Biological Physics - Evolution

Pietro Cicuta and Diana Fusco

Experimental and Theoretical Physics
Part III
Michaelmas 2021

Notes version: v0.05
Release name: Evolving Emergence

Evolution

1

1.1 Concepts in evolution

full credit: Notes on Population Genetics, Graham Coop, UC Davies

Evolution is change over time. Biological evolution is change over time in the genetic composition of populations. We define the genetic composition of a population to be the set of genomes that the individuals in our population carry. While at first this definition of evolution seems at odds with the common textbook view of the evolution of phenotypes (such as the changing shape of the finch beaks over generations) it is genetic changes that underpin these phenotypic changes. The genetic composition of the population can alter due to the death of individuals or the migration of individuals in or out of the population. If our individuals have different numbers of children, this also alters the genetic composition of the population in the next generation. Every new individual born into the population subtly changes the genetic composition of the population. Their genome is a unique combination of their parents' genomes, having been shuffled by segregation and recombination during meioses, and possibly changed by mutation. Population genetics is the study of the genetic composition of natural populations. It seeks to understand how this composition has been changed over time by the forces of mutation, recombination, selection, migration, and genetic drift. To understand how these forces interact, it is helpful to develop simple theoretical models to help our intuition. In these notes we will work through these models and summarize the major areas of population genetic theory. We will also highlight areas in which physics as contributed to create new and more realistic models to describe the evolution of populations. Throughout the course we will see that these simple models yield accurate predictions, such that much of our understanding of the process of evolution is built on these models.

1.1.1 Allele and genotype frequencies

Population genetics emerged from early efforts to reconcile Mendelian genetics with Evolution. Thanks to Mendel and Mendelian genetics (and a lot of prior and subsequent work), we understand that the genome of an individual is formed from two gametes that

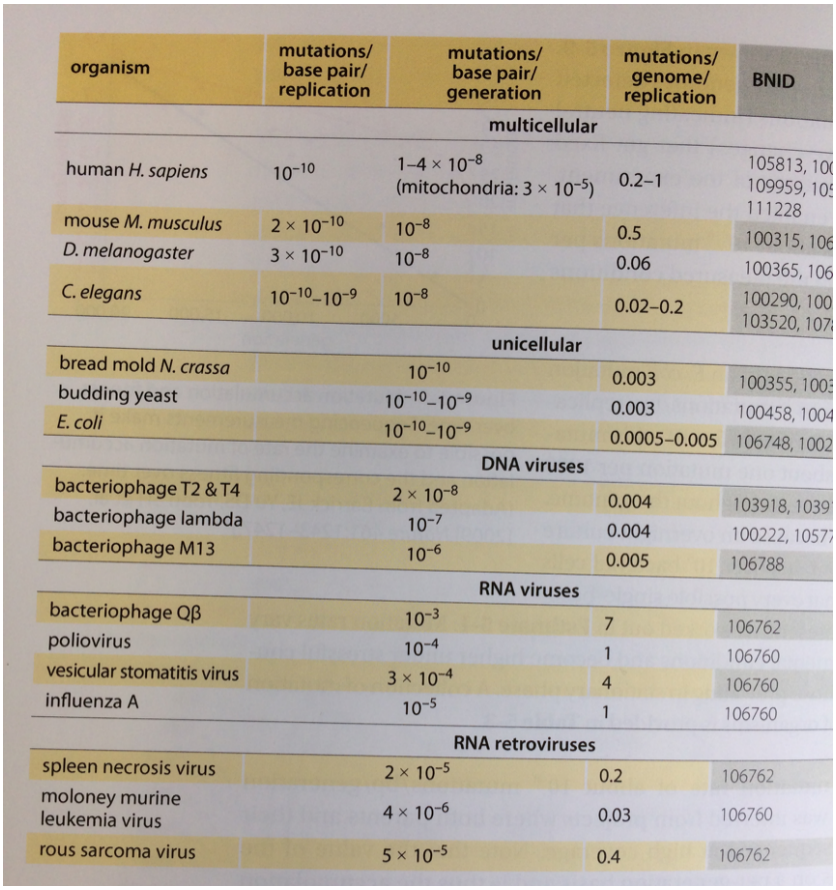
fused to form a zygote. The genomes of each gamete originate from a parental genome through meiosis, in particular the segregation and recombination of the parental genome's two gametes. Loci and alleles are the basic currency of population genetics–and indeed of genetics. Each individual's genetic makeup is defined in their genome. A locus (plural: loci) is a specific spot in the genome. A locus may be an entire gene, or a single nucleotide base pair such as A-T. At each locus, there may be multiple genetic variants segregating in the population—these different genetic variants are known as alleles. For example, at a particular nucleotide site in the genome, a population may segregate for A-T and G-C base pairs (note that due to the complementary nature of DNA, it will suffice to say the site segregates for A and G variants). If all individuals in the population carry the same allele, we say that the locus is monomorphic; at this locus there is no genetic variability in the population. If there are multiple alleles in the population at a locus, we say that this locus is polymorphic.

Note that not all populations are made of diploid individuals. Bacteria carry only one copy of their genome and they reproduce asexually, meaning that the daughter will be genetically identical to mother, modulus errors that occur in the DNA replication step.

1.1.2 Mutation, selection and genetic drift

New alleles can be introduced in a population through genetic mutations. Mutations occur when errors are made in the process of DNA replication. These errors are normally very rare (see Fig. 1.1), since often mutations can result in a protein not folding correctly or not functioning properly. In these cases, mutations are called *deleterious*, as they will negatively affect the growth rate of the individual carrying them, or even impede reproduction alltogether (*lethal mutations*). Why then does nature allow the possibility of mutations? Occasionally mutations can confer big advantages, such as resistance to an antibiotic, or the ability to metabolize a new carbon source. These are called *beneficial* mutations, as they positively affect the growth rate of the carrier. Finally, there are mutations that do not lead to any change in growth rate. These are called *neutral mutations* and are often linked to synonymous substitutions, which do not alter the aminoacid composition of a protein.

