8.592J–HST.452J: Statistical Physics in Biology

Assignment # **1** Due: **9/19/25**

Evolving probabilities

1. *Open Reading Frames:* Assume that the nucleotides A, G, T, C occur with equal probability (and independently) along a segment of DNA.

(a) Of the 4^3 nucleotide triplets, the genetic code assigns a stop sign to TAG, TGA, and TAA. Calculate the probability p_s that a randomly chosen triplet of bases corresponds to a stop signal.

(b) What is the probability for an open reading frame (ORF) of length N, i.e. a sequence of N non-stop triplets followed by a stop codon?

(c) The genome of E-coli has roughly 5×10^6 bases per strand, and is in the form of a closed loop. If the bases were random, how many ORFs of length 600 (a typical protein size) would be expected on the basis of chance. (Note that there are six possible reading frames depending on the starting point and direction.)

*****

(Optional) 2. *ORFs in E. coli:* To compute the actual distribution of ORFs in *E. coli* you will need to download the complete sequence of its genome from
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/
000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz .
This file is also posted on the *Assignments* web-page. (More information can be found at https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2/ .)

(a) Write a program that goes through all consecutive (non-overlapping) triplets looking for stop codons. (Make sure you use the genetic code for DNA in the 5'-3' direction.) Record the distance L between consecutive stop codons. Repeat this computation for the 3 different reading frames (0, +1, +2) in this direction. (You may skip calculations for the reverse strand, that is complementary to the given one and proceeding in the opposite direction.)

(b) Plot the distribution for the ORF lengths L calculated above, and compare it to that for random sequences.

(c) Estimate a cut-off value L_{cut}, above which the ORFs are statistically significant, i.e. the number of observed ORFs with $L > L_{cut}$ is much greater than expected by chance.

*****

3. *Point mutations in DNA:* Since the four nucleotides in DNA have different chemical compositions and energetics, they could mutate at different rates. We shall explore whether, without natural selection at work, such preferential mutation may lead to different compositions of nucleotides.

(a) Consider a simple model in which all *transitions* (i.e. mutations between purines A and G, or between pyramidines T and C) occur with probability q, while *transversions* (i.e. any mutation from a purine to a pyrimidine or vice versa) occur with probability p, in each generation. Write down the 4×4 (Markov) transition matrix, Π_1, that relates the frequencies

of nucleotides ($p_{\rm A}$, $p_{\rm G}$, $p_{\rm T}$, $p_{\rm C}$) from one generation to the next. Note the constraint on q and p that ensures positivity of the transition matrix.

(b) Find the eigenvalues of the transition matrix Π_1. (**Hint:** You should be able to simply guess the eigenvectors by considering the symmetries of the matrix.)

(c) Find the matrix $\Pi_t = \Pi_1^t$, describing the evolution of probabilities after t generations.

(d) Show that in steady state (after many duplications), all nucleotides occur with the same frequency. Estimate the number of generations (as a function of p and q) needed to reach such a steady state.

(e) You should be able to convince yourself that for any model in which mutation rates between pairs of bases are the same in the forward and backward directions, all nucleotides are equally likely in the steady state. However, in the human genome the nucleotides C and G occur less often than A and T. This is partly due to methylation of successive CG pairs which makes them more susceptible to mutations. To mimic this asymmetry, consider an unrealistic model in which transversions from A to C and T to G occur with probability p_+, while the reverse transversions (from C to A or G to T) occur at a higher probability of p_-. (The other transversions occur at rate p, and transitions at rate q as before.) Write the modified transfer matrix corresponding to this model, and obtain the resulting frequencies of nucleotides in steady state.

*****

4. (Optional) *Activation/deactivation reaction:* Many molecules in biology can be made active or inactive through the addition of a phosphate group. The enzyme that adds the phosphate group is usually termed a kinase, while a phosphatase removes this group. Let us consider a case where a finite number N of such molecules within a cell can be exchanged between the two forms at rates a and b, i.e.

$$\mathrm{A} \rightleftharpoons_b^a \mathrm{B} \quad ,$$

where we have folded the probabilities to encounter the enzymes in the reaction rates.

(a) Write down the Master equation that governs the evolution of the probabilities $p(N_A = n, N_B = N - n, t)$.

(b) Assuming that initially all molecules are in state A, i.e. $p(n, t = 0) = \delta_{n,N}$, find $p(n, t)$ at all times. You may find it easier to guess the solution, but should then check that it satisfies the equations obtained before.

*****

5. *Mutation-selection balance.* Consider a population of a fixed number N cells. In each generation, a cell randomly acquires j *additional* mutations, where j is Possion distributed, $p(j) = e^{-\mu}\mu^j/j!$, with average μ. These mutations are mildly deleterious, such that a cell with j mutations has a relative fitness of $f(j) = (1 - s)^j$, with (multiplicative) selection coefficient $0 < s \ll 1$. Consider a steady state of mutations and selection in the system.

(a) Write the recursion relation for the fraction of cells with k mutations, x'_k, after one generation. Consider that cells *first* survive with a probability proportional to their relative

fitness, and *then* acquire new mutations. Remember to normalize by the mean fitness of the population, $\overline{f} = \sum_{i=0} x_i(1-s)^i$.

(b) Find the mean fitness of the population by considering the steady state for cells with zero mutations ($x'_0 = x_0$).

(c) Solve for the steady state distribution, such that $x'_k = x_k$. (Hint: try a Poisson distribution.)

*****

6. (Optional) *Global selection and mutation:* Consider a very large population of individuals characterized by a fitness parameter f, which is assumed to be Gaussian distributed with a mean m and variance σ^2. The population undergoes cyclic evolution, such that at each cycle: *(i)* one half of the population with lower fitness f is removed without creating progeny; *(ii)* the remaining half (with f values in the upper half) reproduces before dying; *(iii)* because of mutations that are *on average neutral* the f values of the new generation is again Gaussian distributed, with mean value and variance reflecting the parents (i.e. coming from the upper half of the original Gaussian distribution).

(a) Relate the mean m_n and variance σ_n of fitness values of the n-th generation to those of the previous ones (m_{n-1} and σ_{n-1}).

(b) What happens to the distribution of fitness after many generations?

(c) Most mutations are deleterious, while at the same time increasing the diversity of the population. To study these effects, assume that at each generation the distribution obtained in (a) above is convoluted with a Gaussian of mean $-\mu$ (thus reducing the mean fitness) and variance s^2 (acting to increase the variance in fitness). Find the recursion relations for m_n and σ_n in this case.

(d) What happens to the fitness distribution at long times in this case?

*****