Thanks to Paul Lewis, Jeff Thorne, and Joe Felsenstein for the use of slides
• Hennigian logic reconstructs the tree if we know \textit{polarity} of characters and there is \textbf{no homoplasy}
• UPGMA infers a tree from a distance matrix:
  – groups based on \textit{similarity}
  – fails to give the correct tree if rates of character evolution vary much
• Modern distance-based approaches:
  – find trees and branch lengths: patristic distances \(\approx\) distances from character data.
  – do \textbf{not} use all of the information in the data.
• Parsimony:
  – prefer the tree that \textbf{requires} the fewest character state changes. Minimize the number of times you invoke homoplasy to explain the data.
  – can work well if if homoplasy is not rare
  – fails if homoplasy very common or \textbf{is concentrated on certain parts of the tree}
• Maximum likelihood
  – computes the probability of the data given a model (tree and branch lengths)
  – computationally expensive
Review Tree Searching

- Hennigian logic **builds** a tree directly from the characters
- UPGMA **builds** a tree from distances
- Parsimony, maximum likelihood, and modern distance methods are *optimality criteria*. We still have to **search** for the best tree.
- Too many trees to enumerate them exhaustively
- We rely on hill-climbing heuristics
Even if we find the optimal tree, we do not know that it is the *true* tree.

How do we assess statistical support?
The bootstrap

\[ \hat{\theta} \text{ (unknown) true value of } \theta \]

empirical distribution of sample

Distribution of estimates of parameters

Bootstrap replicates

Week 7: Bayesian inference, Testing trees, Bootstraps – p.33/54
The bootstrap for phylogenies

Original Data

sequences

Bootstrap sample #1

sequences

sample same number of sites, with replacement

Bootstrap estimate of the tree, #1

Estimate of the tree

Bootstrap sample #2

sequences

sample same number of sites, with replacement

Bootstrap estimate of the tree, #2

(and so on)
Bootstrapping: first step

From the original data, estimate a tree using, say, parsimony (could use NJ, LS, ML, etc., however)
Bootstrapping: first replicate

<table>
<thead>
<tr>
<th>weights</th>
<th>1</th>
<th>2</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>3</th>
<th>1</th>
<th>...</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

From the bootstrap dataset, estimate the tree using the same method you used for the original dataset.

Sum of weights equals $k$ (i.e., each bootstrap dataset has same number of sites as the original)
Bootstrapping: second replicate

<table>
<thead>
<tr>
<th>weights</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>0</th>
<th>...</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

This time the tree that is estimated is different than the one estimated using the original dataset.
Bootstrapping: 20 replicates

Note: usually at least 100 replicates are performed, and 500 is better
Bootstrapping: first step

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

From the original data, estimate a tree using, say, parsimony (could use NJ, LS, ML, etc., however)
# Bootstrapping: first replicate

<table>
<thead>
<tr>
<th>weights</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

Sum of weights equals $k$ (i.e., each bootstrap dataset has same number of sites as the original)

From the bootstrap dataset, estimate the tree using the same method you used for the original dataset.
# Bootstrapping: second replicate

![Image of bootstrapping example]

This time the tree that is estimated is different than the one estimated using the original dataset.

<table>
<thead>
<tr>
<th>weights</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

Note that weights are different this time, reflecting the random sampling with replacement used to generate the weights.
Bootstrapping: 20 replicates

<table>
<thead>
<tr>
<th>1234</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>------</td>
<td></td>
</tr>
<tr>
<td>-**-</td>
<td>75.0</td>
</tr>
<tr>
<td>-***</td>
<td>15.0</td>
</tr>
<tr>
<td>---**</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Note: usually at least 100 replicates are performed, and 500 is better.
200 Million Year Old Fossil
The "Clock Idea"

20% Sequence Divergence in 200 Million Years means 1% divergence per 10 Million Years
"A comparison of the structures of homologous proteins ... from different species is important, therefore, for two reasons. First, the similarities found give a measure of the minimum structure for biological function. Second, the differences found may give us important clues to the rate at which successful mutations have occurred throughout evolutionary time and may also serve as an additional basis for establishing phylogenetic relationships."

From p. 143 of
The Molecular Basis of Evolution

by Dr. Christian B. Anfinsen (Wiley, 1959)
A problem with the "Clock Idea": Rates of Molecular Evolution Change Over Time!!
“Ernst Mayr recalled at this meeting that there are two distinct aspects to phylogeny: the splitting of lines, and what happens to the lines subsequently by divergence. He emphasized that, after splitting, the resulting lines may evolve at very different rates... How can one then expect a given type of protein to display constant rates of evolutionary modification along different lines of descent?”

Molecular Clock

No Clock

(amount of evolution
(substitutions per site)
Assuming a Strict Molecular Clock

LR test statistic = 232
n=15 taxa, n-2 = 13 d.f.
Null (clock) hypothesis rejected


Reasons that the clock might be rejected

1. Rates of evolution vary across lineages can vary over time:
   (a) mutation rates can vary (mutations per cell cycle, mutations per time, number of cell cycles per generation, generation time).
   (b) strength and targets of selection can vary
   (c) population sizes can vary
2. Incorrect models of sequence evolution lead to errors in the estimation of rates
   (a) Almost any error in the model can lead to biases (or higher than needed variance) in detecting multiple hits
   (b) Assumption of a Poisson clock can be wrong – even if we correctly count the number of changes, if we don’t count for over-dispersion (higher than Poisson-variance in the # of substitutions) then we can falsely reject Cutler (2000)
- Penalized likelihood (penalize rates that vary too much)
- Bayesian approaches:
  - model the rate of evolution of the rate of evolution.
  - incorporates prior knowledge of what rates combinations are most likely.