Review of Ancestral Character Estimation and Stochastic Character Mapping

Alex Beams

Department of Mathematics
Simon Fraser University

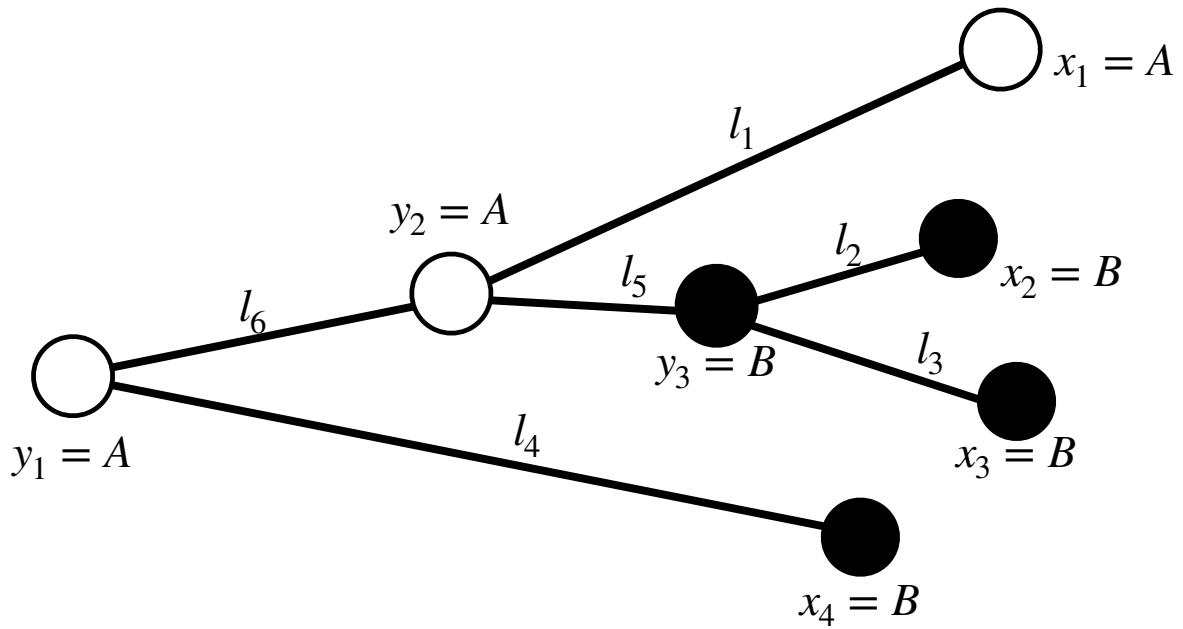
PhylogeogRaphy Workshop

Day 2

Table of Contents

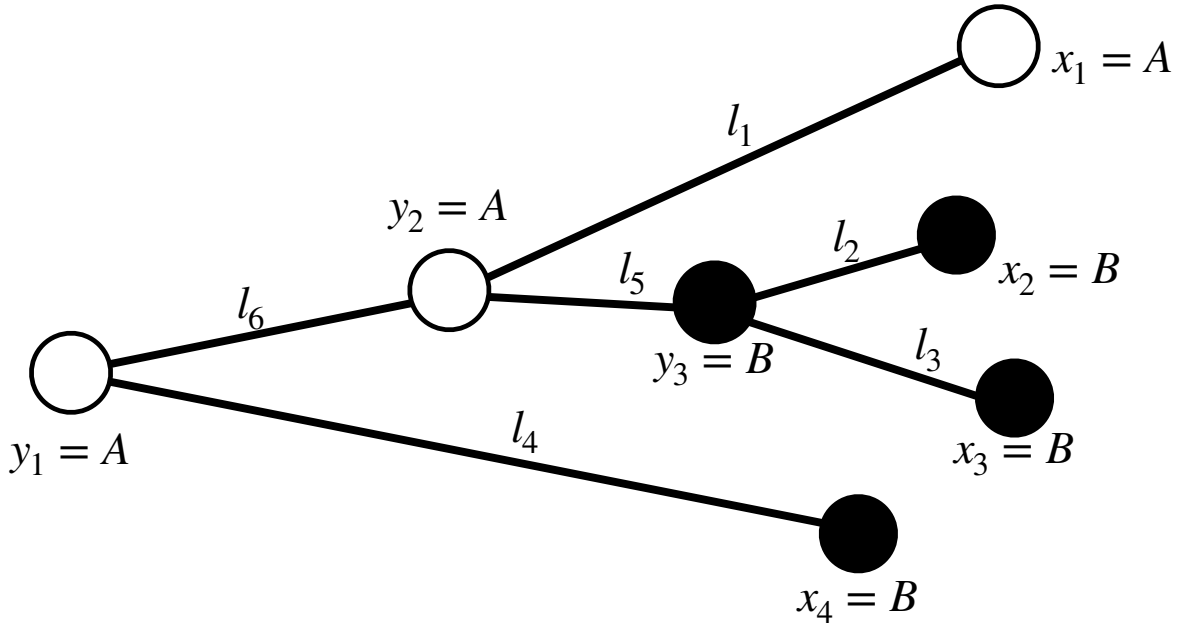
Markov models for discrete states on trees

