Adrián Barreno Sánchez

Assignment 1: Study of the fixation of neutral genes

INTRODUCTION

The neutral theory proposed by Japanese geneticist Motoo Kimura proposes that most genetic substitutions and polymorphisms in DNA and protein sequences within a species are neutral and do not affect the organism's fitness. These neutral mutations are spread and stablished over the population not by Darwinian selective pressure, but by random genetic drift [1, 2].

Genetic drift is a process in which allele frequencies within a population vary only due to demographic stochasticity, that is, because of random sampling from generation to generation. Random drift often occurs after severe reductions in population size, named "bottlenecks", and in founder events, where a population starts to grow from a reduced number of individuals. Genetic drift reduces the diversity within the gene pool of a population and can lead to increased homozygosity [3].

In this work, I will study how two population parameters -*population size* and *bottleneck size*- influence the time to fixation (t_{fix}) and the fixation probability (p_{fix}) of a neutral mutation in a population.

METHODS

In order to study the role of population size and bottleneck size on the fixation of neutral mutations, I created an R function named *mutation_fix* that simulates the evolution of a population of *N* genomes harbouring a randomly inserted neutral mutation in one single individual.

This function defines a 1 x *N* zero matrix and inserts a single mutation (represented as a number *1*) with a uniformly distributed probability over the different individuals of the population. Then, it runs an iterative process that simulates the replication of population over the course of 10,000 days (considered consecutive generations). In each iteration, population is randomly sampled emulating a genetic bottleneck that reduces its size by a constant factor (*bottleneck size*). The probability of being sampled is uniformly distributed over the population and sampled individuals are replaced, so that each genome can appear more than once within the offspring. The resulting offspring genomes are sequentially duplicated until the original population *N* size is restored. If the inserted mutation is fixed (homogenously stablished in the population) or lost, the replication loop stops. This iterative simulation of the evolutionary process is entirely repeated 100,000 times to overcome the low fixation probability for neutral mutations. As a result, the function returns the total number of fixations reached after the stablished number of replicates and the average days required to achieve fixation.

Aiming to identify differences in fixation probabilities and times to fixation, I run this function under 12 varying conditions combining population sizes (*N*=100, *N*=300, *N*=1,000 and *N*=3,000) and bottleneck sizes (*bot*=0.25, *bot*=0.5 and *bot*=1). I stored the output values of interest, total fixations and average time to fixation, in two 3 x 4 matrices, and calculated the probability of fixation for each condition (total fixations / 100,000 replicates). Further, I merged and stacked all data in one single Data Frame object to facilitate storage, analysis and visualization.

Next, I built regression models to identify potential correlations between time to fixation and fixation probability with population size and bottleneck size. To estimate bottleneck size – fixation probability and population size – fixation time relationships linear models were trained, while base 10 logarithmic models were employed to analyse population size – $\log_{10}$(fixation probability) and bottleneck size – $\log_{10}$(fixation time) correlations. R^2 scores and Pearson correlation coefficients were calculated to evaluate each model.

Finally, data was represented in boxplots and scatter plots, together with the regression lines. Data Frame was exported as a *.csv* file and edited and transformed into *.png* format using Python *dataframe_image* library.

	Population.size	Bottleneck.size	Total.Fixations	Average.fixation.time
0	100	0.25	1004	48.09
1	100	0.50	1013	97.07
2	100	1.00	1054	194.28
3	300	0.25	342	145.27
4	300	0.50	329	293.43
5	300	1.00	299	576.28
6	1000	0.25	111	513.86
7	1000	0.50	99	914.12
8	1000	1.00	91	2023.64
9	3000	0.25	24	1754.83
10	3000	0.50	36	2817.94
11	3000	1.00	31	5319.74

Table 1. *Data Frame containing data from the different simulations. Data is coloured following a gradient pattern that reflects their value within each variable.*

RESULTS

Total number of fixations and average time to fixation 100,000 after repeated simulations of the replicative process are represented in Figure 1. Data reveal a strong inverse relationship between the population size and the total number of fixations, which ranges from $\bar{x}$ = 1,024 for *N* = 100 to $\bar{x}$ = 31 for *N* = 3000. Further, within each

population size category, absolute fixation values reflect low dispersion levels -calculated as standard deviation (not shown)-, suggesting that fixation probability might be independent of bottleneck size (**Table 1**).

