Roadmap

- Final exam schedule
- Inferring trees
  - Distance matrix methods
  - Likelihood methods
  - Bayesian methods
- Validating trees

Final exam schedule

- You are welcome to take the final either:
  - Tuesday 3/19 at 10:30 am-12:30 pm
  - Wednesday 3/20 at 2:30 pm-4:30 pm (original schedule time)
- Both in S110, with any luck
- I ask the Tuesday group not to discuss the final until after Wednesday

Four major approaches to phylogeny inference

- Prefer the tree which–
  - Parsimony: explains the data with the fewest mutations
  - Distance: minimizes the difference between observed and expected distances between taxa
  - Likelihood: maximizes the probability of the data
  - Bayesian: maximizes the posterior probability of the data given a prior

- The first two are easier: given a correct mutational model the second two are likely more accurate

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
- Seldom used today:
  - Assumes a molecular clock
  - Behaves badly if clock assumption violated
- Neighbor-joining is a non-clock version that is widely used:
  - Very fast
  - Allows use of a sophisticated mutation model
- UPGMA demonstrates the idea of distance methods in a simple way

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 recap

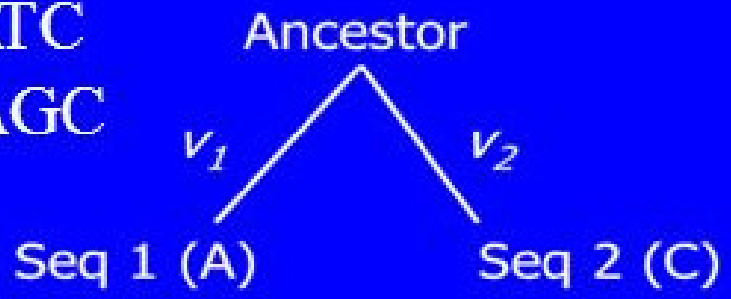
- Advantages
  - Fast
  - Can use a sophisticated mutational model
- Disadvantages
  - Loss of information in converting data to distances
  - Long distances often very noisy
  - Clock-assuming versions tend to be brittle

Maximum likelihood inference

- Basic principle: prefer the tree on which the data are most likely
- Requires:
  - An equation for the chance of changing from one state to another as a function of branch length
  - Sum over all possible states at all interior nodes
- Requires us to search among all possible topology and branch length combos
- (Can generally integrate out branch lengths, but topologies remain a problem)

Seq 1 ATATC

Seq 2 CTAGC



$$L_{(1)} = \pi_A P_{AA}(v_1) P_{AC}(v_2) + \pi_C P_{CA}(v_1) P_{CC}(v_2) + \pi_G P_{GA}(v_1) P_{GC}(v_2) + \pi_T P_{TA}(v_1) P_{TC}(v_2)$$

The Likelihood (i.e. the probability of observing the data) is a sum over all possible assignments of nucleotides to the internal nodes

Mutational models

- For DNA/RNA:
  - Simple symmetrical model (Jukes-Cantor)
  - Transitions differ from transversions
  - Unequal base frequencies
  - Invariant sites
  - Unequal rates per site
- Also possible: codons, amino acids

Maximum likelihood feasibility

- Felsenstein proposed this in the 1960's
- It was COMPLETELY infeasible with 60's technology
- Needed advances in:
  - Computer speed
  - Computer memory
  - Algorithm optimization
- Now feasible for around 50 taxa, but not for really large data sets

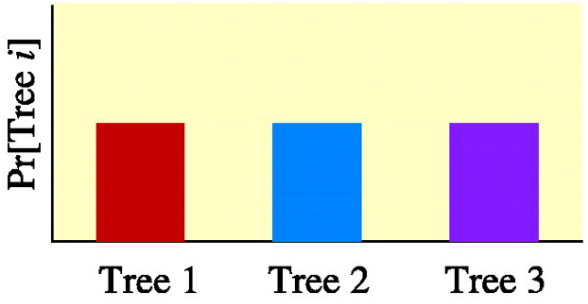
Maximum likelihood

- Pros:
  - Allows complex modeling of mutational process
  - Statistically robust
- Cons:
  - Very, very slow
  - Specifying the mutational model opens it to criticism
  - Yields just one estimate of the best tree with little information about alternatives
  - Search may not find best tree

