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.

- From the genetic table, it is easy to check that $p_s = 3/64$.

(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?

- If the probability of having a stop is p_s, then the probability of having a sequence of DNA long N is

$$P_N = (1 - p_s)^N p_s.$$

(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.)

- Setting $N = 600$, we get

$$P_{600} = 1.4 \times 10^{-14}.$$

Since there are 6 ways of reading the DNA (2 directions, and 3 non-cequivalent points from which to begin the reading), and we have $(5 \times\ 10^6)/3$ triplets each time, we get, if M is the expected number of ORF with length 600:

$$M = 6 \times (5 \times 10^6)/3 \times 1.4\ 10^{-14} = 1.4 \times 10^{-7}.$$

*****

(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.

• Once you have written the code that computes the number of times ORFs of length ℓ appear in the *E.coli* genome, you can plot the obtained distribution along with that based on random distribution of bases obtained in problem 1, as depicted below. The cutoff length

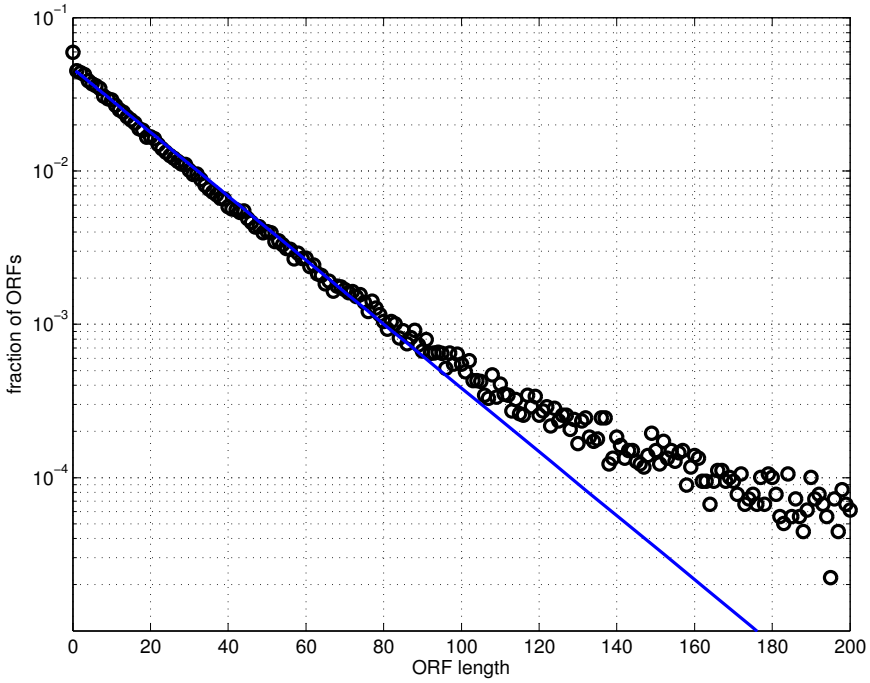

Figure 1: The actual probability distribution of ORFs Ecoli (circles), compared to the theoretical one (blue solid line).

L_{cut} can be estimated from the the graph as $L_{cut} \approx 90$.

*****

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.

• The diagonal element is fixed by the normalization condition to be equal to the sum of the remaining elements on each row, resulting in

$$\begin{pmatrix} P^A_{t+1} \\ P^G_{t+1} \\ P^T_{t+1} \\ P^C_{t+1} \end{pmatrix} = \begin{pmatrix} 1-q-2p & q & p & p \\ q & 1-q-2p & p & p \\ p & p & 1-q-2p & q \\ p & p & q & 1-q-2p \end{pmatrix} \begin{pmatrix} P^A_t \\ P^G_t \\ P^T_t \\ P^C_t \end{pmatrix} .$$

Requiring that the diagonal elements are non-negative, leads to the constraint $q + 2p \leq 1$.

(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.)

