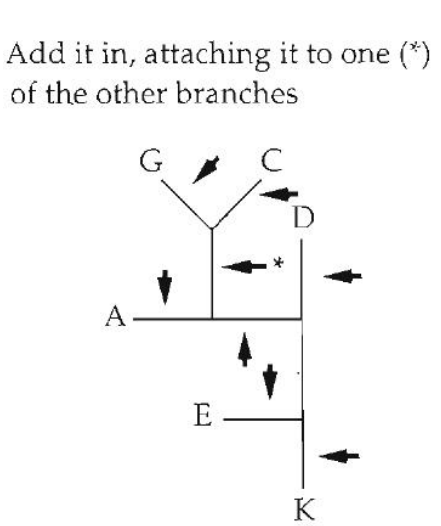
Nearest Neighbor Interchange (NNI)

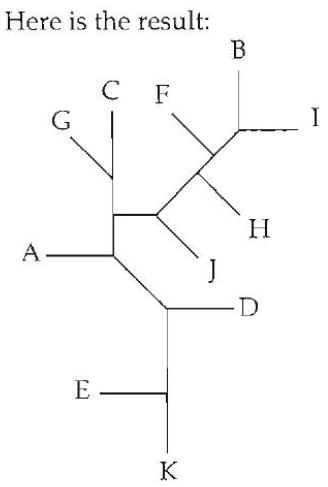






Tree bisection and reconnection (TBR)





Subtree pruning and regrafting (SPR)

Distance between tree spaces changes depending on which moves you use

NNI fastest, least computationally intensive, but the largest distances between trees

TBR slowest, most computnationally intensive, but shortest distances between trees

