Problem set 4: Finite number fluctuations and random walks, models with discrete time

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Simulation of the Moran process

The Moran process is a model of population genetics that aims to describe the evolution of the composition of a population of asexual microorganisms with fixed number of individuals N under natural selection and genetic drift (stochastic fluctuations due to the finite size of the population). We will assume that there are two types of individuals in the population, namely type A (e.g. mutants) and type B (e.g. wild-type organisms).

In the Moran process, time is discrete. At each time step, two events take place simultaneously: one individual reproduces, yielding one offspring identical to itself, and one individual dies, thus leaving N constant. We will consider that fitness corresponds to reproduction rate: the individual that dies will be chosen uniformly at random, while the individual that divides will be chosen proportionally to its fitness. Let us denote by f_A the fitness of A, by f_B the fitness of B and by r their ratio $r = f_A/f_B$. Because at each time step there is a birth and a death event, the total population size remains constant. (Note that no new mutations can occur in this model, and that we consider that the same individual may be chosen for both death and reproduction at the same time step.)

a) Explain why the state of the population is completely described by the number i of A individuals. What values can i take? Upon each time step, how can it vary?

b) Show that the probability for i to increase by 1 upon a given time step is

$$\alpha_i = \frac{ri(N-i)}{N(ri+N-i)}\,. \tag{1}$$

and that the probability for i to decrease by 1 upon a given time step is

$$\beta_i = \frac{i(N-i)}{N(ri+N-i)}\,. \tag{2}$$

c) Similarly, calculate the probability for i to stay constant upon a given time step. Check that these probabilities are normalized.

d) What happens if $r = 1$? What does this correspond to?

e) You will now simulate one discrete time step in the Moran process using Python. For this, take for instance $N = 10$, $i = 4$ and $r = 1.1$, but name them in your program so that you can then change the values easily. First choose whether an A or a B individual is going to die, by drawing a uniformly distributed random number between 0 and 1 that will subsequently be compared to i/N. Next, choose whether an A or a B individual is going to divide, by drawing another uniformly distributed random number between 0 and 1 that will subsequently be compared to $ri/(ri+N-i)$. Then, by performing these two comparisons, find out whether an A or a B divides and whether an A or a B dies. Explain why this procedure is consistent with the Moran model. Finally, update the number i of A individuals according to the outcomes of your two comparisons. Test your simulation several times, printing the outcome of each intermediate step to check that it does the right things.

f) Now that you have simulated one discrete time step of the Moran process, make a loop over such time steps t, for $t = 0$ to $T - 1$, taking for instance $T = 1000$, and use (copy) your code from the previous question within this loop to run the Moran model. Make a table containing the values of i, so that your codes returns the value of i for each time t. Plot i versus t. Do this at least 3 times and plot the resulting plots. What are the final values of i that you obtain? Explain why intermediate outcomes are not possible.

g) Given the result of the previous question, (copy and) modify your code by making your loop run until type A disappears or takes over. Again, plot i versus t. Do this at least 3 times and plot the resulting plots. What do you observe?

h) Now write a new version of your simulation of the Moran process, within a loop over M replicates. Remove the table storing i as a function of t for each separate replicate, and instead, produce a table that just stores the outcome (0 if A disappeared, 1 if A took over) and the final value of t for each replicate. Test that it works for $M = 10$.

2 Fixation probability and fixation time

From now on, we will start from $i = 1$ instead of $i = 4$. The idea is to investigate the fate of a single mutant of type A appearing in a population of B individuals. We keep $N = 10$ and $r = 1.1$ unless otherwise specified.

a) Using the code you wrote in the last question of the previous exercise, with $M = 1000$, calculate the number of times A disappeared and the number of times A took over, and deduce an estimate of the probability p_A that A takes over (fixation probability of A).

b) Also calculate the average over $M = 1000$ replicates of the time t_f it took for the A to either disappear or take over. By restricting to those replicates where A took over, estimate the average time t_{fA} it takes for A to take over. Do the same for the average time t_{fB} it takes for A to disappear. Check that $t_f = p_A t_{fA} + (1 - p_A) t_{fB}$.

c) Embed (a copy of) your previous code in a loop where N (or r) can be varied, with $M = 1000$ replicates for each such value. Modify it so that it returns in a table the estimates of p_A, t_f, t_{fA} and t_{fB} computed over $M = 1000$ replicates for each value of N (or r). Using this code for $r = 1$, starting from $i = 1$, and varying N so that it takes the values $N = 10$, 12, 15, 20, 25, 50, 100, 200, 500, 1000, plot p_A versus N, and then versus $1/N$. Is the result surprising? Comment on the ability of a neutral mutant to take over in a population. Also plot t_f, t_{fA} and t_{fB} versus N and comment.

d) Keeping $N = 100$ constant, but varying r such that $r = 0.8$, 0.9, 1, 1.1, 1.2, plot p_A versus r and comment. Also plot t_f, t_{fA} and t_{fB} versus r and comment.

e) Compare your plot of p_A versus r with the analytical prediction for p_A:

$$p_A = \frac{1 - r^{-1}}{1 - r^{-N}} \,. \tag{3}$$

3 Additional problem: Transcription factor moving on DNA

This problem was previously given as part of the final exam of this class.

