Ancestral Character Estimation

Alex Beams

Department of Mathematics
Simon Fraser University

PhylogeogRaphy Workshop

September 29, 2025

1. Background for ancestral character estimation

2. Markov chains

3. Likelihood for Markov models on trees

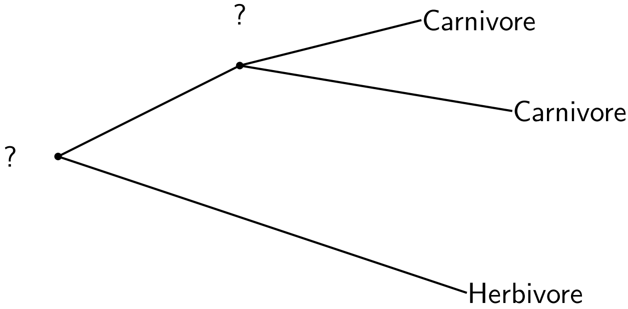
4. Fitting models in R

Table of Contents

1. Background for ancestral character estimation
2. Markov chains
3. Likelihood for Markov models on trees
4. Fitting models in R

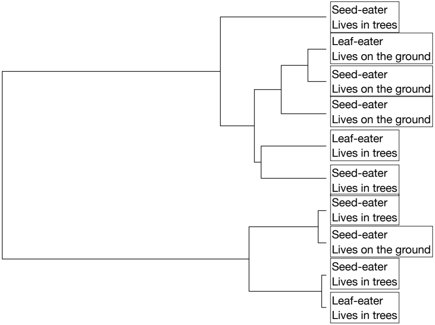
Why do ancestral character estimation?

In evolutionary biology, we might ask: how did a current set of traits evolve?



The historical development of these methods:

Originally, the methods we are about to discuss were developed for understanding **correlated evolution of** ≥ 2 **traits**



Seed-eater
Lives in trees

Leaf-eater
Lives on the ground

Seed-eater
Lives on the ground

Seed-eater
Lives on the ground

Leaf-eater
Lives in trees

Seed-eater
Lives in trees

Seed-eater
Lives in trees

Seed-eater
Lives on the ground

Seed-eater
Lives in trees

Leaf-eater
Lives in trees

Example: Is evolution of diet correlated with habitat?

Traits might be correlated because of common ancestry. Needed to develop methods to account for this

A variety of methods for ancestral reconstruction

A number of different methods to infer ancestral states exist:

- parsimony
- Markov models of discrete state changes
- Brownian motion or Ornstein-Uhlenbeck for continuous traits changing over time (but we won't discuss in this workshop)

Parsimony

Parsimony is intuitive and seems reasonable at first glance

There are some undesirable outcomes from this method

- multiple changes might occur over long periods of time
- parsimony reconstructions may not be unique, and it is also difficult to quantify uncertainty in ancestral inferences

A different approach: using Markov models

If traits evolve in a neutral fashion, then we could model them in the same way that we imagine nucleotides in genomes change over time

For discrete traits, we postulate a rate of change from each trait value to another

Markov chains underlie phylogenetic reconstruction

Nucleotide changes over time are modeled with continuous-time Markov chains .. why not do this with discrete traits as well?

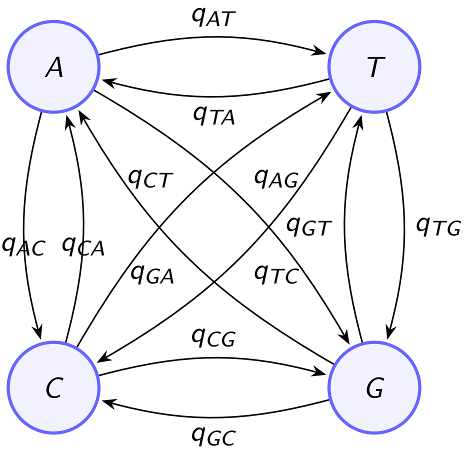


Table of Contents

Essential features of Markov chains

If you worked through *Simulating nucleotide substitution models* on the website, you may have familiarized yourself with some of the mathematics of Markov chains

Briefly, we will discuss some of the essential features together

Transition probabilities

Suppose we have a simple 2-state Markov chain with transitions occurring in continuous time

We let $P_i(t)$ represent the probability that the chain resides in state i at time t

In our simple 2-state case, we can write ordinary differential equations (ODEs) describing rates of change in state probabilities:

$$\frac{dP_1}{dt} = -q_{12}P_1 + q_{21}P_2,$$
$$\frac{dP_2}{dt} = q_{12}P_1 - q_{21}P_2.$$

