Bayesian discrete trait analysis

Alex Beams

Department of Mathematics
Simon Fraser University

PhylogeogRaphy Workshop

Day 3

Table of Contents

A quick recap so far:

We've learned about phylogeographic methods that treat a phylogeny as given:

- discrete trait analysis
  - ace (joint and/or marginal reconstructions)
  - simmap (joint and/or marginal reconstructions + character mapping)
  - BiSSE (???)

A quick recap so far:

Now, we want to examine what happens when we jointly estimate a phylogeny with a model of character change

Then, we will extend this in two ways:

1. multi-type birth death models (this afternoon)
2. Bayesian stochastic search variable selection (tomorrow)

Sorry, structured coalescent fans. The fundamentals of joint estimation are sufficiently important in their own right

Table of Contents

Character evolution likelihood

For ancestral character estimation and stochastic character mapping, we considered fixed trees, $\mathcal{T}$, and defined Markov chains on them to obtain a likelihood for tip states, $\mathbf{x}$,

$$p(\mathbf{x}|\mathcal{T}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\mathcal{T}),$$

where we are marginalizing over unobserved configurations of node states, $\mathbf{y}$

Nucleotide evolution evolution likelihood

We have not discussed phylogenetic reconstruction in this workshop, but at this point we probably should.

Let **x** represent tip states of interest (geography or traits), and now let's introduce **s** to represent our multiple sequence alignment.

Nucleotide substitution models use Markov chains just like the trait evolution models we saw in ace and simmap.

If we let **u** represent the unobserved configuration of ancestral nucleotide states, the sequence likelihood given the tree has the form

$$p(\mathbf{s}|\mathcal{T}) = \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}),$$

except that with a large collection of sites in the genome, we calculate likelihoods at each site and multiply them all together to obtain the likelihood of the sequences

Sequence and character likelihood

We haven't had to worry about this because we have assumed the tree is fixed

However, in practice we will need to estimate the tree alongside our model for trait evolution in BEAST2 (and this will also involve estimating parameters of more complicated BiSSE-ish models)

For now, we just consider the likelihood of sequences and traits jointly:

$$p(\mathbf{s}, \mathbf{x}|\mathcal{T}) = \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}),$$

Joint character and sequence likelihood

BEAST2 can accommodate a discrete trait analysis that models trait evolution independently of sequences, conditional on a tree (and can use BSSVS, which we will discuss tomorrow)[1]:

$$
\begin{aligned}
p(\mathbf{s}, \mathbf{x}|\mathcal{T}) &= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) p(\mathbf{x}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) \sum_{\mathbf{u}} p(\mathbf{x}, \mathbf{y}|\mathcal{T})
\end{aligned}
$$

Note: the $p(\mathbf{x}, \mathbf{y}|\mathcal{T})$ depends on the generator matrix $\mathbf{Q}$ of our geography model (but I'm suppressing notation for readability)

[1] Philippe Lemey et al. "Bayesian phylogeography finds its roots". In: *PLoS computational biology* 5.9 (2009), e1000520.

Joint character and sequence likelihood

BEAST2 can accommodate a discrete trait analysis that models trait evolution independently of sequences, conditional on a tree (and can use BSSVS, which we will discuss tomorrow):

$$
\begin{aligned}
p(\mathbf{s}, \mathbf{x}|\mathcal{T}) &= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) p(\mathbf{x}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) \sum_{\mathbf{u}} p(\mathbf{x}, \mathbf{y}|\mathcal{T})
\end{aligned}
$$

If we try to estimate $\mathcal{T}$ with Maximum likelihood or perform Bayesian inference, this will make use of information contained in both s and x

But more importantly, the idea is that this accommodates uncertainty in $\mathcal{T}$ we develop more realistic estimates for $\mathbf{y}$ and $\mathbf{Q}$

Joint character and sequence likelihood

$$
\begin{aligned}
p(\mathbf{s}, \mathbf{x}|\mathcal{T}) &= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) p(\mathbf{x}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) \sum_{\mathbf{u}} p(\mathbf{x}, \mathbf{y}|\mathcal{T})
\end{aligned}
$$

Question If we omit the BSSVS part (which is like LASSO on the generator matrix for character change[2]), how different is this from just running ace on a posterior collection of trees?

[2] Philippe Lemey et al. "Bayesian phylogeography finds its roots". In: *PLoS computational biology* 5.9 (2009), e1000520.

Joint character and sequence likelihood

$$
\begin{aligned}
p(\mathbf{s}, \mathbf{x}|\mathcal{T}) &= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) p(\mathbf{x}, \mathbf{y}|\mathcal{T}), \\
&= \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) \sum_{\mathbf{u}} p(\mathbf{x}, \mathbf{y}|\mathcal{T})
\end{aligned}
$$

Problem How do we compare/average ancestral character estimates from different trees?

Table of Contents

Bayesian inference in BEAST2

Our joint likelihood of the sequences **s** and the trait **x**, conditional on the tree, $\mathcal{T}$, is

$$p(\mathbf{s}, \mathbf{x}|\mathcal{T}) = \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}),$$

Bayesian inference in BEAST2

Yesterday, we discussed models that generate trees. In BEAST2 parlance, these are called "tree priors":

$$p(\mathcal{T}|\theta).$$

If we incorporate this into our likelihood, we have a likelihood function in the parameters of the tree-generating model:

$$p(\mathbf{s}, \mathbf{x}|\theta) = \int_{\mathcal{T}} \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}) p(\mathcal{T}|\theta),$$

Bayesian inference in BEAST2

For now, we are focusing on the situation where we can treat character evolution independent of sequence evolution, given a tree,

$$p(\mathbf{s}, \mathbf{x}|\theta) = \int_{\mathcal{T}} \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}) p(\mathcal{T}|\theta),$$

$$= \int_{\mathcal{T}} \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}) \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\mathcal{T}) p(\mathcal{T}|\theta)$$

This can incorporate BSSVS as-is[3], but this morning let's start by just using this to jointly estimate $\mathbf{Q}$ and the phylogeny without the BSSVS part

[3] Philippe Lemey et al. "Bayesian phylogeography finds its roots". In: *PLoS computational biology* 5.9 (2009), e1000520.

Bayesian inference in BEAST2

Of course, BEAST2 is Bayesian, and relies on MCMC. We must supply a prior distribution $p(\theta)$ for the tree-generating model (birth-death, coalescent, whatever), as well as a prior distribution $p(\mathbf{Q})$ for the character evolution model, as well as $p(\mathbf{A})$ for the nucleotide evolution model:

$$p(\theta, \mathbf{Q}, \mathbf{A}|\mathbf{s}, \mathbf{x})$$
$$\propto \int_{\mathcal{T}} \sum_{\mathbf{u}} p(\mathbf{s}, \mathbf{u}|\mathcal{T}, \mathbf{A}) \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\mathcal{T}, \mathbf{Q}) p(\mathcal{T}|\theta) p(\theta) p(\mathbf{Q}) p(\mathbf{A})$$

Bayesian inference in BEAST2

The multi-type birth death model, the structured coalescent model, and their variants deal with the situation where sequence evolution and character change are not conditionally independent given the tree:

$$p(\theta, \mathbf{Q}, \mathbf{A}|\mathbf{s}, \mathbf{x})$$
$$\propto \int_{\mathcal{T}} \sum_{\mathbf{u},\mathbf{y}} p(\mathbf{s}, \mathbf{x}, \mathbf{u}, \mathbf{y}|\mathcal{T}) p(\mathcal{T}|\theta) p(\theta) p(\mathbf{Q}) p(\mathbf{A})$$