Project Notes

Ralph Møller Trane

Larget, 2013

Introduction

The posterior probability distribution on a set of phylogenetic trees is a well-defined mathematical object, if we are given a likelihood model, prior distribution, and data. However, it is difficult to use in partice.

Using MCMC methods, we can create samples of trees drawn from the posterior distribution, and base inferences on these posterior samples. Posterior distributions of phylogenetic trees typically are estimated with simple sample relative frequencies.

Large tree space ⇒ many trees not sampled. Even trees with non-zero probabilities might not be sampled. The probability of trees not sampled would be estimated to be 0. Inference based on a smaller tree space.

A posterior sample is summarized using a single consensus tree. Each edge is annotated with the probability that the corresponding clade is a monophyletic group relative to other taxa in the tree. (??) In some applications, the full probability distribution over the set of possible trees is required, or at least an accuracte approximation of this.

For example, in Bayesian concordance analysis (BCA), the first step of an analysis is to calculate separately the posterior probabilities of trees for each of many genes (Ane et al. 2007). If a tree is very probable in several distributions based on single genes, but is unsampled in the distribution of another given gene, the simple estimate of zero for the posterior probability of the tree for the given gene can bias the results in the second stage of the BCA analysis. What is needed to correct this shortcoming is a means to estimate the posterior probabilities of trees that do not appear in samples, but might be relatively probable because they contain clades that are probable.

Conditional clade probability distributions (CCDs) determine a probability distribution on trees that is an accurate estimate of the true posterior distribution, and can be applied to measure the posterior probability of any tree, whether or not it was sampled in the posterior sample.

Motivating Example

Q: Can we use clade probabilities to estimate tree probabilities?

If a tree T_1 is made up of clades $C_2, \ldots, C_11$, then $P(T_1) = P(C_2 \cap \cdots \cap C_11)$. If all clades we're independent, RHS equal to product. However, this is not generally the case. But we could hypothesize that clades in separate regions of the tree may be approximately independent. Or at least approximately conditionally independent. I.e. given a certain split, the two subclades might be independent.

To use this method, we need more details about the tree sample than the clade proportions. For example, to estimate $P(C_6|\ C_5) = P(C_5 \cap C_6)/P(C_5)$, we need to compute proportions of trees that include C_5, and proportions of trees that include C_5 and C_6.

Methods

If including single taxa notes (using that $P(C_i \cap C_j|C) = 1$ if C is a clade consisting of only C_i and C_j), one can write the probability of a tree T as

$$P(T) = \prod_{C \in \text{ all caldes of T}, |C|>1} P(L(C,T) \cap R(C,T)|C).$$

Here, $L(C,T)$ and $R(C,T)$ are the two subclades of C in the tree T.

Computational algorithm utilizing CCDs.

Validation; i.e. why we care

CCD method based on the theoretical assumption of approximate conditional independence of separated subtrees is accurate on a single sample from a single data set. For additional evidence, Larget examine its behavior with repeated samples on the Carnivora data set, on a sample from a uniform distribution over tree space where exact calculations are possible, and for multiple other real data sets.

Additional data sets

- When coverage drops, correlation between CCD and SRF drops.
  - Seems reasonable. With less of the tree space sampled, SRF estimates are "artificially higher". CCD does not suffer from this in the same way.

Discussion

CCD provides additional accuracy when a smaller proportion of the sample space is sampled. This extra accuracy is important if performing analyses such as BCA (Bayesian Concordance Analysis)

Importance sampling

In this particular situation

- use MrBayes to estimate conditional clade distributions.
- use these to estimate tree probabilities
- sample trees using these estimates

In general

Assume we want to calculate the integral $I = \int h(y)dy$. If f is a probability density function of a random variable Y, then $I = \int \frac{h(y)}{f(y)} f(y)dy = E_f(\frac{h(y)}{f(y)})$. Using Monte Carlo integration, we would sample N samples from the density f and then estimate the expectation as $\frac{1}{N}\sum_{i=1}^{N} \frac{h(y)}{f(y)}$, which in turn is also an estimator of I. Law of Large Numbers ensures us that this estimator converges to the desired expectation.

This works well... if we can sample from the density f. But what if we can't? Then, rather than sampling from f we sample from a different probability distribution g. We then estimate I by observing that

$$I = \int h(y)f(y)dy = \int h(y)\frac{f(y)}{g(y)}g(y)dy = \int \frac{h(y)f(y)}{g(y)}g(y)dy = E\left[\frac{h(Y)f(Y)}{g(Y)}\right].$$

So drawing N samples, $Y_1, \ldots, Y_N$, from g, $\hat{I} = \frac{1}{N}\sum_{i=1}^{N} \frac{h(Y_i)f(Y_i)}{g(Y_i)}$, which converges to $E_g\left[\frac{h(Y)f(Y)}{g(Y)}\right] = I$.

We have to be smart about how we choose g. If f/g gets too large, the standard error of $\hat{I}$ coul be infinite. So we want to make sure that g

- has a similar shape to f,

- has thicker tails than f,
- is proportional to |f|: $g(y) \propto |f(y)|$
- is such that we can sample from $g(y)$ with ease

What I should (?) do

- given MrBayes samples, estimate conditional clade distributions
- sample trees

Implement in Julia.

ccdprobs

Inputs: summary of MrBayes .t file produced with BUCKy. This

Output: for each input tree, the estimated tree probability based on the conditional clade probability distribution.

Overall summary of problem

We're given a data set consisting of aligned DNA. We aim at estimating the posterior probability of trees.

- MrBayes gives us a sample of trees (created somehow using MCMC).
- The posterior probability of a given tree can then be found using SRF (simple relative frequencies) or CCD (conditional clade probability distribution).
- The argument is that SRF gives unreliable estimates of the probability of trees that appear once in the sample, or not at all (these will be 0). CCD gives more realistic estimates of non-sampled trees.