One-minute responses

- I liked the phylogenetics lecture today. The MCMC lesson was very helpful. (x5) $I\ recommend\ playing\ with\ the\ MCRobot\ on\ your\ own–it\ is\ a\ great\ teaching\ toy.$

- I am unclear on when it is best/appropriate to use the different tree building methods.

- Still have difficulty grasping trees, but think I will feel better after sitting down with Wikipedia and the recommended paper.

- Phylogeny part interesting, not sure how I would apply it.

One-minute responses

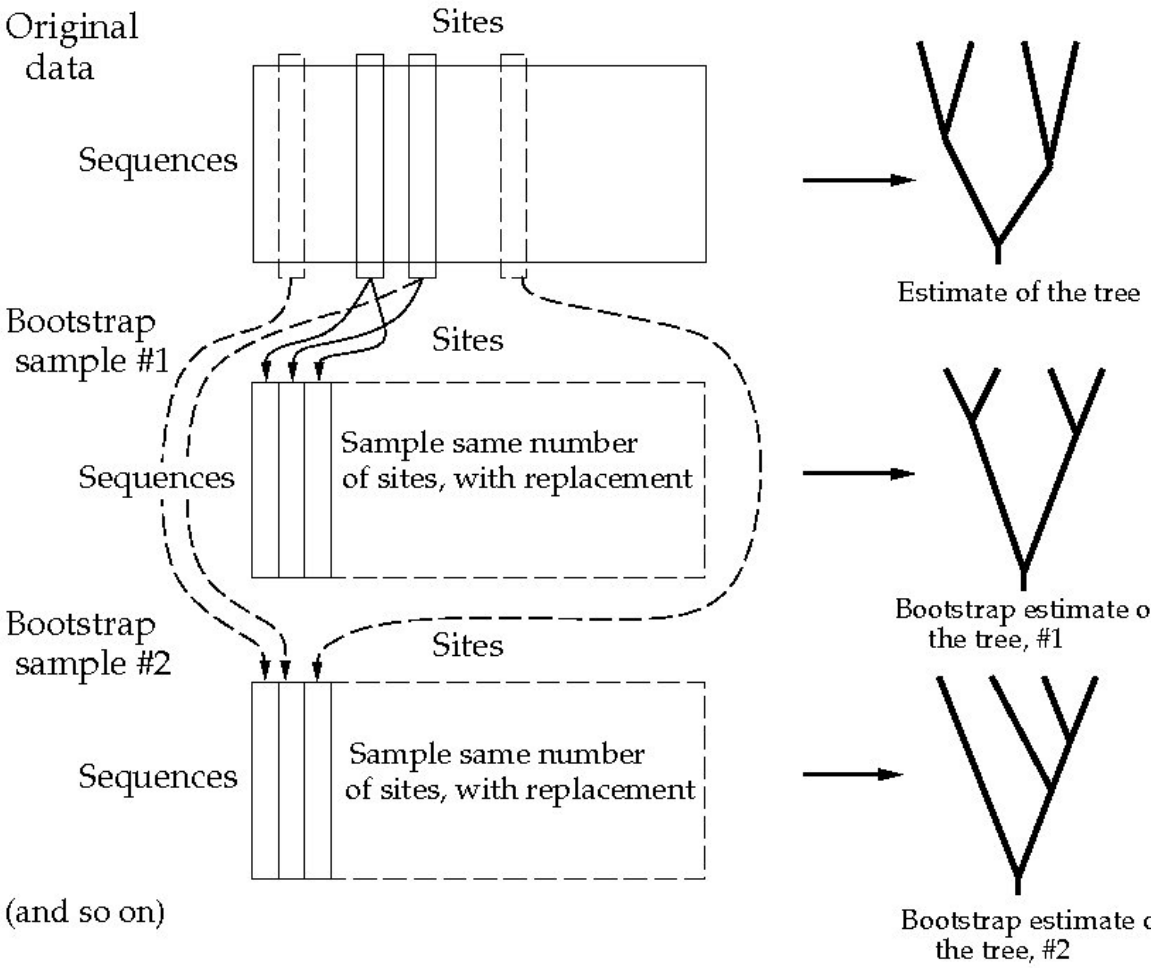
- It seems like using a Markov chain to generate a set of trees would give you a set weighted not only by tree score but also by the size of the "watershed" a tree finds itself in–i.e. a large region of so-so trees would get represented more than an isolated region of very high-scoring trees. Is this true in practice? If so, is it desirable?

