BIOREPS Problem Set #11
The Evolution of DNA Strands

1 Background

In the mid 2000s, evolutionary biologists studying DNA mutation rates in birds and primates discovered something surprising. There were a large number of mutations in DNA of different species over the last several thousand years, which indicats a robust evolution. However going back further, comparing genomes across millions of years indicates a much lower mutation rate. How are these consistent with each other? It turns out the overall rate of mutation is dependent on the timescale. At low timescales, there are rapid fluctuations, but since these fluctuations are essentially random, some of them cancel each other out, leading to an overall slower evolution rate. A good analogue of this system is the stock market. While there are rapid fluctuations by the hour, over course of days or weeks, the overall change is much more gradual [1].

In this problem set, we explore this mutation rate within the context of a single DNA strand. Although different mutation rates have been measured for each base pair transition, we will work in the simplified example, where all of them are equal. We will characterize the mutation rate by the self similarity, defined as the fraction of bases are are the same as in the original strand. In our model, we will find that the self-similarity converges on 25% (the same as a completely random DNA strand) at an exponential rate.

In this context, in the long time approximation the DNA strand can be represented as a binomial random variable, (analogous to a series of coin flips, but with an arbitrary probability of getting heads). The binomial random variable gives the probability distribution of any number of heads, This is a standard distribution in probability, and thus its properties are well understood.

At this point, we are able to characterize the average behaivor in the long-time approximation. Applying the central limit theorem to binomial random variables, if we average over a large number, the distribution will converge to a normal (Gaussian) distribution, centered around the mean of a single binomial random variable, with a standard deviation given by the square root of the variance of that distribution [2].

References

[1] Evolution Runs Faster on Shorter Timescales. Carrie Arnold. *Quanta* 2017. https://www.quantamagazine.org/evolution-runs-faster-on-short-timescales-20170314/

[2] A First Course in Probability. Sheldon Ross.

2 Questions

2.1 Modeling Similarity of a Single DNA Site

a) Consider the model for DNA mutation considered above: at each time-step, a specific site on the DNA strand is chosen with uniform problem (i.e., for *n* sites, the probability of choosing a specific site is $\frac{1}{n}$). Once this site is chosen, if it is currently in state *i*, it transitions to state *j* with probability $\mu_{ij},\ i \neq j$. Here, both *i* and *j* are elements of the discrete state space $E = \{A, G, C, T\}$ representing the four possible base pairs for DNA.

If we assume that these transition probabilities are time-independent and site-independent, write the general transition matrix Ω describing the process of base pair mutation. Then, simplify this matrix to the specific case where all probabilities μ_{ij} are uniform. You should find that your Ω-matrix is a very simple and symmetric.

b) With the transition matrix Ω, we can now begin to solve the master equation, but using a different approach than we learned in class. Our first step is to derive an ordinary differential equation for the time-evolution of the probability of a specific site being a certain base-pair value. From the master-equation formalism, we can write the probability of a site being a specific base-pair value as

$$p_i(t + \Delta t) = p_i(t) - p_i(t)\mu_i i\Delta t + \sum_{j \neq i} p_j(t)\mu_{ji}\Delta t, \tag{1}$$

where $i \in E$. If instead we consider the probability vector $\mathbf{P}(t) = [p_A(t)\ p_G(t)\ p_C(t)\ p_T(t)]^T$, show that equation (1) simplifies to a simple ordinary differential equation in the limit as Δt goes to zero,

$$\frac{d\mathbf{P}(t)}{dt} = \Omega\mathbf{P}(t). \tag{2}$$

c) The equation derived in **b)** is very familiar to us; if quantities were not vectors, the solution would be a simple exponential. Luckily for us, we can solve this vectorized equation in the same way as the non-vectorized case, and obtain the solution

$$\mathbf{P}(t) = \mathbf{P}(0)e^{\Omega t} = \mathbf{P}(0)P(t). \tag{3}$$

However, we now have to deal with what an exponential power means. We can define matrix exponentiation using the Taylor series

$$e^{\Omega t} = \sum_{k=0}^{\infty} \Omega^k \frac{t^k}{k!} = I + \Omega t + \Omega^2 \frac{t^2}{2} + \dots. \tag{4}$$

