One-minute responses

- Q: Are co-speciation and coevolution the same thing?
- Coevolution is any reciprocal evolutionary interaction:
  - Two toxic butterflies converge on the same color pattern
  - Plant and pollinator adapt to each others' needs
- Co-speciation is specifically correlation between speciation patterns in two groups of species
  - Could come from coevolution
  - Could simply come from isolation (when hosts speciate they separate parasites)

Phylogeny methods

- Four major approaches to phylogeny inference
  - Parsimony
  - Distance
  - The statistically complex siblings:
    - Maximum likelihood
    - Bayesian inference

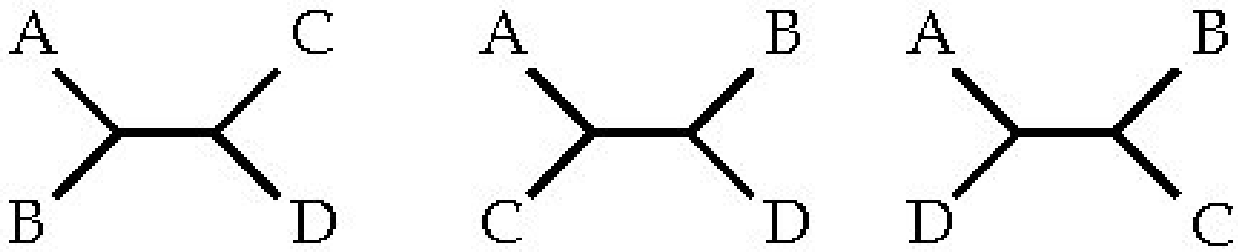
Parsimony methods

- (Philosophical) Principle of Parsimony: Make as few assumptions as possible
- (Phylogenetic) Principle of Parsimony: Prefer the tree requiring the fewest evolutionary changes
- Assumes that changes are *fairly rare* and *evenly distributed*

Parsimony methods

- Advantages of parsimony:
  - No explicit mutational model required
  - Applicable to the widest variety of data–including morphological traits (all we have for fossils)
  - Moderately fast
- Disadvantages:
  - No explicit mutational model possible
  - Long branch attraction
  - Limited ability to put error bars on phylogeny estimate

Practice problem–parsimony

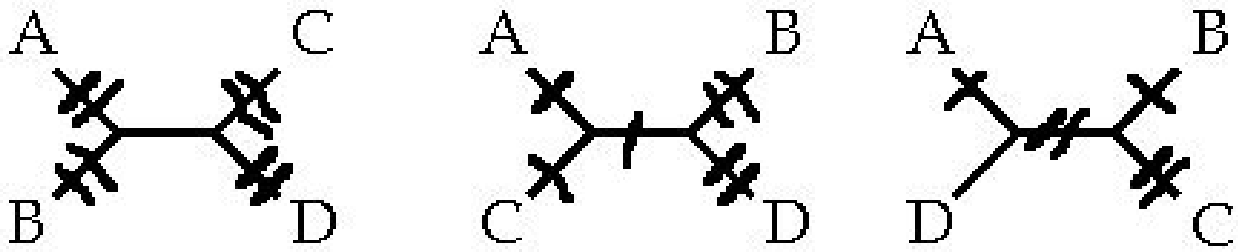


Taxon	1	2	3	4	5
A	A	A	C	G	A
B	T	A	A	T	T
C	T	A	A	G	A
D	A	C	C	G	T

How many changes are needed on each tree topology?

Which topology is preferred by parsimony?

