GPOP PROJECT REPORT

Khanh Nam NGUYEN

December 4, 2024

1 Introduction

GPOP: simulating an evolving population. This project is built in Python programming language with GUI using NICEGUI package. The source code is uploaded at github.

2 Tasks

2.1 Task 1: Genetic drift

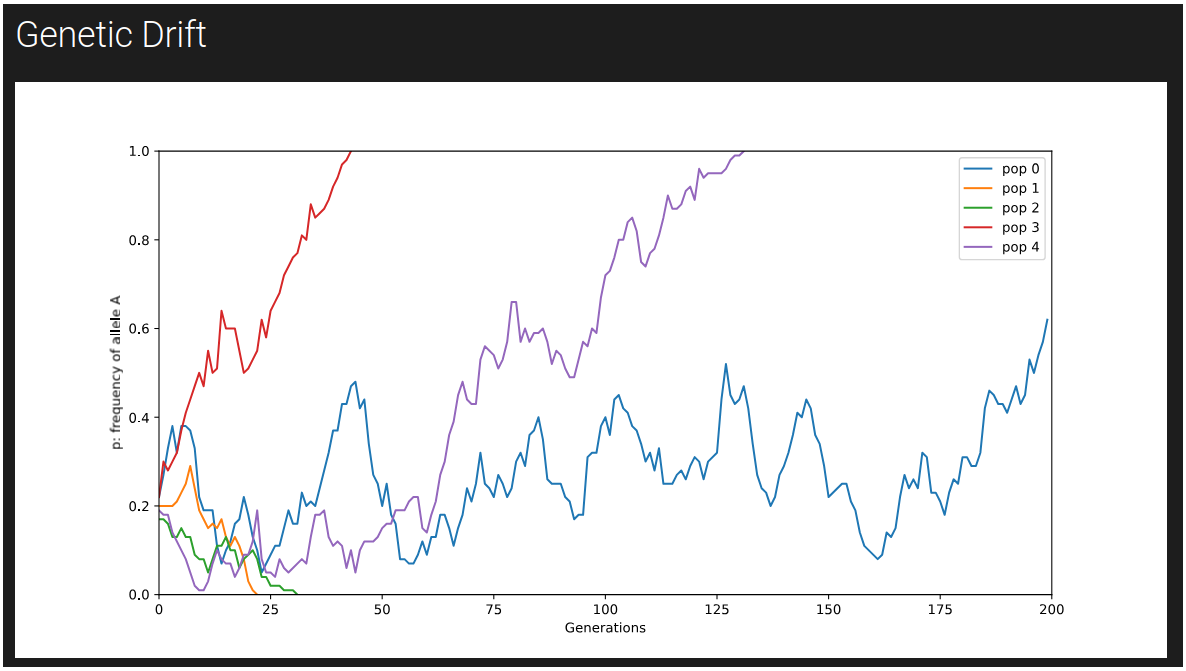

Figure 1: Genetic drift

The fixation probability of allele A with pA = 0.2 over 1000 simulations: 0.14. For pA = 0.2, fixation probability of allele A is in range(0.12, 0.17). That means lowest fixation probability after several times of simulating is 0.12 and the highest is 0.17.

Harmonic mean of actual population size:

$$Ne = (\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{Ni})^{-}1$$

Theoretically, the expected fixation time (number of generations) should follow harmonic mean of actual population size. In this case, with N = 100 and generations = 200 $\rightarrow Ne = 100$. But according to the simulation, the fixation time = 84.36 which is $\approx$ 100.

Genetic drift occurs in all populations of non-infinite size, but its effects are strongest in small populations.