If mutations occur randomly and most of them are deleterious, why don't we observe them routinely in populations? As we mentioned, deleterious mutations confer a *fitness* disadvantage to the individual carrying them, meaning that such individual will be less likely to generate offspring in the next generation. Since these changes are hereditary, the limited offspring will itself be affected by the same disadvantage and therefore replicate even less. Eventually, *selection* will purge these deleterious variants from



organism	mutations/ base pair/ replication	mutations/ base pair/ generation	mutations/ genome/ replication	BNID
multicellular				
human *H. sapiens*	10^{-10}	$1–4 \times 10^{-8}$ (mitochondria: 3×10^{-5})	0.2–1	105813, 10(109959, 10 111228
mouse *M. musculus*	2×10^{-10}	10^{-8}	0.5	100315, 106
D. melanogaster	3×10^{-10}	10^{-8}	0.06	100365, 106
C. elegans	$10^{-10}–10^{-9}$	10^{-8}	0.02–0.2	100290, 100 103520, 107
unicellular				
bread mold *N. crassa*		10^{-10}	0.003	100355, 100
budding yeast		$10^{-10}–10^{-9}$	0.003	100458, 1004
E. coli		$10^{-10}–10^{-9}$	0.0005–0.005	106748, 1002
DNA viruses				
bacteriophage T2 & T4		2×10^{-8}	0.004	103918, 1039
bacteriophage lambda		10^{-7}	0.004	100222, 1057
bacteriophage M13		10^{-6}	0.005	106788
RNA viruses				
bacteriophage Qβ		10^{-3}	7	106762
poliovirus		10^{-4}	1	106760
vesicular stomatitis virus		3×10^{-4}	4	106760
influenza A		10^{-5}	1	106760
RNA retroviruses				
spleen necrosis virus		2×10^{-5}	0.2	106762
moloney murine leukemia virus		4×10^{-6}	0.03	106760
rous sarcoma virus		5×10^{-5}	0.4	106762

Fig. 1.1 Table summarizing typical mutation rates in different organisms (from Physical Biology of the Cell).

the population and we won't be able to observe them. The opposite process occurs for beneficial mutations, which instead tend to spread in the population.

What about neutral mutations? If growth was a purely deterministic process with no source of noise, one would expect neutral mutations to maintain their frequency over time. However, this is not the case, because individuals do not always replicate exactly once at each generation. The stochastic fluctuations associated with the random process of replication come under the name of *genetic drift*. Analogously to selection, genetic drift tends to reduce the population diversity generated by mutations. Unlike selection, though, genetic drift's action is purely random and in many circumstances can act against the purging/enriching process of deleterious/beneficial mutations, respectively. In the following, we will provide a mathematical framework to model genetic drift and then its relationship with selection.

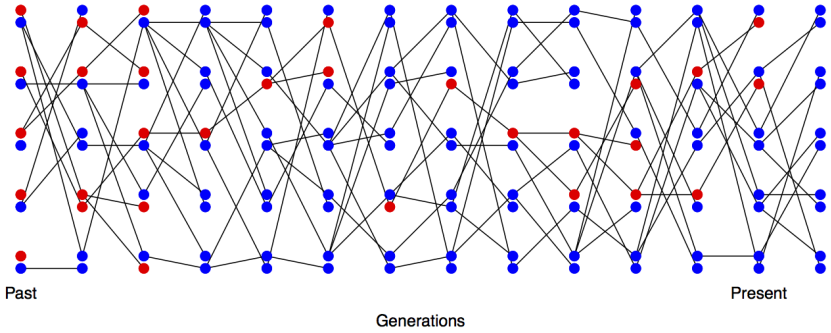


Fig. 1.2 Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals over the generations, with lines showing transmission. In the first generation every individual is a heterozygote.

1.2 Genetic drift and neutral diversity

Various sources of randomness are inherent in evolution. One major source of stochasticity in population genetics is genetic drift. Genetic drift occurs because more or less copies of an allele by chance can be transmitted to the next generation. This can occur because by chance the individuals carrying a particular allele can leave more or less offspring in the next generation. In a sexual population genetic drift also occurs because Mendelian transmission means that only one of the two alleles in an individual, chosen at random at a locus, is transmitted to the offspring.

1.2.1 Loss of heterozygosity due to genetic drift

Genetic drift will, in the absence of new mutations, slowly purge our population of neutral genetic diversity as alleles slowly drift to high or low frequencies and are lost or fixed over time.

Imagine a population of a constant size N diploid individuals, and that we are examining a locus segregating for two alleles that are neutral with respect to each other. This population is randomly mating with respect to the alleles at this locus. See Figure 1.2 to see how genetic drift proceeds, by tracking alleles within a small population.

In generation t our current level of heterozygosity is H_t, i.e. the probability that two randomly sampled alleles in generation t are non-identical is H_t. Assuming that the mutation rate is zero (or vanishing small), what is our level of heterozygosity in generation $t+1$?

In the next generation $t+1$ we are looking at the alleles in the offspring of generation t. If we randomly sample two alleles in generation $t+1$ which had different parental alleles in generation t then it is just like drawing two random alleles from generation t. So the probability that these two alleles in generation $t+1$, that

have different parental alleles in generation t, are non-identical is H_t.

Conversely, if our pair of alleles have the same parental allele in the proceeding generation (i.e. the alleles are identical by descent one generation back) then these two alleles must be identical (as we are not allowing for any mutation).

In a diploid population of size N individuals there are $2N$ alleles. The probability that our two alleles have the same parental allele in the proceeding generation is $\frac{1}{2N}$, the probability that they have different parental alleles is $1 - \frac{1}{2N}$ So by the above argument the expected heterozygosity in generation $t + 1$ is

$$H_{t+1} = \frac{1}{2N} \times 0 + (1 - \frac{1}{2N})H_t.$$

By this argument if the heterozygosity in generation 0 is H_0 our expected heterozygosity in generation t is

$$H_t = (1 - \frac{1}{2N})^t H_0$$

i.e. the expected heterozygosity with our population is decaying geometrically with each passing generation. If we assume that $\frac{1}{2N} \ll 1$ then we can approximate this geometric decay by an exponential decay such that

$$H_t = H_0 \exp(-\frac{t}{2N})$$

1.2.2 Levels of diversity maintained by a balance between mutation and drift

Looking backwards in time from one generation to the next, we are going to say that two alleles which have the same parental allele (i.e. find their common ancestor) in the preceding generation have *coalesced*, and refer to this event as a *coalescent event*.

