1 Developing the Model

Consider a single base pair site on a strand of DNA. The allowed states of this site make up a finite state space $E = \{A, G, C, T\}$. Imagine a finite-state Markov process as a model for the mutation of this site: the site starts in a given state X_0, and in each time step Δt, there is a well-defined probability of transition to state $X_i \in E$.
Given this physical model, we can construct the following transition matrix:

$$\Omega = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_G & \mu_{TG} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}. \tag{1}$$

In (1), we enforce that $\mu_i i$ is set so the columns of Ω sum to one. For our basic model, we will assume that all transitions are equally likely. This allows us to reduce Ω into the simple form

$$\Omega = \begin{pmatrix} -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} \end{pmatrix}. \tag{2}$$

Now that we have the transition matrix (2), we can begin to construct our model. Consider the probability vector $\mathbf{P}(t) = [p_A(t)\ p_G(t)\ p_C(t)\ p_T(t)]^T$. We know from studying the master equation,

$$p_i(t + \Delta t) = p_i(t) - p_i(t)\mu_i i \Delta t + \sum_{j \neq i} p_j(t)\mu_{ji}\Delta t, \tag{3}$$

because probability must be conserved. We can expand this notion to get an expression for $\mathbf{P}(t + \Delta t)$:

$$\mathbf{P}(t + \Delta t) = \mathbf{P}(t) + \Omega\mathbf{P}(t)\Delta t. \tag{4}$$

We can see that equation (4) just represents the system of all equations (3). Dividing both sides of (4) by Δt and taking the limit as Δt goes to zero gives us the differential equation

$$\frac{d\mathbf{P}(t)}{dt} = \Omega\mathbf{P}(t). \tag{5}$$

Equation (5) can be simply solved by direct integration, giving the solution

$$\mathbf{P}(t) = \mathbf{P}(0)e^{\Omega t}, \tag{6}$$

where $e^{\Omega}t$ is given by the definition of a matrix exponential represented by a Taylor series,

$$e^{\Omega t} = \sum_{k=0}^{\infty} \Omega^k \frac{t^k}{k!}. \tag{7}$$

2 Application of the Model: Deriving Long-Run Similarity

Consider $P(t) = e^{\Omega}t$. Using Ω from (2) and (7) gives the following form for $P(t)$:

$$P(t) = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} \\ \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} + \frac{3}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} \\ \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} + \frac{3}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} \\ \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} - \frac{1}{4}e^{-t} & \frac{1}{4} + \frac{3}{4}e^{-t} \end{pmatrix} \tag{8}$$

Plugging this result into equation (6) and taking the limit as t goes to infinity gives the long-run expected probability distribution,

$$\lim_{t\to\infty} \mathbf{P}(t) = \frac{1}{4}\mathbf{P}(0). \tag{9}$$

We can interpret (9) as follows: our DNA site starts at an initial probability distribution $\mathbf{P}(0)$. As time progresses, the probability that this site follows the initial distribution decreases by a factor of four. This means that if we consider not one site, but the entire strand of DNA, the probability that after a long time the mutated strand is identical to the starting strand is one fourth.