Suppose we are given a phylogenetic tree, $\mathcal{T}$, we have information about tip locations, x_i, and we superimpose a continuous-time Markov chain with two states (A and B) on the tree



For a given configuration of internal node states, $\mathbf{y} = (y_1, y_2, ..., y_n)^T$, we can calculate transition probabilities along each branch

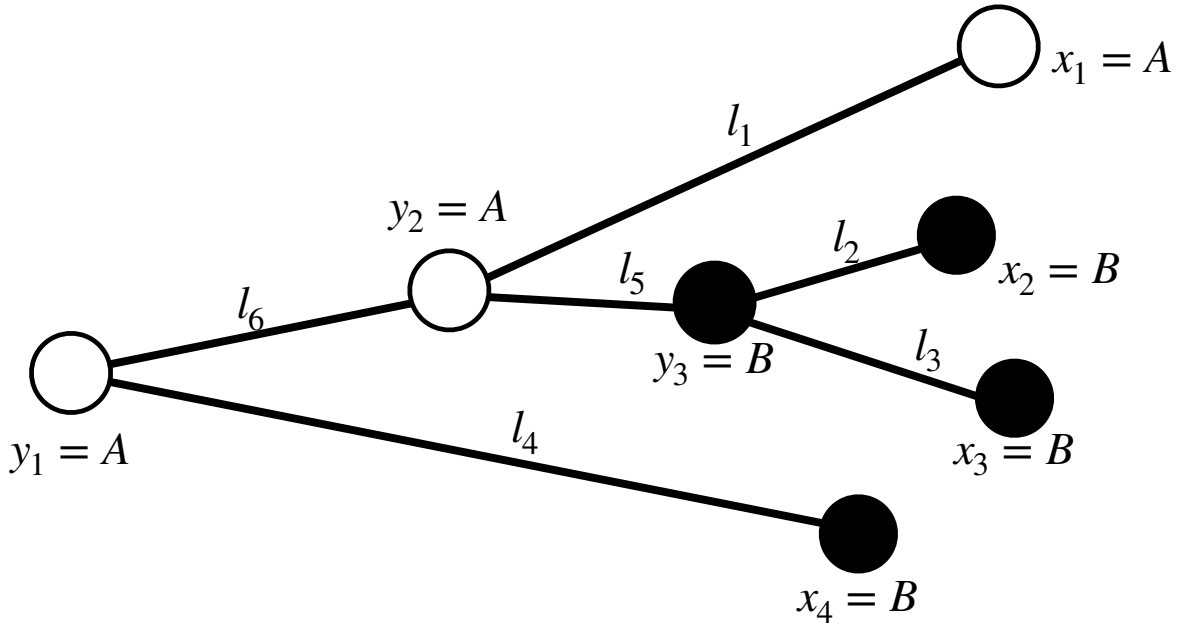
Markov models for geographic movement



Example: transition from $y_1 = A$ to $x_4 = B$:

$$p(x_4|y_1, l_4) = p(y_1)e^{Ql_4}P_0, \text{ with } P_0 = (1, 0)^T.$$

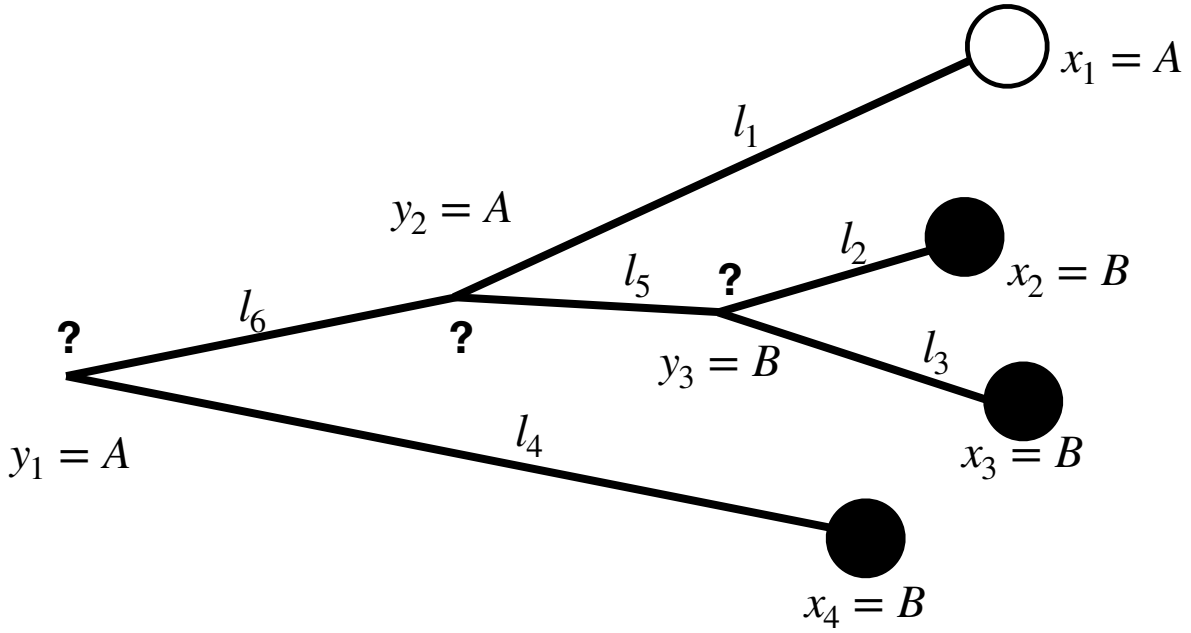
Markov models for geographic movement



The **likelihood** of the model is the probability tip states, $\mathbf{x} = (x_1, x_2, .., x_n)^T$ given the branch lengths l_i and parameters of the Markov model in **Q**:

$$P(\overrightarrow{x}|\mathcal{T}) = \sum_{\mathbf{y}} \pi(y_1)p(x_4|y_1)p(y_2|y_1)p(x_1|y_2)p(y_3|y_2)p(x_2|y_3)p(x_4|y_3)$$

Calculating likelihood



Because don't know the the states of internal nodes, y_i, we have to sum over all possible configurations, **y**:

$$P(\vec{x}|\mathcal{T}) = \sum_{\mathbf{y}} \pi(y_1)p(x_4|y_1)p(y_2|y_1)p(x_1|y_2)p(y_3|y_2)p(x_2|y_3)p(x_4|y_3)$$

Calculating likelihood

Even though there are a lot of conditionals ($p(y_k|y_j)$) in the expression of the likelihood,

$$P(\overrightarrow{x}|\mathcal{T}) = \sum_{\mathbf{y}} \pi(y_1)p(x_4|y_1)p(y_2|y_1)p(x_1|y_2)p(y_3|y_2)p(x_2|y_3)p(x_4|y_3),$$

don't be confused – this is still a marginalization of a joint distribution, having the form

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

The likelihood is **not** of the form $p(\mathbf{x}|\mathcal{T}, \mathbf{y})$

In the likelihood, **x** is the configuration of tip states we observe, and **y** is the configuration of internal node states we do not observe.

Ancestral states as latent variables

The basic form of this likelihood,

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

immediately suggests how we should determine ancestral states, $\mathbf{y}$.