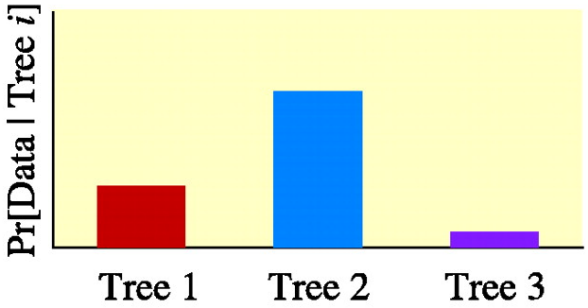
Bayesian phylogenetics

- A disadvantage of likelihood is it tells you $P(D|T)$ when you probably wanted $P(T|D)$
- $P(T|D)$ would involve a denominator which sums over ALL TREES–not feasible
- Bayesian phylogenetics tries to estimate $P(T|D)$ without computing the whole thing

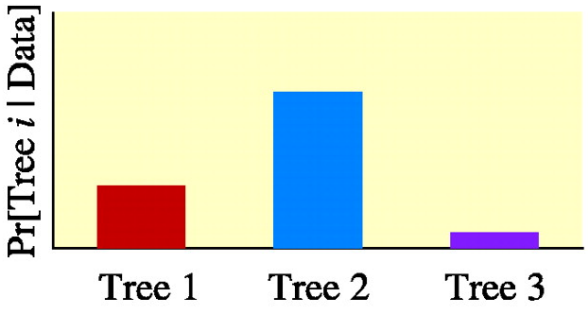
The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable, a priori. However, other information can be used to give some trees more prior probability (e.g., the taxonomy of the group).



The **likelihood** is proportional to the probability of the observations (often an alignment of DNA sequences) conditional on the tree. This probability requires making specific assumptions about the processes generating the observations.



The **posterior probability** of a tree is the probability of the tree conditional on the observations. It is obtained by combining the prior and likelihood for each tree using Bayes' formula.



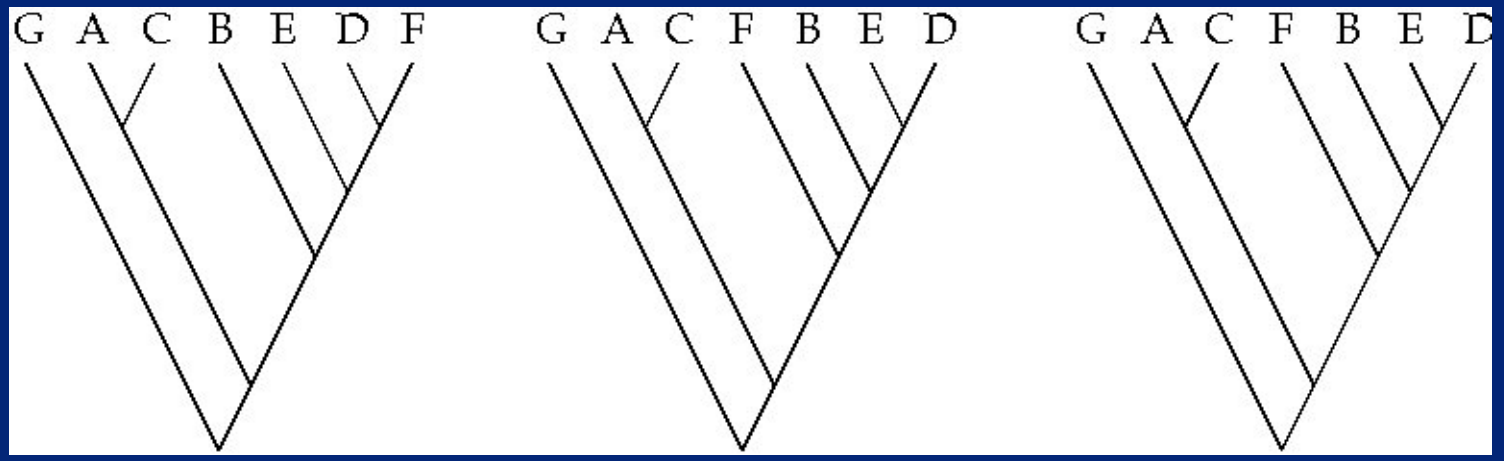
Bayesian phylogenetics

- Establish priors on parameters of interest (tree topology, base frequencies, rate categories, ...)
- Pick a starting tree from the prior
- Iterate:
  - Modify the tree slightly
  - Compute the likelihood of old and new trees
  - Accept the new one proportionate to the likelihoods:
    - Always accept if new tree is better
    - If new tree is worse, proportionally reduced chance of accepting
  - Keep a record of sampled trees
- Consider entire "cloud" of sampled trees as an estimate of the phylogeny