Practice problem–parsimony



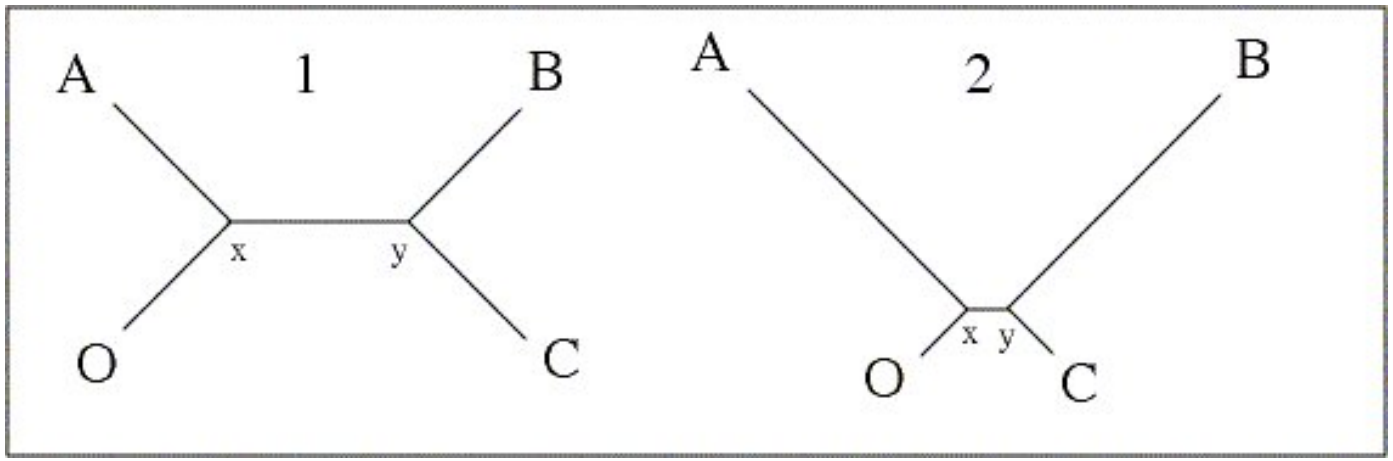
Taxon	1	2	3	4	5
A	A	A	C	G	A
B	T	A	A	T	T
C	T	A	A	G	A
D	A	C	C	G	T

How many changes are needed on each tree topology? *8, 7, 6*

Which topology is preferred by parsimony? *Third topology*

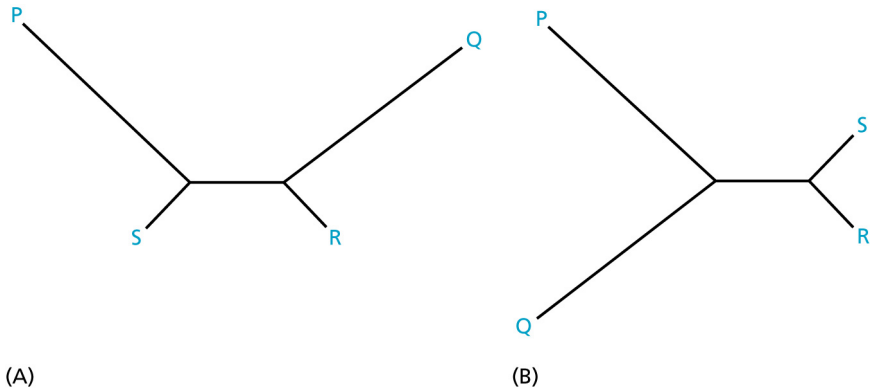
Parsimony methods

- Some trees give inconsistent results with parsimony
- "Inconsistent" means that results get worse with more data
- With infinite data you would be 100% sure to get the wrong answer
- (Research by Joe Felsenstein here at UW)



Long branch attraction

- When the data come from the left-hand tree, parsimony prefers the right-hand tree
- Two convergent changes on the long branches are more likely than a single change on the short branches
- This violates the basic principle of parsimony: prefer the solution with the fewest changes