2.2 Task 2: Coalescent model

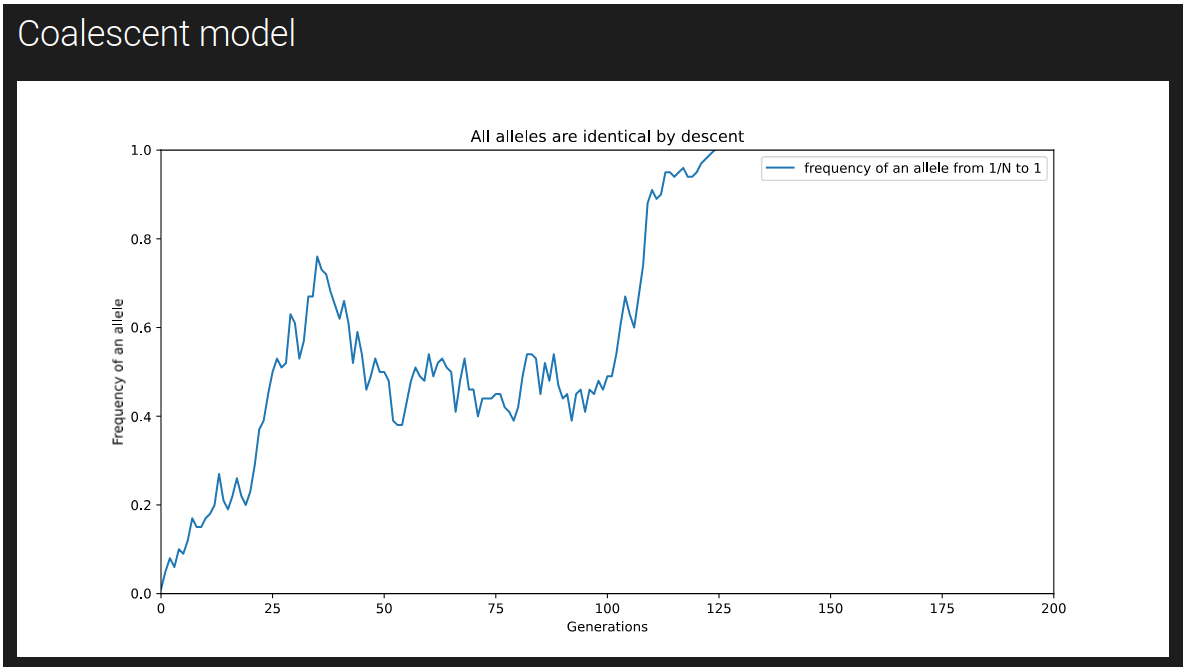

Figure 2: All alleles are identical by descent

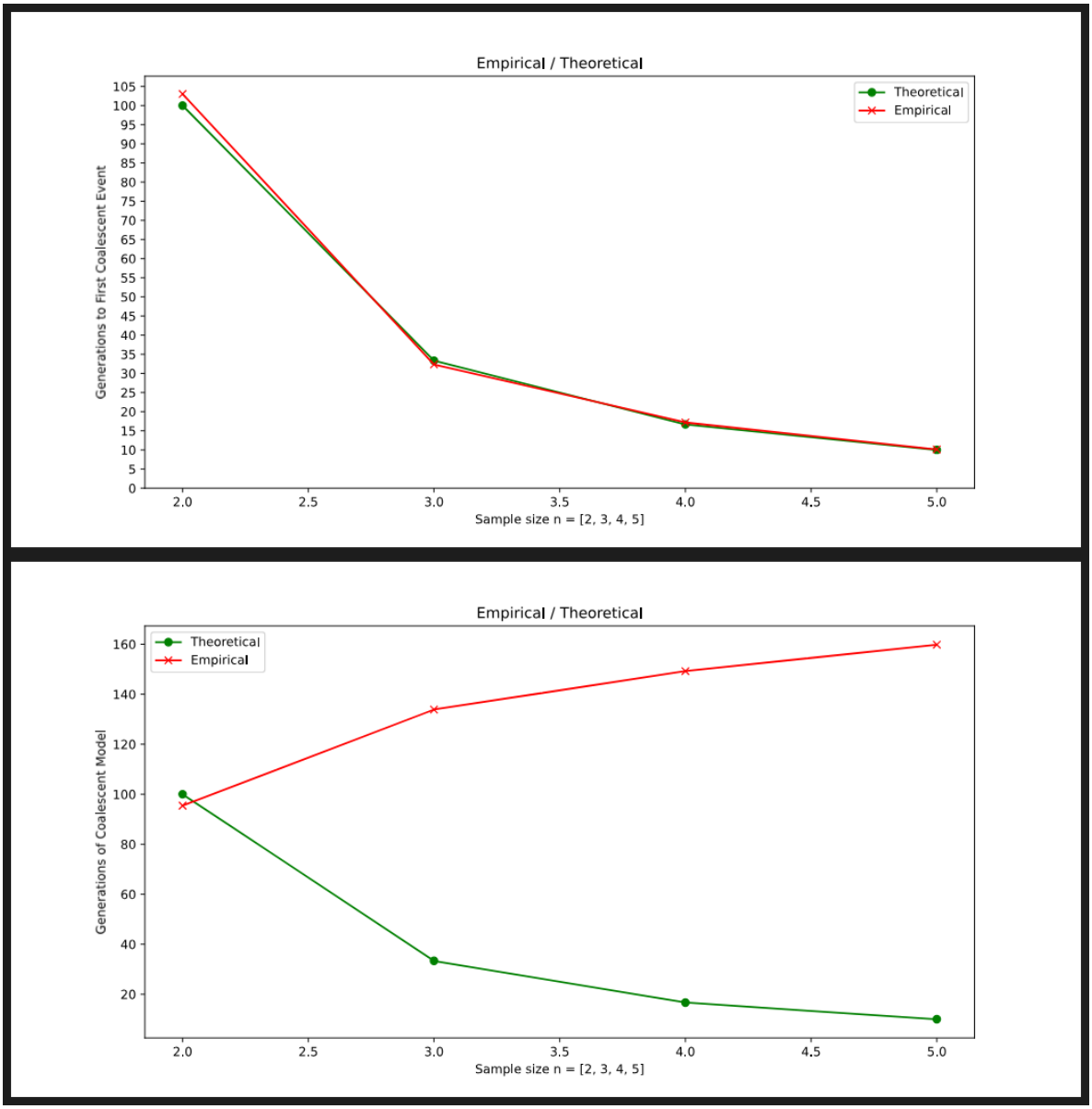

Figure 3: Coalescent model

Tracking a population of N = 100 individuals until all alleles are identical by descent is like tracking genetic drift. Because if all alleles are identical by descent that means the population has only 1 allele left to be parent in the next generation. To do the simulation, trace the highest frequency allele after sampling the population.

Expected generation to the first coalescent event:

$E\{T_n\}$	$\frac{4N}{n(n-1)}$
$E\{T_2\}$	100
$E\{T_3\}$	33.34
$E\{T_4\}$	16.67
$E\{T_5\}$	10

Table 1: Expected generations.

To the first coalescent event, number of generations in empirical experiment and theory are the same. But in case of running the model until the number of lineages = 1, number of generations in

empirical experiment is far away from theory. Because the assumption n $\ll$ N is violated.

