Evolutionary Dynamics: Homework 02

Guido Putignano, Lorenzo Tarricone, Gavriel Hannuna,
Athanasia Sapountzi

February 22, 2024

Problem 1: Sequence Space and Hamming Distance

Consider an alphabet A of size $|A| = B$. For a binary alphabet, one has $A = \{0, 1\}$ and $B = 2$, and for DNA, one has $A = \{A, T, C, G\}$ and $B = 4$. We are studying sequences $S \in A^L$ of length L. Assume sequences are random with a uniform distribution.

(a) How many unique binary and DNA sequences exist for $L = 30$? (1 point)

Solution:

Sequence space A^L has dimension L, and consists of $|A|^L = B^L$ sequences.

For a binary alphabet: $A = \{0, 1\}$, $B = 2$ and $L = 30$

The number of unique binary sequences is $B^L = 2^{30}$.

For DNA: $A = \{A, T, C, G\}$, $B = 4$ and $L = 30$

The number of unique binary sequences is $B^L = 4^{30}$.

(b) What is the average Hamming distance between two random binary sequences? What is the expected Hamming distance for two random DNA sequences? (1 point)

Solution:

By definition, the Hamming distance between two sequences of equal length is the number of positions at which the corresponding symbols are different.

To calculate the average Hamming distance between two random binary sequences, we consider that for binary sequences, there are 2 possible symbols for each position. Thus, there is a 50% possibility of a mismatch. The length of the sequences is L, so the average Hamming distance is $0.5L$.

To calculate the expected Hamming distance for two random DNA sequences, we consider that for DNA sequences, there are 4 possible symbols for each position. Thus, there is a 75% possibility of a mismatch. The length of the sequences is L, so the expected Hamming distance is $0.75L$.

(c) Given a binary sequence of length L, how many sequences exist at a Hamming distance three from it? How many at distance K with $K \leq L$? Repeat the calculation for DNA sequences.

Solution:

To calculate the number of sequences that exist at a Hamming distance three for a given binary sequence of length L, we use the binomial coefficient. The binomial coefficient is defined as the number of ways of choosing k objects out of n without regard to order and is given by $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Since we want to calculate distances of length L with Hamming distance 3 we have $\binom{L}{3} = \frac{L!}{3!(L-3)!}$. The binary alphabet consists of two characters. Thus to have a mismatch, there is a $B - 1 = 2 - 1 = 1$ possible option for each position and there are $1^3 = 1$ possible combinations. So the number of sequences is $\frac{L!}{3!(L-3)!} * 1 = \frac{L!}{3!(L-3)!}$.

Similarly for Hamming distance k we have $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and there are $1^k = 1$ possible combinations. So the number of sequences is $\frac{n!}{k!(n-k)!} * 1 = \frac{n!}{k!(n-k)!}$.

To calculate the number of sequences that exist at a Hamming distance three for a given DNA sequence of length L, we follow the same methodology. to calculate distances of length L with Hamming distance 3 we have $\binom{L}{3} = \frac{L!}{3!(L-3)!}$. The DNA alphabet consists of four characters. Thus to have a mismatch, there are $4 - 1 = 1$ possible options for each position and there are $3^3 = 9$ possible combinations. So the number of sequences is $\frac{L!}{3!(L-3)!} * 9 = \frac{3L!}{2(L-3)!}$.

Similarly for Hamming distance k we have $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and there are 3^k possible combinations. So the number of sequences is $\frac{n!}{k!(n-k)!} * 3^k = \frac{3^k n!}{k!(n-k)!}$.

Note: In the point c we used combinations instead of permutations because it doesn't matter in which order you flip the bits or change the elements in a sequence. All that matters is which positions are affected. Using permutations would imply considering the order, which is not relevant in this case.

Problem 2: Quasispecies

Consider the quasispecies equation with two genotypes 0,1 (i.e., binary sequences of length 1). Let the fitness of genotype 0 be $f_0 > 1$, and the fitness of genotype 1 be $f_1 = 1$. Moreover, genotypes are replicated error-free with probability q.

(a) Write down the mutation-selection matrix W and find its eigenvalues. (2 points)

Solution: By definition the mutation-selection matrix W has the form $(w_{ij})_{ij} = (f_i q_{ij})_{ij}$ that in this case is:

$$W = \begin{bmatrix} qf_0 & (1-q)f_0 \\ (1-q) & q \end{bmatrix}$$

(To ease the notation we will write f_0 as f, because the value f_1 is fixed to 1 and therefore there is no ambiguity)
In order to find the eigenvalues we need to find the roots of the characteristic polynomial of $W - \lambda\mathbb{I}$

$$det|W - \lambda\mathbb{I}| = (qf - \lambda)(q - \lambda) - (1-q)(1-q)f = 0$$

$$q^2 f - \lambda qf - \lambda q + \lambda^2 - [(1 + q^2 - 2q)f] = q^2 f - \lambda qf - \lambda q + \lambda^2 - f - q^2 f + 2qf$$

$$q^2 f - \lambda qf - \lambda q + \lambda^2 - f - q^2 f + 2qf = \lambda^2 + (-qf - q)\lambda + (2qf - f) = 0$$

