One-minute responses

- Nice class–no complaints.
- Your explanations of ML were very clear.
- The phylogenetics portion made more sense to me today.
- The pace/material covered for likelihoods was more difficult than previous lectures' topics, but I think I just need to look it over for longer.
- The bioinformatics portion was pretty fuzzy for me.

Introduction to Phylogenies: Bayesian methods

- Bayes' theorem
- Bayesian phylogeny estimation
- Markov chain Monte Carlo
- Consensus trees

Bayes' theorem

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

- The probability that our hypothesis is correct is:
  - The support given by the data times
  - The prior probability of that hypothesis
  - Normalized by a sum over all hypotheses

Bayes' theorem

- Bayes's theorem is true, but use of it can be controversial
- Often we don't really know $P(H)$, the prior probability
- Is it okay to use a not-quite-correct prior?
- Also, the denominator can be difficult to compute
- Maximum likelihood methods are an attempt to avoid Bayes' theorem....
- What if we bite the bullet and try to use it?

Bayes' theorem

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

- In phylogenetics, $P(D|H)$ is the probability of the data for a given tree hypothesis
- This comes from a mutational model and is straightforward (it's what ML methods maximize)
- $P(H)$ is the prior chance that any given tree is the right one
- This has two components:
  - Topology
  - Branch lengths

Bayesian prior on trees: topology

- We can assume that every topology is equally likely $a\ priori$
- Unfortunately there are two different ways to count them!
- In practice this choice does not seem to matter much
- A really good prior would reflect our prior knowledge of reasonable trees
- No one has even attempted this

Bayesian prior on trees: topology

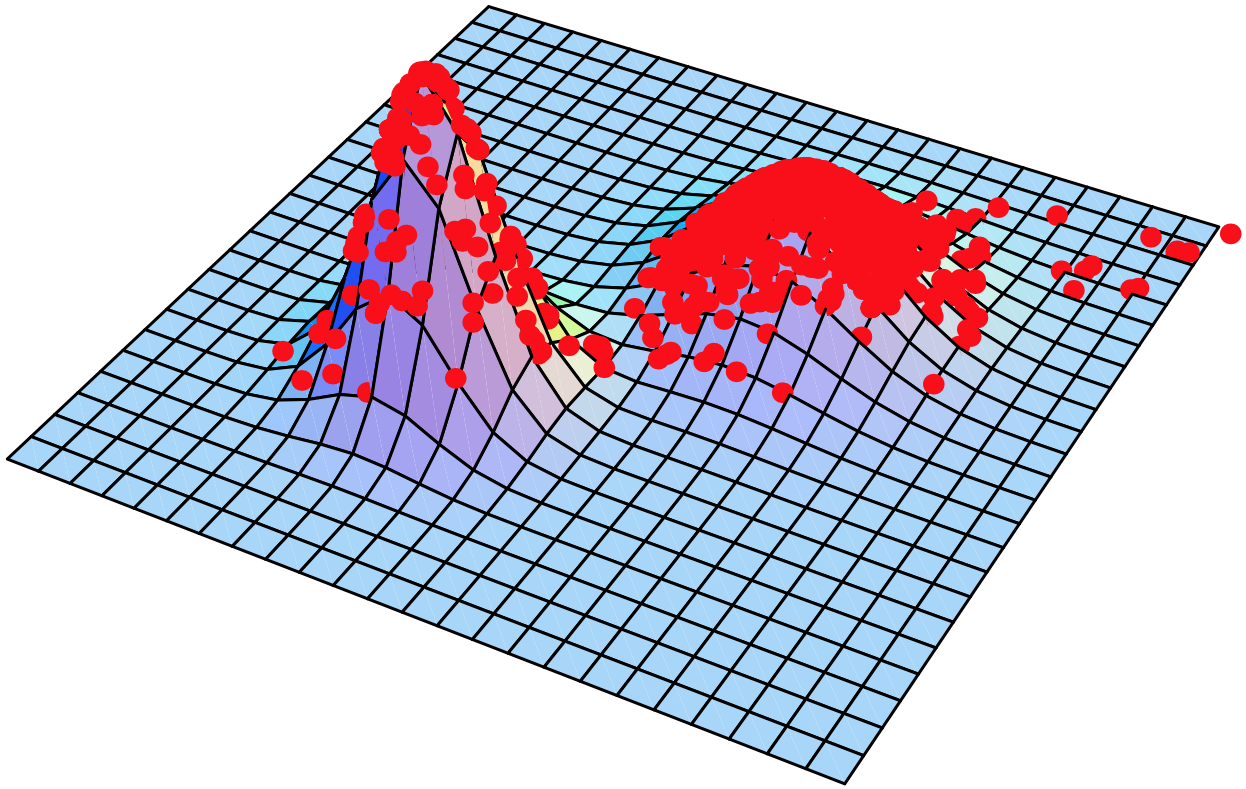
- We assume a minimum and maximum length for branches
- Usually the minimum is 0 and the maximum is some large value
- We assume a flat prior between those boundaries
- Again, this is not a very satisfying representation of our prior knowledge, but no one has been able to do better

