Bioinformatics2022

Doi: 10.93 bioinformatics/xxxx

Advanced Access Publication Date: 21March 2022

Manuscript Category

Software

SIMULATION-Gene: An application to simulate genomes in populations

Ayushi Pathak [1]

[1]Department of Biology, Bioinformatics, Lund University, Biology Building, Solvegatan 35, 223 62

Contact: Ayushi Pathak- ayushipathakofficial@gmail.com, ay4188pa-s@student.lu.se

Received on: 21-03-2022; Revised on: 20-03-2022; Accepted on: xx-xx-xxxx

©2022 Ayushi-Pathak; licensee Info Dept Biology Ltd.

This is an Open Access article distributed under the xxxxxxxxxxxx Licence (http:// ---------) which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Simulation of genomes in population biology can be useful in various fields of biology. Simulations are done in two ways forward and backwards, each having advantages and disadvantages over each other. The tools provide a way to develop data required for research under various evolutionary models. Simulation-Gene is one suck tool that will give you data with various input options that can change in real-time.

Motivation: All the research done so far gives us tools that can be used to create data, but are less interactive. The website of simulation-gene is created using R shiny that is interactive and quick.

Results: The website integrated various models for genomic simulations. The models give a graphical depiction of the model according to the input and data associated with it.

Keywords: Population Genomics, forward and backward simulations, R programming and R shiny

Introduction

Genomic simulations are used to create various data frames for applications in statistical models and various methods in population and genetic studies.[1] This process provides us with the ease of generating bulk data that might be difficult to obtain in normal circumstances.

Genomic Simulations are of two types: forward and backward simulations. Forward simulations are based on the concept that the whole population is simulated from past to present.[2] Hence, it is computationally less efficient. Backward simulations are also known as coalescent-based simulations and are computationally more efficient. These are based on the history of generations with surviving progeny in the present population. It does not take into consideration all the past and ancestral individuals of the populations. [3] However there are some limitations with the backward simulations including a lack of track of ancestral information that might be necessary if the research is focused on the evolutionary process. Coalescent simulations originate from simple genetic concepts such as selection but ignore the evolutionary basis in its base. [4] Also, these are not yet capable of simulating complex human diseases in general.[5]

Implementation

Markov models of DNA mutation

In phylogeny, Markov Models describe the rate of exchange of amino acids or nucleotides. It is a measure of substitutions to know the evolutionary distance between sequences by creating a probabilistic framework of all possible mutation sequences.

Hardy Weinberg Model

Migrations models

The simulator gives an option for five different migration models: island mainland model, island model, one-dimension stepping stone model, general model and model for circular migration. The web application is made to be interactive that changes the output as the input is changed. Each model has different parameters that generate a graph and dataset which can be used for various purposes.

The island mainland model is the most simple version of the population model, which consists of one large mainland population and another small island population. As the population reproduced some portion (fraction) m of individuals move to an island. So population on the island is denoted by 1-m, as m is the per cent the immigrants. The allele frequencies of two populations are defined as p_x: allele frequency on a larger island and p_y: allele frequency on a smaller island.

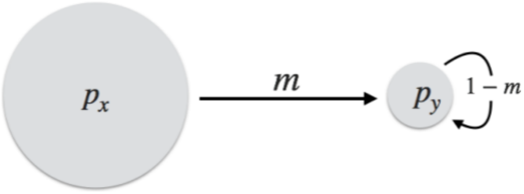

Figure 1 Island Mainland Model

In the next generation, allele frequency is the same in the mainland (p_x), assuming the migration was not large enough to affect the mainland allele frequency. The island frequency changes by :

$$p_{y,t+1} = (1\text{-}m)p_{y,t} + mp_{x,t}$$

The **island model** is a more complex model in which all populations migrate at a constant migration rate. The stability of the system is more.