- *This would be a good thesis question.... It's almost surely true in practice. Whether it's desirable could be assessed by computer simulation. You would have to decide whether the goal of phylogenetic analysis is to find the best tree or to get the best assessment of support for different parts of the tree.*

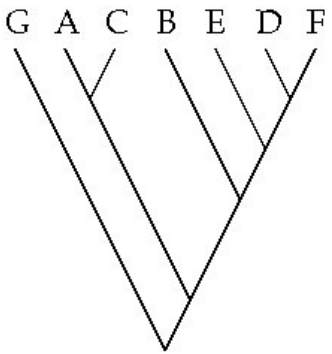
Introduction to Phylogenies: Validation

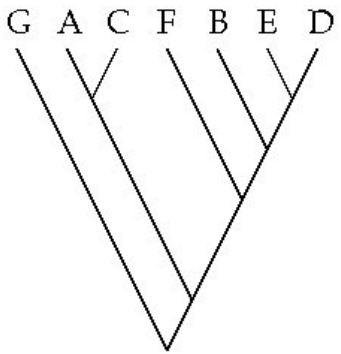
- The Bootstrap
- Bayesian support intervals
- Comparison among methods

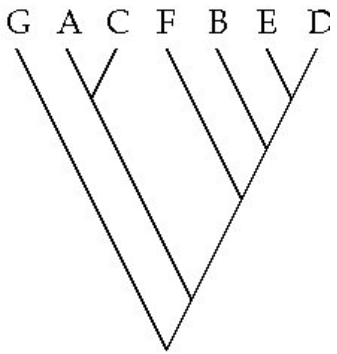
The Bootstrap

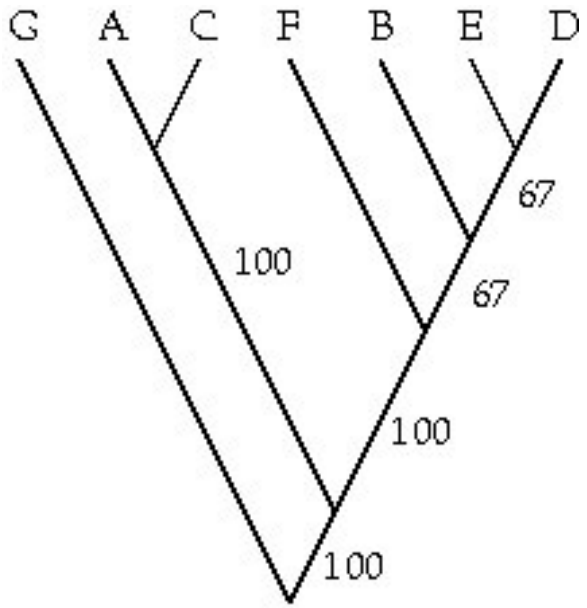


Reminder on consensus trees









The Bootstrap

- Resample the data to make a new data set
- Infer the tree from the new data set
- Repeat 100+ times
- Make a consensus of the resulting trees
- "Bootstrap support" = percentage of times that a grouping appears in the collection of trees

The Bootstrap

Advantages:

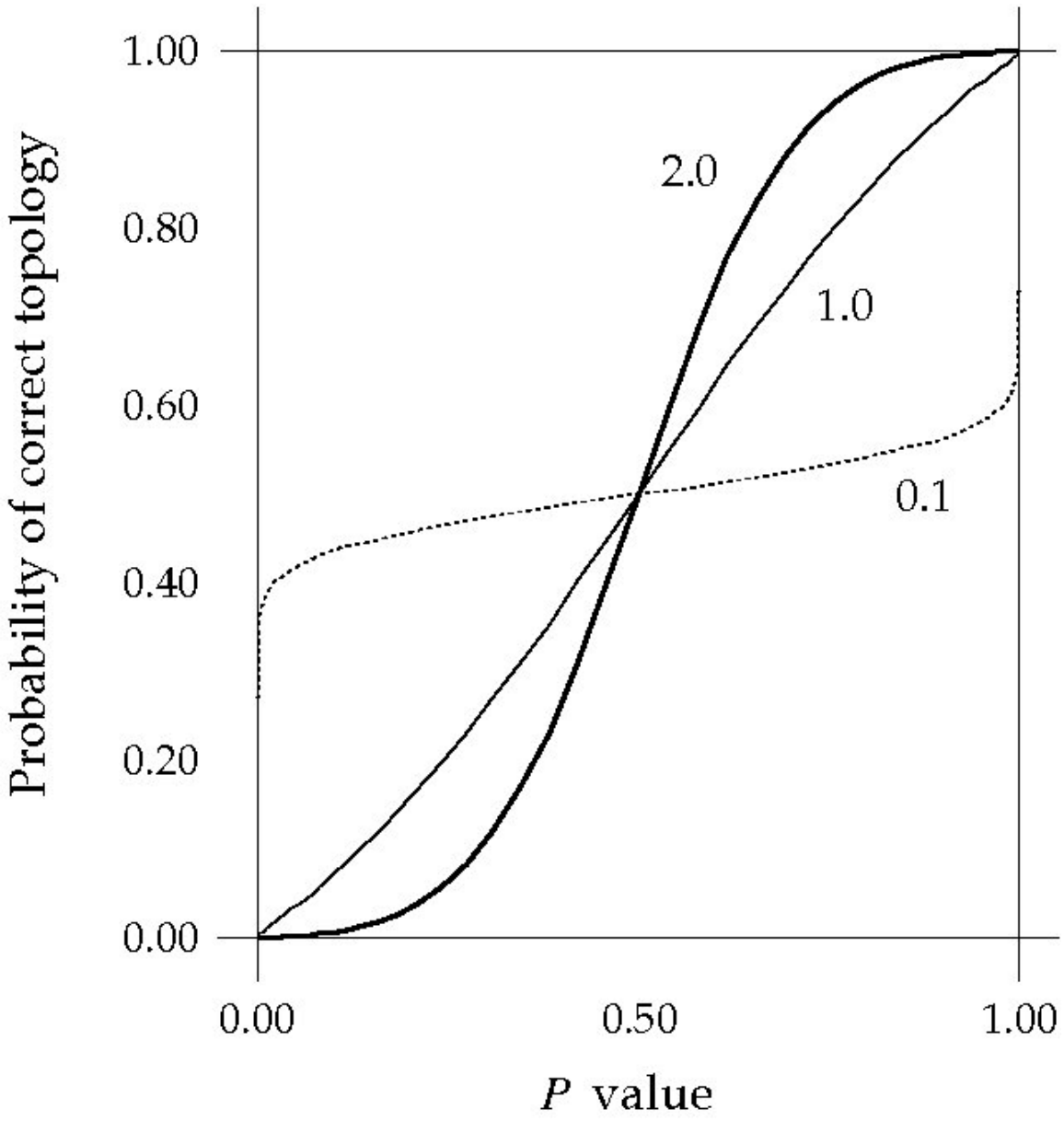
- Can be applied to any phylogeny method
- Shows which features of the tree are broadly supported by the data

Disadvantages:

- Takes 100x as long as a single analysis
- Does not detect flaws in your tree inference (wrong model, etc)
- Bootstrap support values are not easy to interpret

Interpreting the bootstrap support

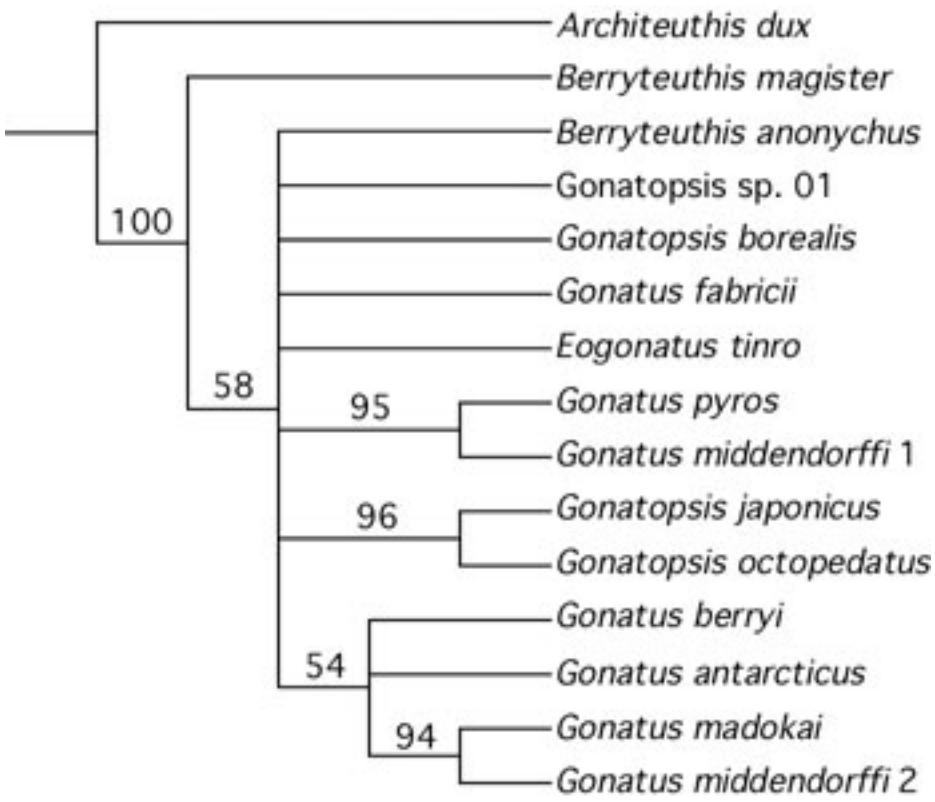
- We'd love it if the bootstrap support were the probability that the group is on the true tree
- Unfortunately it is not