Regarding the average time to achieve fixation in our simulation, we can observe a direct relationship between population size and average time to fixation (ranging from $\bar{x}$ = 113.15 days for N = 100 to $\bar{x}$ = 3298 days for N = 3000). In contrast to fixation probability, the values of fixation time are rather disperse within each population size category, and we can observe an *a priori* direct relationship between bottleneck size and time to fixation.

Aiming to evaluate the potential correlations within our data, I trained different linear and logarithmic (log_{10}) regression models and evaluate those that adapted better to our variables. For training the models, I previously converted total number of fixations into fixation probability (total fixations/ 100,000 replicates).

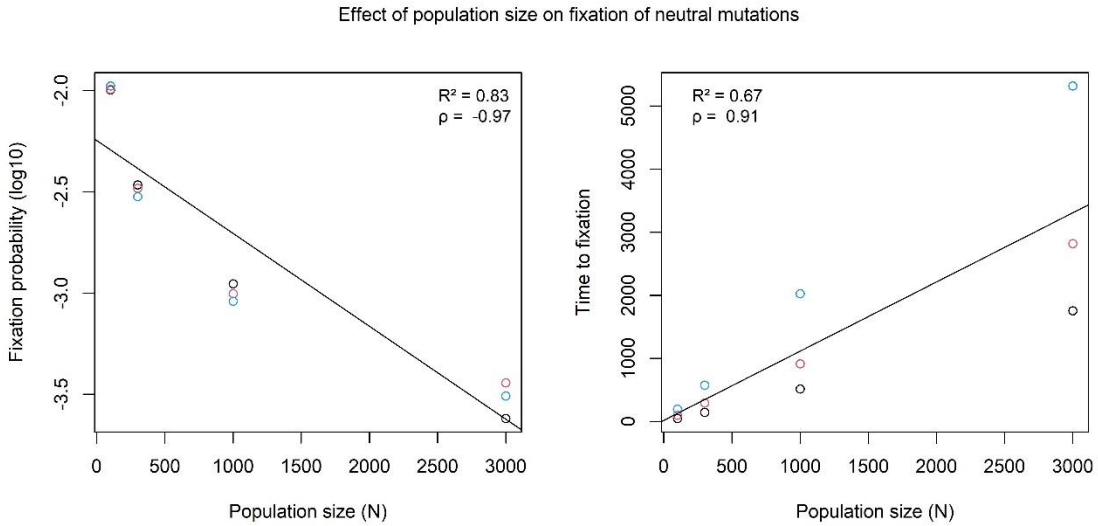

Figure 1. *Scatter plots and correlation lines representing correlation between population size and log_{10} fixation probability, and between population size and time to fixation. Colours represent bottleneck size for each observation (black = 0.25, red = 0.5, blue = 1).*

Our data reflect a strong negative logarithmic correlation between population size and fixation probability (R^2 = 0.84 and ρ = -0.97), and an equally significant positive linear correlation between population size and time to fixation (R^2 = 0.7 and ρ = 0.91) (**Fig. 1**). In contrast, regarding the effect of a genetic bottleneck on the fixation of neutral mutations, no correlation was observed between bottleneck size and fixation probability (R^2 = 0 and ρ = -0.03) (**Fig. 2**). Nevertheless, we can observe a slight logarithmic correlation between the size of the bottleneck and the time of fixation despite its low significance (R^2 = 0.14 and ρ = 0.41). The high dispersion in fixation time values could be the reason of the low R^2 value obtained.