2.3 Task 3: Mutations in the infinite-allele model

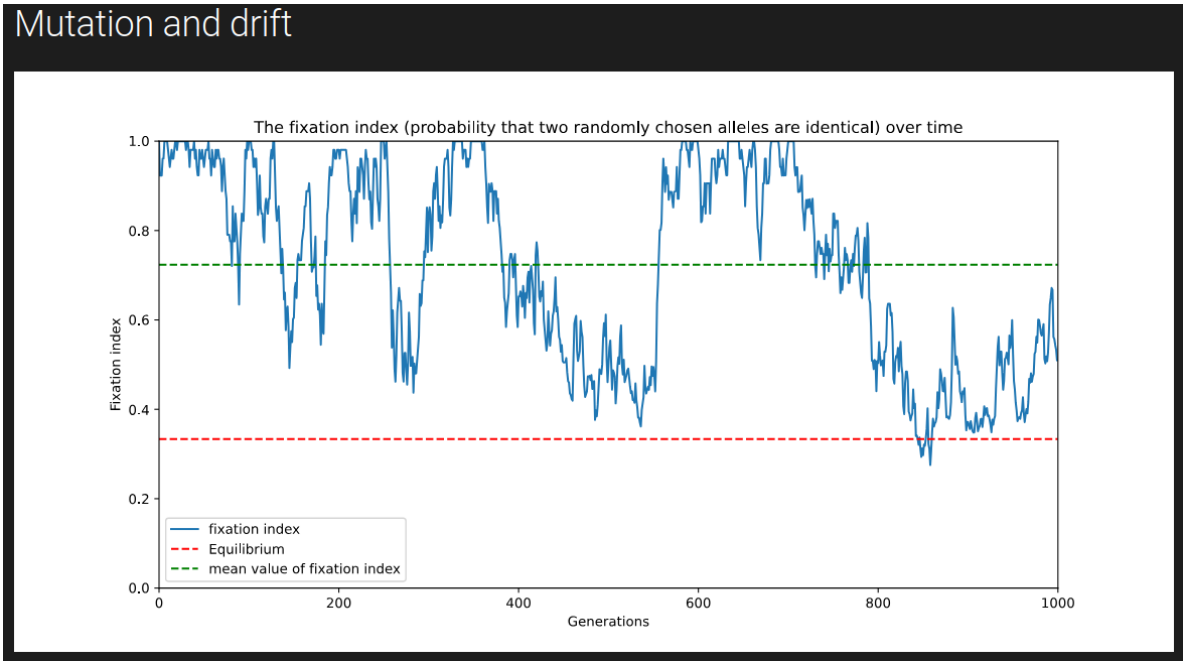

Figure 4: Mutation and drift

The fixation index: the value of $\mathcal{G}$ after one round of random mating and mutation:

$$\mathcal{G}' = (1-\mu)^2[\frac{1}{2N} + (1-\frac{1}{2N})\mathcal{G}]$$

With G and $\mathcal{G}$ are "almost the same" + the total frequency of homozygotes is given by:

$$G = \sum_{i=1}^{k} p_i^2$$

At equilibrium, the probability that 2 alleles different by origin are identical by state is given by:

$$\hat{\mathcal{G}} = \frac{1}{1+2N\mu}$$

For this example, the equilibrium = 0.33 which is smaller the mean value of fixation index = 0.72. Genetic drift increases fixation index $\mathcal{G}$ while mutation decreases $\mathcal{G}$. Run the experiment for several times until 1000 generations. Most of the plots have peaks that reach the equilibrium and then return, with a few times that can surpass the equilibrium. Anyway, equilibrium $\neq$ static $\rightarrow$ new mutations/alleles appear, drift happens.