Bayesian phylogenetics

- Pros:
  - Sophisticated mutation models (same as likelihood)
  - If prior information available, can be used
  - Gives excellent information on the range of good trees, not just single best tree
- Cons:
  - Exposes mutational model to criticism
  - If you stop search too soon, results are too confident (support intervals are too narrow)
  - As slow as likelihood if not slower–unless you stop too soon

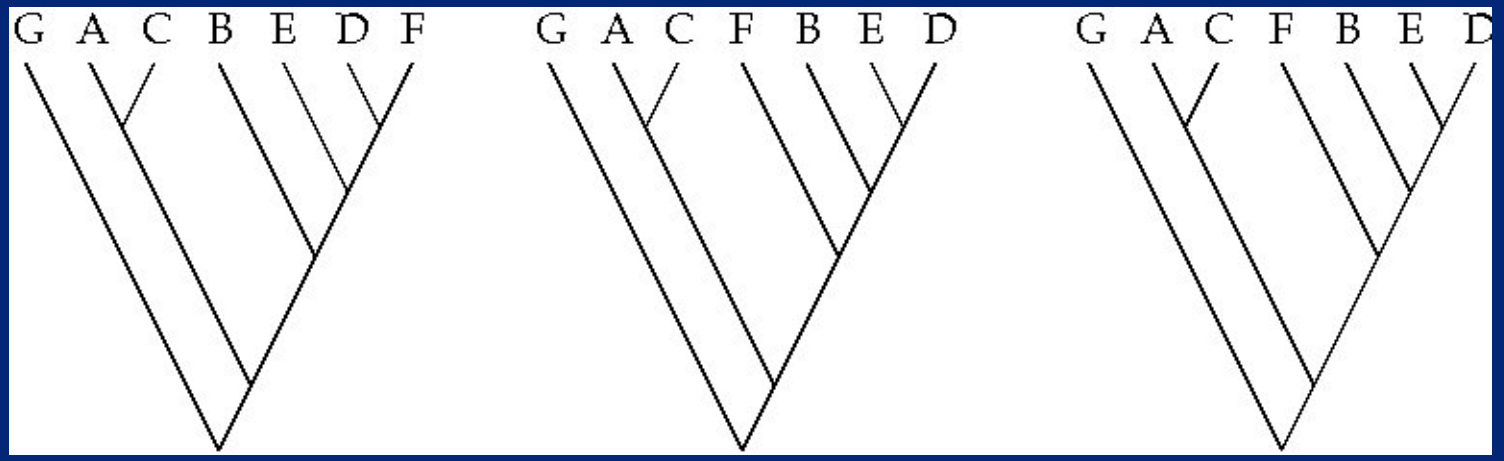
Consensus trees

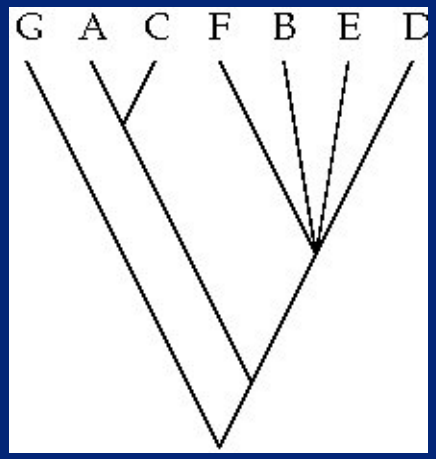


What information is common to all of these trees?

