Evolutionary Dynamics Exercise 2

Minghang Li

October 09, 2024 (America/New_York)

This solution is organized in R Markdown, and the coding part is written in `python` via `reticulate`.

Problem 2: Sqeuence space and Hamming distance

Consider a binary alphabet $\mathcal{A} = \{0, 1\}$ and a DNA alphabet with $\mathcal{A} = \{A, T, C, G\}$. We are studying sequences $S \in \mathcal{A}^L$ of length L.

(a)

> Assume sequences are random with a uniform distribution. What is the average Hamming distance between two random binary sequences? What is the expected Hamming distance for two random DNA sequences?

Hamming distance is defined to be the number of positions at which the corresponding symbols between two sequences are different.

For two random binary sequences, the probability of having a different symbol at a position is $1/2$, and the expected Hamming distance is $L/2$.

For two random DNA sequences, the probability of having a different symbol at a position is $3/4$, and the expected Hamming distance is $3L/4$.

(b)

> Instead of a uniform distribution over the alphabet, we assume 0 has a 2 times higher probability of appearing at a position in the sequence than 1. Furthermore, the probabilities $p(A) = p(T)$ are twice the probabilities $p(C) = p(G)$. What is now the average Hamming distance between two random binary sequences? And the expected average Hamming distance for two random DNA sequences?

Binary

Let x, y be two random binary sequences. As 0 has a 2 times higher probability of appearing at a position in the sequence than 1, we have $P(x_i = 0) = 2/3$ and $P(x_i = 1) = 1/3$.

The probability of having a different symbol at a position is

$$\begin{aligned} P(x_i \neq y_i) &= P(x_i = 0, y_i = 1) + P(x_i = 1, y_i = 0) \\ &= P(x_i = 0)P(y_i = 1) + P(x_i = 1)P(y_i = 0) \\ &= \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} \\ &= \frac{4}{9} \end{aligned}$$

The expected Hamming distance is $4L/9$.

DNA

Let x, y be two random DNA sequences. As $p(A) = p(T)$ are twice the probabilities $p(C) = p(G)$, and $p(A) + p(T) + p(C) + p(G) = 1$, we have $p(A) = p(T) = 1/3$ and $p(C) = p(G) = 1/6$.

The probability of having a different symbol at a position is

$$\begin{aligned} P(x_i \neq y_i) &= P(x_i = A, y_i \neq A) + P(x_i = T, y_i \neq T) + P(x_i = C, y_i \neq C) + P(x_i = G, y_i \neq G) \\ &= 2 \cdot \frac{1}{3} \cdot \frac{2}{3} + 2 \cdot \frac{1}{6} \cdot \frac{5}{6} \\ &= 2 \cdot \left(\frac{2}{9} + \frac{5}{36} \right) \\ &= \frac{13}{18} \end{aligned}$$

The expected Hamming distance is $13L/18$.

(c)

> Given a binary sequence of length L, how many sequences exist at a Hamming distance 2 from it? How many at distance K with $K \leq L$? Repeat the calculation for DNA sequences.

Let's denote Hamming distance as d_H.

Binary

$d_H = k$ means that there are k out of L positions that are different. At each position with different symbols, we need to choose from $|\mathcal{A}| - 1 = 2 - 1 = 1$ symbol to have a mismatch, and that results in $1^k = 1$ combinations.

$$\begin{cases} d_H = 2: & \binom{L}{2} \cdot 1^2 \\ d_H = K: & \binom{L}{K} \cdot 1^K \end{cases}$$

DNA

The rationale is similar, except that at each position we now can choose from $|\mathcal{A}| - 1 = 4 - 1 = 3$ symbol to have a mismatch, and that results in 3^k combinations.

$$\begin{cases} d_H = 2: & \binom{L}{2} \cdot 3^2 \\ d_H = K: & \binom{L}{K} \cdot 3^K \end{cases}$$

Problem 3: 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. The quasispecies equation is defined as

$$\dot{x}_i = \sum_{j=1}^{n} \underbrace{x_j}_{\text{frequency of } j} \overbrace{f_j}^{\text{fitness of } j} \underbrace{q_{ji}}_{\substack{\text{mutation} \\ \text{from } i \text{ to } j}} - \overbrace{\phi}^{\substack{\text{average} \\ \text{fitness}}} x_i, \qquad i = 0, \dots, n$$

In vector notation, this can be written as

$$\dot{x} = xW - \phi x$$

where $W = (w_{ij}) = (f_j q_{ji})$ is the mutation-selection matrix and $x = (x_0, \dots, x_n)$.

(a)

Write down the mutation-selection matrix W and find its eigenvalues.

$$\begin{aligned} W &= \begin{bmatrix} w_{00}, & w_{01} \\ w_{10}, & w_{11} \end{bmatrix} \\ &= \begin{bmatrix} f_0 q_{00} & f_1 q_{10} \\ f_0 q_{01} & f_1 q_{11} \end{bmatrix} \\ &= \begin{bmatrix} f_0 q & (1-q) \\ f_0(1-q) & q \end{bmatrix} \end{aligned}$$

The eigenvalues of W are the solutions to the characteristic equation:

$$\begin{aligned} \det(W - \lambda I) &= 0 \\ \det\left(\begin{bmatrix} f_0 q - \lambda & (1-q) \\ f_0(1-q) & q - \lambda \end{bmatrix} \right) &= 0 \\ (f_0 q - \lambda)(q - \lambda) - f_0(1-q)(1-q) &= 0 \\ f_0 q^2 - \lambda f_0 q - \lambda q + \lambda^2 - (f_0 q^2 - 2 f_0 q + f_0) &= 0 \\ \lambda^2 - q(f_0 + 1)\lambda + f_0(2q - 1) &= 0 \end{aligned}$$

The eigenvalues are

$$\lambda_{1,2} = \frac{q(f_0 + 1) \pm \sqrt{q^2(f_0 + 1)^2 - 4 f_0 (2q - 1)}}{2}$$

(b)

What are the equilibrium points of the population? To which eigenvalue corresponds the non-trivial equilibrium point? *Hint: Perron-Frobenius theorem.*

Let's recall Perron-Frobenius theorem:

Theorem 0.1 (Perron-Frobenius theorem)**.** *Consider a **irreducible non-negative** $n \times n$ matrix M. Matrix M has a positive eigenvalue λ_{max}, such that all other eigenvalues satisfy $|\lambda| < \lambda_{max}$. Furthermore, λ_{max} is simple and the components of the associated eigenvector w are all (strictly) positive, $w_i > 0 \forall i$.*

(c)

Examine the dynamics of the quasispecies equation and confirm the results obtained in **(b)**. Assume that $q = 0.7$ and $f_0 = 1.5$, with the initial condition $(0.9, 0.1)$. Modify your simulation to track the average fitness of the population over time. How does the average fitness evolve as the population approaches equilibrium?

(d)

What is the equilibrium point for $f_0 = f_1 = 1$?

(e)

In the lecture you heard about a concept called the error threshold. Assume for this subtask a wildtype genotype with fitness $f_{wt} > 1$ and all other genotypes have fitness $f_m = 1$. For a large sequence length, you can assume, that once a sequence is mutated, back-mutations are negligibly unlikely. Can you derive a condition based on q and f_{wt} for this error threshold? What happens when the threshold is crossed?