EXERCISE #|

Computational Tools in Evolutionary Biology

Master Degree in Computational Biology 2021-2022
Sandra Alonso Paz

Introduction

What is Neutral Theory?

The neutral theory of molecular evolution was first proposed by Motoo Kimura in 1968, and independently by Jack King and Thomas Jukes in 1969. At the time, studies on genetic sequences were showing that the previous idea which postulated that most of the differences between species were caused by selection on advantageous mutations was actually not true.

The neutral theory instead proposed that the majority of molecular changes, such as in DNA sequence, are caused by random processes acting on selectively neutral mutants, meaning they inferred no advantage or disadvantage.

But, how can these mutations be selected?

By using complex calculations, Kimura showed that the rate of evolution cannot be explained by positive or negative selection because it is too high and that many mutations must instead be neutral. Neutral mutations become widespread by a process called random genetic drift, in which a mutation spreads throughout the population due to chance alone.

Research questions

As it has been mentioned, drift is responsible for some for the changes during the evolution process of a population. And, as we have learned in class, any finite population will always have an extinction and fixation time. With this in mind let's set the bases for the work I will develop:

Time taken by an individual to get fixed (**TFix**) is approximately two times its population size (2N). Consequently, the bigger is the population, the more difficult it is for an individual to get fixed or extinct. Moreover, the probability to get fixed (PFix) is approximately one over the population size ($\frac{1}{N}$). Therefore, as larger is the population, as less probability will have an individual to get fixed.

Thanks to these arguments, we can assume that drift scales with population size.

On the other hand, it has been studied the possible relationships between drift and bottleneck. It is well known that big values of bottleneck, make small changes on the population (most of the individuals will cross it), in consequence, it is more difficult for an individual to get fixed or extinct. Otherwise, small bottlenecks produce several changes in the system. As not every individual will be able to reproduce (could not cross the bottleneck) will become extinct, besides the ones that crossed, which are more likely to get fixed.

In the light of these statements, I will try to solve the following related research questions in this work:

- Are TFix and PFix truly related with bottleneck and population size?
- Which type of relationship are between them?

Estimated values

In order to get the answer to the proposed questions, I have developed a R-notebook (which can be found in this folder). The results of the execution are 4 linear charts and a table with the numeric results:

Population size	Bottleneck	TFix	PFix
100	0.25	49.52637	0.01005
100	0.5	98.46768	0.01052
100	1.0	199.06687	0.01002
300	0.25	162.67059	0.00340
300	0.5	309.45338	0.00311
300	1.0	590.94671	0.00319
1000	0.25	526.68367	0.00098
1000	0.5	953.90291	0.00103
1000	1.0	1959.53000	0.00100
3000	0.25	1494.08571	0.00035
3000	0.5	3234.51351	0.00037
3000	1.0	4768.57143	0.00028

Numeric results table

Green dots represent numeric result of each scenario (population size and bottleneck size). Red line shows the linear smooth which is a function that attempts to fit principal points.

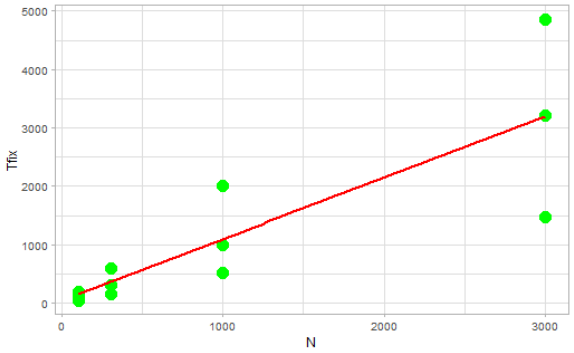

Figure 1

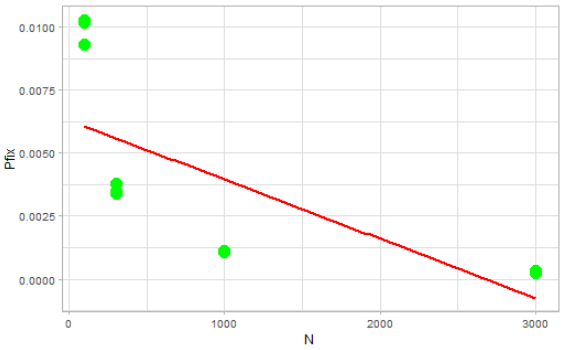

Figure 2

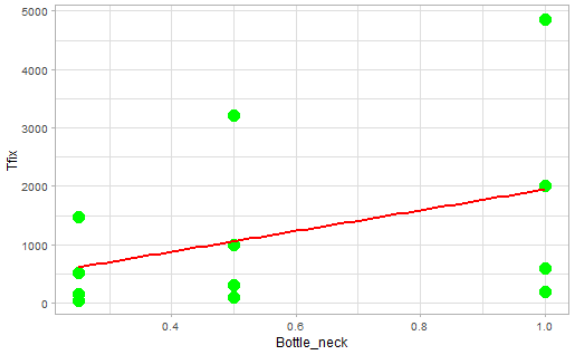

Figure 3

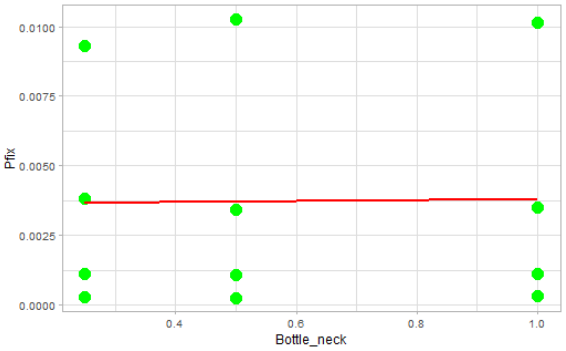

Figure 4

Discussion

Once we have obtained these results it is time to discuss them.