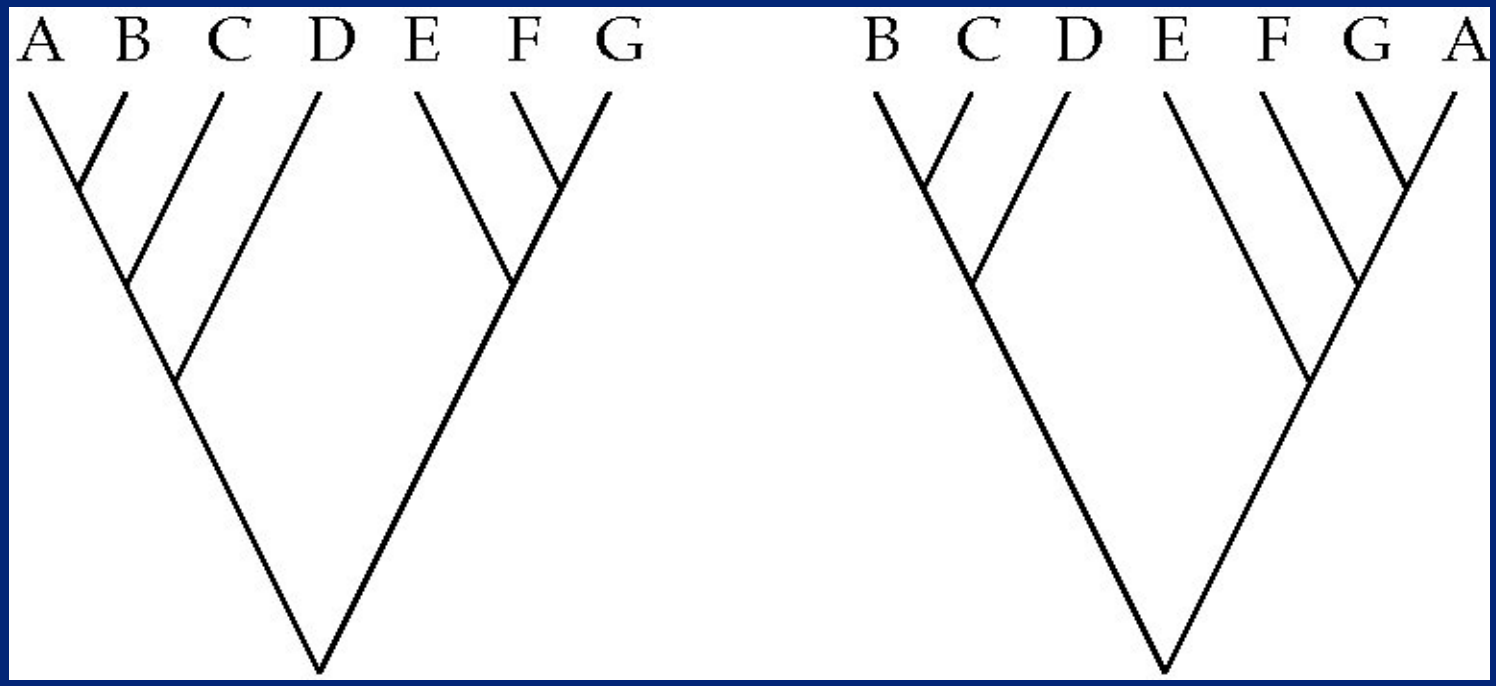
How can we clearly represent that information?