The probability that our pair of randomly sampled alleles have coalesced in the preceding generation is $\frac{1}{2N}$, the probability that our pair of alleles fail to coalesce is $1 - \frac{1}{2N}$.

The probability that a mutation changes the identity of the transmitted allele is μ per generation. So the probability of no mutation occurring is $1 - \mu$. We'll assume that when a mutation occurs it creates some new allelic type which is not present in the population. This assumption (commonly called the infinitely-many-alleles model) makes the math slightly cleaner, and also is not too bad an assumption biologically. See Figure 1.3 for a depiction of mutation-drift balance in this model over the generations.

This model let's us calculate when our two alleles last shared a common ancestor and whether these alleles are identical as a result of failing to mutate since this shared ancestor. For example we can work out the probability that our two randomly sampled alleles coalesced 2 generations in the past (i.e. they fail to coalesce

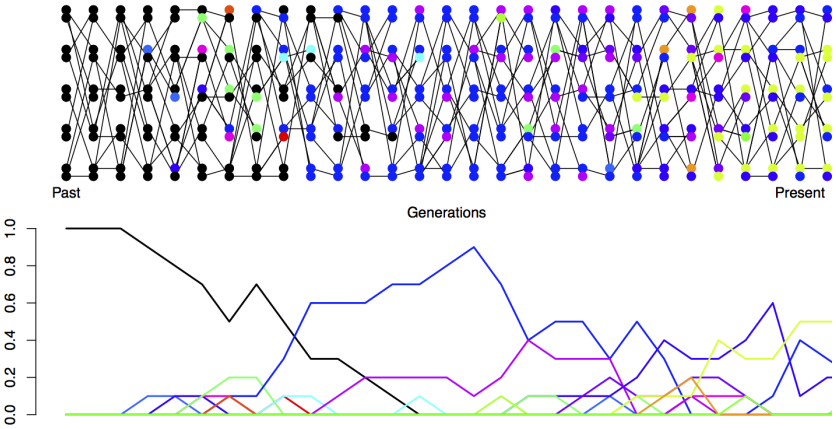


Fig. 1.3 Mutation-drift balance. A diploid population of 5 individuals. In the first generation everyone has the same allele (black). Each generation the transmitted allele can mutate and we generate a new colour. In the bottom plot I trace the frequency of alleles in our population over time.

in generation 1 and then coalescent in generation 2), and that they are identical as

$$(1-\frac{1}{2N})\frac{1}{2N}(1-\mu)^4$$

note the power of 4 is because our two alleles have to have failed to mutate through 2 meioses each.

More generally the probability that our alleles coalesce in generation $t+1$ and are identical due to no mutation to either allele in the subsequent generations is

$$P_{\text{coal}}(t+1)=\frac{1}{2N}(1-\frac{1}{2N})^t(1-\mu)^{2(t+1)}.$$

Over long times, $t\approx t+1$, so we can write

$$P_{\text{coal}}(t+1)=\frac{1}{2N}(1-\frac{1}{2N})^t(1-\mu)^{2t}.$$

This gives us the approximate probability that two alleles will coalesce in the $(t+1)^{th}$ generation. In general, we may not know when two alleles may coalesce: they could coalesce in generation t=1,t=2,..., and so on. Thus, to calculate the probability that two alleles coalesce in any generation before mutating, we can write:

$$P_{\text{coal}}=\sum_{t=1}^{\infty}P_{\text{coal}}(t)$$

If we assume a large population, $\frac{1}{2N}\ll 1$ and small mutation rate $\mu\ll 1$, then we can approximate the geometric decay with an exponential decay,

$$P_{\text{coal}}(t+1)\approx\frac{1}{2N}\exp[-t(2\mu+1/(2N))],$$

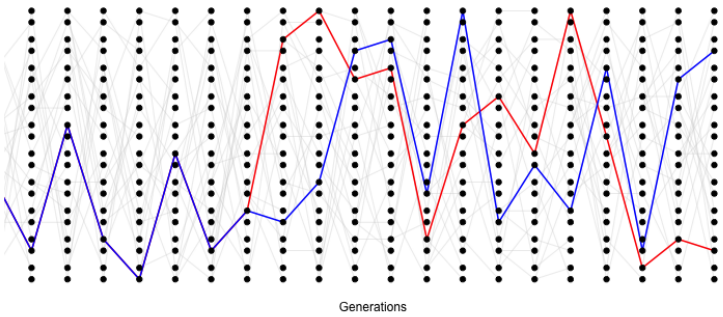


Fig. 1.4 A simple simulation of the coalescent process. The simulation consists of a diploid population of 10 individuals (20 alleles). In each generation, each individual is equally likely to be the parent of an offspring (and the allele transmitted is indicated by a light grey line). We track a pair of alleles, chosen in the present day, back 14 generations untill they find a common ancestor.

and the summation with an integral:

$$P_{\text{coal}} \approx \frac{1}{2N} \int_0^{\infty} \exp[-t(2\mu + 1/(2N))]dt = \frac{1}{1 + 4N\mu}$$

Then, the complementary probability that our pair of alleles are non-identical (or heterozygous) is simply one minus this. This gives the equilibrium heterozygosity in a population at equilibrium between mutation and drift:

$$H = \frac{4N\mu}{1 + 4N\mu}.$$

The component $4N\mu$ is often referred to as θ and represents the population-scaled mutation rate. So all else being equal, species with larger population sizes should have proportionally higher levels of neutral polymorphism.