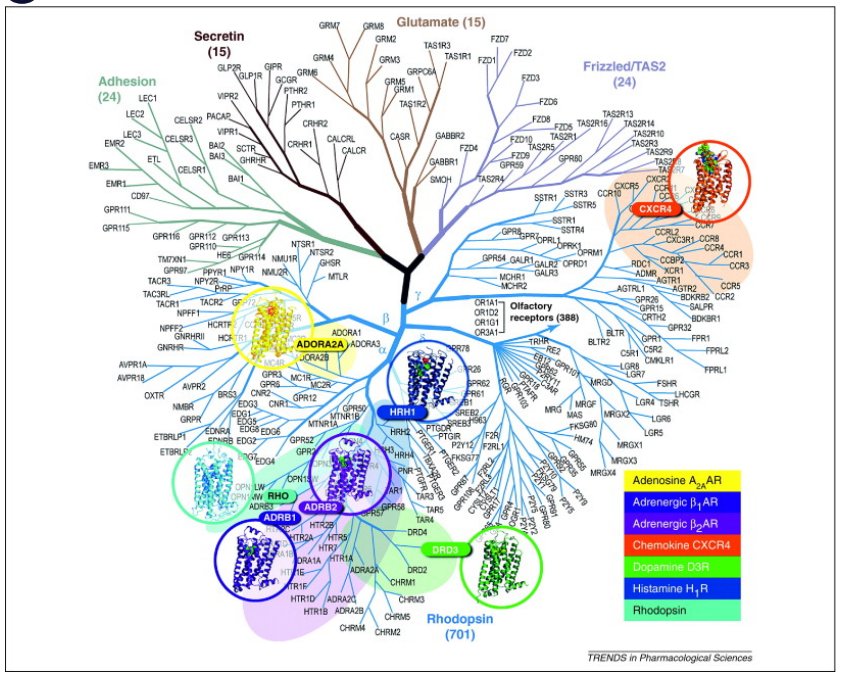


(A) (B)

Betting on your trees

- Ken Rice makes parsimony trees of human G-protein coupled receptors
  - Maximum likelihood much too slow
  - Distance methods didn't perform well
- If they group with:
  - Odor receptors – discard
  - Neurotransmitter receptors
    - spend $2K to validate

G-protein coupled receptor genes



Distance methods

- Transform data into a table of pairwise distances
- Find a tree which fits these distances well
- Different distance methods use different fitting criteria

	Human	Bonobo	Chimp	Gorilla	Orang
Human	–	4	5	8	12
Bonobo	4	–	1	9	14
Chimp	5	1	–	8	14
Gorilla	8	9	8	–	13
Orang	12	14	14	13	–

Distance methods

- For very sparse mutations, counting differences may be good enough
- If some sites have mutated multiple times, this will undercount changes on the longer branches
- Use a mutational model to correct the distances
- Various models available:
  - Transition/transversion bias
  - Unequal base frequencies
  - Rate variation
  - Invariant sites

UPGMA

- UPGMA (Unweighted Pair-Group Method of Analysis) is a simple distance method
- It assumes a molecular clock and is fragile if clock is wrong, so seldom used anymore
- Its non-clocklike sibling Neighbor-Joining performs better and is very widely used
- I teach UPGMA because it illustrates the principles and is easy

UPGMA rules

- Group together the two most similar species
- Divide their distance evenly across the branches leading to them
- Average their distances to all other species
- Rewrite the distance matrix with the new group and distances
- Repeat until tree is finished
- In case of ties, break arbitrarily or draw as three-way split

UPGMA example

	A	B	C	D	E
A	-	5	1	8	9
B	5	-	4	10	11
C	1	4	-	9	9
D	8	10	9	-	2
E	9	11	9	2	-

UPGMA example

	A	B	C	D	E
A	-	5	1	8	9
B	5	-	4	10	11
C	1	4	-	9	9
D	8	10	9	-	2
E	9	11	9	2	-

Group A and C to form AC, with branches of length 0.5

	AC	B	D	E
AC	-	4.5	8.5	9
B	4.5	-	10	11
D	8.5	10	-	2
E	9	11	2	-

UPGMA example

	AC	B	D	E
AC	-	4.5	8.5	9
B	4.5	-	10	11
D	8.5	10	-	2
E	9	11	2	-

Group D and E to form DE, with branches of length 1.0

	AC	B	DE
AC	-	4.5	8.75
B	4.5	-	10.5
DE	8.75	10.5	-

UPGMA example

	AC	B	DE
AC	-	4.5	8.75
B	4.5	-	10.5
DE	8.75	10.5	-

Group B with AC to form ABC, with branches of length 2.25

	ABC	DE
ABC	-	9.625
DE	9.625	-

UPGMA example

	ABC	DE
ABC	-	9.625
DE	9.625	-

Group ABC with DE, with branches of length 4.80

Distance methods

- Advantages:
  - Very fast
  - Can use sophisticated mutational model to obtain distances
  - Can be used for data that are intrinsically distances (DNA annealing temperature, immunological cross-reactivity)
- Disadvantages:
  - Loss of information by reducing data to distances
  - Clocklike versions (UPGMA) are brittle
  - Long distances hard to estimate accurately

