GPOP project: simulating an evolving population

The goal of the project simulate an evolving population in different settings and to confront observations to results derived in the lecture. Your simulation should be based on a Wright-Fisher model with the following assumptions.

- The population is made of N haploid individuals that reproduce asexually.
- Generations are discrete.
- We track one locus in the genome. Each individual carries one allele at this locus.

According to the modeled situation, the number of possible alleles may be infinite or finite, there may be selection or mutation, etc... When specific values of parameters ($N, \mu, s, ...$) are not given, it means that it is up to you to choose them; pick values that you think are relevant. Whenever possible ($\sim$ always), you should compare your results to what was derived in the lecture!

Note: most of the results of the lectures are statistical in nature; they are relevant for an *ensemble* of populations. You will have to repeat each simulation many times and make statistics.

1. Genetic drift

Simulate a population where there are only two possible alleles, both of them neutral. There are no mutations.

- Trace the allele frequencies over time for a few simulations (use $N = 1000$).
- Trace the average diversity as a function of time (probability that the alleles of two randomly picked individuals are different).
- Determine the probability of fixation of one of the alleles as a function of its initial frequency (use $N = 1000$).
- For one of the alleles, measure the average fixation time (for those cases where it eventually fixes) and the average segregation time as a function of N and of the starting frequency x_0. Try to explore a large range of values for N and x_0.

2. Coalescent

Simulate a population of $N = 100$ individuals until all alleles are identical by descent. Select $n = 2, 3, 5, 10$ and 20 individuals from the population, and measure the time to their MRCA. Compare to the coalsecsent theory.

3. Mutations in the infinite alleles model

Use the infinite alleles model: all mutations lead to a new, never seen before allele. Start from an homogenous population with all individuals carrying the same initial allele. Trace the diversity H as a function of time, where H is defined as the probability that two randomly picked individuals carry different alleles. Use different values of N and μ so that you illustrate at least two different interesting regimes.

4. Mutations in the two-alleles model

Use the two-alleles model with mutation rate μ: there are two alleles $\{0, 1\}$, and mutations switch from one to the other. Start from an homogeneous population with all individuals carrying allele 0.

1. Pick a relatively small μ and trace the frequency of allele 1 as a function of time for one simulation as an illustration.
2. Measure the *site frequency spectrum*: the probability to observe allele 1 at frequency x as a function of x. Repeat for different values of N and μ so that you illustrate at least two different interesting regimes.

Note: simulate for long enough for statistical equilibrium to take place. What is the relevant timescale?

5. Population bottleneck

Simulate a population with a time dependent size: $N = 1000$ for 25 generations in a row, then $N = 50$ for 5 generations, then back to $N = 1000$ for the next 25 generations and so on. Perform the same measurements as in the "genetic drift" case.

6. Selection

Simulate a population in the two-alleles model. Alleles 0 and 1 have respective fitnesses 1 and $1 + s$.

1. Study the case where allele 1 is initially present at frequency x_0 and $s > 0$. Trace its frequency as a function of time. Pick values of N, s and x_0 so that selection beats drift most of the time.
2. Now consider the case where allele 1 just appeared and $x_0 = 1/N$. Measure the fixation probability of the allele. Fix N and consider different values of s.

Organization

Code: The code must be sent to me the day before the examination. Choice of language is up to you. I suggest (but it is not strictly necessary) that the main results be in presented in jupyter notebooks (or equivalent) that clearly link the results (plots, ...) with the code used to generate them. You may work on the simulation by groups of maximum **two** people.

Evaluation: The evaluation is based on an *individual* presentation and discussion. The duration of the presentation is ten minutes. The total duration (presentation + discussion) will be around twenty minutes. The discussion can include questions on your simulation and presentation as well as on the lecture material.