Using the definition in equation (4), solve for the matrix $P(t)$ in equation (3). You should find that the solution is another symmetric matrix (*Hint*: write out the first few terms of the Taylor series and sum them together. You should find that the entries in this matrix sum resemble a Taylor series for e^t).

d) With your solution for **c)**, take the limit as *t* goes to infinity to find the long-run behavior of the probability $\mathbf{P}(t)$. Your answer should be simple, yet make very intuitive sence: the probability of a site staying the same is $\frac{1}{4}$. Why does this answer make conceptual sence based on the model?

2.2 Long-Run DNA Similarity Estimates: the Well-Mixed Binomial Approximation

e) While we have a derivation for the similarity of a single DNA site, we really want to model the similarity of the entire DNA strand after mutation. If limit our investigation to long-run behaviors, we argue that we can model the number of mutated DNA sites that are the same base-pair value as the original DNA sequence as a binomial random variable.

As a quick introduction to probability theory and random variables, we denote a random variable *X* that is binomially distributed as $X \sim Binom(n, p)$, which implies that

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i},\ i = 1, \ldots, n,\ \binom{n}{i} = \frac{n!}{(n-i)!i!}. \tag{5}$$

We can think of a binomial random variable with parameters *n* and *p* as the number of sucessful outcomes of *n* trials, where each trial is independent from the others and the probability of success of a single trial is *p*. The easiest example of a binomial random variable is asking the probability of observing i heads when flipping *n* coins.

With this knowledge, why does our argument that the number of similar DNA sites after many mutations have occured can be modeled as a binomial random variable?

f) From our well-mixed binomial model, we can derive analytic estimates of the mean and standard deviation of the similarity in the limit of large number of simulations. First, we need to derive the expectation and variance of a binomial random variable. For a general random variable, these values are given by

$$E[X] = \sum_i x_i p(x_i) = \mu.$$
$$Var(X) = E\left[(X-\mu)^2\right] \sum_i (x_i - \mu)^2 p(x_i) = E\left[X^2\right] - (E[X])^2.$$

Using these definitions, and the definition of the distribution of a binomial random variable from equation (5), show that if $X \sim Binom(n, p)$, then $E[X] = np$ and $Var(X) = np(1-p)$ (*Hint*: There are many ways to derive these equations; one method is to calculate the *k*-th moment of *X*, $E\left[X^k\right]$, and use this to obtain the mean and variance. The identity $i\binom{n}{i} = n\binom{n-1}{i-1}$ may also be helpful.).

g) Now, consider an experiment where we draw from the above binomial distribution many times, where each draw is independent of the others, and create a distribution of the values obtained from each draw. By the Central Limit Theorem from probability theory, in the limit of large sample size, this distribution converges to a normal distribution (i.e., a Gaussian) with the same mean

and variance as the binomial distribution we draw from. Using this approximation, calculate the mean and standard deviation of this normal distribution, using the fact that the standard deviation σ is just the square root of the variance. Then, rescale the distribution by dividing by n to measure similarity on the interval $[0, 1]$, and plug in values $n = 100$ and $p = 0.25$. You should find that the mean of this distribution matches the theoretical long-run similarity of a single site obtained in **d)**.

2.3 Simulating DNA Mutation and Similarity

Now we will simulate the DNA mutation strand as we have done in previous problem sets. It may be helpful to get back to the first problem set to refresh your mind in doing this simulation.

h) Write a simulation that models the self-similarity of a DNA strand as a function of time. The self similarity is simply defined as the fraction of base pairs that are the same as the original DNA Run this simulation for a large number of trials and plot the mean of the self-similarity at each time step. Describe the notable features. Do they make sense given the mathematical modeling we have done above?

Simulation Hints: Start with a random distribution of base pairs. At each time step, choose a random base pair, and generate a random number to determine if it mutates. At the end of each time step, use a loop to count the number of base pairs that are the same as the original, and divide by total base pairs to find the self-similarity. The self-similarity values, the original DNA strand, and the current DNA strand are the only things that should be stored between time steps. If you store the current DNA strand at each time step, and attempt to calculate the self-similarity afterwards, you will quickly run out of space.

