Announcements

1. Final exam: Friday April 21, 2006, 8-10AM here
2. Paper / project: You should all have datasets at this point
3. Final Debate next Wednesday - For Whiting: Daynika Schade & Steven Jackson, Against: Nicole Webster, Pat Wise, Jessica Zelazo

Bayesian Phylogenetic Inference

Outline

1. Introduction, history
2. Advantages over ML
3. Bayes’ Rule
4. The Priors
5. Marginal vs Joint estimation
6. MCMC
7. Posteriors vs Bootstrap values

Bayesian Phylogenetic Inference - MCMC

MCMC relies on marginal estimation (volume under the posterior-probability curve is estimated) rather than joint estimation (seeking the highest point)

This volume is estimated using a search procedure analogous to exploring tree-space for the optimal tree(s)

MCMC explores parameter space (not tree-space) and MCMC is not seeking just the top (optimal) portion but instead will describe the entire space

Parameter space

MCMC seeks the area under the probability surface

MCMC can start anywhere

Bayesian Phylogenetic Inference - MCMC

Posterior Probabilities are determined by integrating the product of the prior probabilities and the model over all possible parameter values

The likelihood functions are too complex to integrate analytically

MCMC is used instead - Markov chain Monte Carlo
dates from the work of Metropolis (1953) – an algorithm that does not seek the maximum likelihood parameter values
Bayesian Phylogenetic Inference - MCMC

MCMC is much like our mountaineer in the fog analogy - except the mountaineer is trying to map the entire mountain range rather than just find the tallest peak!

The MCMC algorithm follows a set of rules and takes “steps” by making changes to the parameters including the topology and the branch lengths.

Bayesian Phylogenetic Inference - MCMC

If the changes improve the likelihood the ‘step’ is taken

If the changes would make the likelihood worse the step is only taken under certain conditions
- when the ratio between the current likelihood and the new (worse) likelihood is larger than a number drawn at random

This allows the MCMC algorithm to traverse valleys - although it can still get “stuck” in suboptimal regions of parameter space.

Bayesian Phylogenetic Inference - MCMC

The MCMC chain visits regions of parameter space in proportion to their probabilities

Samples of trees and parameters are taken from the chain at regular intervals
- typically every 100 (or rarely 1000) steps

The longer the chain is allowed to run the closer the approximation becomes.

Bayesian Phylogenetic Inference - MCMC

The sample includes trees with low, moderate, and high likelihood scores - but proportionally more high scoring trees.

The frequency a tree appears in the sample of a well-run MCMC search is an estimate of its posterior probability.

eg a MCMC run of 2,000,000 steps (or generations) that is sampled every 100 steps will yield a final sample of 20,000 trees.

Bayesian Phylogenetic Inference - MCMC

These 20,000 trees can be used to build a majority rule consensus tree (after the “burn-in” trees have been discarded - see next slides)

The proportion a clade is represented in the sample is its posterior probability.

The MCMC chain starts sampling in a region with very poor posterior probabilities because starting values are set arbitrarily or chosen randomly.
History or trace plot plot of $-\ln L$ over 2 million MCMC generations

Bayesian Phylogenetic Inference - MCMC

All values should be estimated from post-burnin samples of the MCMC chain

Typically it is safe to use 20% as a cut-off for how many trees to discard as burnin (eg sample 5000 trees & discard the first 1000 as burnin)

However, examination of the history plot is a more accurate method to assess convergence

Bayesian Phylogenetic Inference - MCMC

Another solution to prevent getting stuck in suboptimal space:

Metropolis-coupled Markov chain Monte Carlo (MCMCMC or MC²)

Several chains are run simultaneously

One cold chain which is sampled and multiple (usually 3) heated chains that are not sampled but act as scouts - looking for better regions of parameter space (what does this remind you of?)
Bayesian Phylogenetic Inference - MCMC

Even with MCMC one should still do several independent runs and compare the results.

Prepare a regression plot of the posterior probabilities for each clade between different searches - should be highly correlated.

Also note: The trees visited by the MCMC algorithm are highly auto-correlated

- this is why we don’t sample every tree
- this is why we skip lots of trees and sample every 100 (or rarely every 1000) trees.

Bayesian Phylogenetic Inference - MCMC

Another issue is how many generations should one specify?