Our best estimate for the ancestral state configuration is the particular $\mathbf{y}^*$ that makes the largest contribution to $p(\mathbf{x}|\mathcal{T})$

This is the **joint reconstruction** of ancestral states

This is a hard problem. Various approximate methods exist, and ace uses a two-pass approximate joint reconstruction approach by default[1]

[1] Tal Pupko et al. "A fast algorithm for joint reconstruction of ancestral amino acid sequences". In: *Molecular biology and evolution* 17.6 (2000), pp. 890–896.

Table of Contents

Marginal reconstruction of ancestral states

Marginal ancestral character estimation[2] proceeds by calculating, for each internal node y_k, the expression

$$p(\mathbf{x}, y_k|\mathcal{T}) = \sum_{y_1,y_2,..,y_{k-1},y_{k+1},\cdots,y_n} p(\mathbf{x}, y_1, y_2, .., y_{k-1}, y_k, y_{k+1}, \ldots, y_n|\mathcal{T}).$$

The value of y_k that corresponds to the largest $p(\mathbf{x}, y_k|\mathcal{T})$ is the **marginal ancestral character estimate** for node y_k.

[2] Dolph Schluter et al. "Likelihood of ancestor states in adaptive radiation". In: *Evolution* 51.6 (1997), pp. 1699–1711, Joseph Felsenstein. "Inferring phylogenies". In: *Inferring phylogenies*. 2004, pp. 664–664.

Pruning

How do we calculate the $p(\mathbf{x}, y_k|\mathcal{T})$? Recall from our example that the likelihood can be calculated using the *pruning* (or, dynamic programming) method:

$$\sum_{\vec{y}} \pi(y_1)p(x_4|y_1)p(y_2|y_1)p(x_1|y_2)p(y_3|y_2)p(x_2|y_3)p(x_4|y_3) =$$

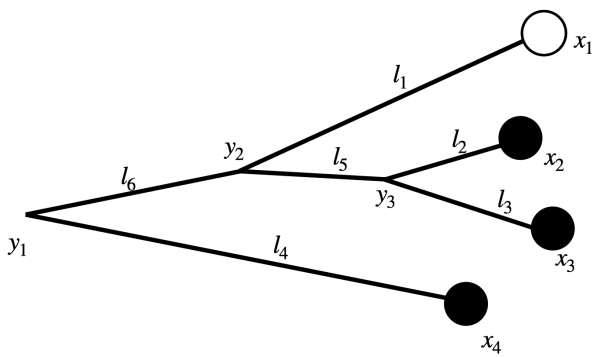
$$\sum_{y_1} \pi(y_1)p(x_4|y_1)\left(\sum_{y_2}(p(y_2|y_1)p(x_1|y_2)\left(\sum_{y_3} p(y_3|y_2)p(x_2|y_3)p(x_3|y_3)\right)\right)$$

The calculation is carried from right to left, starting with the marginalization in y_3, then in y_2, then in y_1

Pruning

The form of the pruned likelihood indicates that we calculate from the tips toward the root. First, we calculate the likelihood of the tree descending from node y_3,

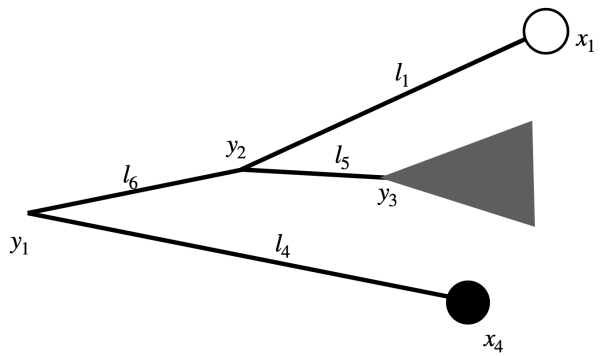
$$L_{y_3}(z) = p(x_2|y_3 = z, l_2)p(x_3|y_3 = z, l_3) :$$



Pruning

The notation $L_{y_k}(z)$ refers to the likelihood of the tree descending from node y_k, given that y_k has state z.