1.2.3 The effective population size

In practice populations rarely conform to our assumptions of being constant in size with low variance in reproduction success. Real populations experience dramatic fluctuations in size, and there is often high variance in reproductive success. Thus rates of drift in natural populations are often a lot higher than the census population size would imply. To cope with this population geneticists often invoke the concept of an effective population size N_e. In many situations (but not all), departures from model assumptions can be captured by substituting N_e for N.

Specifically the effective population size N_e is the population size that would result in the same rate of drift in an idealized constant population size as that observed in our true population

(following our modeling assumptions). If population sizes vary rapidly in size, we can (if certain conditions are met) replace our population size by the harmonic mean population size. Consider a diploid population of variable size, whose size is N_t t generations into the past. The probability our pairs of alleles have not coalesced by the generation t^{th} is given by

$$\prod_{i=1}^{t}(1-\frac{1}{2N_i}).$$

Note that this simply collapses to our original expression $(1-\frac{1}{2N})^t$ if N_i is constant in time. If $\frac{1}{N_i}$ is small, then $1-\frac{1}{2N_i} \approx \exp(-1/2N_i)$, then

$$\prod_{i=1}^{t}(1-\frac{1}{2N_i}) \approx \Pi_{i=1}^{t}\exp(-1/2N_i) = \exp(-\sum_{i=1}^{t}1/2N_i).$$

So the variable population coalescent probabilities are still of the same form but the exponent has changed. Comparing the exponent in the two cases we see

$$t/2N = \sum_{i=1}^{t}1/2N_i$$

so that we can define the effective population size as

$$N_e = \frac{1}{1/t\sum_{i=1}^{t}1/N_i}$$

Many processes can affect the value of the effective population size, such as reproductive success, population structure, etc... We will see more examples later in the chapter.

1.2.4 The coalescent and patterns of neutral diversity

Thinking back to our calculations we made about the loss of neutral heterozygosity and equilibrium levels of diversity, you'll note that we could first specify what generation a pair of sequences coalesce in, and then calculate some properties of heterozygosity based on that. That's because neutral mutations do not affect the probability that an individual transmits that allele, so don't affect the way in which we can trace ancestral lineages back. As such it will often be helpful to consider the time to the common ancestor of a pair of sequences, and then think of the impact of that on patterns of diversity. The probability that a pair of alleles have failed to coalesce in t generations and then coalesce in the $t+1$ generation back is

$$\frac{1}{2N}(1-\frac{1}{2N})^t \approx \frac{1}{2N}e^{-t/2N}$$

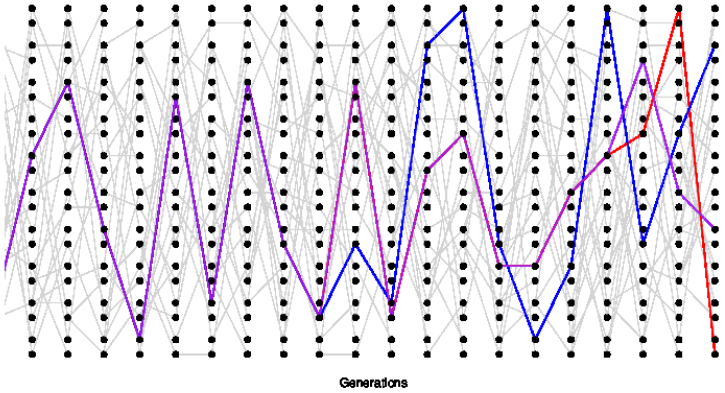


Fig. 1.5 A simple simulation of the coalescent process for three lineages. We track the ancestry of three modern-day alleles, the first pair (blue and purple) coalesce four generations back their are then two independent lineages we are tracking, this pair then coalesces twelve generations in the past. Note that different random realizations of this process will differ from each other a lot.

thus the coalescent time of a pair of sequences T_2 is approximately exponentially distributed with a rate $1/(2N)$. The mean coalescent time of a pair of a pair of alleles is $2N$ generations.

Conditional on a pair of alleles coalescing t generations ago there are $2t$ generations in which a mutation could occur. Thus the probability of our pair of alleles are separated by j mutations since they last shared a common ancestor is

$$P(j|T_2 = t) = \binom{2t}{j} \mu^j (1-\mu)^{2t-j}$$

i.e. mutations happen in j generations, and do not happen in $2t - j$ generations. Assuming that $\mu \ll 1$ and $j \ll 2t$ then we can approximate the equation above with a Poisson distribution

$$P(j|T_2 = t) = \frac{(2\mu t)^j e^{-2\mu t}}{j!}.$$

As our expected coalescent time is $2N$ generations (which follows from the expected value of exponential distributions), the expected number of mutations separating two alleles drawn at random from the population is

$$E(j) = 2\mu E(t) = 4N\mu = \theta$$

If experimentally we observe a given number of differences between two alleles in the sample, this formula allows to estimate the population size.

Usually we are not just interested pairs of alleles, or the average pairwise diversity, we are interested in the properties of diversity

in samples of a number of alleles drawn from the population. To allow for this instead of just following a pair of lineages back until they coalesce, we can follow the history of a sample of alleles back through the population.

When we sample i alleles there are $i(i-1)/2$ pairs, thus the probability that no pair of alleles coalesces in the preceding generation is

$$(1-\frac{1}{2N})^{\binom{i}{2}} \approx (1-\frac{\binom{i}{2}}{2N})$$

while the probability of any coalescing is $\approx \frac{\binom{i}{2}}{2N}$.

