Using Coalescent Simulations to Estimate Effective Population Sizes

MATH795-Simulation Modeling with Applications - *Piyush Agarwal*

Computer Simulations occupy a central place in evolution and population genetics. This is evident from large number of programs developed (almost 40)[1], to carry out simulations under variety of assumptions regarding demography, mutation, recombination, migration and life history. They have gained prominence because genetic models have been analytically intractable,and modelling recombination has been very challenging.
Effective Population Size(N_e) is a key parameter in pop. genetics[2], it can be defined as per requirements, but generally measure the rate of genetic drift and inbreeding, and in some simplified settings correspond to the number of breeding individuals in a population.

There are two categories of simulators in this field, backward or **coalescent** simulators and forward simulators. Coalescent Simulators take a lineage approach where a sample of copies is followed back in time to the most recent common ancestor, whereas forward simulators follows the life-cycle of individuals. Since forward simulators are computationally heavy and require an initial genotype to start off, I want to focus on coalescent simulators which are faster and don't require any real data.

I have a two stage plan for the project. Initially, I want to focus on **MS** developed by Richard Hudson in 2002 [3]. It's said to be the most classical simulation algorithm working under the infinite sites model of mutation and the neutral theory. The entire code is written in C and I plan to rewrite the code in Python. This is aimed more at my own personal learning. The second part of the project is to use MS to estimate the effective population size N_e. The idea is to simulate a big tree for some big sample, say $n = 1000$ knowing N_e, and then using the coalescent time of random tip pairs to predict N_e. In a standard coalescent model with no selection and recombination, the coalescent time is an exponential random variable with mean $\frac{N_e}{\binom{i}{2}}$ where i is the number of current lineages. I hope to use this relation to identify N_e and work out the details more clearly in time.

References:

1. Hoban, S., Bertorelle, G. Gaggiotti, O. Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet 13, 110–122 (2012). https://doi.org/10.1038/nrg3130

2. Wang, J., Santiago, E. Caballero, A. Prediction and estimation of effective population size. Heredity 117, 193–206 (2016). https://doi.org/10.1038/hdy.2016.43

3. Richard R. Hudson(May, 2007) MS- a program for generating samples under neutral models