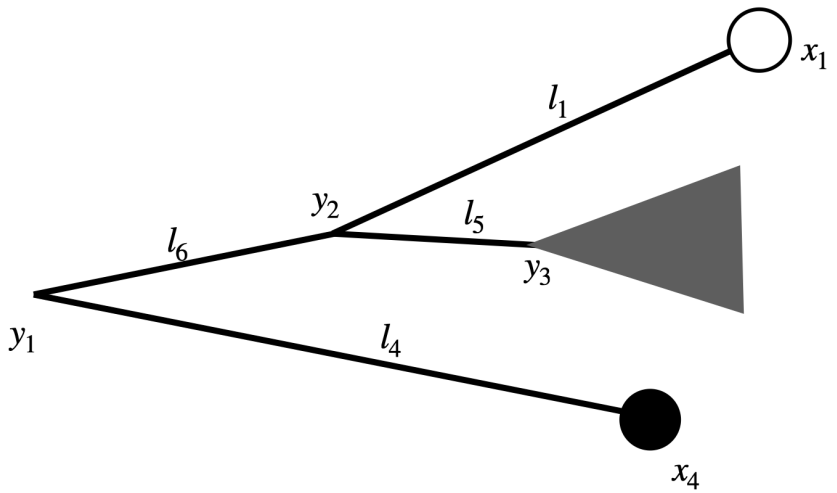
So we have as many terms $L_{y_3}(z)$ as there are possible values, z, for the node y_3



Pruning

Once we have $L_{y_3}(z)$, we can calculate $L_{y_2}(z)$:

$$L_{y_2}(z) = p(x_1|y_2 = z, l_1)\left(\sum_w p(y_3 = w|y_2 = z, l_5)L_{y_3}(w)\right)$$

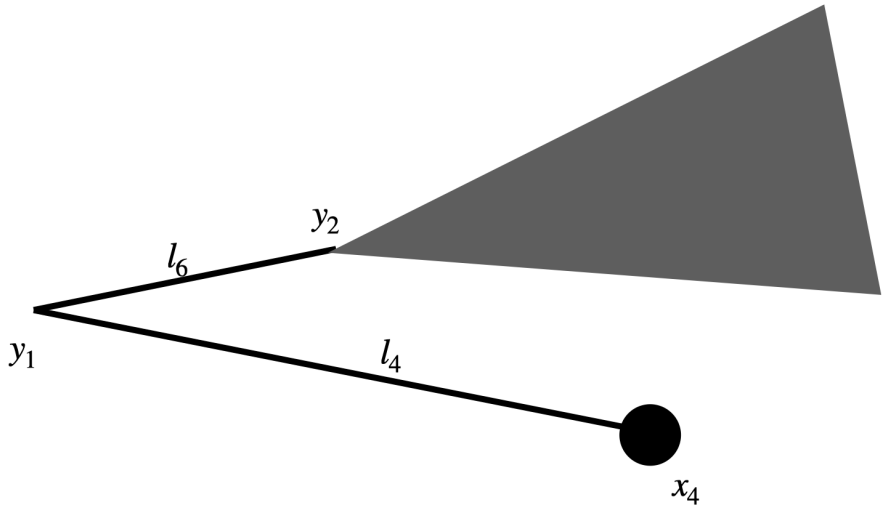


Pruning

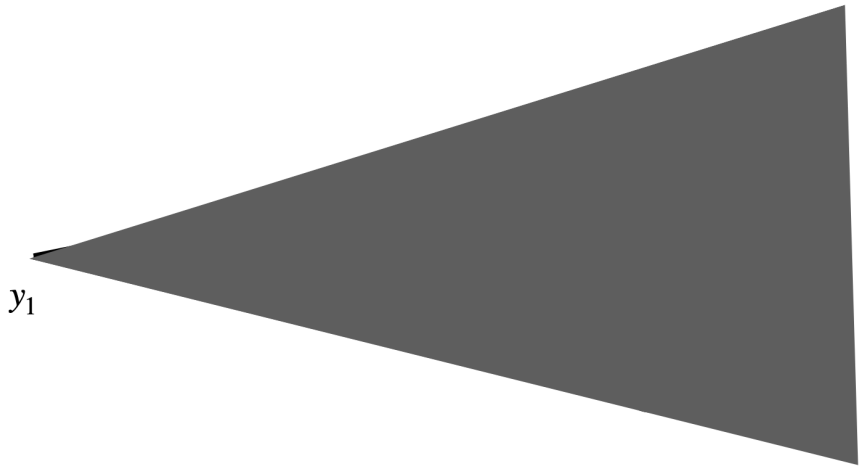
Finally, we get to $L_{y_1}(z)$:

$$L_{y_1}(z) = p(x_4|y_1 = z, l_4)\left(\sum_u p(y_2 = u|y_1 = z, l_6)L_{y_2}(u)\right),$$

but we have contributions for each of the particular values of $y_1 = z$.