Strict consensus



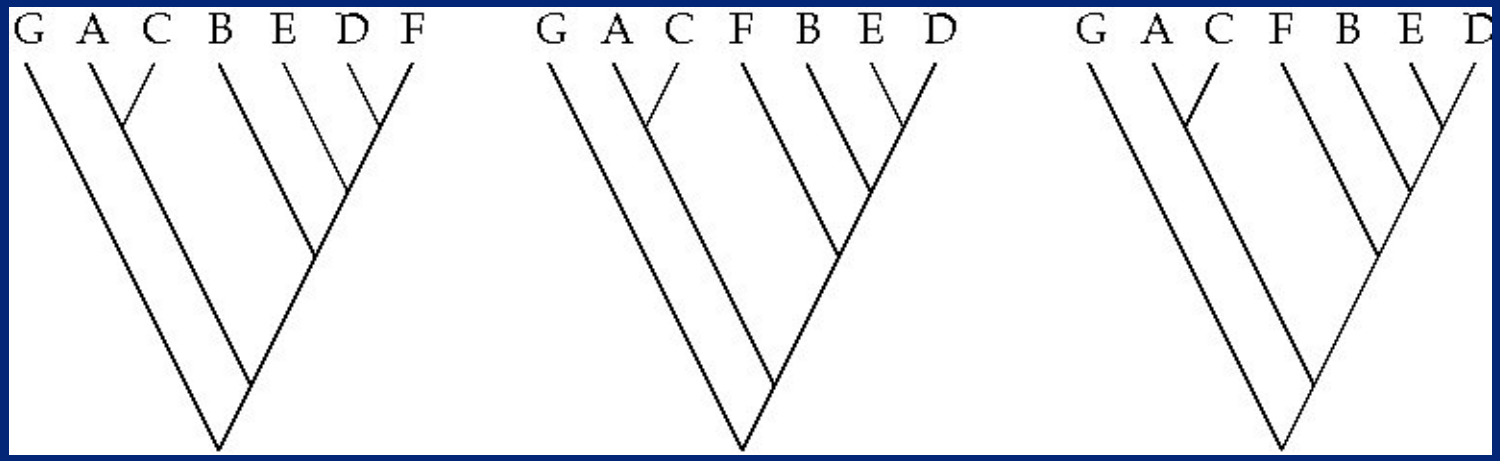


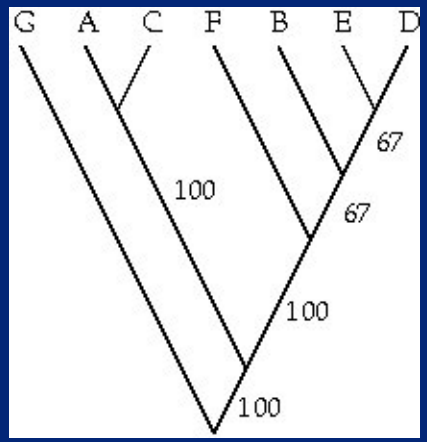
Strict consensus has problems



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

Majority-rule consensus



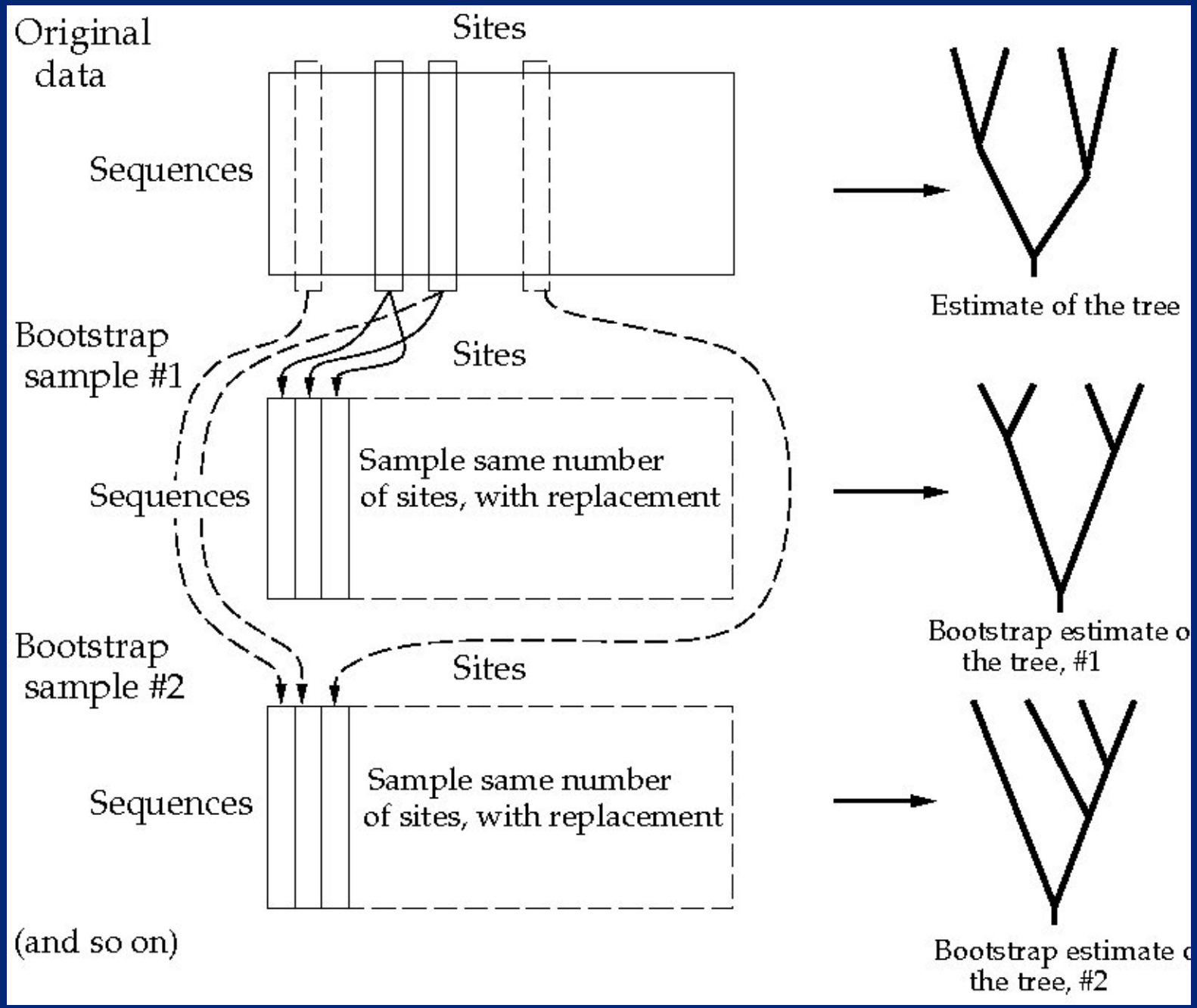


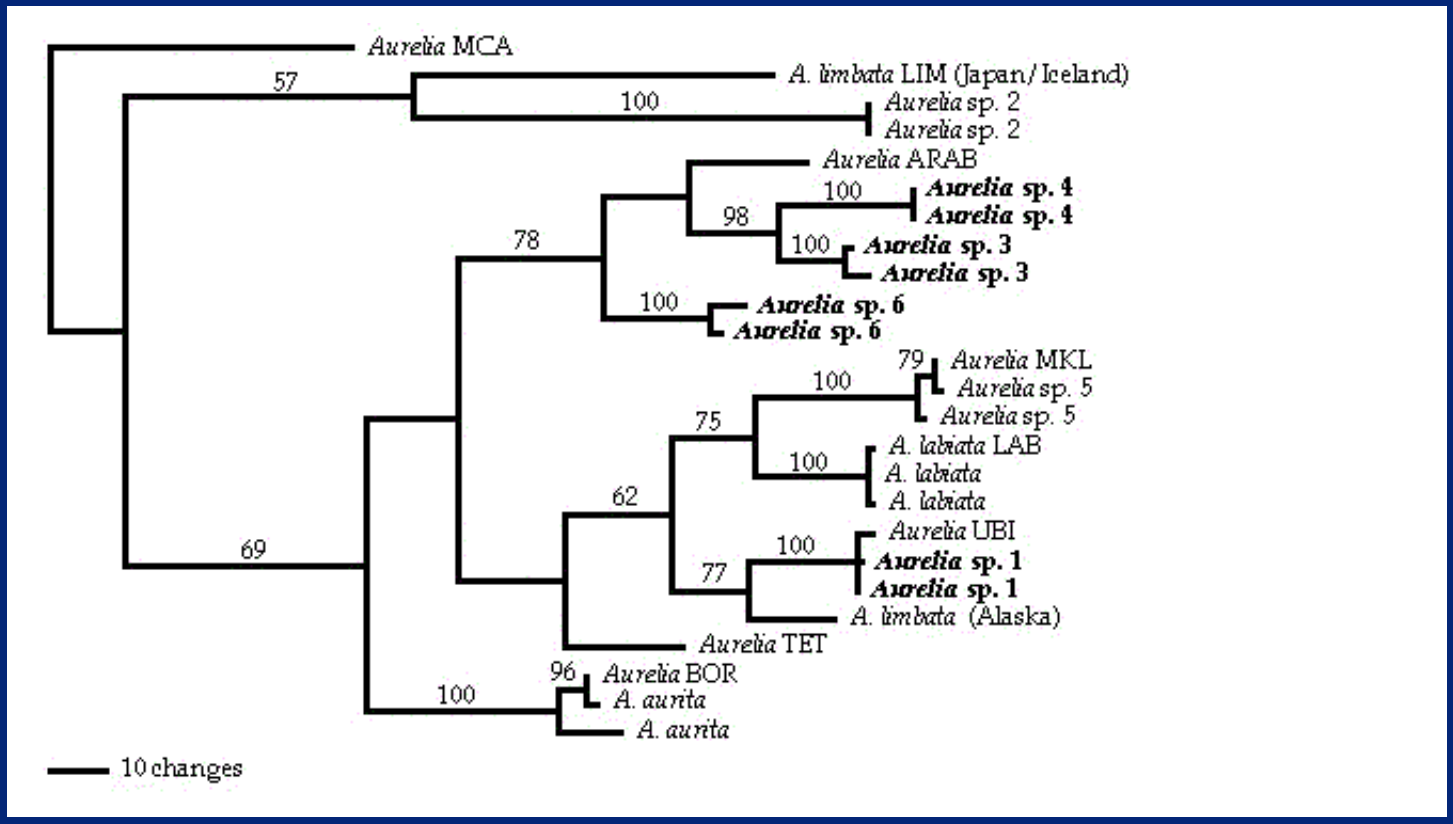
Expanded majority-rule consensus

- Assemble all groups with $> 50\%$ support
- These can always fit on the same tree–why?
  - (pigeonhole principle)
- Then start with the most popular groups that are below 50%, and add them if they are compatible with the existing tree
- This resolves the whole tree, but can include relationships that are very poorly supported
