Stochastic dynamical modeling in biology

- Homework 8 solutions -

Andrew Powers

---

1 Exercise 1

In this homework we were tasked with solving the master equation of a homogeneous CTMC for the nucleotide sequence of a neutrally evolving gene. We are told to consider the HKY nucleotide substitution model with independent sites, this can be found in the book on section 6.1.3. The parameters of the model will be $\mu = 2 \times 10^{-3} yr^{-1}$, $\sigma_A = 0.4, \sigma_C = 0.15, \sigma_G = 0.25, \sigma_T = 0.2$, and $\kappa = 5$. We are going to track a singlue nucleotide as it changes over time.

1. We are supposing that we start with and Adenine (A). We have a state space $S = \{A, C, G, T\}$. We want to determine the probability distribution of the the four possible states over time, from 0 until 5000 years. Since S is non-numeric, we can enumerate our states. This gives us $S = \{A : 1, C : 2, G : 3, T : 4\}$. Doing this allows us to take advantage of computational resources like Matlab. We are starting with A as our initial state, i.e. $X_0 = A$. However, we can write this into a $p(t)$ vector, because $A \to 1$:

$$p(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{1}$$

This is our initial probability at time 0. Also, we know from the definition of the problem above that our CTMC is homogeneous. This means that our $\mathbb{Q}$ matrix is independent of time. Which allows us to use the master equation:

$$p(t) = e^{t\mathbb{Q}} \cdot p(0) \tag{2}$$

However, we need to calculate the transition rate matrix, $\mathbb{Q}$. We can get the equation for that by looking at 6.1.3 in the book. Since we are basing our model off of the HKY nucleotide substitution model we can get $\mathbb{Q}$ by ...

$$\mathbb{Q} := \beta\mu\widetilde{\mathbb{Q}} \tag{3}$$

We are given μ in the problem. However, we can calculate β with this formula ...

$$\beta := \frac{1}{2\kappa(\sigma_A\sigma_G + \sigma_C\sigma_T) + 2(\sigma_A + \sigma_G)(\sigma_C + \sigma_T)} \tag{4}$$

We can also calculate $\widetilde{\mathbb{Q}}$ by filling in this matrix ...

$$\widetilde{\mathbb{Q}} := \begin{pmatrix} * & \sigma_A & \kappa\sigma_A & \sigma_A \\ \sigma_C & * & \sigma_C & \kappa\sigma_C \\ \kappa\sigma_G & \sigma_G & * & \sigma_G \\ \sigma_T & \kappa\sigma_T & \sigma_T & * \end{pmatrix} \tag{5}$$

Now that we have everything defined we can use (Eq. 2) to calculate the probabilities of our states as we move forward in time.