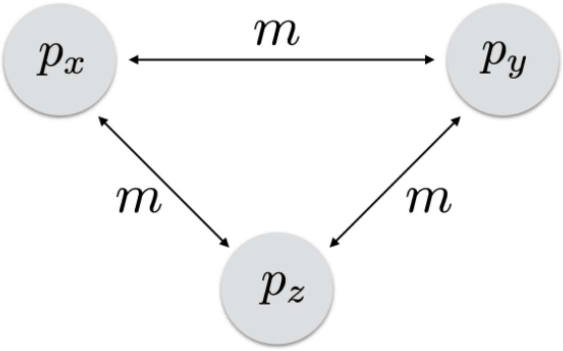

Figure 2 Island Model

Every population share the same rate of migration and hence has a unified allele frequency denoted by $\bar{p}$. If the initial allele frequency is given by $p_{\circ}$, allele frequency at time t is given by :

$$P_t = \bar{p} + (p_{\circ} - \bar{p})(1-m)^t$$

Stepping stone models are more realistic models that were designed by Kimura and Weiss. The models take into consideration the connectivity of the population by taking an infinite length of the population, all having the same migration rate m.

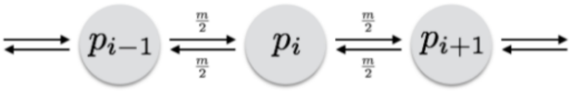

Figure 3 Stepping Stone Model

Other parameters are :

-p_i: frequency of a population on the present generation

-m/2: migrants entering in population from one side

-m_∞: background migration

-p_∞: allele frequency of background migration

-η_i: stochastic change in allele frequencies in each generation

$$P_{i,t+1}=(1-m_1-m_\infty)*p_I+m/2(p_{I-1,t}+p_{i+1,t})+m_\infty \bar{p}+\eta_i$$

The **general population model** helps to broaden the approach as it can be applied to all types of connectivity models and then estimate allele frequency. In this model, the allele frequencies and migration rates can be different which allow personalising the model more.

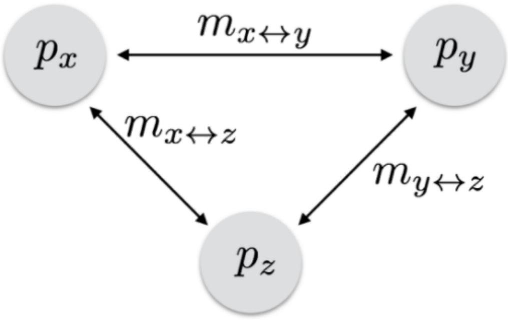

Figure 4 General Model

The equation for 3 population system is given by:

$$P_{y,t+1}=m_{x\leftrightarrow y}\, p_{x,t} + m_{y\leftrightarrow z} p_{z,t}+[1-(m_{x\leftrightarrow y}+m_{y\leftrightarrow z)})]p_y$$

Results

The web application was used making R shiny and is available at the github profile: AyushiPathak (Ayushi Pathak) (github.com). The dashboard consists of two working spaces the side bar and main panel to work on. The side bar has 3 different of models types available. The working space provides the various options to create a flexible model. The graph and data is observed as soon as the values are added and both changes as soon as the input values are changed.

Conclusions

Simulation-Gene has some interesting characteristics of creating genomic models that are very flexible with the inputs. It also uses the real time edit so modification is easy and fast. But still, there are some limitations regarding the accuracy and website crash. The website is stable If the number of generations is approximately 10,000. Beyond that the website has high chances to crash. A lot of improvement is need which can be done with more time. The material on the internet is also very limited and I never saw anyone working with R programming for models like these, making it unique in its own way.

Acknowledgements

I would like to thank our teacher Eran Elhaik for giving this opportunity that turned into a great learning experience. I have truly gained so much knowledge regrading this filed and created a spectrum of ideas that can be implemented on this project.

Reference

1. Riggs, K. *et al.* On the application, reporting, and sharing of in silico simulations for genetic studies. *Genet. Epidemiol.* **45**, 131–141 (2021).

2. Yang, X. *et al.* AdmixSim: A Forward-Time Simulator for Various Complex Scenarios of Population Admixture. *Front. Genet.* **11**, 1474 (2020).

3. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**, 380–390 (2002).

4. M Anisimova, R. N. Z. Y. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics* **164**, 1229–1236 (2003).

5. Peng, B., Amos, C. I. & Kimmel, M. Forward-Time Simulations of Human Populations with Complex Diseases. *PLoS Genet* **3**, e47 (2007).