40 2025-12-03 | Week 15 | Lecture 40

40.1 The Jukes Cantor Model

This lecture presents a simplified (but real!) application of homogeneous first order linear differential equations to mathematical biology.

The graphics for this lecture were taken from Bob Thompson's tutorial page at ***https://treethinkers.org/jukes-cantor-model-of-dna-substitution/***

A genome can be thought of as a squence of letters, called nucleotides:

`CACAGGAAGCATTTATATAGAGGTAATAATTAAATTTAACTTTTTATTATTCTCTATGCCCAAGAGAATGTG...`

For simplicity, let us assume an asexually reproducing species. In that case, in each generation, the whole genome is passed from parent to child. When that happens, each nucleotide has a small chance $\mu > 0$ of mutating to a different letter.

We regard μ as the mutation rate per generation. For example, if the genome consists of 1000 base pairs, and $\mu = .02$, then on average you expect to see about 2 mutations per generation. (Realistic values are quite different: for humans, $\mu \approx 2 \times 10^{-8}$.)

In this lecture I will introduce the ***Jukes-Cantor*** model of DNA evolution. This makes several assumptions

- There are four nucleotides `A,T,C,G`
- When a mutation occurs, the letter is equally like to change to each of the other three nucleotides (e.g., `A` mutates, then it is equally likely to be `C`, `T` or `G` after the mutation).

The second assumption is not realistic, but more realistic variants can be described.

We will measure time t in "generations". To be precise, we should have t measured only as integer values $0, 1, 2, \ldots$, but for mathematical simplicity it's easier if we assume t is a continuous variable.

Let

$$Y(t) = \begin{bmatrix} y_{\mathtt{A}}(t) \\ y_{\mathtt{C}}(t) \\ y_{\mathtt{T}}(t) \\ y_{\mathtt{G}}(t) \end{bmatrix}$$

where

$$y_{\mathtt{A}}(t) = \text{ proportion of sites with letter } \mathtt{A} \text{ at time } t$$

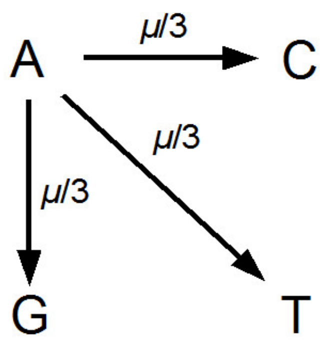
and $y_{\mathtt{C}}, y_{\mathtt{G}}$ and $y_{\mathtt{T}}$ defined similarly.

Question: Suppose $\mu = 10^{-8}$. What are the proportions when after $t = 1,000,000$ generations? Assume that at time $t = 0$, the proportions are

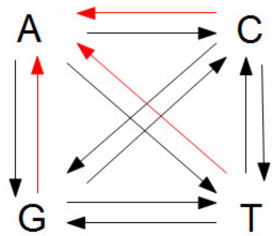
$$Y(0) = \begin{bmatrix} 1/6 \\ 2/6 \\ 1/6 \\ 2/6 \end{bmatrix}$$

This is an initial value problem.

To answer it, we'll need to think about how the proportions change over time. Over time, we lose `A`'s due to mutations. The rate of loss is μ per generation, and it gets divided up equally as `A`'s become `C`, `G`, and `T` at equal rates. Here is a diagram representing this flow:



But we also have flow *into* the A's, as the other letters sometimes mutate into A's:



Each red arrow represents a flow *into* A of rate $\mu/3$.

Combining the flow into and out of A, we have:

$$y'_{\mathtt{A}}(t) = \underbrace{-\mu y_{\mathtt{A}}(t)}_{\text{total flow out of }\mathtt{A}} + \underbrace{\frac{\mu}{3}y_{\mathtt{C}}(t) + \frac{\mu}{3}y_{\mathtt{G}}(t) + \frac{\mu}{3}y_{\mathtt{T}}(t)}_{\text{total flow into }\mathtt{A}}$$

Similarly,

$$\begin{aligned} y'_{\mathtt{C}} &= \frac{\mu}{3}y_{\mathtt{A}} - \mu y_{\mathtt{C}} + \frac{\mu}{3}y_{\mathtt{G}} + \frac{\mu}{3}y_{\mathtt{T}} \\ y'_{\mathtt{G}} &= \frac{\mu}{3}y_{\mathtt{A}} + \frac{\mu}{3}y_{\mathtt{C}} - \mu y_{\mathtt{G}} + \frac{\mu}{3}y_{\mathtt{T}} \\ y'_{\mathtt{T}} &= \frac{\mu}{3}y_{\mathtt{A}} + \frac{\mu}{3}y_{\mathtt{C}} + \frac{\mu}{3}y_{\mathtt{G}} - \mu y_{\mathtt{T}} \end{aligned}$$

