Evolutionary Dynamics: Homework 11

Guido Putignano, Lorenzo Tarricone, Gavriel Hannuna,
Athanasia Sapountzi

February 22, 2024

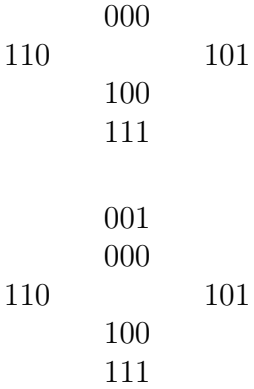
Problem 1: Risk polynomial and escape for a three-locus genotype

HIV can develop resistance to antiviral drugs by accumulating mutations. Assume that resistance to therapy is developed by accumulating three specific mutations, i.e. the mutant $1 = 111$ is the only resistant genotype. We consider the following order constraints and genotype lattices according to which the virus can evolve:

(i) $\emptyset$

(ii) $\{1 < 2, 2 < 3\}$

(iii)



(iv)

(a) Draw the Hasse diagram and genotype lattice for the order constraints (i) and (ii) (tutorial exercise).

(b) Write down the order constraints and corresponding Hasse diagrams defined by the lattices (iii) and (iv) (tutorial exercise).

(c) Count the number of chains from 0 to 1 for all four posets (tutorial exercise).

(d) Write down the risk polynomial for the four posets (tutorial exercise).

(e) Suppose that, after intervention, all variants except the escape mutant 1 have fitness 0.9. Assume a mutation rate of $u = 3 \cdot 10^{-5}$. Compute the critical population size of all posets and order them by critical population size. What connection do you observe between the evolutionary constraints and the critical population size? (2 points)

Solution

The critical population size N^* is defined as $1/\xi_0$, so we must calculate ξ_0.

$$\xi_0 = \xi_1 f_0 \Pi_{e \in \epsilon} \mu_e R(G, f)$$

Where μ_e is the mutation rate for event e. Since we are in a 3-dimensional space, $\Pi_{e\in\epsilon}\mu_e = (3 * 10^{-5})^3$. $R(G, f)$ is the risk polynomial, which we should evaluate for all the previous posets. ξ_g is the probability of escape, so $\xi_1 \approx 1$, while f is the fitness, which is given to be 0.9 for all mutants other than mutant 111. The final equation will be:

$$\xi_0 = 0.9 \cdot (3 * 10^{-5})^3 \cdot R(G, f)$$

For the empty set $\{\emptyset\}$, we will have 13 chains and the risk polynomial will be:

$$R = 1+f_{100}+f_{010}+f_{001}+f_{011}+f_{101}+f_{110}+f_{100}(f_{101}+f_{110})+f_{010}(f_{011}+f_{110})+f_{001}(f_{011}+f_{101})$$

$$R = 1 + 0.9 \cdot 6 + 0.9^2 \cdot 6 = 11.26$$

So $\xi_0 = 0.9 \cdot (3 * 10^{-5})^3 \cdot 11.26 = 2.74 \cdot 10^{-13}$ and $N^* = 1/\xi_0 = 3.65 \cdot 10^{12}$. Now, doing similar calculations with all of the risk polynomials that we calculated during the tutorial, we will get the results listed in Table 1

Set	N^*
$\{1 < 2, 2 < 3\}$	$1.14 \cdot 10^{13}$
$\{1 < 2, 1 < 3\}$	$7.74 \cdot 10^{12}$
$\{1 < 2\}$	$5.85 \cdot 10^{12}$
$\{\emptyset\}$	$3.65 \cdot 10^{12}$

Table 1: Critical population size for all posets

The table unmistakably indicates that as the evolutionary constraints on a population decrease, the critical population size also diminishes, meaning escape requires fewer individuals. This is rather intuitive, since the more freedom of evolution we have the more possible pathways to escape we can reach.