1.00
0.80
0.60
0.40
0.20
0.00
Probability of correct topology
2.0
1.0
0.1
0.00
0.50
1.00
P value

Interpreting the bootstrap support

- The relationship between bootstrap support and probability is complex
- Generally, when the bootstrap is high it is conservative
  - Bootstrap of 85% often means 90-95% chance of correctness
  - Many researchers will take a bootstrap of 80% seriously
- When the bootstrap is low, it's non-conservative
  - Bootstrap of 15% may mean a 5% chance of correctness
  - We probably don't care just how bad a bad group is!

Bootstrap: A live example



The web page which showed this graphic says: "This data suggests that $Gonatus$ is polyphyletic. $Gonatopsis\ borealis$ groups more closely with $Berryteuthis$ than other species of $Gonatopsis$." Is this conclusion justified?

Other resampling methods

- Jackknife–delete some of the sites randomly
- Similar behavior to bootstrap
- Lots of controversy on how many sites to delete
- Bootstrap currently has better practical acceptance

Approximate likelihood ratio tests

- Likelihood methods offer an approximate test of whether a branch exists (has non-zero length)
- These can point out uncertain areas on a tree
- Because the methods are approximate, they are not fully accepted by many researchers

Bayesian support

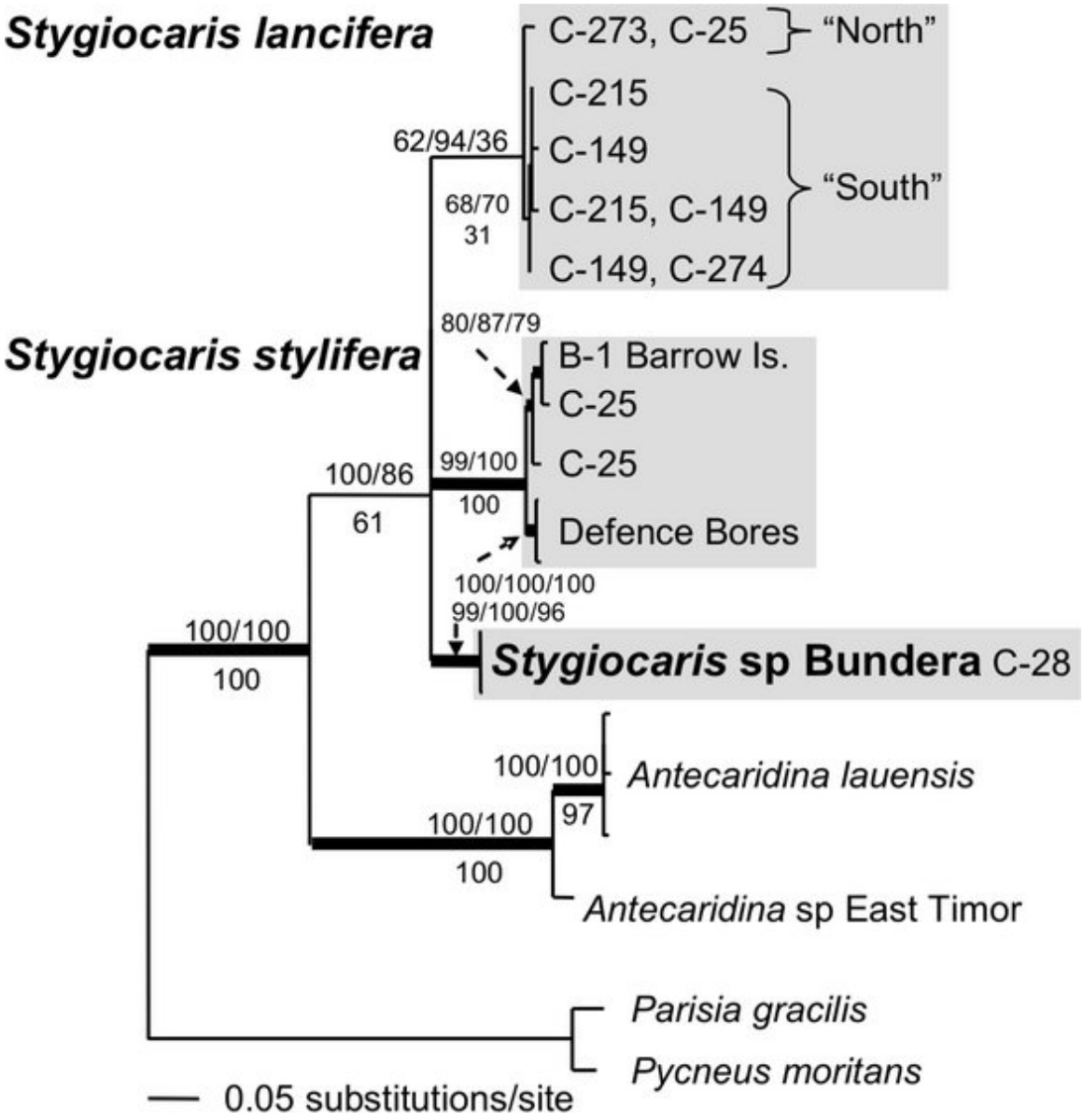
- Bayesian methods visit many trees in proportion to $P(D|H)P(H)$
- A consensus of the resulting trees can be used to assess support
- 85% of the trees visited contain a given grouping–that grouping has 85% Bayesian support