• The tranfer Matrix Π_1, being real and symmetric, has a complete basis of orthonormal vectors. The eigenvalues are: $\lambda = 1,\ 1-4p,\ 1-2p-2q,\ 1-2p-2q$. The corresponding normalized eigenvectors are

$$\frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \qquad \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}.$$

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

• A brute-force approach to the problem (not too lengthy, by the way) would be as follows. First, knowing the eigenvalues, we write down the diagonal form of the evolution matrix $\Pi^{(d)}_t$. Next, knowing the eigenvectors, we can build the matrix T that diagonalizes Π_1, i.e.

$$T = \begin{pmatrix} -1/\sqrt{2} & 1/2 & 1/2 & 0 \\ 1/\sqrt{2} & 1/2 & 1/2 & 0 \\ 0 & 1/2 & -1/2 & 1/\sqrt{2} \\ 0 & 1/2 & -1/2 & -1/\sqrt{2} \end{pmatrix}.$$

Finally, we compute

$$\Pi_t = T^{-1}\Pi^{(d)}_t T.$$

A much more elegant way is to realize that the *structure* of the evolution matrix Π_t does not change with time, i.e. it can be written as

$$\Pi_t = \begin{pmatrix} 1-q_t-2p_t & q_t & p_t & p_t \\ q_t & 1-q_t-2p_t & p_t & p_t \\ p_t & p_t & 1-q_t-2p_t & q_t \\ p_t & p_t & q_t & 1-q_t-2p_t \end{pmatrix}.$$

Also, the eigenvalues of Π_t will be $\lambda_t = 1,\ 1-4p_t,\ 1-2p_t-2q_t,\ 1-2p_t-2q_t$. At the same time, they should be equal to the corresponding eigenvalues of the original matrix Π_1 raised to the t-th power. Namely,

$$1-4p_t = (1-4p)^t, \qquad 1-2p_t-2q_t = (1-2p-2q)^t,$$

so that

$$p_t = \frac{1}{4}\left[1-(1-4p)^t\right], \qquad q_t = \frac{1}{4}\left[1+(1-4p)^t\right] - \frac{1}{2}(1-2p-2q)^t.$$

(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.

• For large t, the only surviving component of the initial vector is the one on the eigenvector corresponding to the largest eigenvalue $\lambda = 1$. Since the eigenvector corresponding to the eigenvalue $\lambda = 1$, is

$$\begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix},$$

the asymptotic solution will have an equal probability for all the bases. The time scale(s) for reaching steady state are set by the subdominant eigenvalues of the matrix, all with real non-negative values less than 1. This is also clear by contemplating the equations for p_t and q_t in the previous part. We see that p_t reaches its asymptotic value of $1/4$ when $(1-4p)^t \ll 1$, i.e. on a time scale of $t_1 = -1/\ln(1-4p)$. The decay of q_t to its asymptotic limit also involves a second time scale of $t_2 = -1/\ln(1-2p-2q)$.

(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.

• In general, for any real symmetric, normalization preserving transfer matrix Π_1, there exists an orthonormal basis of eigenvectors, with the largest eigenvalue equal to unity. The Frobenius theorem constrains the other three eigenvalues to be less than unity. Reasoning as before, for large t, the only surviving component is the projection on the eigenvector of the eigenvalue 1, and this eigenvector has again the form:

$$\begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}.$$

For asymmetric transition rates,

$$\begin{pmatrix} P^A_{t+1} \\ P^G_{t+1} \\ P^T_{t+1} \\ P^C_{t+1} \end{pmatrix} = \Pi_1 \cdot \begin{pmatrix} P^A_t \\ P^G_t \\ P^T_t \\ P^C_t \end{pmatrix},$$

with

$$\Pi_1 = \begin{pmatrix} 1-q-p-p+ & q & p & p_- \\ q & 1-q-p-p_- & p_+ & p \\ p & p_- & 1-q-p-p_+ & q \\ p_+ & p & q & 1-q-p-p_- \end{pmatrix}.$$

Again, we look for the eigenvalues, and find $\lambda = 1,\ 1-2p-p_--p_+,\ 1-2p-2q,\ 1-2q-p_--p_+$. As before, for large N the solutions will be approximated by the eigenvector corresponding to the eigenvalue 1, which is

$$\begin{pmatrix} P^A_t \\ P^G_t \\ P^T_t \\ P^C_t \end{pmatrix} = \frac{1}{2+2\alpha} \begin{pmatrix} \alpha \\ 1 \\ \alpha \\ 1 \end{pmatrix},$$

where $\alpha = \frac{p_-+q}{p_++q}$. Note that for $\alpha > 1$ we need $p_- > p_+$.

*****

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 $p(N_A = n, N_B = N-n, t)$.