In matrix form,

$$\begin{bmatrix} y'_{\mathtt{A}} \\ y'_{\mathtt{C}} \\ y'_{\mathtt{T}} \\ y'_{\mathtt{G}} \end{bmatrix} = \begin{bmatrix} -\mu & \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & -\mu & \frac{\mu}{3} & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & -\mu & \frac{\mu}{3} \\ \frac{\mu}{3} & \frac{\mu}{3} & \frac{\mu}{3} & -\mu \end{bmatrix} \begin{bmatrix} y_{\mathtt{A}} \\ y_{\mathtt{C}} \\ y_{\mathtt{T}} \\ y_{\mathtt{G}} \end{bmatrix}$$

Therefore, letting

$$Q = \mu \begin{bmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{bmatrix},$$

the intial value problem we need to solve is

$$\begin{cases} Y' = QY \\ Y(0) = \begin{bmatrix} 1/5 \\ 2/5 \\ 1/5 \\ 2/5 \end{bmatrix} \end{cases}$$

To solve this, we first observe that Q has two eigenvalues:

- $\lambda = 0$, with eigenvector $\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$
- $\lambda = -\frac{4}{3}\mu$, with three linearly indendent eigenvectors:

$$\begin{bmatrix}1\\-1\\1\\-1\end{bmatrix}, \begin{bmatrix}1\\1\\-1\\-1\end{bmatrix}, \begin{bmatrix}1\\-1\\-1\\1\end{bmatrix}$$

Therefore there are 4 pure expoenential solutions

$$Y_1(t) = \begin{bmatrix}1\\1\\1\\1\end{bmatrix} e^{0t}, \quad Y_2(t) = \begin{bmatrix}1\\-1\\1\\-1\end{bmatrix} e^{-\frac{4}{3}\mu t}, \quad Y_3(t) = \begin{bmatrix}1\\1\\-1\\-1\end{bmatrix} e^{-\frac{4}{3}\mu t}, \quad Y_4(t) = \begin{bmatrix}1\\-1\\-1\\1\end{bmatrix} e^{-\frac{4}{3}\mu t}$$

The general solution is

$$Y(t) = c_1 \begin{bmatrix}1\\1\\1\\1\end{bmatrix} + c_2 \begin{bmatrix}1\\-1\\1\\-1\end{bmatrix} e^{-\frac{4}{3}\mu t} + c_3 \begin{bmatrix}1\\1\\-1\\-1\end{bmatrix} e^{-\frac{4}{3}\mu t} + c_4 \begin{bmatrix}1\\-1\\-1\\1\end{bmatrix} e^{-\frac{4}{3}\mu t} \tag{38}$$

Using the initial condition

$$Y(0) = \begin{bmatrix}1/6\\2/6\\1/6\\2/6\end{bmatrix}$$

gives the system

$$\begin{cases} c_1 + c_2 + c_3 + c_4 = 1/6 \\ c_1 - c_2 + c_3 - c_4 = 2/6 \\ c_1 + c_2 - c_3 - c_4 = 1/6 \\ c_1 - c_2 - c_3 + c_4 = 2/6 \end{cases}$$

And this has solution $c_1 = \frac{1}{4}$, $c_2 = -\frac{1}{12}$, $c_3 = c_4 = 0$. Therefore the solution the initial value problem is

$$\begin{aligned} Y(t) &= \frac{1}{4}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} - \frac{1}{12}\begin{bmatrix}1\\-1\\1\\-1\end{bmatrix} e^{-\frac{4}{3}\mu t} \\ &= \begin{bmatrix}\frac{1}{4} - \frac{1}{12}e^{-\frac{4}{3}\mu t}\\ \frac{1}{4} + \frac{1}{12}e^{-\frac{4}{3}\mu t}\\ \frac{1}{4} - \frac{1}{12}e^{-\frac{4}{3}\mu t}\\ \frac{1}{4} + \frac{1}{12}e^{-\frac{4}{3}\mu t}\end{bmatrix} \end{aligned}$$

These are the proportions of `A`,`T`, `C`, and `G` at time t. As $t \to \infty$, the proportions all approach 1/4.

To answer our original problem, if $\mu = 10^{-8}$ and $t = 10^6$ generations, then

$$y_{\texttt{A}}(10^6) = \frac{1}{4} - \frac{1}{12}e^{-\frac{4}{3}\cdot 10^{-2}} \approx .168$$

and

$$y_{\texttt{C}}(10^6) = \frac{1}{4} + \frac{1}{12}e^{-\frac{4}{3}\cdot 10^{-2}} \approx .332$$

These aren't very far off from the initial values $1/6 \approx .1667$ and $1/3 \approx .3333$.