Maximum-likelihood methods

- Begin with an explicit model of evolution
- Evaluate each candidate tree:
  - How probable are the data given this tree and model of evolution?
  - What are the best branch lengths on this tree to explain these data?
- Can't try all possible trees, so heuristics used to find good trees
- Developed in this department by Joe Felsenstein around 1981

Maximum-likelihood methods

- Advantages:
  - Can use sophisticated mutational models
  - Gives approximate error bars for branch lengths
  - Makes full use of all information in the data

- Disadvantages:
  - Exposes its mutational model, which can then be criticized (they are always oversimplifications)
  - Extremely slow

Bayesian methods

- Begin with an explicit model of evolution
- Wander among possible trees in proportion to their fit to the data
- Result is a cloud of trees
- To assess any given feature, count how often it appears in the cloud
- Example: Where is root of human mtDNA tree?

Bayesian methods

- Advantages
  - Can use sophisticated mutational models
  - Excellent error bars (which parts of the tree can we trust?)
  - Makes full use of all information in the data
- Disadvantages
  - Exposes its mutational model, which can then be criticized
  - If the search is cut too short, the answer is overly certain
  - As slow as likelihood, maybe slower

What are the methods good for?

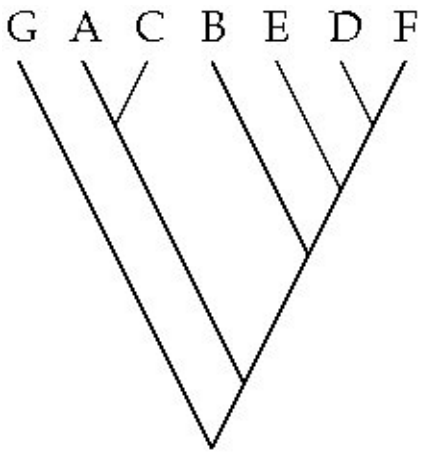
- Some data force a given method:
  - Biometric measurements – use parsimony
  - Immunological cross-reaction distances – use distance method
- Likelihood and Bayesian methods are powerful and accurate, but:
  - Require a detailed model of the mutational process
  - Too slow for big data sets

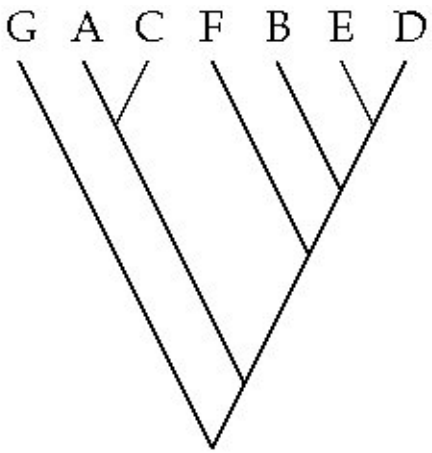
Consensus trees

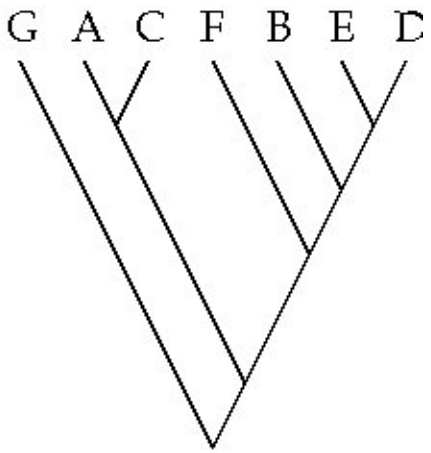
What information is common to all of these trees?

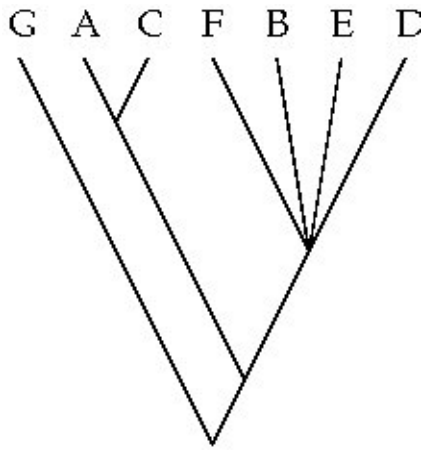
How can we clearly represent that information?