- Almost all software produces this kind (no one wants a half finished tree)

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


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
78
98
100
100
100
79
100
75
100
62
69
100
77
96
100
10 changes

Bootstrap

- Consider a problem data set:

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

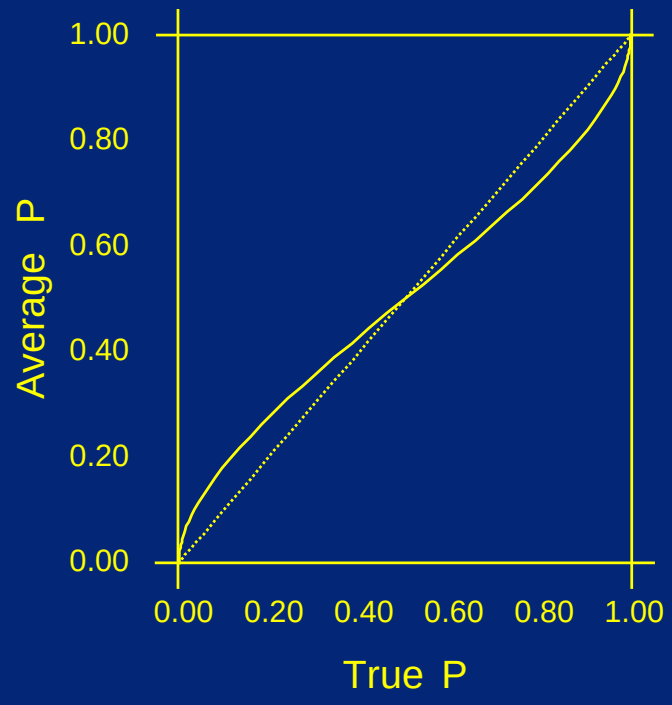
- Many of the resampled data sets will have 50-50 or 49-51 instead of 51-49.
- The human+chimp branch will not get strong bootstrap support
- This correctly reflects the poor signal of the data

Bootstrap

- Bootstrap assesses how sensitive your results are to random fluctuation in the data
- Does not detect violations of your assumptions
- 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
- These are only approximate for finite data
- Seldom used, I think because poorly understood

Bayes vs. bootstrap

- 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

Two hazards of phylogeny

- Garbage in, garbage out:
  - Long pieces of autosomal DNA
  - Misaligned sequences
  - Non-homologous traits
- Gene tree not necessarily the same as the species tree
  - Paralogs
  - Incomplete lineage sorting (ancestral polymorphism)
  - Horizontal gene transfer
  - Hybrid species

Friday

- Leftover phylogenetics
- Within-population inference using the coalescent