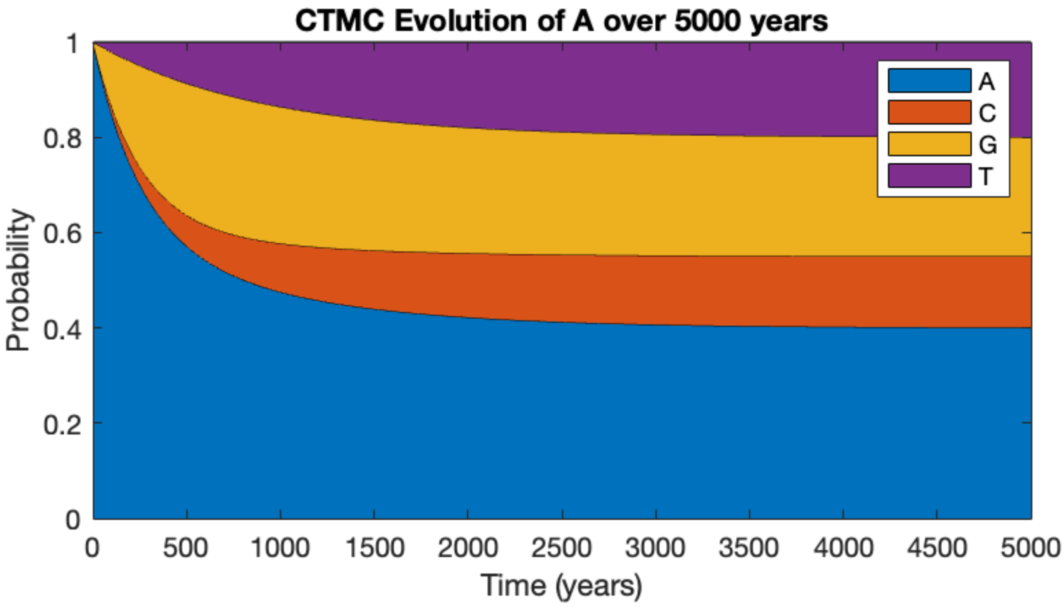

Figure 1: Nucleotide probabilites starting at Adenine as we move 5000 years into the future

We can see from the figure above (Fig. 1)that as we move forward in time Adenine (A) becomes the most probable nucloetide to be present after 5000 years. Then Cytosine, Guanine, and finally Tyrosine. It seems that the probability distribution converges in the long term. The final probabilites are ...

$$p(5000yr) = \begin{pmatrix} 0.4007 \\ 0.1495 \\ 0.2505 \\ 0.1993 \end{pmatrix} \tag{6}$$

2. We were asked to repeat the same process but with a new Cytosine (C) as the starting point this time. This will allow us to determine if we reach the same probability distribution at time T (5000 years) in the future.

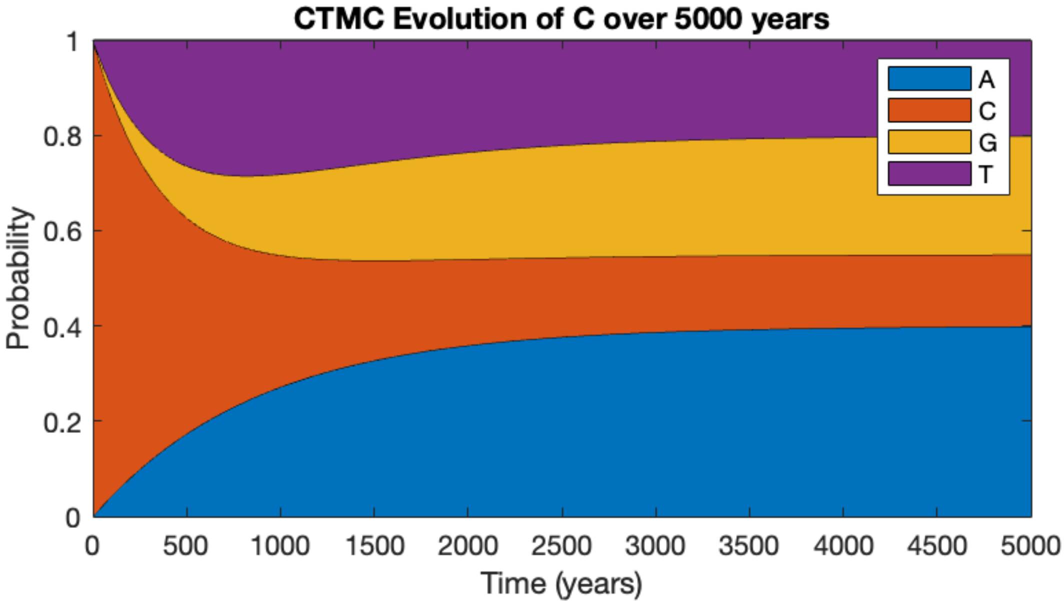

Figure 2: Nucleotide probabilities starting at Cytosine as we move 5000 years into the future

Looking at the graph above (Fig. 2) and comparing it with Fig. 1 we can see that again the same pattern is repeated. Adenine, Cytosine, Guanine, and then Tyrosine. However, We can see that at the beginning Cytosine was dominant for the first 100 years or so. If we look at the probability distribution at 5000 years we see and almost identical probability distribution that we saw above.

$$p(5000yr) = \begin{pmatrix} 0.3987 \\ 0.1509 \\ 0.2492 \\ 0.2012 \end{pmatrix} \tag{7}$$

In conclusion it seems that regardless of X_0 being A or C we converge to the same probability distribution in the long term.