Strict consensus

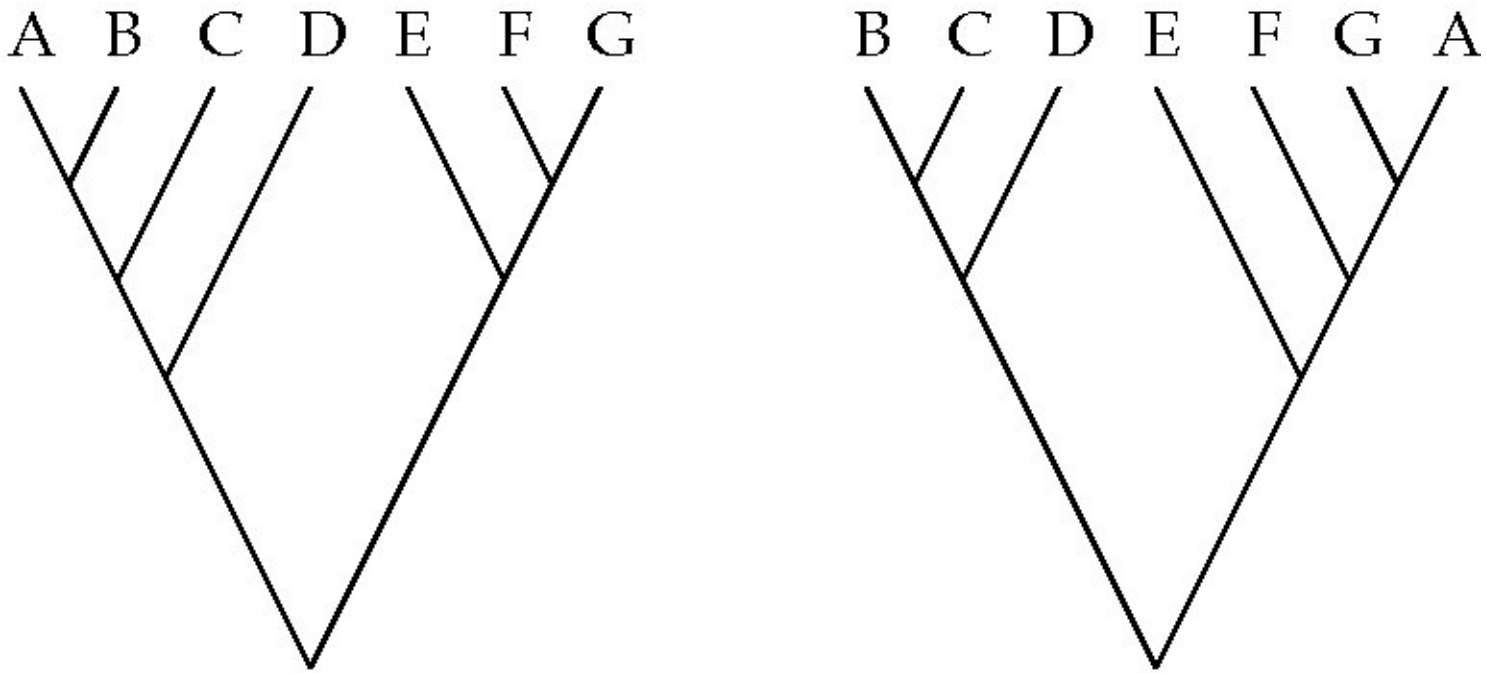






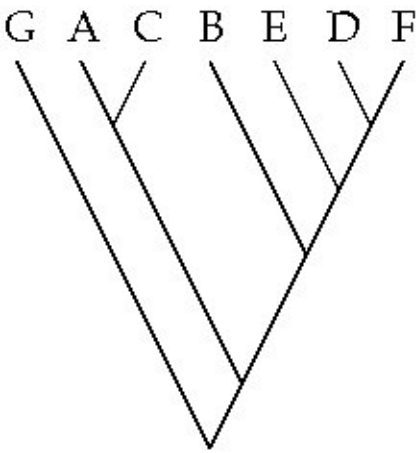


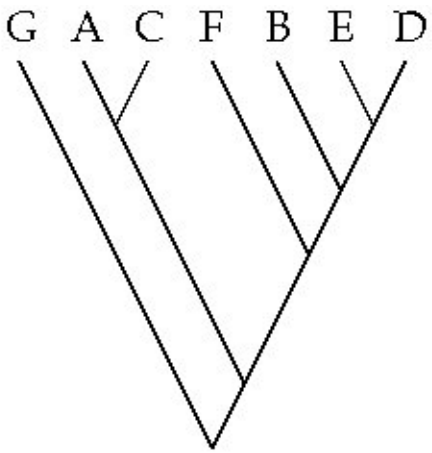
Strict consensus has problems

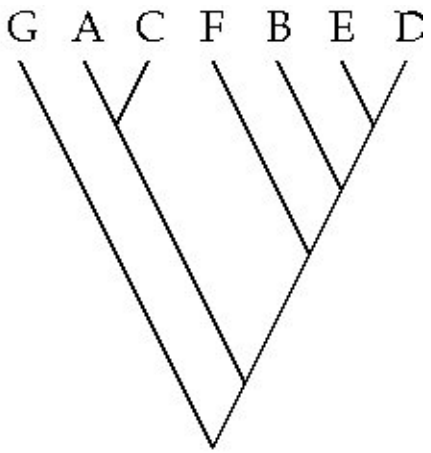


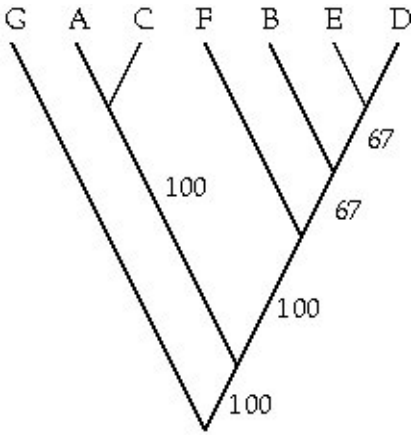
These trees appear similar, but their strict consensus is a "star" tree with no structure

Majority-rule consensus









Validating phylogenies

- Agreement among methods increases our confidence in our phylogeny

- However, consider this data set:

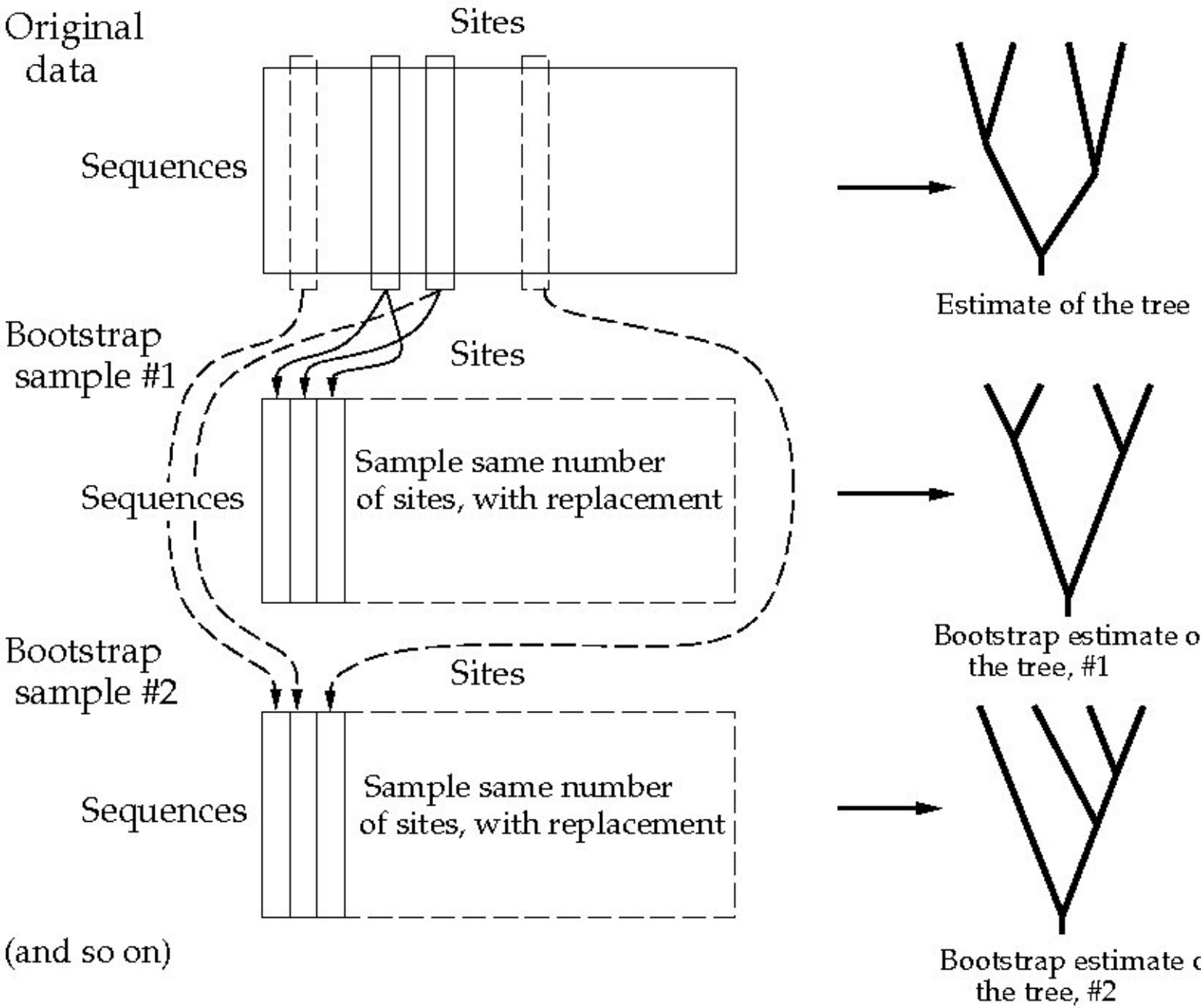
Sites supporting human+chimp	51
Sites supporting gorilla+chimp	49

