Stochastic effects on the genetic structure of populations

Level 1 module in "Modelling course in population and evolutionary biology"

(701-1418-00)

Module author: Roger Kouyos

Course director: Sebastian Bonhoeffer
Theoretical Biology
Institute of Integrative Biology
ETH Zürich

1 Introduction

The genetic structure of natural populations is strongly affected by random genetic drift: random effects can destroy the genetic diversity built up by mutation, counteract the effect of selection, and build up statistical associations between different loci. Therefore, a proper understanding of most questions in evolutionary biology requires random effects to be taken into account.

Understanding random effects is also important for many practical questions. For instance, the fixation of drug-resistance mutations in disease-causing organisms (from either newly arising or pre-existing mutations) depends crucially on genetic drift. Consider newly arising mutations: in a very large population, drug resistance mutations are generated in every generation, and stochasticity plays only a minor role. In a small population, however, such mutations are generated only with a small probability, and their time of emergence is therefore random. (This randomness has been evoked to explain the fact that the time to the emergence of drug resistance varies largely from patient to patient treated with HIV infection (Nijhuis *et al.* 1998)).

The importance of stochastic effects depends on many model parameters such as the population size and the mutation rate. In this module, simple population genetic simulations are

used to disentangle these different factors and to obtain a clearer understanding of the role of genetic drift.

2 Developing the models

2.1 The basic model

In the simple model we consider just one locus with alleles A and a. The two alleles have the frequencies f_A and $f_a = 1 - f_A$. For simplicity we assume that generations are discrete. In every generation mutation and selection occur. In principle, we should simulate mutation and selection stochastically. In order to simplify the model, we consider an approximation in which mutation and selection are both deterministic and the finite population size is taken into account by sampling N genomes (N is the population size) after mutation and selection have occurred. (You are welcome to develop a model in which the mutation and selection steps themselves are stochastic and to investigate when the above approximation works.).

In detail, the mutation, selection and sampling steps look as follows:

- Mutation: A fraction m of all a alleles mutate to A and vice versa. Therefore the frequency of A after mutation reads:

$$f_A' = (1 - m)f_A + m(1 - f_A) \tag{1}$$

- Selection: The two alleles have fitnesses $w_a = 1$ and $w_A = 1 - s$, where s is the selection coefficient. After selection A has the frequency:

$$f_A'' = f_A' \frac{w_A}{\bar{w}} = f_A' \frac{1 - s}{(1 - s)f_A' + (1 - f_A')} , \tag{2}$$

  where $\bar{w}$ denotes mean fitness in the population.

- Sampling: We obtain the number of A alleles after sampling by drawing a random number N_A according to the binomial distribution with parameters N and f_A''; this corresponds to drawing successively N offspring alleles from a gene pool in which a fraction f_A'' of the alleles is A and the rest is a. `R` has a function `rbinom()` that draws the random number for us. The frequency of A after sampling is then $f_A''' = N_A/N$.

In summary, we simulate the frequency changes for A in one generation by performing the mutation, selection and sampling steps above and then setting $f_A = f_A'''$. (Notice that we do not have to track f_a explicitly because we always have $f_a = 1 - f_A$). Does this model describe a haploid asexual, or a diploid sexual population?

The model has been implemented in the `R` script *stochasticStart.R* that can be downloaded from the course website. Download and run the script (the script produces a plot of the time course of f_A). Try out different parameter values and check how they affect the shape of the

time course. Start with a neutral model (i.e. $s = 0$) and vary N and m. Then introduce selection. You should further consider for some cases the deterministic limit: either set N to a very large value or eliminate the sampling step from the simulation. Check whether and how the former approach converges to the latter as you increase N. Important note: because the simulation includes stochastic steps, each run will be different. To get a reliable idea on the possible range of behaviours, you need to run several simulations with the same parameter set, or even formally inspect the distribution of the outcomes. Hint: to be able to run simulations with varied parameters conveniently, try to wrap a function around the basic simulation code. The call to the function should look something like

```
sim(initialFrequency=0.1,numberOfGenerations=10000,selectionStrength=0,
mutationRate=0.01,populationSize=100,plotTimeCourse=TRUE)
```

and it should return the time steps and the frequency of the A allele at each time step (e.g. in a data frame). The last argument is intended as a switch whether you wish also to plot the time course of the simulation.

2.2 The extended model: What is the effect of recombination?

Next we extend the model to two loci and two alleles; i.e. we now have the genotypes ab (wildtype), aB and Ab (single mutants) and AB (double mutant). This allows us to investigate how recombination affects the course of evolution. I recommend to read the review by Otto and Lenormand (Otto and Lenormand 2002), in order to get an understanding of this field.

Considering two loci requires some minor changes in the model:

- Mutation: Consider the genotype ab. With probability $(1-m)^2$ it does not mutate; with probability $(1-m)m$, aB or Ab mutates to ab (one locus mutates while the other does not); with probability m^2, AB mutates to ab (both loci have to mutate). Thus, the frequency of ab after mutation reads

$$f_{ab}{}' = (1-m)^2 f_{ab} + (1-m)m(f_{Ab} + f_{aB}) + m^2 f_{AB} \ . \tag{3}$$

  For the other three genotypes the formulas can be derived in a similar way. Do it yourself.