Bayes' theorem

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

- We have $P(D|H)$ (from mutational model)
- We have $P(H)$ (from prior)
- How can the enormous summation be handled?

Markov chain Monte Carlo (MCMC)



Markov chain Monte Carlo (MCMC)

- Monte Carlo means solving a problem by simulation
- A Markov chain is a simulation that moves around on its surface in steps
- In our case, we move from genealogy to genealogy, being guided by $P(D|H)P(H)$

Show MCRobot here

This is the MCRobot program of Paul Lewis.

It's available at:

http://www.eeb.uconn.edu/people/plewis/software.php

Unfortunately it requires Windows, but if you have the chance I encourage you to try it out.

Bayesian phylogenetics

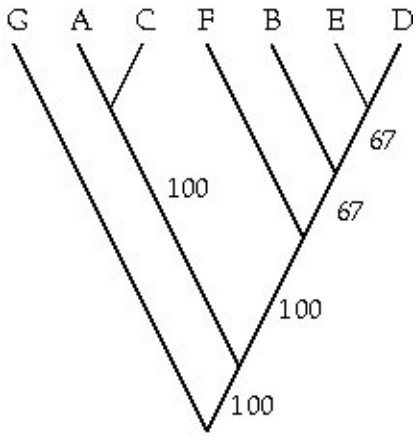
- Search the space of possible trees, guided by $P(D|H)P(H)$
- Can report the best tree ever found
- A better option is often making a consensus tree of all trees found

Consensus trees

Here are three trees which don't entirely agree:

Consensus trees

Here is their majority-rule consensus:



Consensus trees

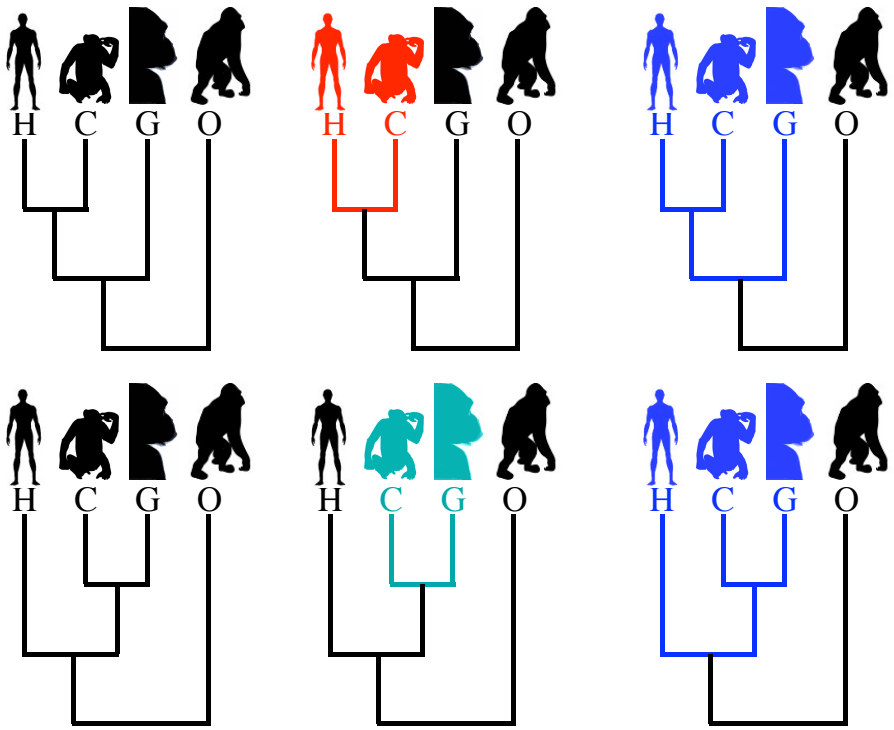
Majority-rule consensus:

- Include all groupings with more than 50% support
- All of these must fit on a single tree (why?)
- Finish up the tree, if necessary, using the best groupings among the less than 50% category
- Branches are often labeled with their support

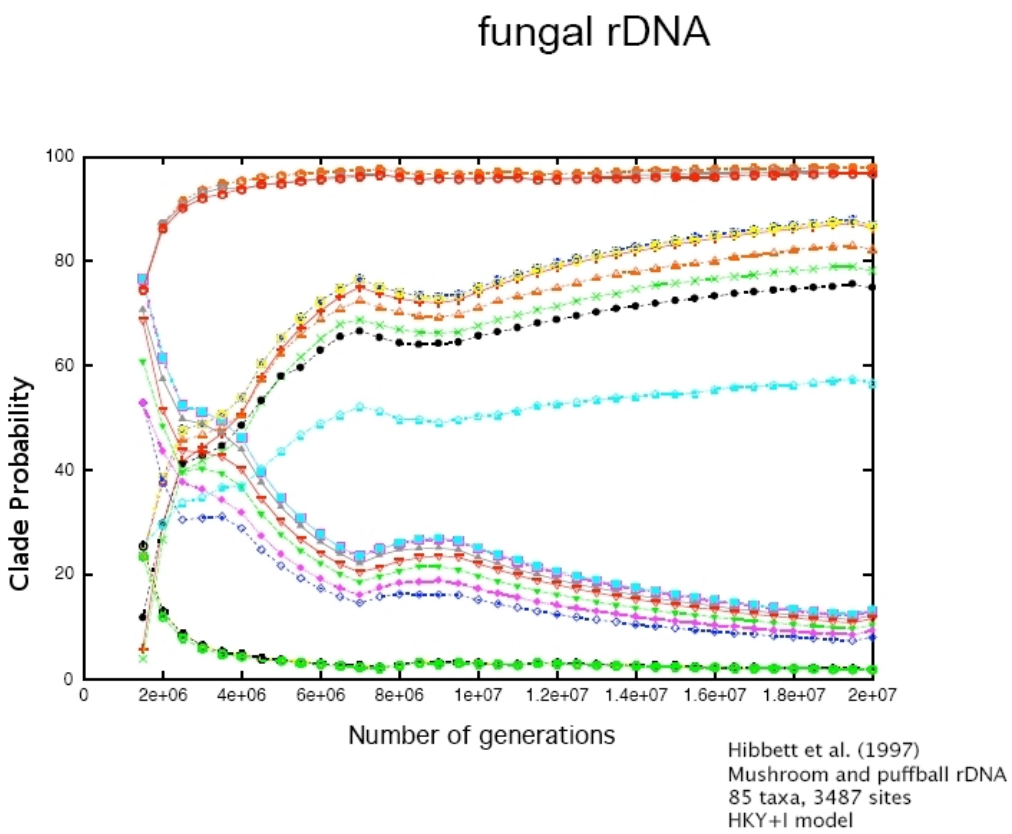
Strengths and weaknesses of Bayesian approach