• This is identical to the mutating population studied in class, and the probabilities satisfy the same equation, i.e.

$$\frac{dp(n,t)}{dt} = a(n+1)p(n+1,t) + b(N-n+1)p(n-1,t) - anp(n,t) - b(N-n)p(n,t)\,,$$

for $0 < n < N$, and with boundary terms

$$\frac{dp(0,t)}{dt} = ap(1,t) - bNp(0,t), \quad \text{and} \quad \frac{dp(N,t)}{dt} = bp(N-1,t) - aNp(N,t)\,.$$

(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.

• We know that this problem is equivalent to independently evolving binary systems. From the example solved in class, we know that starting from $p_A(0) = 1$, the probability to stay in state A decays as

$$p_A(t) = \frac{b}{a+b} + e^{-(a+b)t}\frac{a}{a+b} \quad , \quad \text{and} \quad p_B(t) = 1 - p_A(t).$$

At any time t, we thus expect a binary distribution

$$p(n,t) = \frac{N!}{n!(N-n)!} p_A(t)^n p_B(t)^{N-n} \quad .$$

All that remains is to check that these probabilities indeed satisfy the equations in part (a). We start by directly calculating the time derivative

$$\begin{aligned}
\frac{dp(n,t)}{dt} &= \frac{N!}{n!(N-n)!}\Big[-anp_A(t)^{n-1}p_B(t)^{N-n}e^{-(a+b)t} + a(N-n)p_A(t)^n p_B(t)^{N-n-1}e^{-(a+b)t}\Big] \\
\frac{dp(n,t)}{dt} &= \frac{N!}{n!(N-n)!} a p_A(t)^{n-1}p_B(t)^{N-n-1}e^{-(a+b)t}\big[Np_A(t) - n\big] .
\end{aligned}$$

In the last step we use the fact that $p_A(t) + p_B(t) = 1$. Expression for the time derivative must be the same if we calculate it using the Master equation derived in part (a):

$$\begin{aligned}
\frac{dp(n,t)}{dt} &= a(n+1)p(n+1,t) - anp(n,t) + b(N-n+1)p(n-1,t) - b(N-n)p(n,t) \\
\frac{dp(n,t)}{dt} &= \frac{N!}{n!(N-n)!} a p_A(t)^n p_B(t)^{N-n-1}\Big((N-n)p_A(t) - np_B(t)\Big) \\
&\quad + \frac{N!}{n!(N-n)!} b p_A(t)^{n-1} p_B(t)^{N-n}\Big(np_B(t) - (N-n)p_A(t)\Big) \\
\frac{dp(n,t)}{dt} &= \frac{N!}{n!(N-n)!} p_A(t)^{n-1}p_B(t)^{N-n-1}\big[Np_A(t) - n\big]\big(ap_A(t) - bp_B(t)\big) \\
\frac{dp(n,t)}{dt} &= \frac{N!}{n!(N-n)!} a p_A(t)^{n-1}p_B(t)^{N-n-1}e^{-(a+b)t}\big[Np_A(t) - n\big] .
\end{aligned}$$

We have thus demonstrated that the guessed binary distribution for $p(n, t)$ satisfies the Master equation. It is easy to check that this solution also satisfies the boundary equations.

*****

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, $\overline{f} = \sum_{i=0} x_i(1-s)^i$.

• The surviving fraction with j mutations is equal to $x_j(1-s)^j/\overline{f}$, with $\overline{f} = \sum_{i=0} x_i(1-s)^i$. As the (properly normalized) transition probability from $(k-j)$ to k mutations is $e^{-\mu}\mu^{k-j}/(k-j)!$, we arrive at the recursion relations

$$x'_k = \sum_{j=0}^{k} \frac{x_{k-j}(1-s)^{k-j}}{\bar{f}} \times \frac{\mu^j}{j!} e^{-\mu} \,.$$

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

• The only term in the above recursion sum for $k = 0$, corresponds to $j = 0$, leading to

$$x'_0 = \frac{x_0}{\bar{f}} \times e^{-\mu} \quad \Longrightarrow \quad \bar{f} = e^{-\mu} \quad \text{for } x'_0 = x_0 \,.$$

Interestingly, this mean fitness is independent of s.