Pruning



To obtain the tree likelihood, we have to combine these, and usually this is done by setting

$$L_{y_1} = \sum_{y_1} \pi(y_1) L_{y_1}(z).$$

Pruning

$$L_{y_1} = \sum_{y_1} \pi(y_1) L_{y_1}(z).$$

In ace, the distribution π is chosen to be the stationary distribution of the Markov chain (which ensures a detailed balance condition holds throughout the tree whenever the Markov model is reversible)

The $\pi(y_1)$ is necessary for the tree likelihood to be interpreted as a marginalization of a joint distribution; there is no Bayesian interpretation here (or at least, there does not need to be)

Pruning

$$L_{y_1} = \sum_{y_1} \pi(y_1) L_{y_1}(z).$$

In the phytools functions (e.g. make.simmap), the distribution π can be the stationary distribution of the Markov chain like ace, but can also be the "FitzJohn root prior"[3], or a user-supplied distribution

For reversible Markov chains (irreducible, aperiodic, and positive-recurrent), selecting π to be the stationary distribution ensures detailed balance throughout the phylogeny

[3] Richard G FitzJohn, Wayne P Maddison, and Sarah P Otto. "Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies". In: *Systematic biology* 58.6 (2009), pp. 595–611.

Marginal ancestral state estimation

$$L_{y_1} = \sum_{y_1} \pi(y_1) L_{y_1}(z).$$

The individual $L_{y_1}(z)$ terms in the likelihood suggest a way to carry out **marginal ancestral character estimation**.

Just to reiterate: $L_{y_1}(z)$ is the likelihood of the tree conditional on the root taking on state z

The value of z that makes the largest contribution to L_{y_1} is the most likely state. We select it as our **marginal** estimate for the root state.

Marginal ancestral state estimation

So, we can find a marginal ancestral character estimate for the root of the tree, but what about all of the other nodes?

Whenever the Markov model of trait change is reversible (it usually is), the tree can actually be re-rooted at any node. So we just re-root, and then carry out the pruning algorithm each time[4].

In this way, it is possible to calculate

$$p(\mathbf{x}, y_k|\mathcal{T}) = \sum_{y_1,y_2,..,y_{k-1},y_{k+1},...,y_n} p(\mathbf{x}, y_1, y_2, .., y_{k-1}, y_k, y_{k+1}, ..., y_n|\mathcal{T}),$$

for each y_k. The value of y_k that gives the largest value for this expression is the marginal estimate.

[4] Joseph Felsenstein. "Inferring phylogenies". In: *Inferring phylogenies*. 2004, pp. 664–664.

Joint vs. Marginal ancestral state estimation

Marginal character estimation is certainly tractable, but does not always produce the same estimates a joint reconstruction would

For more information about the standard two-pass approach to approximate a joint reconstruction, read Pupko et. al.[5]

In the ace function in R, there is an optional command "marginal" that defaults to FALSE, so that ace carries out an approximate joint reconstruction. Setting marginal=TRUE calculates ancestral states using the marginal reconstruction approach outlined here

[5] Tal Pupko et al. "A fast algorithm for joint reconstruction of ancestral amino acid sequences". In: *Molecular biology and evolution* 17.6 (2000), pp. 890–896.

Summarizing up to now:

Defining Markov models on trees allows us to model trait changes over time.

