07/10/2024

GPOP: simulating an evolving population

The **objective of the project** is the simulation of a simple population-genetics model to observe the influence of genetic drift, mutations, selection and population structure on the evolution of the population. The simulations will be done using a **clonal version of the Wright-Fisher model**:

- The population is formed by N haploid individuals, each one carrying one allele for a given locus.
- The population evolves with constant population size and discrete non-overlapping generations.
- Each individual in generation t+1 is a copy of a randomly selected individual in generation t.

According to the situation to be studied, the selection can be fitness dependent, or mutations between parent and offspring may appear with some mutation rate.

1) **Genetic drift**
   Study a finite population without selection and mutation, but initialized with two different alleles A and B having allele frequencies p and $q = 1 - p$. Simulate the population to (a) trace the allele frequencies over time; (b) determine the fixation probability of allele A in dependence of p; (c) estimate the expected fixation time (number of generations) in dependence of N. Simulation results are to be compared to the analytical results derived in the lectures.
   Note 1: The analytical results in the lectures concern probabilities and expectation values. To enable a comparison, the evolution of the population has to be simulated several times, with identical initialisation but different realizations of the random parent selections, to estimate empirical fractions and averages.
   Note 2: When comparing to analytical results, you have to consider that the simulated population is a clonal population of N alleles, not a population of diploid individuals with $2N$ alleles.

2) **Coalescent model**
   Simulate a population of $N = 100$ individuals until all alleles are identical by descent. Select several random samples of size $n = 2,3,4,5$ from the population and measure the number of generations to the first coalescent event. Are these consistent with the analytical expressions for the expectations $\mathbf{E}(T_n)$? Note again in this analysis, that our population now has N alleles and not $2N$.
   Note: The mathematical derivations were done for $1 < n << N$. You will observe deviations in particular when using n being comparable in size to N.

3) **Mutations in the infinite-allele model**
   Start from an initially homogeneous population of N identical alleles. Simulate the population as described before, but with a mutation rate μ measured in mutations / allele / generation.

Each mutation leads to a new allele. Trace the fixation index (probability that two randomly chosen alleles are identical) over time. Is the result coherent with the balance between genetic drift and mutation derived in the lectures?

4) **Selection**
   Introduce selection into the population, considering a two-allelic locus with alleles A of fitness 1, and B of fitness 1+*s.* The fitness is realized by selecting a parent allele with a probability proportional to fitness from the parent population. Follow the fraction of allele B over time.

5) **Clonal interference**
   Use the setting of Task 4 to observe the evolution of a population with a three-allelic locus. Allele A with fitness 1 is initially dominant (p_A = 0.79), allele B with fitness 1.05 less frequent (p_B = 0.2) and allele C with fitness 1.1 rare (p_C = 0.01). Follow the allele frequencies over time (and for various *N*), and use the results to explain the concept of clonal interference.
   Note: You may need to vary the population size to see the effect of clonal interference.

6) **Population structure**
   Divide the population of Task 1 now in 10 equally large sub-populations. Simulate the system with a complete separation of the sub-populations: parents are selected only in the same subpopulation, and no mutation, selection or migration exists. How does the evolution of the population change as compared to Task 1?

7) **Migration**
   Use the population of Task 6, but introduce a migration of a fraction *m* = 0.1 of each population towards and from randomly chosen subpopulations (i.e. the subpopulations exchange individuals, but remain of the same size). Estimate the expected fixation time and compare to Task 1.

Evaluation: The project will be evaluated in a final oral examination. Two grades will be given, one for the presentation of the results (12min presentation), a second for the discussion (13min discussion). In the discussion, questions about the project **and** the more general contents of the lectures will be asked. Simulation results will be put into the mathematical context presented in the lectures.
The code (Jupyter or Colab notebook preferentially) and the pdf of the presentation have **to be sent the day before the examination**. The project can be elaborated in groups of two students, but the oral examination is individual. Please bring your computer with working code and presentation to the examination to avoid unpleasant surprises.