1 Establishing the Argument

Consider the DNA mutation model discussed earlier. The model outlines the following physical steps:

1. A site on the DNA strand is chosen at random among the n possible sites with uniform probability ($P(X = i) = \frac{1}{n}$ for any $1 \leq i \leq n$).
2. For the chosen site i, the new base pair value for the site is chosen from the state-space $E = \{G, A, T, C\}$ with uniform probability.
3. The time step is incremented by one, and steps 1 and 2 are repeated. This process continues until a final time is reached.

With this process in mind, imagine letting this process run for a long period of time, after which we look at the new mutated DNA strand and compare it to the original DNA strand site by site to see how many sites have the same base pair states between the two strands. Because we waited a long time before this comparison, there have been so many mutations that we can consider the entire strand as a "well-mixed", or homogeneous, system. With this argument, we can think of the probability of a particular site being the same between the two strands as a weighted coin flip, with $\frac{1}{4}$ chance of being successful (the sites are the same) and a $\frac{3}{4}$ chance of failure (the sites are different). Expanding this to the entire strand, the probability of the two strands being identical can be thought of as a series of n such weighted coin flips. This interpretation lends itself to the application of the binomial distribution.

2 Properties of the Binomial Distribution

A binomial random variable, denoted $Binom(n, p)$ represents n independent trials, each with a probability of success of s. A random variable X that is binomially distributed ($X \sim Binom(n, p)$), can therefore take integer values between zero and n, where the probability for X to take on a particular value i is given by

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}. \tag{1}$$

From (1), we can derive two important properties for any random variable; the mean and the variance. The k-th moment of X is given by

$$\begin{aligned}
E\left[X^k\right] &= \sum_{i=0}^{n} x^k p(x). \\
&= \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}. \\
&= \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}. \\
&= np \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}, \text{ since } i\binom{n}{i} = n\binom{n-1}{i-1}. \\
&= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j}, \ n = i-1. \\
&= npE\left[(Y+1)^{k-1}\right],
\end{aligned}$$

where $X \sim Binom(n-1, p)$.
From this calculation, we can easily obtain an equation for the mean of X (defined as the first moment):

$$E[X] = npE[1] = np. \tag{2}$$

We can also calculate the variance of X:

$$Var(X) = E\left[X^2\right] - (E[X])^2 = np((n-1)p+1) - n^2p^2 = np(1-p). \tag{3}$$

3 The Central Limit Theorem and Standard Deviation of Repeated Simulations

With (2) and (3), we are in a position to calculate the distribution of a multi-sample simulation of our model. As previously stated, the distribution of the similarity between original and mutated DNA strands in our model obeys a binomial distribution after a many time steps. Imagine now repeating this simulation many times to obtain a large sample size, so that you can calculate the distribution of similarities. By the Central Limit Theorem from probability theory, such a distribution converges to a normal distribution (otherwise known as a Gaussian) for a large enough sample size as long as each sample is independent (one sample does not alter the probability of another) and each sample follows the same distribution. Moreover, the mean and variance of the normal distribution are the same as the mean and variance of the distribution of each sample.

Therefore, because each sample $X \sim Binom(n,p)$ with mean np and variance $np(1-p)$, the normal distribution obtained after many samples is $\mathcal{N}(np, np(1-p))$.

To obtain the average similarity measured on the interval $[0,1]$, we can take the mean of the normal distribution and divide it by the number of sites to rescale it. This gives a mean similarity $\mu = p$, which matches the long-run similarity obtained from the Markov model from earlier in the text.

To obtain the standard deviation of the similarity, once again measured on the interval $[0,1]$, we can simply use the common method of taking the square root of the variance, then dividing by the number sites to rescale it. This gives a theoretical standard deviation of $\sigma = \frac{\sqrt{np(1-p)}}{n}$.