2.4 Task 4: Selection

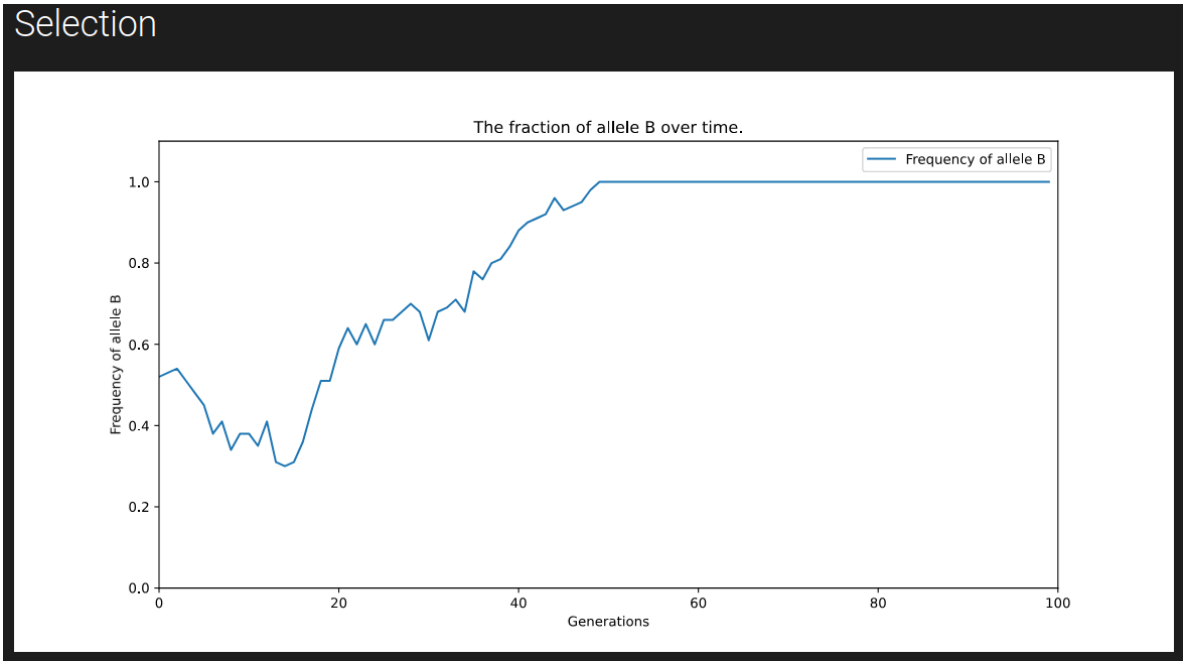

Figure 5: Enter Caption

Because allele B has higher fitness than allele A in population, so after generations, allele B will be fixed. That also means allele A will be removed out of population. Fitness-increasing alleles become more common in the population,