$$\lambda_{1,2} = \frac{qf + q \pm \sqrt{q^2f^2 + q^2 + 2fq^2 - 8qf + 4f}}{2}$$

(b) To which eigenvalue corresponds the non-trivial equilibrium point? (1 point) Hint: Perron-Frobenius theorem.
Solution: Perron-Frobenius theorem applies in this setting as we have a matrix with all non-zero entries (at least as long as $q \neq \{0, 1\}$) and this implies that no higher power of this matrix will have zeros in its entries. In practice, this can also be seen by assessing that the graph relative to W is strongly connected (i.e. there exists just one connected component). The theorem therefore ensures that we will have a positive real eigenvalue with a modulus bigger than all the other eigenvalues (in this case of the second eigenvalue). The positivity of $qf - q$ allows us to conclude that the eigenvalue with the + is gonna be the biggest in modulus and therefore the one to which will correspond a non-trivial solution of the system.
Eigenvalue with the non-trivial equilibrium point

$$\lambda_1 = \frac{qf + q + \sqrt{q^2f^2 + q^2 + 2fq^2 - 8qf + 4f}}{2}$$

(c) Examine the dynamics of the quasispecies equation and confirm the results obtained in (b). Assume that $q = 0.6$ and $f_0 = 1.5$, and initial condition $(0.75, 0.25)$. (1 point)
Solution: Here is the analysis of our quasispecies system, with the values of $q = 0.6$ and $f_1 = 1.5$. The blue curve represents the population count of genotype 0, while the orange curve represents the one of genotype 1

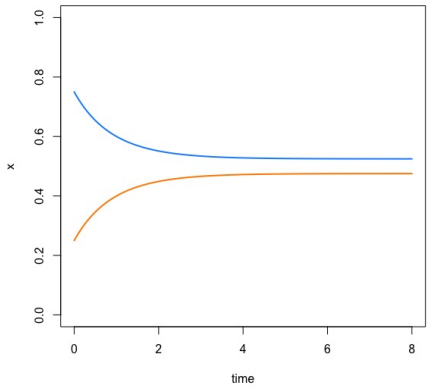

Figure 1: Normalized numerical solution of the Quasispecies equation

When plugging in the values of f = 1.5 and q = 0.6 we will find the following equation

$$\lambda_{1,2} = \frac{0.9 + 0.6 \pm \sqrt{(0.36 * 2.25) + 0.36 + 2(1.5 * 0.36) - 8 * 0.9 + 6}}{2}$$

$$\lambda_{1,2} = \frac{1.5 \pm \sqrt{1.05}}{2}$$

$$\lambda_1 = 1.26$$

$$\lambda_2 = 0.24$$

These eigenvalues were also found in our numerical simulation. The equilibrium points of the population curves of genotypes 0 and 1 can be found at $x = 0.525$ and $x = 0.475$, and are confirmed by Figure 1.

(d) What is the equilibrium point for $f_0 = f_1 = 1$? (1 point):
Solution: for $f_0 = f_1 = 1$ the matrix W becomes

$$W = \begin{bmatrix} q & (1-q) \\ (1-q) & q \end{bmatrix}$$

And the corresponding characteristic polynomial of which we want to find the roots becomes

$$(q-\lambda)(q-\lambda) - (1-q)(1-q) = \lambda^2 - 2q\lambda - [1 + q^2 - 2q] = \lambda^2 - 2q\lambda - 1 + 2q = 0$$

This has solution $\lambda_1 = 2q - 1$ and $\lambda_2 = 1$. We know average fitness is the largest eigenvalue and that the solution x^* is the corresponding normalized eigenvector. Here it's clear to see how the biggest eigenvalue (excluding the degenerate cases $q = \{0, 1\}$) is 1 and this also coincides with the average between $f_0 = 1$ and $f_1 = 1$. Plugging back this eigenvalue in the matrix gives the system of equations

$$W - 1\mathbb{I} = \begin{bmatrix} (q-1) & (1-q) \\ (1-q) & (q-1) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The system has of course one degree of freedom and yields the relationship $x = y$. The unique vector in the probability simplex respecting this condition is $x^* = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ which is our desired solution. Interestingly, we can notice how when the two genotypes have the same fitness and there is mutation we have a steady state where the two types coexist with the same ratios

(e) Calculate the equilibrium point in the limit of low mutation rate ($q \approx 1$). (1 point) **Solution**: Analogously to what we have done in the preceding point we now have the following mutation-selection matrix:

$$W = \begin{bmatrix} f & 0 \\ 0 & 1 \end{bmatrix}$$

As we are dealing with a diagonal matrix we can directly read the eigenvalues on the main diagonal, therefore $\lambda_1 = 1$ and $\lambda_2 = f$. Given our initial assumption of $f > 1$, we see that this is the largest (real positive) eigenvalue that will dictate the behaviour of the system. Imposing a system analogous to the one described above will produce the following condition:

$$y(1 - f) = 0 \longrightarrow y = 0$$

Now the only solution in the simplex becomes $x^* = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ which again makes sense as with no mutation we witness the phenomenon of the survival of the fittest.