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

• Let us try a solution of the form $x'_k = x_k = e^{-\theta}\theta^k/k!$ in the recursion relation:

$$x'_k = e^{-\theta}\frac{\theta^k}{k!} = \frac{1}{\bar{f}} \sum_{j=0}^{k} e^{-\theta} \frac{[\theta(1-s)]^{k-j}}{(k-j)!} \times \frac{\mu^j}{j!} e^{-\mu} = \frac{e^{-\theta-\mu}}{\bar{f}} \times \frac{[\theta(1-s)+\mu]^k}{k!} \,,$$

where we have used the binomial summation formula in the last expression. We can now see that the two sides of the equation agree with the previous choice of $\bar{f} = e^{-\mu}$ as long as

$$\theta = \frac{\mu}{s} \,.$$

.

*****

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}).

• After selection acts on the n-th generation, the normalized distribution is

$$p_n(x) = \begin{cases} \sqrt{\dfrac{2}{\pi\sigma_n^2}} \exp\left[-\dfrac{(x-m_n)^2}{2\sigma_n^2}\right] & x \geq m_n \\ 0 & x < m_n \end{cases} \tag{0.1}$$

The mean value of this distribution is

$$m_{n+1} = m_n + \sqrt{\frac{2}{\pi\sigma_n^2}} \int_0^\infty y \exp\left[-\frac{y^2}{2\sigma_n^2}\right] dy = m_n + \sqrt{\frac{2}{\pi}}\sigma_n, \tag{0.2}$$

where we made a change of variables $x = y + m_n$ for the sake of convenience. It is easy to verify that this change of variables preserves the variance and thus

$$\sigma_{n+1}^2 = < x^2 > - < x >^2 |_{n+1} = < y^2 > - < y >^2 |_{n+1} = \left(1 - \frac{2}{\pi}\right)\sigma_n^2. \tag{0.3}$$

Thus, the mean an variance satisfy the recursion relations

$$m_{n+1} = m_n + \sqrt{\frac{2}{\pi}}\sigma_n, \tag{0.4}$$

$$\sigma_{n+1} = \sqrt{1 - \frac{2}{\pi}}\ \sigma_n. \tag{0.5}$$

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

• From the above recursion relations, one can see that the width of the distribution vanishes as $n \to \infty$. Also,

$$m_{n+1} = m_n + \sqrt{\frac{2}{\pi}}\sigma_n = m_{n-1} + \sqrt{\frac{2}{\pi}}(\sigma_{n-1} + \sigma_n) = ... = m_0 + \sqrt{\frac{2}{\pi}}\sum_{i=0}^{n}\sigma_i. \tag{0.6}$$

Plugging in the recursion relation for σ_n, we find as $n \to \infty$

$$m_\infty = m_0 + \sqrt{\frac{2}{\pi}}\sigma_0 \sum_{i=0}^{\infty}\left(\sqrt{1 - \frac{2}{\pi}}\right)^i = m_0 + \frac{\sqrt{2/\pi}}{1 - \sqrt{1 - \dfrac{2}{\pi}}}\sigma_0 \simeq m_0 + 2\sigma_0. \tag{0.7}$$

(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.

• The Fourier transform of a convolution is the product of the Fourier transformed components. As the Fourier transform of a Gaussian is another Gaussian, the generating function of the probability of fitness after convolution takes the simple form

$$\langle e^{-ikx} \rangle = \exp\left(-ikm_n - k^2\sigma_n^2/2\right) \times \exp\left(+ik\mu - k^2 s^2/2\right) . \tag{0.8}$$

Clearly this describes a Gaussian distribution of mean $m_n - \mu$ and variance $\sigma_n^2 + s^2$. The recursion relations of part (a) are thus modified to

$$m_{n+1} = m_n + \sqrt{\frac{2}{\pi}}\sigma_n - \mu \,, \tag{0.9}$$

$$\sigma_{n+1}^2 = \left(1 - \frac{2}{\pi}\right) \sigma_n^2 + s^2 \,. \tag{0.10}$$

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

• Due to the additional s^2 in the recursion relation for variance, it no longer goes to zero at long times, instead converging to a value of $\sigma_\infty^2 = \pi s^2/2$. After the variance has converged to this value, the mean continues to evolve as

$$m_{n+1} = m_n + s - \mu \,. \tag{0.11}$$

If $s > \mu$, the fitness distribution continues to move towards larger value, taking advantage of the occasional good mutations. However, if $\mu > s$, the mostly deleterious mutations overwhelm the population and fitness continues to decrease.

*****