Report Neutral Drift Modelling

Beatriz Brajal Varejão

August 21, 2022

1 Introduction

My project intends to simulate stochastically the neutral genetic drift in a finite population with constant size. For that I used as basis the Moran Model without selection. The simulations were made in Pyhton scripts in the Spyder environment, and were used the packages numpy and matplotlib.

2 Theoretical Basis

2.1 Moran Model

The Moran Model consists on a stochastic model to simulate evolution in a finite population. In my scripts I simulated a case of neutral drift, in which the alleles A and B have the same fitness. In the model we assume:

- a finite population with population size fixed (N)
- discrete generations (in other words, generations can overlap)
- a very low mutation rate, so no new mutants will appear in the simulations
- assexual reproduction

In each iteration one individual in the population will be randomly chosen to reproduce. Then, one individual in the population will be randomly chosen to be replaced by the offspring of the first individual. Assuming the amount of i individuals with the A allele in the population and $N - i$ individuals with the B allele, we can conclude that the probability that an A individual replaces a B, increasing to $i + 1$ the amount of A individuals in the population, will be:

$$P_{i,i+1} = \frac{i}{N} \times \frac{N-i}{N}$$

In the same way, the probability that a B individual replaces an A individual, decreasing to $i - 1$ the amount of A individuals in the population, will be:

$$P_{i,i-1} = \frac{N-i}{N} \times \frac{i}{N}$$

And the probability that a B individual replaces a B individual or an A individual replaces an A individual, without any change in the amount of A individuals, will be:

$$P_{i,i} = \left(\frac{i}{N}\right)^2 + \left(\frac{N-i}{N}\right)^2$$

We can realize that $P_{i,i+1} = P_{i,i-1}$, in other words, in one iteration, the probability that the amount of A individuals increases is the same as the probability that the amount of A individual decreases in the population. It's easy to realize that this system will have only two steady states:

- when the frequency of the A allele in the population is 0
- when the frequency of the A allele in the population is 1

Therefore, it's not possible to have a coexistence of the two alleles, we will always face an elimination of one of the alleles eventually. Intuitively, we can think in the beginning all of the N alleles will have the same probability of fixating, so the probability that the allele A fixates, considering the initial amount of i individuals with the A allele in the population will be:

$$\rho_A = \frac{i}{N}$$

A formal math explanation of this result can be found in the book Evolutionary Dynamics of Martin Novak [Nov06].

3 Stochastic Simulations

3.1 Script

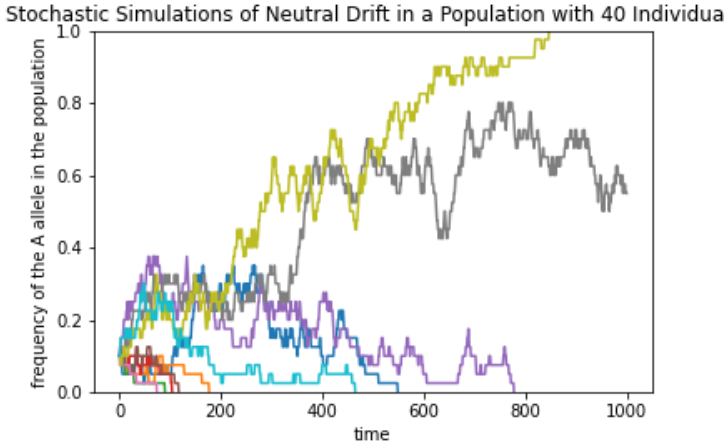
In the first script, named "moranmodel.py", a stochastic simulation of the Moran Modal in a case of a neutral genetic drift is runned in a population of a constant size N with initial amount of i individuals with the A allele. In the script are runned n different simulations with the same initial condition, and each simulation is runned in a maximum time of t_{max} or until the A allele has reached the 0 or 1 frequency, achieving stability. The script generates a graph of the frequency of the A allele in the population $\times$ the time in each simulation.