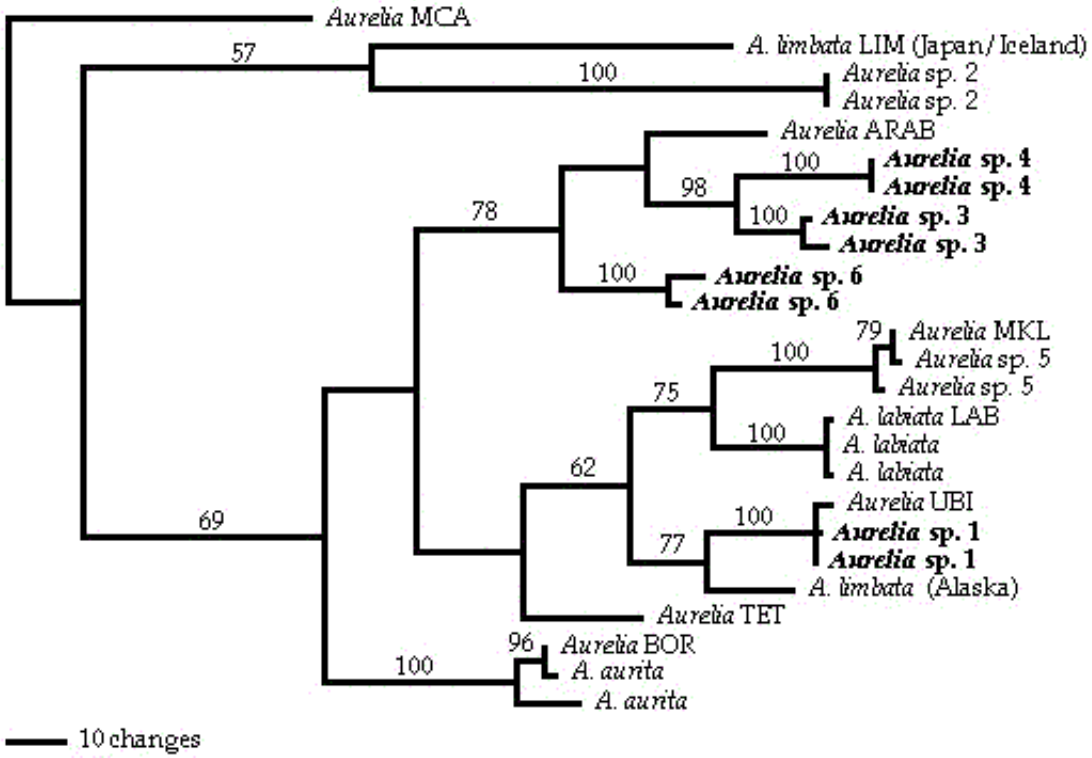
- All phylogeny methods will prefer human+chimp

