

# Markov chains over a discrete state space acting in discrete time steps

## Usage

This document provides background and derivations for the demo at <http://phylo.bio.ku.edu/mephytis/disc-state-disc-time-Markov/index.html> <sup>1</sup>.

The widget lets you control the probability of switching states ( $s$  in the notes below) and the number of observations to add every time that you click on the blue button.

You can then generate a series of observations and note how the likelihood estimate  $\hat{s}$  tends to approach the true value of  $s$  you used in your simulation as you add more data.

## Background

The set of possible values that a stochastic process can visit is the “state space” of the model. So a model of DNA evolution would have 4 states:  $\mathcal{S} = \{A, C, G, T\}$ .

A first order Markov process is one in which the probability distribution of the next state depends only on the current state – not the previous history of the process. Markov processes can be formulated as working on variables that have continuous or discrete state spaces. They can also be formulated to work with a discrete or continuous index for moving from one state to the next step. Often this indexing variable is thought of as time. Here, we’ll explore a discrete-state, discrete-time Markov chain. Imagine that we have 4 states (I’ll label them  $A, C, G, T$ ), and the probability of switching states at each step is constant and equal no matter what the “source” state or the “destination” state is. When the states are DNA bases this model is called the **Jukes Cantor 1969 model**, and it is usually described in continuous time. Here we’ll imagine that we had some design (such as long-term experimental evolution study) that we can observe at even time intervals. So we’ll see time in steps.

We might want not know the probability of changing to a different state in one time step, so we can use a parameter. Let’s call this parameter  $s$ . If  $x_i$  denotes the state of the process at time step  $i$ :

$$\begin{aligned}\mathbb{P}(x_i \neq x_{i+1}) &= s \\ \mathbb{P}(x_{i+1} = C \mid x_i = A) &= s/3 \\ \mathbb{P}(x_{i+1} = G \mid x_i = A) &= s/3 \\ \mathbb{P}(x_{i+1} = T \mid x_i = A) &= s/3 \\ \mathbb{P}(x_{i+1} = A \mid x_i = A) &= 1 - s\end{aligned}$$

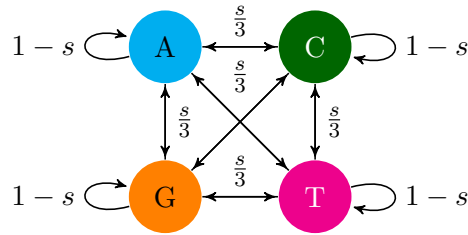
and we could write similar formulas for the other starting states. We can view the probability transition more compactly using the transition probability matrix or graph:

There is no tendency to stay in one state longer than any other, so if you were to sample this process for a huge number of time steps (and  $s > 0$ ) then we’d expect to spend approximately equal time in each of the 4 states. In the language of Markov chains, we’d say that the equilibrium frequency of each state is  $1/4$  (or  $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ , in notation).

---

<sup>1</sup>source code at <https://github.com/mtholder/mephytis>

$$\mathbf{P} = \begin{pmatrix} 1-s & s/3 & s/3 & s/3 \\ s/3 & 1-s & s/3 & s/3 \\ s/3 & s/3 & 1-s & s/3 \\ s/3 & s/3 & s/3 & 1-s \end{pmatrix}$$



The probability rules above describe the transitions between adjacent steps of the chain, but we also need to know what state the process starts in. It is common in molecular evolution to assume that the process has been running for a long time. Thus it is reasonable to say that each of the 4 states are equally likely for the first state in the chain.

## Inference

Imagine that we have a sequence of observations at 10 evenly spaced time points for the state at a particular site:

$$\begin{array}{l} i = 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \\ x_i = G \quad G \quad C \quad C \quad C \quad C \quad C \quad C \quad C \quad A \end{array} \quad \text{or (more compactly) } X = GGCCCCCCA.$$

**Note:** we typically use the Jukes Cantor model when comparing sequence at the same site from different species – here we are imagining that we saw the sequence as it was evolving in a series of steps down an unbranched ancestor-descendant lineage. This is not the typical case!

How could we estimate what value of  $s$  generated the sequence of observations?

We have a probabilistic model, so using a likelihood based procedure should be the most powerful approach (see [Wikipedia on the Likelihood Principle](#)).

So if our sample size is  $n = 10$ , we can write the likelihood as

$$\mathbb{P}(X | s) = \mathbb{P}(x_0) \prod_{i=1}^{n-1} \mathbb{P}(x_i | s, x_{i+1})$$

Substituting  $1/4$  for the start state, and  $1-s$  for every time that state does not change and  $s/3$  for every time the base changes between adjacent states:

$$\begin{aligned} L(s) = \mathbb{P}(X = GGCCCCCCA | s) &= \left(\frac{1}{4}\right) (1-s) \left(\frac{s}{3}\right) (1-s)(1-s)(1-s)(1-s)(1-s)(1-s) \left(\frac{s}{3}\right) \\ &= \left(\frac{1}{4}\right) (1-s)^7 \left(\frac{s}{3}\right)^2 \end{aligned}$$

We could graph the likelihood as a function of  $s$  and find the point that maximizes the likelihood. Or we could use calculus to find the point where the likelihood profile is at its maximum by finding the point where the derivative of the likelihood is 0. It is easier to find the maximum on the log-likelihood scale:

$$\ln L(s) = \ln\left(\frac{1}{4}\right) + 7\ln(1-s) + 2\ln\left(\frac{s}{3}\right)$$

$$\begin{aligned}
&= \ln\left(\frac{1}{4}\right) + 7\ln(1-s) + 2\ln(s) - 2\ln(3) \\
\frac{d \ln L(s)}{ds} &= \frac{-7}{1-s} + \frac{2}{s}
\end{aligned}$$

We usually denote the maximum likelihood estimator of a parameter with a cute little hat, so  $\hat{s}$  is our MLE. Because the likelihood is continuous the MLE will be the value of  $s$  when the curve is flat (the derivative is 0) or at one of the endpoints of the feasible range for  $s$  ( $s = 0$  or  $s = 1$  in our case because  $s$  is a probability). So, we find:

$$\begin{aligned}
\frac{-7}{1-\hat{s}} + \frac{2}{\hat{s}} &= 0 \\
\frac{2}{\hat{s}} &= \frac{7}{1-\hat{s}} \quad 2 - 2\hat{s} &= \\
7\hat{s} &= 2 - 2\hat{s} \\
\hat{s} &= \frac{2}{9}
\end{aligned}$$

Intuitively, the likelihood is maximized when  $s$  is equal to the number of changes of state divided by the number of observations that could possibly have shown a change of state. If we had put in a generic  $n_d$  for the number of transitions where the states differed and  $n_s$  for the number of transitions where the state stayed the same, we would have found that  $\hat{s} = \frac{n_d}{n_s+n_d}$ .