3.2 Results

In the scripts were defined the following parameters:

- $N = 40$
- $i = 4$
- $t_{max} = 1000$
- $n = 10$

It's really important that the t_{max} is defined as a high number so we can see the fluctuation of the allele frequency in a long period of time. The n should be a number not too big or not too small, so we are able to differ each simulation frequencies and also see a significant number of simulations to realize the different possible paths that the population can go through.
I also defined a low population size so it would be possible to see significant variabilities in the frequencies of the A allele during each simulation, (ideally, it would be possible to see the extinction and the fixation of the A allele in the same graph). Another important point is that, according to theory, we would expect a frequency of fixation of 0.1 (since $i = 4$ and $N = 40$). In other words, in each 10 stochastic simulations run, we would expect that in 1 the A allele ends up fixating in the population in average.
See below one example of a graph obtained by the script:



Since we are dealing with stochastic simulations, it's important to realize that in every time the script will generate a different graph, with different frequencies and different times.

4 Fixation Frequency

4.1 Script

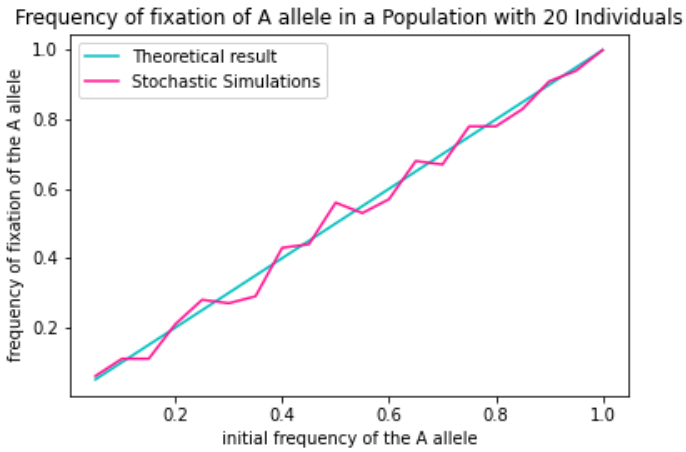
In the second script, named "fixationfrequency.py", a stochastic simulation of the Moran Modal in a case of a neutral genetic drift is runned in a population of a constant size N with initial amount of i individuals with the A allele. However, in this time the algorithm does not care about the frequencies fluctuation in each simulation, but if in the end of a simulation the A allele was fixed or not. Aditionally, the algorithm will be runned in $iter_i$ different initial amounts of A, ranging from 0 to N. For each initial quantity of A, the algorithm will run n different stochastic simulation during , returning the frequency of fixation (the number of times that the allele A reached frequency of 100% under the number of times the algorithm was run). In the end, the script plots a graph with the initial frequencies of the allele A in the population x the frequency of fixation of the allele A in the simulations. The algorithm also plots in the same graph the result given by theory, which is that the fixation rate of an allele A with initial amount of i in a population with constant size N will be $\frac{i}{N}$, in average.

4.2 Results

In the scripts were defined the following parameters:

- $N = 20$
- $iter_i = 20$
- $t_{max} = 1000$
- $n = 100$

It's really important that n and t_{max} are defined as high numbers, so we can have a good approximation of the frequency of fixation of the A allele in each inital condition. A high t_{max} value will be responsible for guaranteeing that each simulation will end in a steady state (frequency of A in the population of 0 or 1) and the n value assures that we are taking a good average of the fixation frequency. The $iter_i$ parameter should be close to N, so the graph plotted can have many different initial conditions. The population size N should also be small, in order to reach the steady state in the end of each iteration, avoiding a miscalculation of the fixation frequency.
See below one example of a graph obtained by the script:



We can realize that the algorithm provided a good approximation of the expected fixation frequency according to theory.

References

[Nov06] Martin A. Novak. *Evolutionary Dynamics: Exploring the Equations of Life*. 2006.