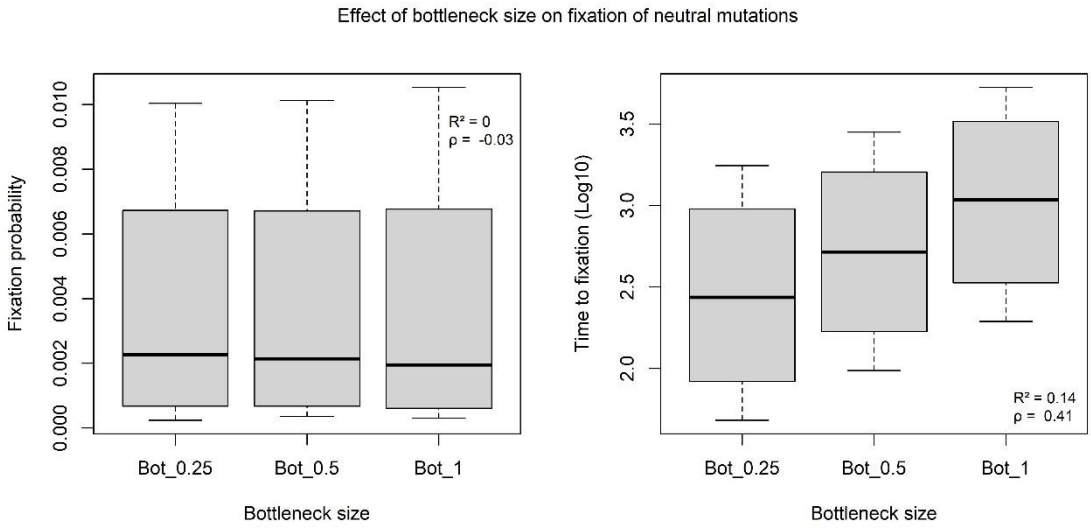

Figure 2. *Boxplots representing the effect of bottleneck size on fixation probability and log_{10} time to fixation.*

DISCUSSION

In this study, I simulated the evolution of a population to analyse the effect of two population parameters -population size and bottleneck size- on the fixation probability of neutral mutations and the time required for fixation to occur. On view of the results, we can conclude that population size strongly influences both the fixation rate of neutral mutations and the average time, whereas the outcome of modulating the genetic bottleneck size is more complex. Interestingly, while bottleneck size does not seem to influence fixation rate, it correlates with the time to fixation of neutral mutations, even though data dispersion might bias the regression model.

This fact might appear counterintuitive; considering that small population size seems to favour genetic drift and fixation of neutral mutations, we could have speculated that small bottleneck sizes, which cause a reduction in the offspring, might increase the probability of a neutral mutation being fixed. However, this is not what we observed. The reason for this might be that small bottlenecks lead to reduced genetic diversity in the offspring, decreasing the fixation probability of genes with low frequency within the population.

As I mentioned before, there might be a potential bias in the regression models created by the high dispersion in the values of fixation time. We could address this issue by considering in these models the external variables responsible, at least partially, for the dispersion. For instance, in the right plot in **Figure 1**, we could build 3 independent models to describe the correlation between population size and time to fixation considering the bottleneck size as well. In that case, we would also need to replicate the whole experiment several times to achieve statistical significance. A similar approach could be followed for bottleneck size – fixation time model shown in **Figure 2**.

The experimental setup involves several assumptions and limitations that need to be taken into consideration. First, we assume that population remains constant during the whole evolutionary process, something that does not reflect the behaviour of real populations. Instead, we could consider that populations grow or reduce in size over consecutive generations, what could alter the effect of genetic drift in a dynamic manner. Moreover, we are considering the whole population model as a discrete-time system, where generations are represented as non-overlapping discrete steps. This is not realistic for most organisms; thus, we could perhaps consider a more complex dynamic model that incorporates births and deaths to our system in a continuous manner. Last, we are analysing the spread and fixation of an isolated neutral mutation in a single-gene genome population. It would be interesting to include more genes and neutral mutations within our population and study how the fixation probability and time of each mutation relate with each other.

REFERENCES

1. https://en.wikipedia.org/wiki/Neutral_theory_of_molecular_evolution
2. https://en.wikipedia.org/wiki/Motoo_Kimura
3. https://www.apsnet.org/edcenter/disimpactmngmnt/topc/PopGenetics/Pages/GeneticDrift.aspx

LINK TO SCRIPT

https://drive.google.com/drive/folders/1aK57NXDjb64cu022_gwqqeU8ZIHvOpUe?usp=share_link