- Selection: No change, i.e.

$$f_{xy}{}'' = f_{xy}{}' \frac{w_{xy}}{\bar{w}} \tag{4}$$

- Recombination: The new step. One can rewrite the genotype frequencies as

$$\begin{aligned} f_{ab} &= p_a p_b + D \\ f_{Ab} &= p_A p_b - D \\ f_{aB} &= p_a p_B - D \\ f_{AB} &= p_A p_B + D \ , \end{aligned} \tag{5}$$

where p's denote allele frequencies and $D = f_{ab}f_{AB} - f_{Ab}f_{aB}$ is linkage disequilibrium. Recombination does not affect the allele frequencies; it only reduces the linkage disequilibrium by a factor $(1 - r)$, where r is the recombination rate. Therefore the frequencies after recombination are

$$\begin{aligned} f_{ab}''' &= f_{ab}'' - rD \\ f_{Ab}''' &= f_{Ab}'' + rD \\ f_{aB}''' &= f_{aB}'' + rD \\ f_{AB}''' &= f_{AB}'' - rD \end{aligned} \tag{6}$$

- Sampling: The multinomial distribution (implemented by `rmultinom()`) provides the natural extension of the binomial distribution when more than two genotypes have to be sampled.

The extended model has been implemented in the `R` script *stochasticExtended.R* that can be downloaded from the course website.

In this part of the module, we try to understand how recombination affects the fixation of beneficial mutations. We therefore consider the following setting. *Fitness values*: $w_{ab} < w_{Ab} = w_{aB} < w_{AB}$, i.e. for both loci, the wild-type allele has the lower fitness (the assumption $w_{Ab} = w_{aB}$ has been made for simplicity, and you are free to drop it). *Initial conditions*: We start with frequencies $(f_{ab}, f_{Ab}, f_{aB}, f_{AB}) = (1, 0, 0, 0)$, i.e. with a population consisting entirely of the unfit wild-type. This situation can be regarded as an oversimplification of the emergence of drug resistance mutations against two different drugs (the resistance mutations are encoded by A and B): the population starts with the double sensitive wild-type ab and the fittest genotype is the double resistant mutant AB. These conditions are implemented in the `R` script. Figure 1 shows a typical outcome of such a simulation: first the single mutants rise in frequency (they are less fit than the double mutants but mutation generates them at a higher rate, thus they have a head start), but eventually the double mutants take over. Plot simulations for different parameter values.

3 Exercises

3.1 Basic exercises

Eb1. a) Consider the basic model without selection; i.e. both alleles, A and a, have the same fitness. In this exercise, you will investigate how stochastic effects affect diversity. A good measure for diversity is given by $d = 2f_af_A = 2f_a(1 - f_a)$; if two genomes are drawn randomly from the population, d gives the probability that they are different. To which values does d converge in a deterministic model (i.e. without sampling) after a large number of generations? We term this value the "steady-state diversity". Plot, for different fixed mutation rates, the steady-state diversity d as a function of the population size also

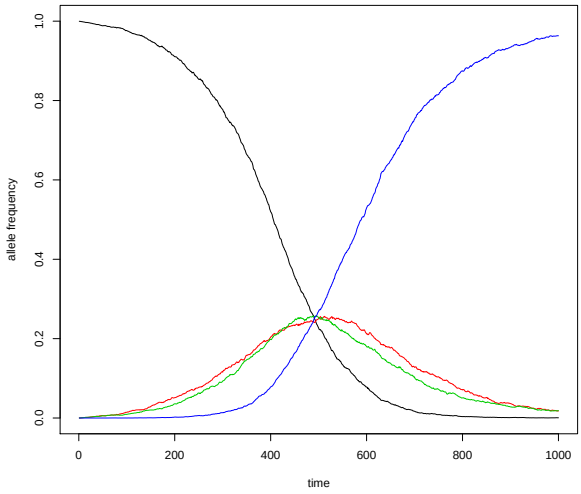

Figure 1: The emergence of drug resistance mutations against two different drugs. Genome frequencies (ab–black; Ab–red; aB–green; AB–blue) as a function of time. Parameters: $m = 10^{-4}, (w_{ab}, w_{Ab}, w_{aB}, w_{AB}) = (1, 1.01, 1.01, 1.012), r = 0.1, N = 10^4$.

in the stochastic model[a]. When is a population size large (with respect to diversity); i.e. when does diversity approach its values of the deterministic model?

b) An alternative way to estimate the diversity is to look directly at the distribution of allele frequencies. You can visualize this distribution by making a histogram of the allele frequencies at different time points. If the frequency is close to 0.5 most of the time, diversity is high; if the frequency is close to 0 or 1 most of the time, diversity is low. Create such histograms for different population sizes; do you find qualitative differences for very large and very small populations? (Hint: use the `R` function `hist(x)` to make a histogram of a vector x; `hist` divides the range of x into intervals – usually of equal length – and then makes a bar plot according to the number of elements of x that fall into each interval. Note that in our example, x is the vector containing the allele frequencies at different time points.)