Population size

Firstly, I will analyze the linear chart graphs:

The first one shows the relationship between Tfix and population size. The line represented shows that as larger populations, the higher Tfix, so the longer it takes to get fixed. In addition, it can be seen how the line has a slope close to 45º from which we can deduce that these two variables have the same growth rate.

The second graph shows the relationship between population size and Pfix. In this case we can appreciate a huge difference with the first graph. If we interpret the line, we will see that with a higher number of individuals in the population, less possibilities will have to get fixed. This is why the principal line has a negative slope.

Now we have done a general study with the charts, let's focus on the numeric results. For instance, the first case: N = 100 and 3000:

Population size	Bottleneck	TFix	PFix
100	0.25	49.52637	0.01005
100	0.5	98.46768	0.01052
100	1.0	199.06687	0.01002
3000	0.25	1494.08571	0.00035
3000	0.5	3234.51351	0.00037
3000	1.0	4768.57143	0.00028

In the light of these results, it seems obvious that Tfix increases at the same time it does the population size. Moreover, if we keep the eye on it, we will discover that their growth rate is proportional ($\frac{100}{3000} = 0.0333$ almost the same to $\frac{49.52}{1494.08} = 0.0331$). In addition, we can see that population is proportional to Tfix and follows the mathematical rule: $Tfix = 2 * population\ size * bottleneck\ size$.

On the other hand, Pfix decreases as population increases. This is because as many individuals have a population, they have less possibilities to get fixed. In addition, this relationship also follows a mathematical rule: $Pfix = \frac{1}{population\ size}$. ($\frac{1}{100} = 0.010$, almost the same to its Pfix which is 0.01005).

This study allows us to assure that TFix and Pfix are very related to population size due its proportionality.

Bottleneck size

Firstly, I will analyze the linear chart graphs:

The tirth graph shows the relationship between Bottleneck and Tfix. As in the first graph mentioned above, the line shows that as higher bottleneck value, the higher it will take an individual to get fixed.

Finally, the last graph represents the relationship between bottleneck and Pfix. This plot is the one with less variance on its dots. This does not mean that axes variables are not related, but although the bottleneck has changed, Pfix only seems to low a bit its value.

Now we have done a general study, let's focus on one case. For instance, the first case: N = 100:

Population size	Bottleneck	TFix	PFix
100	0.25	49.52637	0.01005
100	0.5	98.46768	0.01052
100	1.0	199.06687	0.01002

This table shows the behavior of the model as we increase the bottleneck percentage. Focusing on the fixation time (TFix) it is appreciable an increase in its value as Bottleneck also increases. It seems to double its value as the bottleneck does. Therefore, we can assume that they are not just related, but also, they are proportional.

On the other hand, PFix seems to have different behavior. As we knew, as bigger values of bottleneck, as difficult it gets for individuals to get fixed. At first sight, this rule is followed by very distant values like 25%-100%. However, we can see that the highest value of PFix is for the 50% bottleneck value. This behavior is repeated throughout all the scenarios.

This study allows us to assure that TFix is very related to Bottleneck due its proportionality. On the same hand, Pfix also seems to be related to the bottleneck size although it does not vary a lot.

Conclusions

Once we have done the discussion, I will conclude by answering the research questions.

Yes, population size and bottleneck size are highly related to Pfix and Tfix. And their relationship follows the mathematical rule proposed on the research question:

- $Pfix = \frac{1}{N}$
- $Tfix = 2 * N$

However, I would like to highlight that it is mandatory to consider bottlenecks in the Tfix formula to obtain more accurate results.

Limitations of the computer work code

Large runtime

As code is quite complex, it takes too long to run (for N = 3000 takes almost 20 minutes). This might be a problem if we try to test it with larger populations and higher bottlenecks.

Not 100% realistic model

As it has been postulated, population size and bottleneck scales with TFix and PFix but in real life other parameters like mutations should be taken into account. Moreover, the model considers an immortal population but individuals can die during the experiment and this should also be kept in mind.

In addition, the reproduction which has been taken into account to create this model is the asexual one. Mainly because the complex implementation of the sexual reproduction. Thinking about it, this could be implemented by using tree structures and recursive functions but still seems to be a difficult task.

Possible extensions

Improve included plots.

I would have liked to implement a plot to show the evolution of fixations and extinctions (as the one seen in class). In this way, it would be much more visual and therefore easier to see how, as bottleneck and population size vary, the number of individuals which reach the fix changes.

Run the model with real parameters.

In my opinion, we cannot verify the accuracy of the model until it is tested in a real scenario or compared with real data. One possible extension is to search data and use it to run the model. Then we will be able to analyze its behavior and surely after this test, some new questions will be proposed and therefore this will allow us to continue with the investigation.

Implement some of the limitations of the computer work

Since I thought about how a sexual model could be implemented, a possible extension would be to develop this model and create a visual interface which allows the user to change from one model to another. From this view users will be able to set parameters and see the evolution of the model.

Bibliography:

Moodle

Neutral Theory and their authors

R documentation