The more the better: some rules of thumb -
1. For data exploration (not publication) do as many as can be done in 0.5-1.5 days (overnight)
2. For final results do as many as possible (at least over a weekend - 3 nights)
3. Run multiple chains and compare results:
   - if they differ then none of the chains were run long enough

MCMC - Summary of important issues

1. Convergence - when the MCMC run has burned in, aka stationarity
2. Mixing - evidence the MCMC run isn’t getting stuck in one portion of parameter space (we want good mixing)
3. MCMCMC - multiple chains run simultaneously to help prevent the cold, sampled, chain from getting stuck
4. MCMC generation count - longer is better (see rules of thumb), 0.5 million steps is a good place to start but 2 million, or more, is recommended
5. Sampling the MCMC run - a minimum of every one tree per 100 steps
6. Comparing different runs - plot a regression of the posterior probabilities for each branch - should be highly correlated if both runs sampled the same region of parameter space

Bayesian Phylogenetic Inference

Outline
1. Introduction, history
2. Advantages over ML
3. Bayes’ Rule
4. The Priors
5. Marginal vs Joint estimation
6. MCMC
7. Posterior vs Bootstrap values

Bayesian Phylogenetic Inference - Priors

Priors can matter - their influence is under investigation - “noninformative” priors may have an influence on the posteriors - huge Bayesian literature on this subject

**Bayesian Phylogenetic Inference - Priors**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Current Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tranking</td>
<td>Decay/Fixed</td>
</tr>
<tr>
<td>Rate</td>
<td>Gamma(Fixed)</td>
</tr>
<tr>
<td>Aminoacids</td>
<td>Exponential/Fixed</td>
</tr>
<tr>
<td>Rhyme</td>
<td>Fixed</td>
</tr>
<tr>
<td>Nucleotides</td>
<td>Exponential/Fixed</td>
</tr>
<tr>
<td>St are</td>
<td>Exponential/Fixed</td>
</tr>
<tr>
<td>Rate</td>
<td>Uniform/Exponential/Fixed</td>
</tr>
<tr>
<td>St are</td>
<td>Uniform/Fixed</td>
</tr>
<tr>
<td>Rate</td>
<td>Uniform/Exponential/Fixed</td>
</tr>
<tr>
<td>St are</td>
<td>Uniform/Exponential/Fixed</td>
</tr>
<tr>
<td>Rate</td>
<td>Fixed</td>
</tr>
<tr>
<td>St are</td>
<td>Uniform/Fixed</td>
</tr>
<tr>
<td>Rate</td>
<td>Uniform/Exponential/Fixed</td>
</tr>
<tr>
<td>St are</td>
<td>Uniform/Fixed</td>
</tr>
<tr>
<td>Rate</td>
<td>Uniform/Exponential/Fixed</td>
</tr>
<tr>
<td>St are</td>
<td>Uniform/Fixed</td>
</tr>
</tbody>
</table>

**Posterior Probabilities and Bootstrap Values**

Recall that PP are easily interpreted as:

- **The probability the branch is true** given the model, the priors, the branch lengths and the data

Bootstrap values are less easily interpreted...

And using Maximum Likelihood they are difficult to obtain for large datasets (if a good ML search takes you 1 day, bootstrapping will take you 100 days - or you’ll need a parallel processor with 100 CPUs - a “super cluster”)

---

**Bootstrap - interpretation**

Huestenbeck & Rannala (2004) list 3 common interpretations:

1. Probability that a clade is correct (accuracy)
2. Robustness of the results to perturbation (repeatability / precision)
3. Probability of incorrectly rejecting a hypothesis of monophyly (1-P): probability of getting that much evidence if, in fact, the group did not exist

---

**Posterior Probabilities and Bootstrap Values**

It was noticed repeatedly that Posterior Probabilities tend to be higher than Bootstrap Values.

It was also noticed that there was no clear /strong relationship between the two values - correlations were weak (but strongest between ML-BS & BI-PP)

The higher PP expected from theory - we knew that BS are biased downward - perhaps PP are less biased and thus more likely to correspond to true measures of accuracy?

---

**Posterior Probabilities and Bootstrap Values**

“We show that Bayesian posterior probabilities are significantly higher than corresponding nonparametric bootstrap frequencies for true clades, but also that erroneous conclusions will be made more often. These errors are strongly accentuated when the models used for analyses are underparameterized. When data are analyzed under the correct model, nonparametric bootstrapping is conservative. Bayesian posterior probabilities are also conservative in this respect, but less so.”