Transcription factors are proteins that can regulate the expression of genes by binding to specific DNA sequences, called binding sites. They need to find these binding sites on DNA. For this, they can either move in three dimensions in the cytoplasm, or bind non-specifically to DNA and move along it in one dimension. Here, we will focus on the second type of motion.

a) The binding energy between a transcription factor and a binding site on DNA is typically $\Delta E = 10 k_{\mathrm{B}} T$. Give the expression of the ratio of the probabilities of the bound state (state 1) where the transcription factor is bound to the binding site, and the unbound state (state 2). What is the name of the probability distribution you used to express this ratio?

b) Compute the value of this ratio. Comment: how much more difficult or easy is unbinding compared to binding for the transcription factor?

c) Let us now consider a transcription factor that binds non-specifically to DNA (i.e. it binds weakly to all DNA regions that are not a binding site for it). How do you expect the non-specific binding energy to compare to the specific one, $\Delta E = 10 k_B T$?

For simplicity, we will assume that our transcription factor is on a segment of DNA, delimited on the left side by the end of the chromosome (if this boundary is reached, the transcription factor detaches from DNA) and on the right side by a specific transcription factor binding site (see Fig. 1). We will also discretize the system, and assume that there are N non-specific binding sites on the DNA segment. The transcription factor can jump left or right to the next non-specific binding site. We will denote by α the probability that it jumps right, and by $1-\alpha$ the probability that it jumps left. We assume all non-specific binding sites to be identical, and thus the value of α does not depend on the site i for $1 \leq i < N$. If site N is reached, the transcription factor necessarily binds to the specific binding site in $N+1$. We assume for simplicity that the transcription factor can only detach if it reaches the end of the chromosome, and that once it reaches the specific binding site, it cannot move anymore.

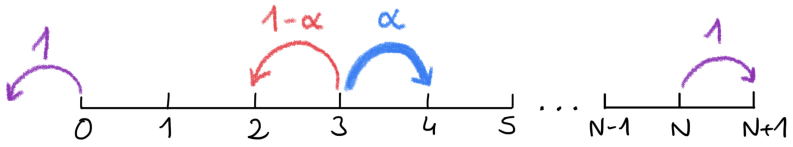

Figure 1: **Schematic of the model.** The black horizontal segment represents DNA, and each vertical tick numbered by i between 1 and $N-1$ represents an identical non-specific binding site. For all of them (for instance site number 3 on the schematic), the probability to jump to the next site on the right is α and the probability to jump to the next site on the left is $1-\alpha$. Site 0 is the end of the chromosome, and a transcription factor arriving here falls off DNA with probability 1. Site N is adjacent to the specific binding site located in $N+1$, and a transcription factor arriving at site N jumps to the specific binding site with probability 1.

d) What is the name of the motion of the transcription factor on DNA? What model seen in the lectures is this problem analogous to?

e) Assume that we start from one transcription factor at site i with $1 \leq i < N$. After a sufficiently long time, what can happen to it? Justify your answer.

f) Consider the first jump of a transcription factor starting at site i with $1 \leq i < N$. Enumerate all the possibilities for this first jump and give their probabilities.

g) Denoting by ρ_i the probability that the transcription factor finally binds to the specific binding site when it starts at site i with $1 \leq i < N$, write an equation relating ρ_i to ρ_{i-1} and ρ_{i+1}. Justify your answer.

h) What are the values of ρ_0 and ρ_N? Explain your answers.

i) Introducing $y_i = \rho_i - \rho_{i-1}$, express y_i as a function of ρ_1 and α.

j) By considering $\sum_{i=1}^{N} y_i$, find the expression of ρ_1 as a function of α and N for $\alpha \neq 1/2$. Same question when $\alpha = 1/2$. Comment on the results: what is the impact of α?
Hint: for $x \neq 1$,

$$\sum_{i=0}^{N-1} x^i = \frac{1-x^N}{1-x}\,. \tag{4}$$