When there are i alleles the probability that we wait until the $t+1$ generation before any pair of alleles coalesce is

$$\frac{\binom{i}{2}}{2N}(1-\frac{\binom{i}{2}}{2N})^t \approx \frac{\binom{i}{2}}{2N}\exp(-\frac{\binom{i}{2}}{2N}t)$$

thus the waiting time T_i to the first coalescent event in a sample of i allels is exponentially distributed with a rate $\frac{\binom{i}{2}}{2N}$. When a pair of alleles first find a common ancestral allele some number of generations back further into the past we only have to keep track of that common ancestral allele for the pair. Thus when a pair of alleles in our sample of i alleles coalesce, we then switch to having to follow $i-1$ alleles back. Then when a pair of these $i-1$ alleles coalesce, we then have to follow $i-2$ alleles back. This process continues until we coalesce back to a sample of two, and from there to a single most recent common ancestor (MRCA).

$$T_{MRCA} = \sum_{i=2}^{n} T_i.$$

As our coalescent times for different i are independent, the expected time to the most recent common ancestor is

$$E(T_{MRCA}) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2N}{\binom{i}{2}}.$$

Using the fact that $\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$ with some rearrangement we can write

$$E(T_{MRCA}) = 4N(1-\frac{1}{n}).$$

Mutations fall on lineages of the coalescent genealogy. These mutations affect all descendants of this lineage, and under the infinitely-many-sites assumption, create a new segregating site for each new mutation. The mutation process is a Poisson process, and the longer a particular lineage branch, the more mutations that can accumulate on it. The total number of segregating sites in the genealogy is thus a function of the total amount of time in the genealogy of the sample, or the sum of all the genealogy branch lengths, T_{tot}. Since our coalescent genealogies are bifurcating (only

two lineages coalesce at once), our total amount of time in the genealogy is:

$$T_{tot} = \sum_{i=2}^{n} iT_i$$

as when there are i lineages each contributes a time T_i to the total time. Taking the expectation of the total time in the genealogy

$$E(T_{tot}) = \sum_{i=2}^{n} i\frac{2N}{\binom{i}{2}} = \sum_{i=2}^{n} \frac{4N}{i-1} = \sum_{i=1}^{n-1} \frac{4N}{i},$$

so our expected total amount of time in the genealogy scales linearly with our population size. Our expected total amount of time is also increasing with sample size but is doing so very slowly. To see this more carefully we can see that for large n

$$E(T_{tot}) = \sum_{i=1}^{n-1} \frac{4N}{i} \approx 4N \int_{1}^{n-1} \frac{di}{i} = 4N \log(n-1)$$

We saw above that the number of mutational differences between a pair of alleles that coalescence T_2 generations ago was Poisson with a mean of $2\mu T_2$. A mutation that occurs on any branch of our genealogy will cause a segregating polymorphism in the sample (making our infinitely-many-sites assumption). Thus if the total time in the genealogy is T_{tot} there is T_{tot} generations for mutations. So the total number of mutations segregating in our sample S is Poisson with mean μT_{tot}. Thus the expected number of segregating in history a sample of size n is

$$E(S) = \mu E(T_{tot}) = \sum_{i=1}^{n-1} \frac{4N\mu}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}.$$

Thus we can use this formula to derive another estimate of the population scaled mutation rate, by setting our observed number of segregating sites in a sample S equal to this expectation.

1.2.5 The fixation of neutral alleles

It is very unlikely that a rare neutral allele accidentally drifts up to fixation; more likely, such an allele will be eventually lost from the population. However, populations experience a large and constant influx of rare alleles due to mutation, so even if it is very unlikely that an individual allele fixes within the population, some neutral alleles will fix.

An allele which reaches fixation within a population is an ancestor to the entire population. In a particular generation there can be only single allele that all other alleles at the locus in later generation can claim as an ancestor. A neutral locus, the actual allele does not affect the number of descendents that the allele has (this follows from the definition of neutrality: neutral alleles don't

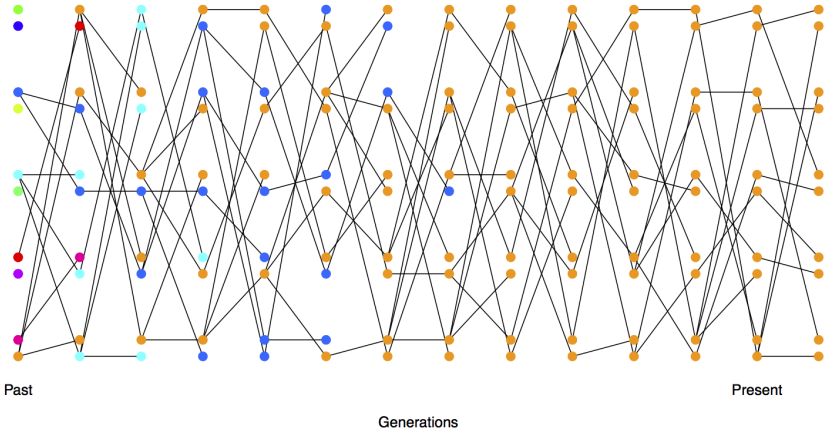


Fig. 1.6 Each allele initially present in a small diploid population is given a different colour so we can track their descendants over time. By the 9th generation all of the alleles present in the population can trace their ancestry back to the orange allele.

leave more or less descendents on average). An equivalent way to state this is that the allele labels don't affect anything; thus the alleles are exchangeable. As a consequence of this, any allele is equally likely to be the ancestor of the entire population. In a diploid population size of size N, there are $2N$ alleles all of which are equally likely to be the ancestor of the entire population at some later time point. So if our allele is present in a single copy, the chance that it is the ancestor to the entire population in some future generation is $\frac{1}{2N}$.

An allele newly arisen mutation only becomes a fixed difference if it is lucky enough to be the ancestor of the entire population. How long does is take on average for such an allele to fix within our population? We've seen that it takes $4N$ generations for a large sample of alleles to all trace their ancestry back to a single most recent common ancestor. Thus it must take roughly $4N$ generations for a neutral allele present in a single copy within the population to the ancestor of all alleles within our population. This argument can be made more precise, but in general we would still find that it takes $\approx 4N$ generations for a neutral allele to go from its introduction to fixation with the population.

1.3 Introducing selection effects

Natural selection occurs when there are differences between individuals in fitness. We may define fitness in various ways. Most commonly, it is defined with respect to the contribution of a phenotype or genotype to the next generation. Differences in fitness can arise at any point during the life cycle. For instance, different genotypes or phenotypes may have different survival probabilities from one stage in their life to the stage of reproduction (viability),

or they may differ in the number of offspring produced (fertility), or both. Here, we define the absolute fitness of a genotype as the expected number of offspring of an individual of that genotype.