Eb2. Next consider the case that allele A is selected against; i.e. $w_A = 1 - s$ and $w_a = 1$. In an infinite population, the frequency f_A converges towards the mutation-selection balance value $\approx \mu/s$; check this result by simulating the model without the sampling step. Then, investigate what the frequency distribution looks like for different finite population sizes. Is there a qualitatively different pattern for very small and very large populations?

Eb3. a) Now introduce periodic bottlenecks into your simulations; i.e. the population size should drop every 100 (or any other number) generations to a few individuals. How do such bottlenecks affect the distribution of allele frequencies at a neutral or at a selected

[a] `R`-hint: you can do this by programming a loop over different values of the population size. Within the loop, run a simulation with the given population size and record steady-state diversity (you are advised to implement a test that it is indeed steady-state). Then plot the vector of d values against the vector of population sizes.

locus? Is there a qualitative difference between the effect of bottlenecks on neutral and selected loci?

b) Having done a): What problems do bottlenecks involve for the estimation of ("effective") population sizes (see for instance Kouyos *et al.* 2006)? Notice that effective population sizes are often estimated by comparing the observed diversity at neutral loci with the diversity predicted from the Wright-Fisher model (this is how our model with $s = 0$ is usually called in the literature).

3.2 Advanced/additional exercises

Ea1. For finite populations, recombination can potentially accelerate the fixation of beneficial alleles at different loci (e.g. resistance mutations against different drugs). This phenomenon is called the Fisher-Muller effect (see for example Althaus and Bonhoeffer 2005). Explore in your simulations for which parameters (N and m) there is such an effect. Assume for this exercise that mutations affect fitness multiplicatively, i.e. $w_{ab} = 1,\ w_{Ab} = 1 + s,\ w_{aB} = 1 + s,\ w_{AB} = (1 + s)^2$ (we make this assumption to avoid the generation of statistical associations between genes by fitness interactions; see next exercise). Compare the average time to fixation of the double mutant (e.g. take the time until the double mutant has reached a frequency of 90%) with and without recombination for different population sizes. Note: because stochastic effects are involved, run each simulation with several repetitions and record the average outcome (e.g., the time to fixation of a mutant).

Ea2. Stochastic effects are not the only factor that influences the effect of recombination. An additional effect is due to fitness interactions between different loci (see for example Bretscher *et al.* 2004). Compare the relative importance of these two factors in the following way: consider first the effect of recombination in the multiplicative model for different population sizes (this is done in Ea1); then add fitness interactions to the model and look how they change the outcome of the simulation. (Fitness interactions mean that mutations at different loci affect fitness in a non-multiplicative way; e.g. $w_{ab} = 1,\ w_{Ab} = 1 + s,\ w_{aB} = 1 + s,\ w_{AB} = (1 + s)^2 + \epsilon$).

Ea3. Use the two-allele model to simulate a specific (but often important) situation in the emergence of drug resistance. The first mutations that appear under therapy are often such that they abrogate much of the drug effect, but decrease the viability of the pathogen. These *primary resistance mutations* thus increase fitness in the presence of the drug, but decrease fitness in its absence. Then under continued drug treatment *compensatory mutations* often arise at other loci, which restore the viability/fitness of the pathogen, sometimes near wild-type levels. Parameterize the two-allele model to reflect this situation. Model the effect of intermittent therapy (turning treatment on and off). What courses of evolution do you observe?

Ea4. You may try to build a fully stochastic model of population genetics. Keep track of the (integer) number of individuals instead of frequencies, and implement each step of the generation cycle (reproduction, mutation, selection, etc) with randomness involved. Start

with the simpler case of an asexual population, then introduce sexual reproduction, which allows for recombination but requires you to implement mate choice, as well.

4 Recommended reading

1. ALTHAUS, C. L., and S. BONHOEFFER, 2005. Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1. J Virol 79: 13572-13578.

2. BRETSCHER, M. T., C. L. ALTHAUS, V. MÜLLER and S. BONHOEFFER, 2004. Recombination in HIV and the evolution of drug resistance: for better or for worse? Bioessays 26: 180-188.

3. KOUYOS, R. D., C. L. ALTHAUS and S. BONHOEFFER, 2006. Stochastic or deterministic: what is the effective population size of HIV-1? Trends in Microbiology 14: 507-511.

4. NIJHUIS, M., C. A. BOUCHER, P. SCHIPPER, T. LEITNER, R. SCHUURMAN et al., 1998. Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. Proc Natl Acad Sci U S A 95: 14441-14446.

5. OTTO, S. P., and T. LENORMAND, 2002. Resolving the paradox of sex and recombination. Nature Reviews Genetics 3: 252-261.