Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Notes

OXFORD

Subject Section

ftprime: Recording the pedigree: efficient simulation of whole genomes

Corresponding Author [1,*], Co-Author [2] and Co-Author [2,*]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

To use genomic data for inference and prediction it is often necessary to obtain whole-genome information from individual-based simulations, but the computational burden of endowing each simulated individual with an entire genome can be substantial. In this note we describe how to both (a) dramatically reduce this burden and (b) efficiently record the entire history of the population. We do this by simulating only those loci that may affect reproduction (those having non-neutral variants), and recording the entire history of genetic inheritance in an efficient data structure, on which neutral mutations can be quickly placed afterwards. *make more clear data structure was already developed? refer to 'tree sequence' by name?* The algorithm is implemented in python, and is designed to be easily used by any forwards-time simulation software. **Availability and implementation:**
Contact:
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years the increase in computing power – gradual only in comparison to the increase in our ability to sequence genomes – has made it possible to simulate the evolution of whole genomes of realistically-sized populations. This major milestone promises to remove the field's reliance on approximations of unknown applicability. Thus far, much of our understanding of how genomes evolve comes from models of one or a few linked loci (**?**), and/or results derived under the assumption of neutrality. These are certainly important, but researchers today have widely differing views on the practical importance of widespread selection (**???**). Thus far what little analytical progress has been made on models of ubiquitous selection are hard to check due to a lack of realistic simulation.

Simulations of organisms endowed with functional, varying genomes interacting with each other and spatially and temporally varying environments will not only allow us to develop and test our understanding of evolution, but also to produce quantitative predictions of population dynamics – for instance, the spread of insecticide resistance and behavior modification of mosquitos.

The most commonly used simulation methods to date rely on neutrality, as this assumption coupled with random mating allows the use of coalescent methods (**?**). These drastically reduce the amount of required computation, because they exploit a time-reversal duality of such models to *only* simulate the portion of history that determines the genomes in the modern population. How does this work? Suppose that to the *population pedigree* – the entire history of parent-offspring relationships of an entire population going back to a remote time – we add information encoding the genetic outcomes of each ancestral meiosis – who inherited which parts of which parental chromosomes. This embellished graph is known as the *ancestral recombination graph*, or ARG (**?**). Combined with ancestral genotypes and the origins of new mutations, it completely specifies the genomic sequence of any individual in the population at any time. However, much less than the entire ARG is needed to specify relationships between any given set of samples – only those portions of it from which those samples have actually inherited, back to their most recent common ancestors. For instance, *add diagram of simple example?* The assumptions of coalescent theory – random mating and neutrality – imply that the stochastic law of this random set forms a Markov process looked at *backwards* in time, and can hence be simulated without reference to the unnecessary remainder. Although this point was in principle known since **?**, only recent algorithmic advances made it possible to actually simulate the process correctly, without approximation, across whole genomes (**?**).

These efficient methods are not available if either key assumption of coalescent theory – neutrality or random mating – are broken. (Extensions

can be made for only a few loci under selection (**??**).) Forwards-time simulation, however, is burdened by recording and passing on entire genomes, which typically have tens to hundreds of millions of varying sites.

2 Methods and Implementation

3 Results and Discussion

4 Conclusion

Acknowledgements

Funding