The likelihood function $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}|\mathcal{T})$ is calculated using transition probabilities of the Markov chain, and is a marginalization over unobserved ancestral states, $\mathbf{y}$

After fitting a model with Maximum Likelihood, we can estimate ancestral states $\mathbf{y}$ with an (approximate) joint approach, or a marginal one (which might be close to the joint estimate)

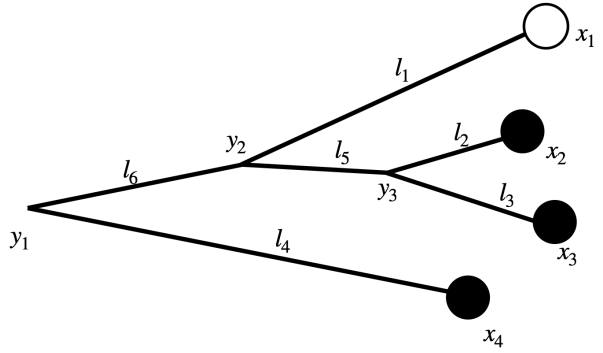
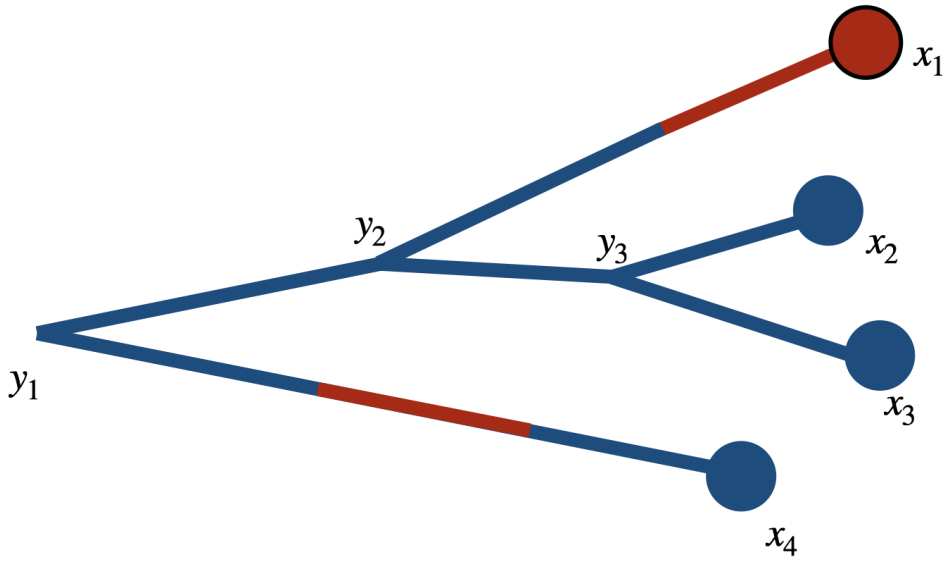


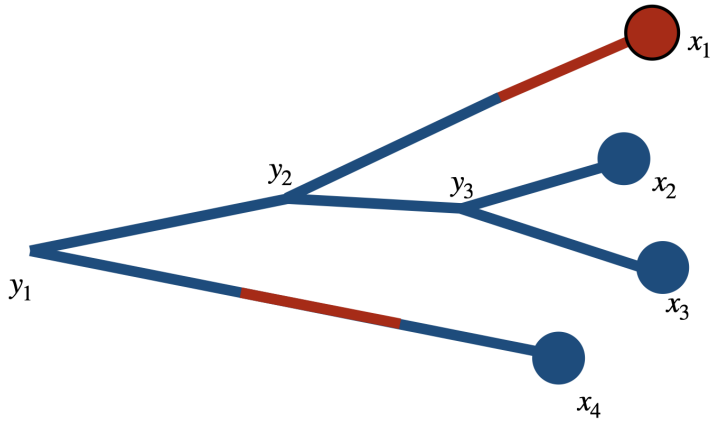
Table of Contents

Stochastic character mapping

In addition to estimating ancestral states at particular nodes, we can estimate individual changes along branches of a phylogeny



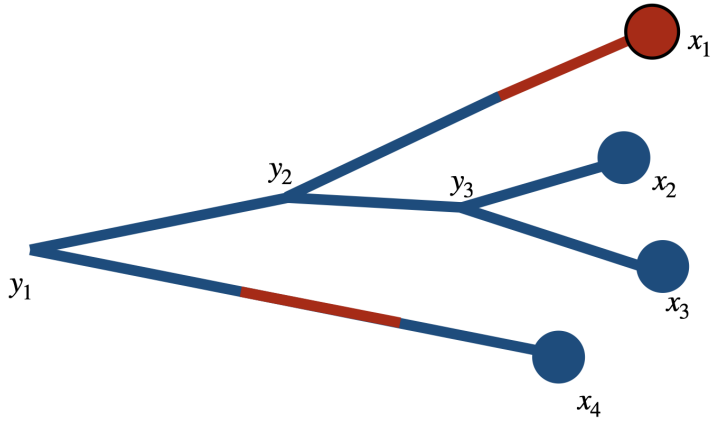
Stochastic character mapping



This proceeds in two (actually, three) steps:

0. Fit a Markov model (using ace from ape, or FitMk from phytools) to obtain a generator, $\mathbf{Q}$
1. Use the Markov model to sample ancestral states at internal nodes (could be a joint or marginal reconstruction)
2. Simulate the Markov model in continuous time along each branch, ensuring that boundary conditions at nodes are satisfied

Sampling node configurations



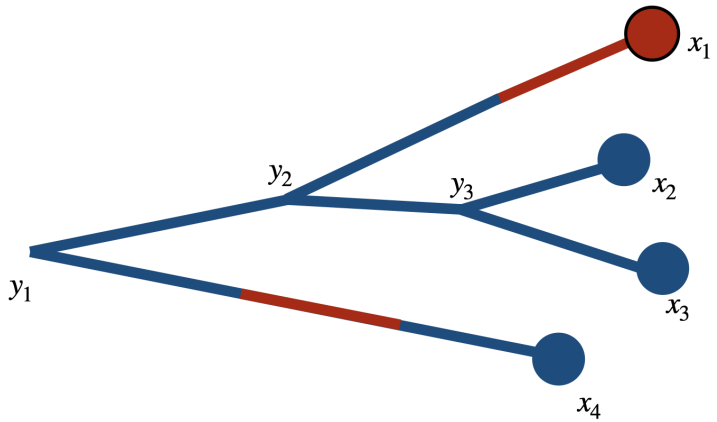
To sample ancestral state configurations in a marginal approach, proceed as follows:

For each internal node y_k, re-root the tree at y_k, and then

I use pruning to calculate $L_{y_k}(z)$ for each possible state z

II sample a realization of Z using these as sampling weights

Sampling node configurations

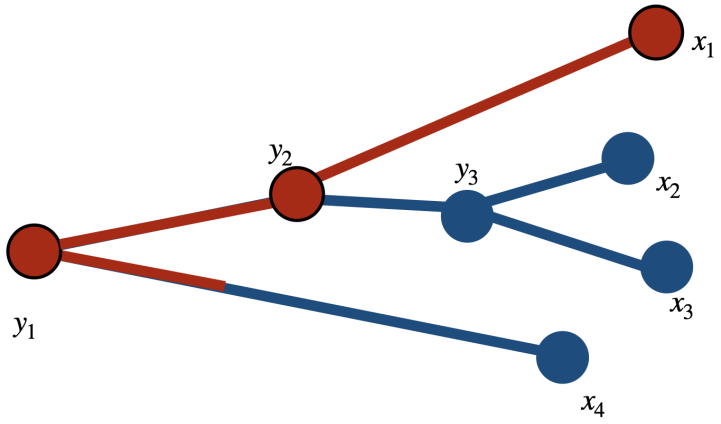


In principle, one could sample joint configurations of ancestral states. For each configuration of ancestral states $\mathbf{y}$,

I. calculate $p(\mathbf{x}, \mathbf{y}|\mathcal{T})$ for each joint configuration, $\mathbf{y}$
II. sample $\mathbf{Y}$, using these as sampling weights

(This is hard, and it does not seem like this is the procedure that any method actually uses in practice)

Simulate state changes along branches



Once an internal node configuration has been sampled through a marginal or joint approach, simulate the Markov chain along each branch

Trajectories of the Markov chain on any branch are conditionally independent of each other, given the boundary conditions supplied by the sampled node states

Stochastic mapping vs. ancestral character estimation

Stochastic mapping is not a replacement for ancestral character estimation – they are complementary

Need to do all of the ancestral character estimation calculations to do stochastic character mapping

Stochastic character mapping helps us count the number of trait changes in a phylogeny. Ancestral character estimation does not directly give this to us