SPR intermediate

Branch and bound

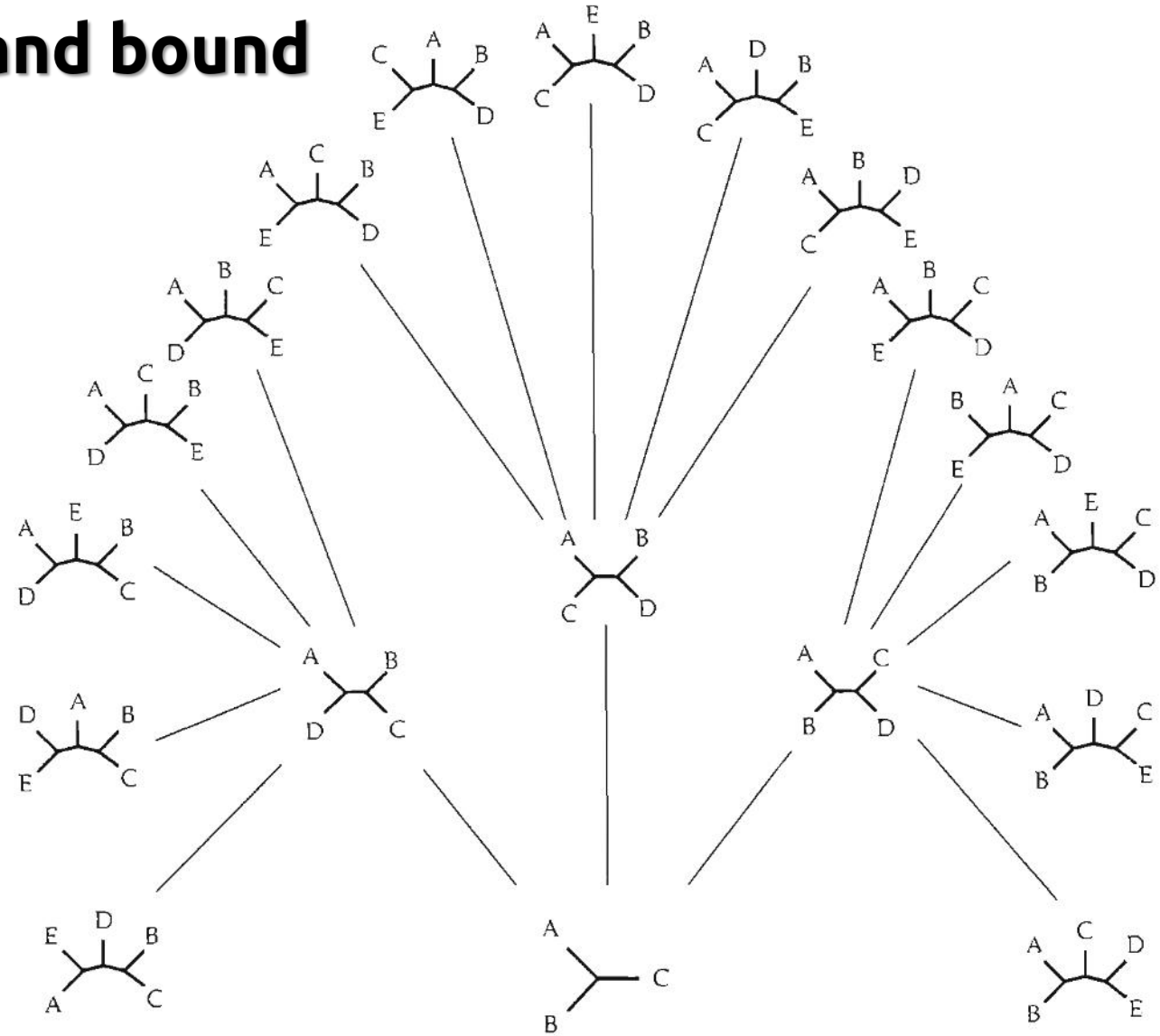


Figure 5.3: Search tree for most parsimonious tree in a five-species case.

Branch and bound

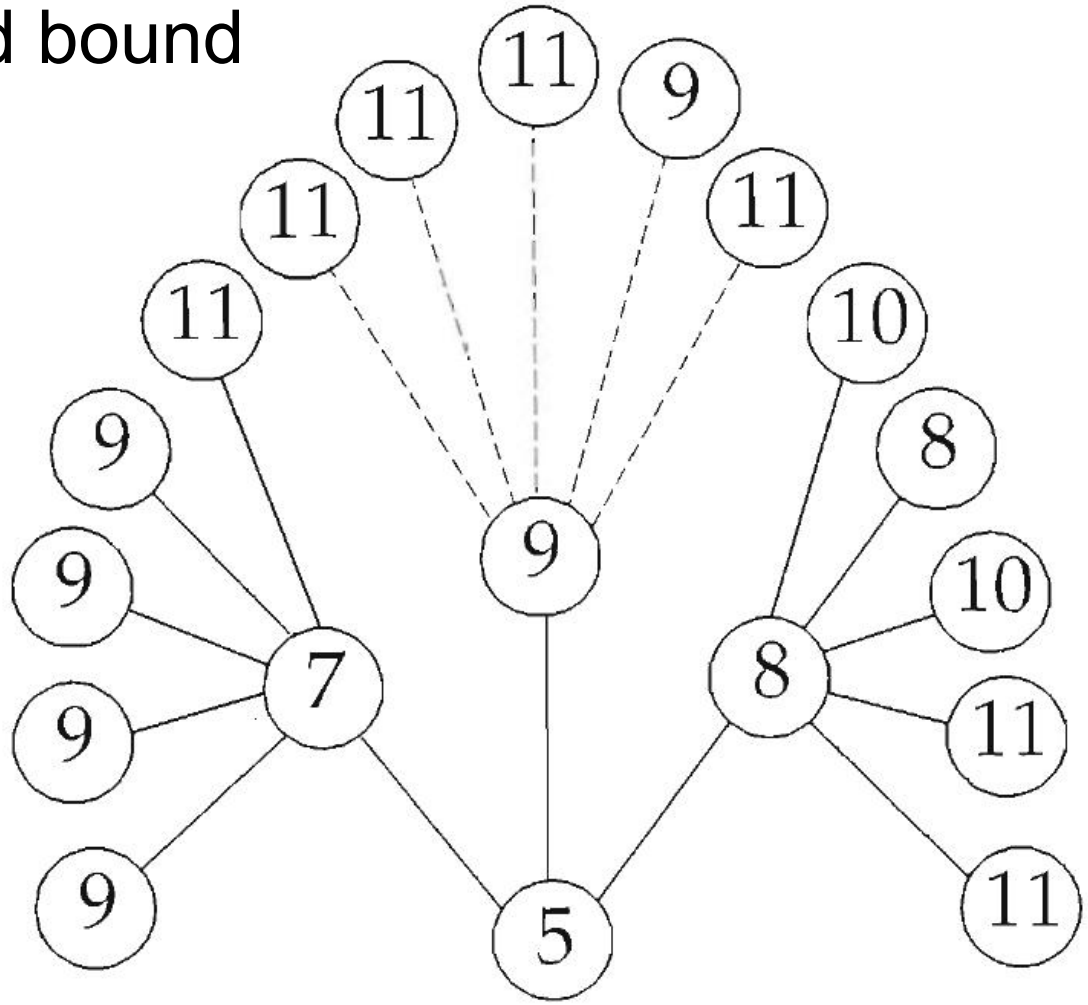


Figure 5.4: Search tree for most parsimonious tree for five species, using the data of Table 1.1. Trees are shown in Figure 5.3. Dashed lines are those not traversed by a branch and bound method. The species names in the data set correspond to labels A through E in Figure 5.3.

Software: PAUP*, TNT, Mesquite, others...