1.3.1 Haploid selection model

We start out by modelling selection in a haploid model, as this is mathematically relatively simple. Let the number of individuals carrying alleles A_1 and A_2 in generation t be P_t and Q_t. Then, the relative frequencies at time t of alleles A_1 and A_2 are $p_t = P_t/(P_t + Q_t)$ and $q_t = Q_t/(P_t + Q_t) = 1 - p_t$. Further, assume that individuals of type A_1 and A_2 on average produce W_1 and W_2 offspring individuals, respectively. We call W_i the absolute fitness. Therefore, in the next generation, the absolute number of carriers of A_1 and A_2 are $P_{t+1} = W_1 P_t$ and $Q_{t+1} = W_2 Q_t$, respectively. The mean absolute fitness of the population at time t is

$$\overline{W_t} = W_1 p_t + W_2 q_t,$$

i.e. the sum of the fitness of the two types weighted by their relative frequencies. Note that the mean fitness depends on time, as it is a function of the allele frequencies, which are themselves time dependent.

The frequency of allele A_1 in the next generation is then given by

$$p_{t+1} = \frac{P_{t+1}}{P_{t+1} + Q_{t+1}} = \frac{W_1}{\overline{W_t}} p_t.$$

Importantly, this equation tells us that the change in p only depends on a ratio of fitnesses. Therefore, we need to specify fitness only up to an arbitrary constant. As long as we multiply all fitnesses by the same value, that constant will cancel out and the equation will hold. Based on this argument, it is very common to scale absolute fitnesses by the absolute fitness of one of the genotypes, e.g. the most or the least fit genotype, to obtain relative fitnesses. Here, we will use w_i for the relative fitness of genotype i. If we choose to scale by the absolute fitness of genotype A_1, we obtain the relative fitnesses $w_1 = W_1/W_1 = 1$ and $w_2 = W_2/W_1$.

Without loss of generality, we can therefore write

$$p_{t+1} = \frac{w_1}{\overline{w}} p_t,$$

dropping the dependence of the mean fitness on time in our notation, but remembering it. The change in frequency from one generation to the next is then given by

$$\Delta p_t = p_{t+1} - p_t = \frac{w_1 p_t}{\overline{w}} - p_t = \frac{w_1 - w_2}{\overline{w}} p_t q_t.$$

Assuming that the fitnesses of the two alleles are constant over time, the number of the two allelic types τ generations after time

t are $P_{t+\tau} = W_1^\tau P_t$ and $Q_{t+\tau} = W_2^\tau Q_t$, respectively. Therefore, the relative frequency of allele A_1 after τ generations past t is

$$p_{t+\tau} = \frac{p_t}{p_t + (w_2/w_1)^\tau q_t}.$$

Rearranging the equation and setting $t = 0$, we can work out the time τ for the frequency of A_1 to change from p_0 to p_τ. First, we write

$$p_\tau = \frac{p_0}{p_0 + (w_2/w_1)^\tau q_0}$$

and rearrange to obtain

$$\frac{p_\tau}{q_\tau} = \frac{p_0}{q_0}(\frac{w_1}{w_2})^\tau.$$

Solving for τ gives

$$\tau = \log(\frac{p_\tau q_0}{q\tau p_0}) / \log(\frac{w_1}{w_2}).$$

In practice, it is often helpful to parametrize the relative fitnesses w_i in a specific way. For example, we may set $w_1 = 1$ and $w_2 = 1 - s$, where s is called the selection coefficient. Using this parametrization, s is simply the difference in relative fitnesses between the two alleles. The equation becomes

$$p_{t+\tau} = \frac{p_t}{p_t + q_t(1-s)^\tau}.$$

If $s \ll 1$ (weak selection), then we can approximate

$$p_{t+\tau} = \frac{p_t}{p_t + q_t e^{-s\tau}}.$$

This equation takes the form of a logistic function. That is because we are looking at the relative frequencies of two 'populations' (of alleles A_1 and A_2) that are growing (or declining) exponentially, under the constraint that $p+q = 1$. Moreover, the equation for the time τ it takes for a certain change in frequency to occur becomes

$$\tau = -\log(\frac{p_\tau q_0}{q\tau p_0}) / \log(1-s).$$

If we assume again $s \ll 1$, we can write

$$\tau \approx \frac{1}{s}\log(\frac{p_\tau q_0}{q\tau p_0}).$$

One particular case of interest is the time it takes to go from an absolute frequency of $1/N$ to near fixation in a population of size N. In this case, we have $p_0 = 1/N$, and we may set $p_\tau = 1 - 1/N$, which is very close to fixation. Then, plugging these values, we obtain

$$\tau \approx \frac{2}{s}\log(N).$$

1.4 Interplay between selection and genetic drift

1.4.1 Stochastic loss of strongly selected alleles

Even strongly selected alleles can be lost from the population when they are sufficiently rare. This is because the number of offspring left by individuals to the next generation is fundamentally stochastic. A selection coefficient of $s = 1\%$ is a strong selection coefficient, which can drive an allele through the population in a few hundred generations once the allele is established. However, if individuals have on average a small number of offspring per generation the first individual to carry our allele who has on average 1% more children could easily have zero offspring, leading to the loss of our allele before it ever get a chance to spread. To take a first stab at this problem let's think of a very large haploid population, and in order for this population to stay constant in size we'll assume that individuals without the selected mutation have on average one offspring per generation. While individuals with our selected allele have on average $1 + s$ offspring per generation. We'll assume that the distribution of offspring number of an individual is Poisson distributed with this mean, i.e. the probability that an individual with the selected allele has i children is

$$P_i = \frac{(1+s)^i e^{-(1+s)}}{i!}.$$

Consider starting from a single individual with the selected allele, and ask about the probability of eventual loss P_L of our selected allele starting from this single copy. To derive this we'll make use of a simple argument (derived from branching processes). Our selected allele will be eventually lost from the population if every individual with the allele fails to leave descendants.