2.5 Task 5: Clonal interference

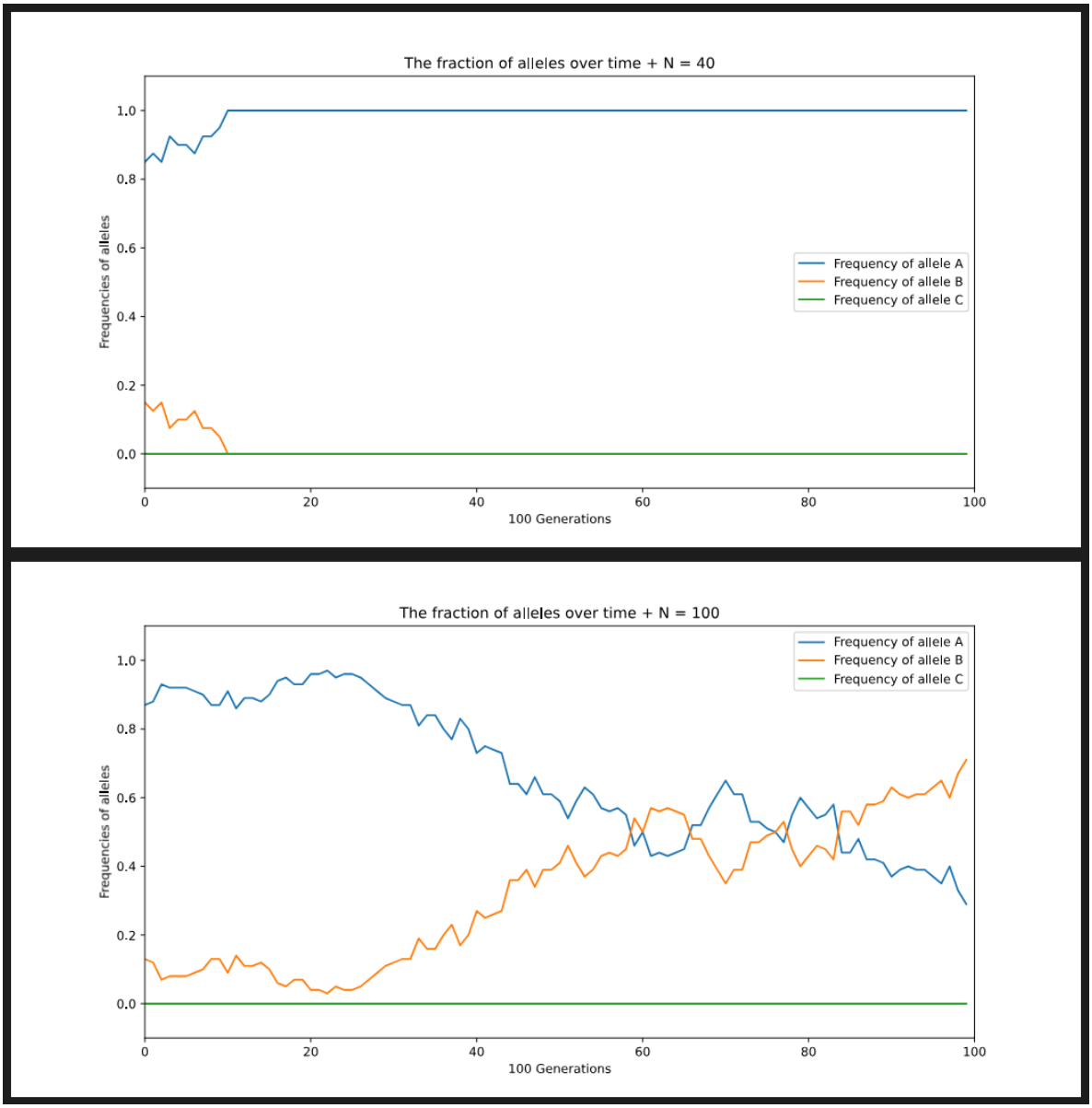

Figure 6: Clonal interference with 100 individuals

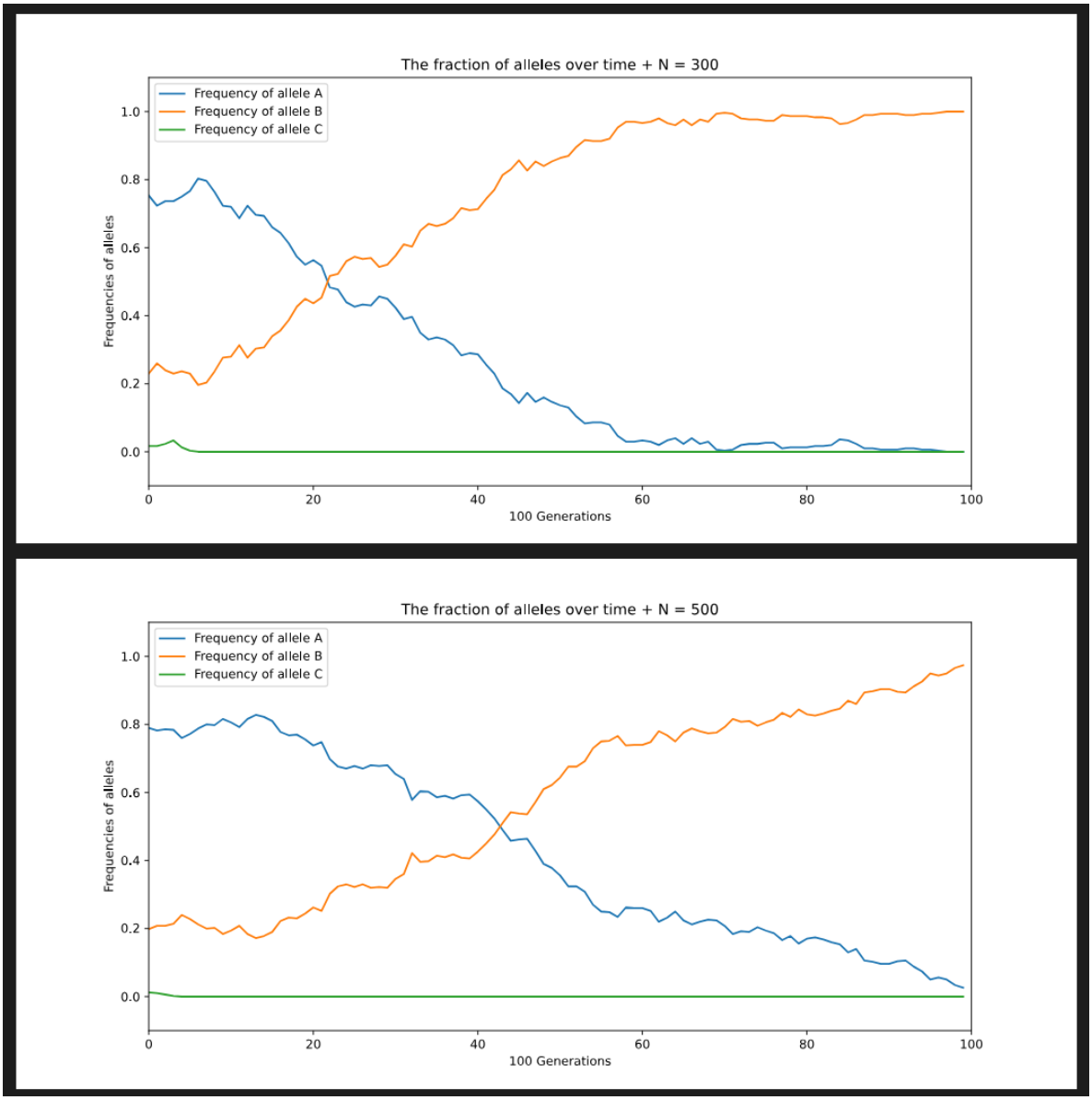

Figure 7: Clonal interference with 500 individuals

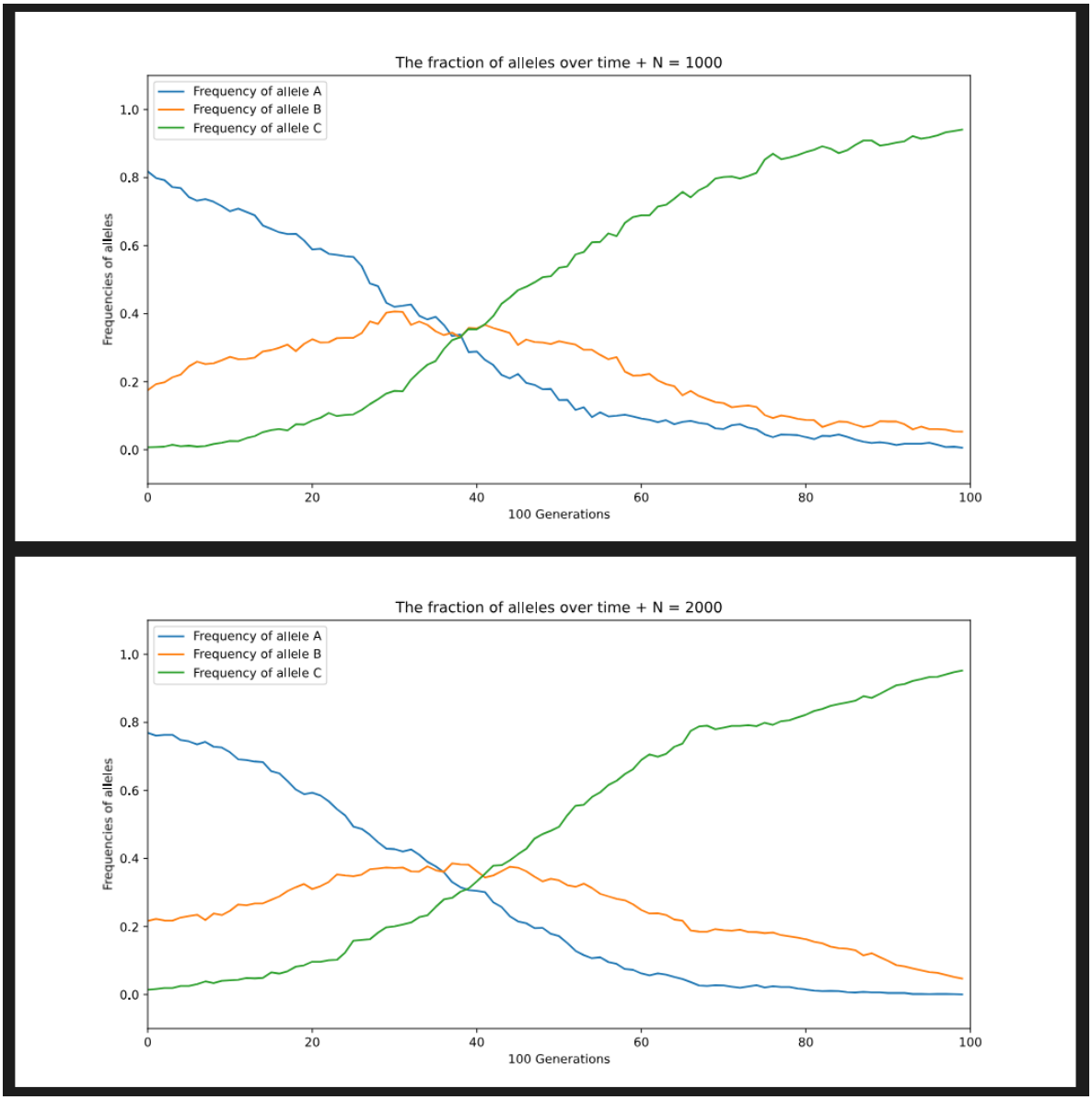

Figure 8: Clonal interference with 1000 individuals

Allele	Initial frequency	Fitness
A	pA = 0.79	1
B	pB = 0.20	1.05
C	pC = 0.01	1.1

Table 2: A population with a three-allelic locus

In figure 6, the number of alleles for these simulations is ≤ 100. Frequency of allele C always goes down until 0. Allele A and allele B share the chance to be fixed in the population, but allele A only gets fixed with small N (usually appears in N = 40).

In figure 7, the with number of alleles is increased up to ≤ 500. I always observer that allele B is fixed in population, only a few times that allele C could be fixed with N = 500. At this step, allele A has no chance to be fixed anymore. The chance of being fixed in population is taken by allele B (mostly) or allele C.