Transition probabilities

$$\frac{dP_1}{dt} = -q_{12}P_1 + q_{21}P_2,$$
$$\frac{dP_2}{dt} = q_{12}P_1 - q_{21}P_2.$$

In the R demo for nucleotide substitution models, we show how to solve this sytem by hand. There are two key steps

- recognize that $P_1 + P_2 = 1$ to eliminate a variable
- use an integrating factor to solve the one-dimensional ODE that is left

Transition probabilities

In the demo, we see that solutions to the system

$$\frac{dP_1}{dt} = -q_{12}P_1 + q_{21}P_2,$$
$$\frac{dP_2}{dt} = q_{12}P_1 - q_{21}P_2,$$

have the form

$$P_1 = P_{1,0}e^{-(q_{12}+q_{21})t} + \frac{q_{21}}{q_{12}+q_{21}}\left(1 - e^{-(q_{12}+q_{21})t}\right),$$
$$P_2 = P_{2,0}e^{-(q_{12}+q_{21})t} + \frac{q_{12}}{q_{12}+q_{21}}\left(1 - e^{-(q_{12}+q_{21})t}\right).$$

Transition probabilities

$$P_1 = P_{1,0} e^{-(q_{12}+q_{21})t} + \frac{q_{21}}{q_{12}+q_{21}} \left(1 - e^{-(q_{12}+q_{21})t}\right),$$
$$P_2 = P_{2,0} e^{-(q_{12}+q_{21})t} + \frac{q_{12}}{q_{12}+q_{21}} \left(1 - e^{-(q_{12}+q_{21})t}\right).$$

Examining the behavior of this solutions, we verify that $P_i(0) = P_{i,0}$ (satisfies initial conditions)

We also immediately see the long-term behavior: $\lim_{t\to\infty} P_i(t) = \frac{q_{ji}}{q_{ij}+q_{ji}}$

Stochastic rate matrix

If we go back to the original ODEs,

$$\frac{dP_1}{dt} = -q_{12}P_1 + q_{21}P_2,$$
$$\frac{dP_2}{dt} = q_{12}P_1 - q_{21}P_2,$$

we can re-write this system as

$$\frac{dP}{dt} = QP,$$

where

$$Q = \begin{pmatrix} -q_{12} & q_{21} \\ q_{12} & -q_{21} \end{pmatrix}$$

and $P = (P_1, P_2)^T$.

Stochastic rate matrix

Writing Q like this:

$$Q = \begin{pmatrix} -q_{12} & q_{21} \\ q_{12} & -q_{21} \end{pmatrix},$$

is using the convention that q_{ij} is the transition rate from state i to state j.

Stochastic rate matrix

Sometimes, the entries of Q are written like this,

$$Q = \begin{pmatrix} -q_{21} & q_{12} \\ q_{21} & -q_{12} \end{pmatrix},$$

so that q_{ij} is the transition rate into state i from state j.

Pay careful attention to this when working in R!

Stochastic rate matrix

$$Q = \begin{pmatrix} -q_{12} & q_{21} \\ q_{12} & -q_{21} \end{pmatrix},$$

A stochastic rate matrix like this always has the property that rates in columns sum to zero

Thus, all stochastic rate matrices have a nullspace, i.e., a zero eigenvalue, making them singular

Stochastic matrices also have negative terms on the diagonal. The sum of the diagonal equals sum of the eigenvalues – so we know that the sum of the eigenvalues is negative

Stochastic rate matrix

$$Q = \begin{pmatrix} -q_{12} & q_{21} \\ q_{12} & -q_1 \end{pmatrix},$$

In this example, the nullspace of Q, $\mathcal{N}(Q)$, is spanned by the vector

$$P^* = \begin{pmatrix} q_{21} \\ q_{12} \end{pmatrix}.$$

If we normalize this vector, we get

$$P^* = \begin{pmatrix} \frac{q_{21}}{q_{12}+q_{21}} \\ \frac{q_{12}}{q_{12}+q_{21}} \end{pmatrix}.$$

Stochastic rate matrix

$$P^* = \begin{pmatrix} \frac{q_{21}}{q_{12}+q_{21}} \\ \frac{q_{12}}{q_{12}+q_{21}} \end{pmatrix}.$$

This vector is the same one that describes the long-term behavior of the system

$$P_1 = P_{1,0}e^{-(q_{12}+q_{21})t} + \frac{q_{21}}{q_{12}+q_{21}}\left(1 - e^{-(q_{12}+q_{21})t}\right),$$
$$P_2 = P_{2,0}e^{-(q_{12}+q_{21})t} + \frac{q_{12}}{q_{12}+q_{21}}\left(1 - e^{-(q_{12}+q_{21})t}\right).$$

Stochastic rate matrix

This is no coincidence. The normalized vector P^* satisfying

$$\frac{dP^*}{dt} = QP^* = 0$$

corresponds to the **stationary distribution** of the Markov chain

Regardless of the initial conditions, in the long term limit, the probability of finding the Markov chain in a particular state is described by the stationary distribution

Master equation

In general, can always obtain the probabilities of a Markov chain by solving the equation

$$\frac{dP}{dt} = QP.$$

This is called the **Master equation**, or sometimes the **Kolmogorov equation**, the **Chapman-Kolmogorov equation**, or the **Forward Kolmogorov equation**

We may use these terms throughout the workshop.