(1) In the first generation with probability P_0 our individual leaves no copies of itself to the next generation, in which case our allele is lost

(2) Alternatively it could leave one copy of itself to the next generation (with probability P_1), in which case with probability P_L this copy eventually goes extinct

(3) It could leave two copies of itself to the next generation (with probability P_2), in which case with probability P_L^2 both of these copies eventually goes extinct

(4) More generally it could leave could leave k copies ($k > 0$) of itself to the next generation (with probability P_k), in which case with probability P_L^k all of these copies eventually go extinct

summing over these probabilities we get

$$P_L = \sum_{k=0}^{\infty} P_k P_L^k = e^{-(1+s)} \left(\sum_{k=0}^{\infty} \frac{(P_L(1+s))^k}{k!} \right) = e^{(1+s)(P_L-1)}$$

The probability of escaping loss $P_F = 1 - P_L$ then satisfies

$$1 - P_F = e^{-P_F(1+s)}.$$

If we consider a small selection coefficient $s \ll 1$ such that $P_F \ll 1$ and expanded the exponential we obtain

$$1 - P_F = 1 - P_F(1+s) + P_F^2(1+s)^2/2$$

which results into $P_F = 2s$. Thus even an allele with a 1% selection coefficient has a 98% probability of being lost when it is first introduced into the population by mutation.

We can also adapt this result to a diploid setting. Assuming that heterozygotes for the 1 allele have $1 + (1-h)s$ children, the probability of allele 1 is not lost, starting from a single copy in the population, is

$$P_F = 2(1-h)s$$

for $h > 0$.

1.4.2 The interaction between genetic drift and weak selection

For strongly selected alleles, once the allele has escaped initial loss at low frequencies, their path will be determined deterministically by their selection coefficients. However, if selection is weak the stochasticity of reproduction can play a role in the trajectory an allele takes even when it is common in the population. To see this lets think of our simple Wright-Fisher model. Each generation we allow a deterministic change in our allele frequency, and then binomially sample two alleles for each of our offspring to construct our next generation. So the expected change in our allele frequency within a generation is given just by our deterministic formula. To make things easy on our self lets assume an additive model, i.e., $h = 1/2$, and that $s \ll 1$ so that $\overline{w} \approx 1$. This gives us

$$E(\Delta p) = \frac{s}{2} p(1-p)$$

while our variance in allele frequency is given by

$$Var(\Delta p) = Var(p' - p) = Var(p') = \frac{p'(1-p')}{2N} \approx \frac{p(1-p)}{2N}$$

this variance in our allele frequency follows from the fact that we are binomially sampling $2N$ new alleles in the next generation from a frequency p' and assuming that $p' \approx p$, since $s \ll 1$.

To get our first look at the relative effects of selection vs drift we can simply look at when our change in allele frequency caused selection within a generate is reasonably faithfully passed across the generations. In particular if our expected change in frequency is much great than the variance around this change, genetic drift will play little role in the fate of our selected allele (once the allele is not too rare within the population). When does selection dominates over genetic drift? This will happen if $E(\Delta p) \gg Var(\Delta p)$ when $Ns \gg 1$. Conversely any hope of our selected allele following its deterministic path will be quickly undone if our change in allele frequencies due to selection is much less than the variance induced by drift. So if $Ns \ll 1$ then drift will dominate the fate of our allele.

To make further progress on understanding the fate of alleles with selection coefficients of the order $1/N$ requires more careful modeling. However, we can obtain the probability that under our diploid model, with an additive selection coefficient s, the probability of allele 1 fixing within the population starting from a frequency p is given by

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}.$$

(try to prove this as an exercise). A new allele will arrive in the population at frequency $p = 1/(2N)$, then its probability of reaching fixation is

$$\pi(1/2N) = \frac{1 - e^{-s}}{1 - e^{-2Ns}}.$$

if $s \ll 1$ but $Ns \gg 1$, the $\pi(1/2N) \approx s$, which nicely gives us back our result that we obtained above. To recover our neutral result we can take the limit $s \to 0$ to obtain our neutral fixation probability $1/2N$.

1.4.3 The fixation of slightly deleterious alleles

We can see that weakly deleterious alleles can fix, especially in small populations. To understand how likely it is that deleterious alleles accidently reach fixation by genetic drift, lets assume a diploid model with additive selection (with a selection coefficient of $-s$ against our allele 2). If $Ns \gg 1$ then our deleterious allele (allele 2) can not possibly reach fixation. However, if Ns is not large then

$$\pi(1/2N) \approx \frac{s}{e^{2Ns} - 1}$$

1.5 Beyond well-mixed populations: the role of space

In many realistic scenarios, a population is not well-mixed, where each individual feels the same environmental conditions, but it's

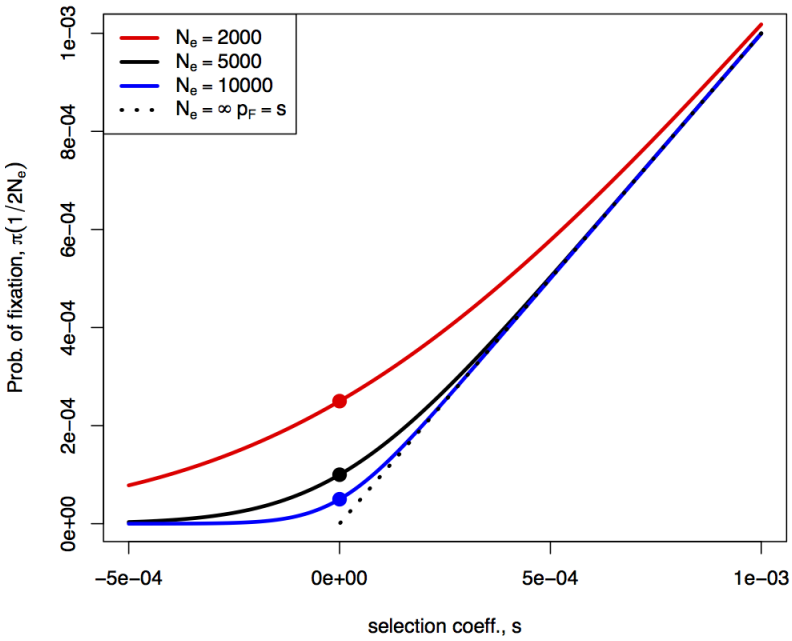


Fig. 1.7 The probability of the fixation of a new mutation with selection coefficient s ($h = 1/2$) in a diploid population of effective size N_e. The dashed line gives the infinite population solution. The dots give the solution for $s \to 0$, i.e. $1/2N_e$.

distributed in space. In the following, we will present different models that have tried to deal with the issue of space and how to include it when describing evolutionary dynamics.