- However, the data do not support either tree very strongly

Bootstrap

- The bootstrap is a general method for validating any type of phylogeny inference
- It answers the question: How sensitive are our conclusions to small variations in the data?
- Felsenstein's paper announcing bootstrap is #41 on "most cited papers of all time"


Original data
Sites
Sequences
Estimate of the tree
Bootstrap sample #1
Sites
Sample same number of sites, with replacement
Sequences
Bootstrap estimate o the tree, #1
Bootstrap sample #2
Sites
Sample same number of sites, with replacement
Sequences
Bootstrap estimate o the tree, #2
(and so on)


Aurelia MCA
A. limbata LIM (Japan/ Iceland)
Aurelia sp. 2
Aurelia sp. 2
Aurelia ARAB
Aurelia sp. 4
Aurelia sp. 4
Aurelia sp. 3
Aurelia sp. 3
Aurelia sp. 6
Aurelia sp. 6
Aurelia MKL
Aurelia sp. 5
Aurelia sp. 5
A. labiata LAB
A. labiata
A. labiata
Aurelia UBI
Aurelia sp. 1
Aurelia sp. 1
A. limbata (Alaska)
Aurelia TET
Aurelia BOR
A. aurita
A. aurita
57
100
98
100
78
100
79
100
75
100
62
69
100
77
96
100
10 changes

Bootstrap

- Consider our problem data set:

Sites supporting human+chimp	51
Sites supporting gorilla+chimp	49

- Many of the resampled data sets will have 50-50 or 49-51 instead of 51-49.

- The human+chimp branch will not get strong bootstrap support

- This correctly reflects the poor signal of the data

Bootstrap

- Bootstrap assesses how sensitive your results are to random fluctuation in the data
- Does **not** detect violations of your assumptions
- Example:
  - Method assumes a clock, but data are not clocklike
  - Original tree is systematically wrong
  - Bootstrap trees are systematically wrong too!

What do bootstrap values mean?

- Bootstrap values were originally interpreted as percent chance the branch was real
- This was disproven in the 1990's by computer simulation
- High values underestimate support; low values overestimate it

What do bootstrap values mean?

- There is no simple way to go from bootstrap value to percent support
- The relationship depends on number of tips and shape of tree
- Most people use a rough rule of thumb that 85% is a pretty good bootstrap and 65% is a definitely poor one
- It's best to publish the actual values and let readers draw their own conclusions

Other methods of validation

- Maximum likelihood algorithms come with built-in estimates of confidence
- Unfortunately these are only approximate for finite sized data sets
- Many researchers present bootstraps instead because they are more generally understood

Other methods of validation

- Bayesian "cloud of trees" can be treated like a bootstrap sample
- They answer different questions:
  - Bootstrap: would a slightly different data set prefer a different tree?
  - Bayesian support: would a slightly different tree fit this data set almost as well?
- It is easier to see that these are different than to understand how to use each one appropriately!
- If "cloud" is too small, results will be overly certain

Garbage in, garbage out

- No sensible tree exists when:
  - A species arose by hybridization of two other species
  - Genes have been exchanged between distantly related species
  - Different genes in the genome have different histories due to recombination and reassortment
- The programs will still run and a tree will be produced!
- Hybrids often move toward the bottom of the tree, or may cluster with one or the other parent
- Ideally we'd infer a tangled graph, but this problem is HARD

One-minute responses

- Tear off a half-sheet of paper
- Write one line about the lecture:
  - Was anything unclear?
  - Did anything work particularly well?
  - What could be better?
- Leave at the back on your way out