Suggested Simulation Values: 100 base pairs, 2000 timesteps, 10000 trials. Transition rate between and base pair and any mutation: 10^{-3} per timestep.

i) You should find that your self-similarity appears to have an exponential distribution. To test this hypothesis, fit your curve to the standard exponential form:

$$P(t) = A + Be^{-kt} \tag{6}$$

Do the values for A, B, and k that you find make sense? Explain.

j) Plot the standard deviation as a function of time, and describe any notable features. With this plot, estimate the asymptotic (late time) standard deviation, and compare to the value you calculated in part **g)**, with appropriate values plugged in.

2.4 Mean First Passage Time to Varying Levels of Self-Similarity

Let us now observe this system through a different lens; consider the varying levels of self-similarity as the possible states for the DNA strand. If we have a DNA of fixed length N, then the discrete state space becomes $\{0, 1, 2, \ldots, N\text{-}1, N\}$ where state i would correspond to having i

base pairs identical to the DNA before any mutations. We can now consider the rates of jumping from i to i+1 and i to i-1 with respect to the mutation rate μ. In order to jump from i to i+1, we would need one of the $(N-i)$ incorrect base pairs to mutate to the correct base pair. Since there are 3 possible mutations that could occur, the rate of one incorrect base pair mutating correctly is $\frac{\mu}{3}$. Considering there are $N-i$ incorrect base pairs, the total rate from i to i+1 is $\frac{(N-i)\mu}{3}$. Now, in order to jump from i to i-1 we need one of the i correct base pairs to mutate to any of the other 3 incorrect base pairs. Thus this total rate occurs as $i\mu$. From this we can see that our transition matrix entries will be:

$$\Omega_{i+1,i} = \frac{(N-i)\mu}{3}$$
$$\Omega_{i-1,i} = i\mu$$
$$\Omega_{i,i} = -(\Omega_{i+1,i} + \Omega_{i-1,i}) = -\frac{(N+2i)\mu}{3}$$
$$\Omega_{else} = 0$$

And thus our transition matrix becomes:

$$\Omega = \begin{bmatrix} -\frac{N\mu}{3} & \mu & 0 & 0 & \dots & 0 & 0 \\ \frac{N\mu}{3} & -\frac{(N+2)\mu}{3} & 2\mu & 0 & \dots & 0 & 0 \\ 0 & \frac{(N-1)\mu}{3} & -\frac{(N+4)\mu}{3} & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & -\frac{(N+2(N-2))\mu}{3} & (N-1)\mu & 0 \\ 0 & 0 & \dots & 0 & \frac{2\mu}{3} & -\frac{(N+2(N-1))\mu}{3} & N\mu \\ 0 & 0 & \dots & 0 & 0 & \frac{\mu}{3} & -N\mu \end{bmatrix}$$

k) Consider a DNA in its original state (i = N). Using the above transition matrix, for N = 100 and μ = $10^{-3}\frac{mutations}{timestep}$, calculate the mean first passage time from i = 100 to i_f = 90, 80, 70, 60, 50, 40, 30, 20, 10, and 0 (minimum self-similarity). What trends in the MFPT do you notice for the DNA to get to increasingly dissimilar states?
Hint: $-1 = \sum_{i'=0}^{100} \Omega_{i',i}\tau_{i'\to i_f}$ for $i \neq i_c$ and $\tau_{i_f \to i_f} = 0$. To use this, first create Ω' by the following rules: $\Omega'_{j,i} = \Omega_{j,i}$ for $i \neq i_c$ and $\Omega'_{j,i_c} = \delta_{j,i_c}$. Next, create $x = [-1, ..., -1, 0, -1, ... -1]$ where the i_c entry of x is 0 and the length is N+1. Then let $\tau = [\tau_{0\to i_f}, \tau_{1\to i_f}, ..., \tau_{100\to i_f}]^T$, and finally solve $\Omega'^T\tau = x$ for $\tau_{100\to i_f}$. Then repeat this process for the various i_f.