Problem 2: Probability of extinction on distributive lattices (4 points)

In the lecture, we defined the risk of escape ξ_g given a genotype lattice G as the probability of reaching genotype 1 starting from one individual of type g. This probability satisfies (approximately) the following recurrence equation

$$\xi_g = f_g \sum_{h \supset g} \xi_h u_{gh}. \quad (1)$$

The goal is to solve the recurrence to show that

$$\xi_g = \xi_1 f_g \prod_{e \in E \setminus g} \mu_e \cdot P_{g1}(f). \quad (2)$$

First, derive equation (2) from (1) in the case where there is no other chain from g to 1 than $g \subset 1$. Then, assuming equation (2) has been shown in the case where no chain starting at g and ending at 1 has length more than $k-1$, show that the equation holds if it the chain may have length k. Finally, conclude that

$$\xi_0 = \xi_1 f_0 \prod_{e \in E} \mu_e \cdot R(G, f),$$

where $R(G, f)$ is the risk polynomial.

Hint: You may use without proof that $P_{g1} = \sum_{g=g_0 \subset g_1 \subset \cdots \subset g_k=1; k=1,\ldots,n} \frac{f_{g1}}{f_{g2}} \cdots \frac{f_{g_{k-1}}}{f_{g_k}}$.

Solution: The explanation of many of the terms are coming from the paper: Beerenwinkel N, Eriksson N, Sturmfels B (2006). Evolution on distributive lattices. J Theor Biol 242:409420.

Given the recurrence equation:

$$\xi_g = f_g \sum_{h \supset g} \xi_h u_{gh} \tag{1}$$

This equation states that the risk of escape, ξ_g, for a genotype g is given by the fitness, f_g, of genotype g times the sum over all genotypes h that are direct successors of g (i.e., $h \supset g$) of the product of the risk of escape, ξ_h, for those genotypes and the mutation rate, u_{gh}, from g to h.

In the specific case where the only path from g to 1 is $g \subset 1$, this implies that $h = 1$ is the only successor of g. Therefore, the summation over h reduces to just one term where $h = 1$. The equation simplifies to:

$$\xi_g = f_g \xi_1 u_{g1} \tag{2}$$

Now, considering the equation:

$$\xi_g = \xi_1 f_g \prod_{e \in E \setminus g} \mu_e \cdot P_{g1}(f) \tag{3}$$

Here, μ_e represents the mutation rates associated with each edge e in the path from g to 1 excluding g, and $P_{g1}(f)$ is the generating function for all chains from g to h = 1 in G.

Since there is no other chain from g to 1 other than $g \subset 1$, the product term simplifies to just the mutation rate along the edge from g to 1, which is u_{g1}. Also, $P_{g1}(f)$ simplifies to the fitness of g, f_g, since it's the only intermediate fitness value.

Therefore, equation (3) simplifies to:

$$\xi_g = \xi_1 f_g u_{g1} \tag{4}$$

Here we demonstrated that those equations can be related to each other.

Given that the element in the mutation matrix $\mathbf{U} = (u_{gh})_{g,h \in \mathcal{G}}$ is defined by

$$u_{gh} = \begin{cases} \prod_{e \in h \setminus g} \mu_e & \text{if } h \in N(g) \\ 0 & \text{otherwise} \end{cases}$$

The generalisation of the equation can be obtain through an inductive approach.

Assume for any chain from g to 1 with length at most $k - 1$, the following holds:

$$\xi_g = \xi_1 f_g \prod_{e \in E \setminus g} \mu_e \cdot P_{g1}(f) \tag{5}$$

Consider a chain from g to 1 of length k. Let g' be an intermediate genotype such that $g \subset g' \subset 1$ and the chain from g' to 1 has length $k-1$. By the induction hypothesis:

$$\xi_{g'} = \xi_1 f_{g'} \prod_{e \in E \setminus g'} \mu_e \cdot P_{g'1}(f) \tag{6}$$

Substitute this into the original recurrence equation:

$$\xi_g = f_g \sum_{h \supset g} \xi_h u_{gh} \tag{7}$$

Focusing on the term involving g':

$$\xi_g = f_g \xi_{g'} u_{gg'} \tag{8}$$

Simplifying:

$$\xi_g = f_g \left(\xi_1 f_{g'} \prod_{e \in E \setminus g'} \mu_e \cdot P_{g'1}(f) \right) u_{gg'} \tag{9}$$