- Strengths:
  - Similiar to likelihood in power, robustness, and consistency
  - Gives information about which parts of tree are well supported
- Weaknesses:
  - Vulnerable to bad choice of priors
  - Because it searches among whole trees, user may easily stop too soon
  - Search can be extremely difficult for some data sets

Clade probabilities



Progress toward assessing clade probabilities

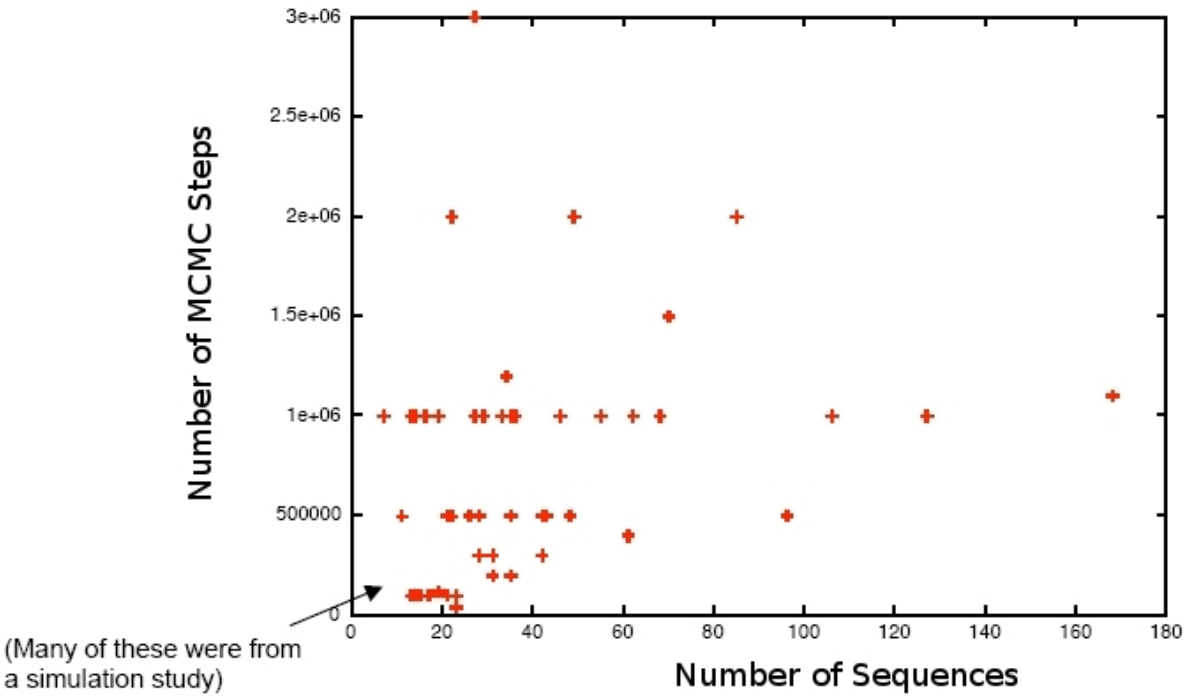


Courtesy of David Swofford

How long are people running their chains?

Literature search for chain lengths used with MrBayes:

- Molecular Biology and Evolution (17 papers)
- Molecular Phylogenetics and Evolution (33 papers)
- Taxon (4 papers)



Courtesy of David Swofford, circa 2004

Summary

- Bayesian methods are quite new
- They are similar to likelihood methods but add a prior
- They must use MCMC to search among possible trees
- Their most powerful output is not a single tree but a set of possible trees
- Consensus tree algorithm converts this set of trees into a useful diagram
- Bayesian methods are vulnerable to too-short MCMC searches

Summary

Other uses of Bayesian methods besides finding the best tree:

- Clade support:
  - are bats monophyletic?
  - does the human mtDNA tree root in Africa?
  - are pandas closer to bears or raccoons?
- Model parameters:
  - transition/transversion ratio
  - base frequencies
  - proportion of sites which do not vary