In figure 8, now the number of alleles in population is thousands. For now, allele C always has the power to be fixed. It wins all the time that I run the simulation. Imagine this is a long race track

where 3 horses A, B and C participate with its horsepower (represent for *Fitness*). The result of this race always has the same scenario: only one winner is allele C, no second prize, allele A and B always lose but the important thing is allele A always loses to allele B based on the fitness of the 2 alleles. The result of this simulation can be predicted based on fitness, of course with big enough N (in thousands).

According to the lecture:

$$\frac{q(t)}{p(t)} = \frac{\lambda_B}{\lambda_a}.\frac{q(0)}{(p(0)}, assumethat : \lambda_B > \lambda_A$$

That means frequency of allele which has higher fitness (λ) grows exponentially when compared to frequency of small fitness allele. This can explain why allele C in the simulation always is fixed with big enough N.

But why does allele C need a number of N which is big enough? This phenomenon slows down the fixation of individual beneficial alleles because one allele must compete others to dominate the population. This typically leads to the loss of diversity in population (winner takes all, no second prize).

Also in large population size, it generates more beneficial alleles leads to a higher overall number of allele in each generation. Beneficial alleles often compete more intensely because many lineages with similar fitness advantages coexist.

In smaller populations size, the genetic drift dominates over the selection causing beneficial alleles to be lost before they can be fixed because the populations has fewer beneficial alleles.