Therefore:

$$\xi_g = \xi_1 f_g \prod_{e \in E \setminus g} \mu_e \cdot P_{g1}(f) \tag{10}$$

By induction, the equation holds for any chain length, including the chain starting at $g = 0$ and ending at 1:

$$\xi_0 = \xi_1 f_0 \prod_{e \in E} \mu_e \cdot R(G, f), \tag{11}$$

where $R(G, f)$ is the risk polynomial defined as:

$$R(G, f) = \sum_{g = g_0 \subset g_1 \subset \cdots \subset g_k = 1; k = 1, \ldots, n} f_{g1} f_{g2} \ldots f_{gk-1} \tag{12}$$

Problem 3

a Show that the column vector of mutant equilibrium probabilities $\vec{x} = (x_1, \ldots, x_m)^T$ is given by

$$\vec{x} = x_0 (I - FU)^{-1} F U_0$$

where F is the diagonal matrix $F = \text{diag}\left[\frac{1}{1-w_1}, \frac{1}{1-w_2}, \ldots, \frac{1}{1-w_m}\right]$, U is the mutation matrix $U = \{u_{ij}\}$, and U_0 is the column vector $U_0 = (u_{10}, u_{20}, \ldots, u_{m0})^T$.

Solution: Given the equilibrium condition, we can define the vector of first derivatives $\frac{\partial x_i}{\partial t} = 0$ for all i (or equivalently $\mathbf{x}' = \mathbf{0}$). We can now rewrite the given equation using the definitions provided in the text:

$$\mathbf{x}' = \mathbf{0} = -\mathbf{F^{-1}x} + \mathbf{Ux} + x_0 \mathbf{U_0}$$

Now, multiply all terms on the left by $\mathbf{F}$ to obtain:

$$\mathbf{0} = -\mathbb{I}\mathbf{x} + \mathbf{FU}\mathbf{x} + x_0\mathbf{FU_0}$$

If we now group the terms in $\mathbf{x}$ and bring them on the left-hand side we have:

$$(\mathbb{I} - \mathbf{FU})\mathbf{x} = x_0\mathbf{FU_0}$$

Given that the matrix $(\mathbb{I} - \mathbf{FU})$ it's clearly square and with a determinant different from zero it will be invertible and we can therefore multiply both sides (again on the left) by this inverse. This will yield the desired result.

b When the wild type is dominant ($x_0 \approx 1$), we can rewrite $\vec{x} \approx FU_0 + FUFU_0 + FUFUFU_0 + \ldots$. This can be rewritten as $x_i = \sum_{q:0\to i} v(q)$, sum over all paths $q : 0 = k_1 \to k_2 \to \ldots \to k_m = i$ connecting 0 to i, and $v(q) = u_{k_2k_1}u_{k_3k_2}\ldots u_{kmk_{m-1}}f_{k_2}f_{k_3}\ldots f_{km}$, with $f_i = \frac{1}{1-w_i}$. Show that, in the limit of small, uniform mutation rate u, this can be written as $x_i = ud_{0i}f_iR(G_i, f)$, where d_{ij} is the Hamming distance between i and j, G_i is the sublattice spanning all paths from 0 to i, and $R(G_i, f)$ is the risk polynomial.

Solution: We explicitly write the given expression as:

$$x_i = \sum_{q:0\to i} u_{k_2,k_1}u_{k_3,k_2}...u_{m,m-1} \cdot f_{k_2}f_{k_3}...f_{k_m=i}$$

. In order to arrive at the desired result we can notice some things

- Whatever the chain that we are considering the last element will always be the genotype i. This means that the term $f_{k_m=i}$ will be present in each summand and therefore we can factor it out of the sum
- If the mutation rate is uniform we can forget about the indices (we are therefore path independent and we can also factor them out. In addition, the number of these u factors will be always the Hamming distance between the wildtype and the genotype i of interest. This last statement is true exactly because the uniform probability of mutating from genotype A to genotype B (without back mutations) amounts to counting the mismatches in genotype positions (let's say these sum up to n). If we then ask for the probability of mutation in these selected loci in the genome we will have (thanks to the assumed independence) u^n. In this case, we have n as the number of mismatches between these binary sequences and therefore the Hamming distance.
- If we now look at what remains in the sum we have $\sum_{q:0\to i} f_{k_2}f_{k_3}...f_{k_{m-1}}$ and this corresponds exactly to how we have defined $\mathcal{R}(\mathcal{G}_i, \mathbf{f})$