**Posterior Probabilities and Bootstrap Values**

Simulations indicate
But both subject to error:
**Errors:**
- Type I - failure to reject a false hypothesis
- Type II - rejection of a true hypothesis


**Posterior Probabilities and Bootstrap Values**

Huelsenbeck & Rannala (2004) indicate that past simulations of data to test Bayesian methods have not been done properly
- a proper simulation study compares data that agrees with the assumptions of the model and contrasts this “best case scenario” with data that violate one or more assumptions
- These authors simulated best-case data for Bayesian methods by using the priors to generate the trees and parameter space in which to simulate data


**Posterior Probabilities and Bootstrap Values**

Best case scenarios - the data fit the model perfectly

**Posterior Probabilities and Bootstrap Values**

Huelsenbeck & Rannala (2004) wanted to answer the question
What does the posterior probability of a phylogenetic tree mean?

Data evolved using GTR+G but analyzed with too simple model

When the model is too simple (underfit) the PP are too high - error of accepting false hypotheses

Even JC69 does better than GTR if ASRV is accommodated

Data evolved using JC69 but analyzed with too complex model

When the model is too complex (overfit) the PP are only slightly too low - error of rejecting true hypotheses

Thus - better to be too complex than too simple - use at least GTR+G to be safe

### Posterior Probabilities and Bootstrap Values

#### Credible sets of trees

Bayesian inference provides a collection of trees assembled by their probabilities to total 95%

If the assumptions of the model are not violated this set has a ~95% probability of containing the true tree

---

What about a comparison with bootstrap values?

**Underfit model**

**Correct model**

PP far more biased due to underfitting than BS

---

**Recommendation**: use the most complex model available: GTR+G for Bayesian Inference (or GTR+I+G) [can't use AIC anyhow]
Posterior Probabilities and Bootstrap Values

Priors - noninformative vs informative

Example: 3 urns
urn 1 - full of coins, 80% are biased coins
urn 2 - full of coins, 10% are biased coins
urn 3 - empty and always has been

You are given a coin and asked to come up with a prior probability for it being from urn 1, urn 2, or urn 3

Obviously the prior for urn 3 is zero!

You might assign a P = 50% for urn 1 and urn 2, then flip your coin (gather data) & use Bayes’ Rule to obtain a posterior probability - see how the data change your prior (modify your beliefs in light of new evidence)

Posterior Probabilities and Bootstrap Values

Example - 5 taxa (A-E) yield 105 bifurcating rooted trees

A clade of 3 taxa (A,B,C) is in only 9 of the 105 trees
A clade of 2 taxa (A,B) is in 15 of the 105 trees

Prior to seeing any data we would expect a clade of (A,B) to be 1.7 times as likely as a clade of (A,B,C)

Although the trees are all equiprobable a priori, all the clades aren’t

Low BS values for mid-sized clades - an artifact? Saturation?

Brandley et al. (2006) found the answer -

Clade priors are only a problem if you have very few data

-which is obviously more of a problem for morphological data!

Terms - from lecture & readings

MCMC  
parameter space  
steps / generations  
convergence  
stationarity  
burnin  
history or generation plot  
mixing  
MCMC  
cold and heated chains  
Type I error  
Type II error  
clade priors  
Bayes factors

Study questions

What is the major difference between what the MCMC algorithm is doing and what a traditional tree searching (eg ML) algorithm does?

Why should one run MCMC searches for a long long time?

What can be done to help prevent getting stuck in suboptimal space?

Why do we throw out up to 20% of the initial samples taken from a MCMC search?

Why do we sample only every 100th tree from a MCMC run?

What are three common interpretations of bootstrap values? What is the interpretation of a posterior probability?

Study questions

Why would we expect PP to be more strongly correlated with ML-BS values than with MP-BS values?

Erxleben et al. (2003) made some important statements about PP and BS - what were two of their key findings?

One of Erxleben et al.’s findings was studied in depth by Huelsenbeck & Rannala (2005) - which finding was this and what were the conclusions and recommendations from this study?

What question did these authors set out to answer and what was the answer?

Picket & Randle (2005) noted an interesting and (if true) disturbing point regarding all branch support measures - what was it? They provided no correction for BS & Jackkning but what correction did they suggest for PP?