2.6 Task 6: Population Structure

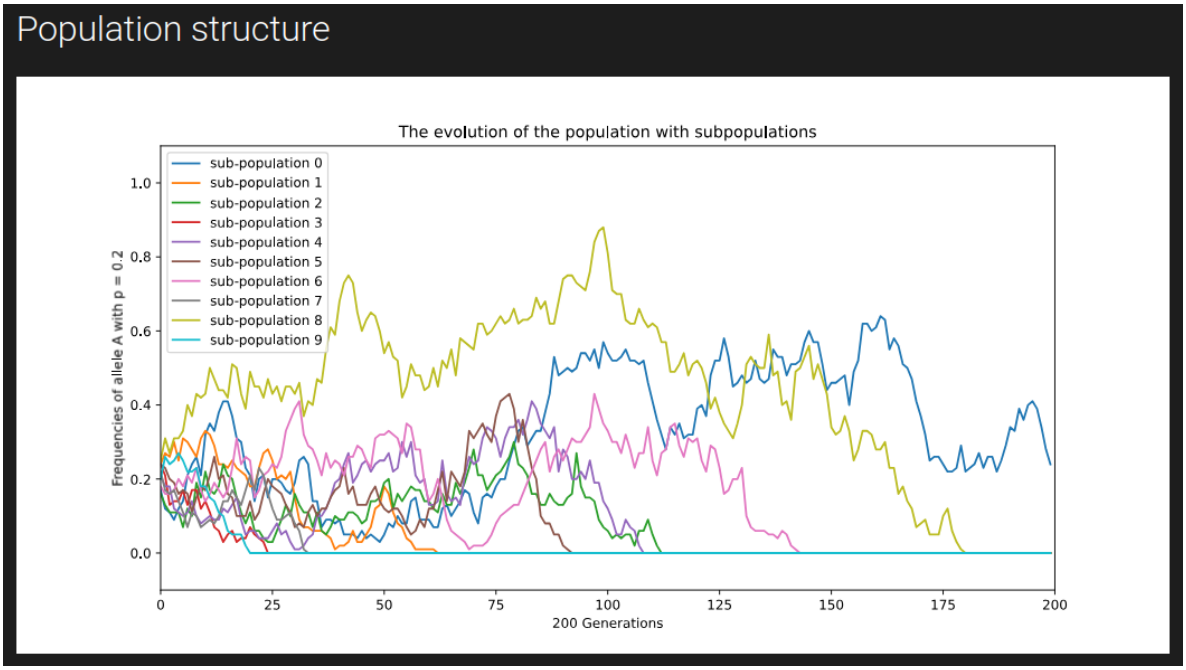

Figure 9: Population structure

When dividing the population into equal sub-populations, the ratio between allele A and allele B in sub-populations is not the same compared to Task 1. For some cases, when dividing the population into sub-populations, some sub-populations are fixed already cause it contains only 1 allele, not 2.

Within each sub-population, the genetic drift stills happen inside of each sub-population. Each sub-population evolves independently can retain global diversity longer. The fixation occurs faster than in Task 1 cause of smaller effective population sizes.

2.7 Task 7: Migration

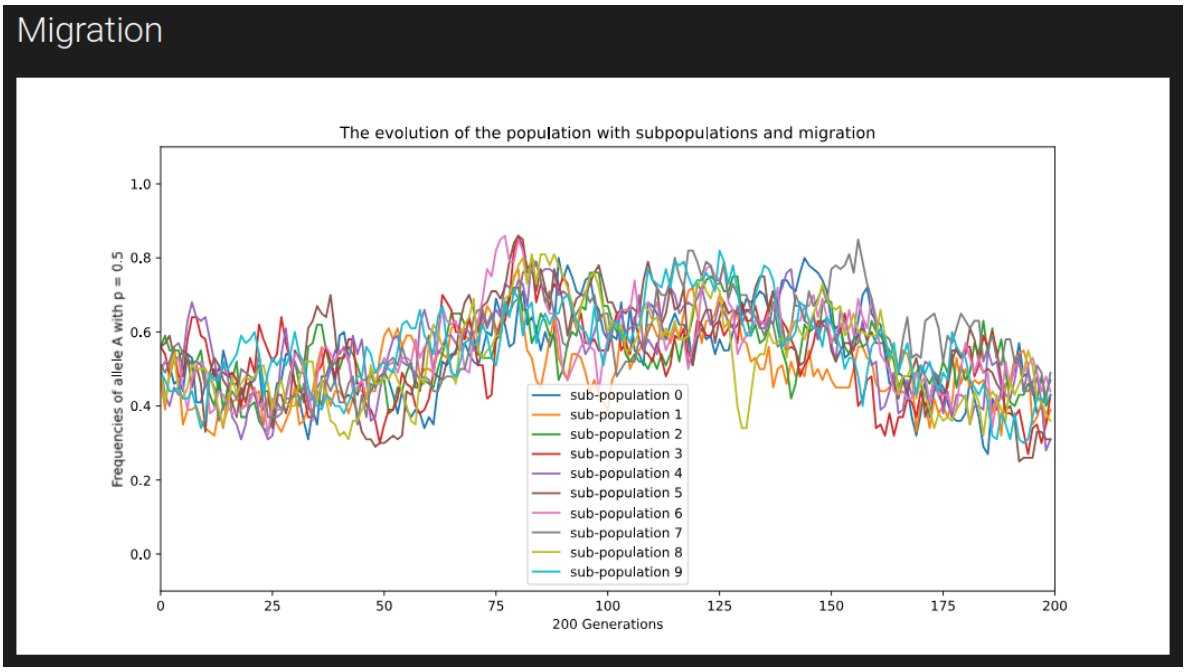

Figure 10: Migration

There is no fixation in this simulation (at least until 200^{th} generation). The migration changes the distribution of genetic diversity among populations, by modifying allele frequencies.

In this case, for small migration rate m $\ll 1$ and according to the equilibrium:

$$F_{eq} = \frac{1}{1 + 2N_e m}$$

which means $F_{eq} \ll 1 \rightarrow$ little differentiation between sub-populations, little fixation within sub-population.