Bootstrap versus Bayesian support

- Both methods try to measure our confidence in the tree
- They disagree, sometimes strongly. Why?
  - Bootstrap is biased
  - Bayesian support depends on a thorough search and is too narrow if the search didn't succeed
  - They answer slightly different questions

Bootstrap versus Bayesian support

- Bootstrap: If the data were slightly different, would they mostly support the same tree?
- Bayesian support: If the tree was slightly different, would it be nearly as likely to produce these data?
- I have no intuition on how these questions differ, but in practice they do differ!



Likelihood bootstrap/Bayesian support/Parsimony bootstrap

Overview of phylogeny methods

Parsimony: Prefer the tree requiring the fewest mutational events

- Advantages:
  - Conceptually simple
  - Applicable to sequence or morphological data; no formal model required
  - Medium speed (painful above 80-100 species)

- Disadvantages:
  - Cannot easily make use of modeling information
  - Risky when evolutionary rates vary among lineages
  - Not consistent–certain tree shapes are troublesome ("long branch attraction")

Overview of phylogeny methods

Distance: Prefer the tree whose distances best match the distance matrix

- Advantages:
  - Can use sophisticated mutational models
  - Applicable to any data for which distances can be developed
  - Only possible method for data which are intrinsically distances
  - Fast to extremely fast (thousands of species)

- Disadvantages:
  - Some loss of information in converting data to distances
  - Fast versions do not search tree-space thoroughly
  - Vulnerable to incorrect models
  - Not suitable for data where every column should have its own model (i.e. many morphological traits)

Overview of phylogeny methods

Likelihood: Prefer the tree on which the data are most probable

- Advantages:
  - Can use sophisticated mutational models
  - Full use of data
  - Robust and powerful
  - Likelihood-ratio tests allow hypothesis testing

- Disadvantages:
  - Extremely slow (painful above 30-40 species)
  - Vulnerable to incorrect models
  - Not suitable for data where every column should have its own model (i.e. many morphological traits)

Overview of phylogeny methods

Bayesian: Prefer the tree of highest probability given model and prior

- Advantages:
  - Can use sophisticated mutational models
  - Full use of data
  - Robust and powerful
  - Naturally provides information on precision of estimate

- Disadvantages:
  - Extremely slow if adequate search is made
  - Vulnerable to wrong priors as well as wrong models
  - Not suitable for data where every column should have its own model (i.e. many morphological traits)
  - Very new method, still not fully understood

What should I use?

- Your data may determine the answer:
  - Morphological data suggest parsimony
  - Measured distances suggest distance methods
  - More options for DNA, RNA, protein, microsatellites
- Speed matters if your data set is very large
- If good mutational models are available, use a method that can take advantage of them
- OPINION: A less powerful method with bootstrapping (or Bayesian support) is preferable to a more powerful method without validation

Should I try multiple methods?

- This never hurts if you can afford it
- However, agreement among methods is weak evidence of correctness
- A data set with 51% support for one tree and 49% support for another tree may choose the first tree with any method
- Bootstraps or Bayesian support will reveal that the evidence is weak
- Trees with highly unequal evolutionary rates may get the same WRONG answer with several methods
- Some journals do demand multiple methods