Analyzing the Master eqution

Whenever the matrix Q does not depend on t, we can easily solve the Master equation,

$$\frac{dP}{dt} = QP.$$

The general solution is

$$P(t) = P_0 e^{Qt},$$

where P_0 is an initial condition capturing the probability the chain resides in a particular state at time $t = 0$, and e^{Qt} is the matrix exponential

This is true for an arbitrary (but finite) collection of states

Calculating transition probabilities

The general solution is

$$P(t) = e^{Qt} P_0,$$

where P_0 is the vector of initial state probabilities at time 0, and e^{Qt} is the matrix exponential operation defined by

$$e^{At} = I + At + \frac{1}{2}!(At)^2 + \frac{1}{3!}(At)^3 + \ldots \tag{1}$$

In R, the function expm (from the package of the same name) can calculate this numerically. So, in practice, we just need the stochastic rate matrix, Q, and an initial condition P_0

Does a stationary distribution always exist?

There is always a 0 eigenvalue of the matrix Q

This means rows/columns are not independent, so we can always find a vector in the nullspace of Q, $\mathcal{N}(Q)$

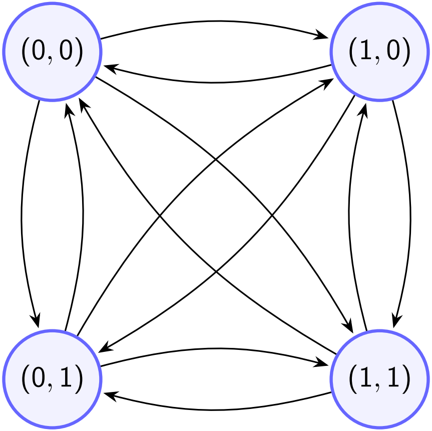
If $\mathcal{N}(Q)$ is spanned by a single vector, then the stationary distribution is unique because of the requirement that $\sum_i P_i = 1$ (assuming some other requirements on Q like aperiodicity and positive recurrence)

Properties guaranteeing a unique stationary distribution

Positive recurrence: the expected time for the Markov chain to return to any particular state is finite

Irreducibility: for a continuous-time Markov chain, there is a positive probability of transitioning from one state to any other (the state space cannot be decomposed into disjoint sets having transitions only amongst themselves)

Back to the original motivation: modeling correlated evolution of two traits



$(x,y) = (\text{trait 1 value}, \text{trait 2 value})$

Markov models for binary trait evolution

We will use Markov models on trees for the next several days to understand phylogeography

$$Q = \begin{pmatrix} -\lambda_{12} & \lambda_{21} \\ \lambda_{12} & -\lambda_{21} \end{pmatrix}$$

Can we use these models to understand evolutionary histories?

Table of Contents

1 Background for ancestral character estimation

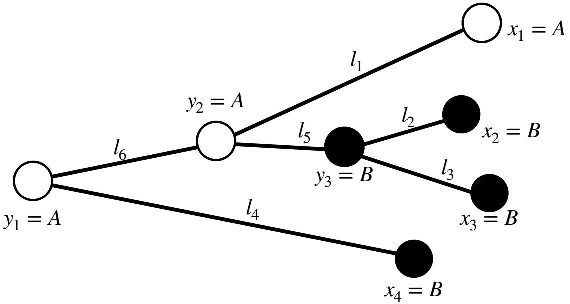
2 Markov chains

3 Likelihood for Markov models on trees

4 Fitting models in R

Markov models for geographic movement

We are given a phylogenetic tree, $\mathcal{T}$, we have information about tip locations, x_i, and we consider a continuous-time Markov chain defined on the tree describing location changes

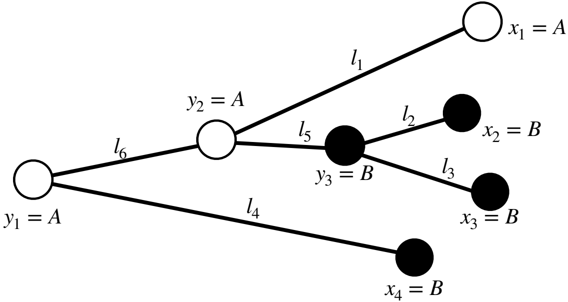


If we knew the locations of internal nodes, y_i, we could calculate the probability of observing the tip states, x_i:

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

Markov models for geographic movement

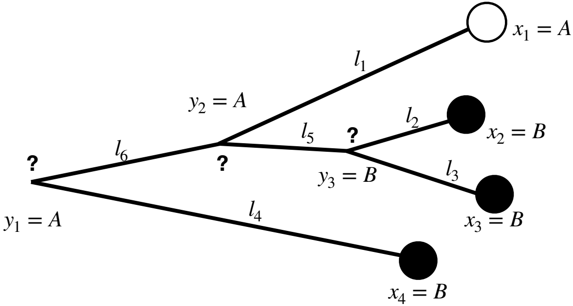
We are given a phylogenetic tree, $\mathcal{T}$, we have information about tip locations, x_i, and we consider a continuous-time Markov chain defined on the tree describing location changes



If we knew the locations of internal nodes, y_i, we could calculate the probability of observing the tip states, x_i:

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

Calculating likelihood



We don't actually know the the locations of internal nodes, y_i, so we have to sum over all possible configurations:

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

Pruning

We can economize on computations by “pruning”:

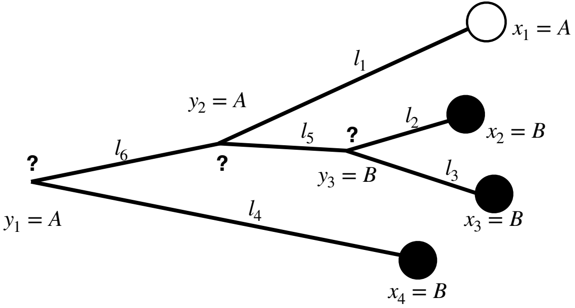
$$\sum_{\vec{y}} \pi(y_1)p(x_4|y_1)p(y_2|y_1)p(x_1|y_2)p(y_3|y_2)p(x_2|y_3)p(x_4|y_3) =$$

$$\sum_{y_1}(\pi(y_1)p(x_4|y_1)\times$$

$$\sum_{y_2}(p(y_2|y_1)p(x_1|y_2)\times$$

$$\sum_{y_3}(p(y_3|y_2)p(x_2|y_3)p(x_3|y_3))))$$

Can fit to data using Maximum Likelihood



The transition probabilities $p(y_j|y_i)$, $p(x_k|y_j)$ depend on parameters, θ, of a continuous-time Markov chain that we specify; if we interpret $P(\overrightarrow{x}|\mathcal{T},\theta)$ as a function of θ, then we can fit the Markov model to the data $(\mathcal{T}, \overrightarrow{x})$ using Maximum Likelihood

Markov models for binary trait evolution

In the original analysis, Pagel compared two Markov models:

If the traits evolve in a correlated manner, state changes between the four combinations of states are modeled with a 4 × 4 Markov generator (up to 12 parameters):

$$Q = \begin{pmatrix} -\sum_{j\neq 1}\lambda_{1j} & \lambda_{12} & \lambda_{13} & \lambda_{14} \\ \lambda_{21} & -\sum_{j\neq 2}\lambda_{2j} & \lambda_{23} & \lambda_{24} \\ \lambda_{31} & \lambda_{32} & -\sum_{j\neq 3}\lambda_{3j} & \lambda_{34} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & -\sum_{j\neq 4}\lambda_{4j} \end{pmatrix}$$

Markov models for binary trait evolution

If the traits evolve independently of each other, state changes between the four combinations of states are modeled with two independent Markov chains, each with a 2 × 2 Markov generator (4 parameters):

$$Q^{(1)} = \begin{pmatrix} -\lambda_{12}^{(1)} & \lambda_{21}^{(1)} \\ \lambda_{12}^{(1)} & -\lambda_{21}^{(1)} \end{pmatrix} \qquad Q^{(2)} = \begin{pmatrix} -\lambda_{12}^{(2)} & \lambda_{21}^{(2)} \\ \lambda_{12}^{(2)} & -\lambda_{21}^{(2)} \end{pmatrix}$$

If the more complicated model fits significantly better after accounting for the larger number of parameters, that suggests trait evolution is correlated.

Table of Contents

1 Background for ancestral character estimation

2 Markov chains

3 Likelihood for Markov models on trees

4 Fitting models in R

Working in R

We will use the R package ape to simulate and fit these models to example datasets (and real ones) Make sure you have the following R packages installed:

- ape
- phytools
- any other packages to make errors go away

Working in R

In the practical, you will

- Fit Markov models to trees with tip states
- Simulate fitted Markov models
- Explore some example datasets