1.5.1 Island model

The most classical example of spatially distributed population is the so-called "island model". The model is based on dividing the population into subpopulations that live in different "islands" or spatially distinct locations. Often some degree of migration is allowed between islands so that subpopulations can exchange individuals. In these models, often the islands have different conditions, therefore mutations that are neutral or beneficial in some islands can be deleterious in others. Because of migration there can be a constant influx of deleterious mutations creating a migration-selection balance which is very similar to the mutation-selection balance. In these models, often migration between islands can be treated very similarly to mutations within one population.

1.5.2 Some theory of spatial distribution of alleles frequencies under deterministic models of selection

Imagine a continuous haploid population spread out along a line. An individual disperses a random distance Δx from its birthplace to the location where it reproduces, where Δx is drawn from the probability density $g()$. To make life simple we will assume that $g(\Delta x)$ is normally distributed with mean zero and standard deviation σ i.e. migration is unbiased an individuals migrate an average distance of σ.

The frequency of allele 2 at time t in the population at spatial location x is $q(x,t)$. Assuming that only dispersal occurs, how does our allele frequency change in the next generation? Our allele frequency in the next generation at location x reflects the migration from different locations in the proceeding generation. Our population at location x receives a contribution $g(\Delta x)q(x+\Delta x,t)$ of allele 2 from the population at location $x+\Delta x$, such that the frequency of our allele at x in the next generation is

$$q(x,t+1) = \int_{-\infty}^{\infty} g(\Delta x)q(x+\Delta x,t)d\Delta x.$$

To obtain $q(x+\Delta x,t)$, let's take a Taylor expansion

$$q(x+\Delta x,t) = q(x,t) + \Delta x\partial_x q|_{x,t} + 1/2\Delta x^2\partial_x^2 q|_{x,t} + ...$$

then

$$q(x,t+1) = q(x,t) + \partial_x q|_{x,t}\int_{-\infty}^{\infty}\Delta x g(\Delta x)d\Delta x + 1/2\partial_x^2 q|_{x,t}\int_{-\infty}^{\infty}\Delta x^2 g(\Delta x)d\Delta x$$

Remembering that $g()$ has zero mean and variance σ^2, we find that

$$q(x,t+1) = q(x,t) + \frac{\sigma^2}{2}\partial_x^2 q.$$

In the limit of continuous time, we get

$$\partial_t q = \frac{\sigma^2}{2}\partial_x^2 q$$

This is a diffusion equation, so that migration is acting to smooth out allele frequency differences with a diffusion constant of $D = \sigma^2/2$. This is exactly analogous to the equation describing how a gas diffuses out to equal density, as both particles in a gas and our individuals of type 2 are performing Brownian motion (blurring our eyes and seeing time as continuous)

We will now introduce fitness differences into our model and set the relative fitnesses of allele 1 and 2 at location x to be 1 and $1+s\gamma(x)$, respectively. To make analytical progress in this model we'll have to assume that selection isn't too strong i.e. $s\gamma(x) \ll 1$

for all x. The change in frequency of allele 2 obtained within a generation due to selection is

$$q'(x,t) - q(x,t) \approx s\gamma(x)q(x,t)(1-q(x,t)),$$

which described logistic growth of the favoured allele. Putting our selection and migration together we find

$$\partial_t q(x,t) = s\gamma(x)q(x,t)(1-q(x,t)) + D\partial_x^2 q(x,t).$$

This is a reaction-diffusion equation, which describes the spreading of a beneficial allele in a population. If $\gamma(x) = 1$, this is known as the Fisher-Kolmogorov-Petrovski-Piskinov (FKPP) deterministic equation. Importantly, this equation is satisfied by a travelling wave solution of characteristic speed $v_F = 2\sqrt{D/s}$, which is called the Fisher velocity.

1.5.3 Emergence of resistance in an antibiotic gradient

We will now use the determinstic FKPP equation to derive the probability of resistant mutations arising in an antibiotic gradient (for exersice compare this probability to the well-mixed case). In class, we saw an experiment on *E. coli* growing over a large agar plate made of slabs with step-wise increasing antibiotic concentration. Let's focus on one particular step of the plate and determine the probability that the wild-type population will be able to develop resistance to the next antibiotic concentration.

We assume that the wild-type population density $c(x,t)$ (rescaled by the carrying capacity K) is described by the reaction-diffusion equation

$$\partial_t c(x,t) = D\partial_x^2 c + s_{WT}(x)c - a_{WT}(x)c^2, \tag{1.1}$$

where $a_{WT}(x)$ is the local wild type birth rate, b_{Wt} is the local antibiotic-induced death rate, $s_{WT}(x) = a_{WT}(x) - b_{WT}(x)$ is the local net growth rate of the wild type.

The model ensures that the steady-state local population density c_{SS} depends explicitly on the local death rate $b_{WT}(x)$ when the death and birth rate profiles change sufficiently slowly in space,

$$c_{SS}(x) = 1 - \frac{b_{WT}(x)}{a_{WT}(x)}. \tag{1.2}$$

Given a single resistance mutation arises in the population at position x, its probability $u(x,t)$ to survive for a time t can be derived by modeling the mutant lineage as a branching random walk.

Let $u_x(t)$ denote the probability that a mutation born at lattice site x survives for time t. Denote by $a(x)$ and $b(x)$ the local birth and death rate, respectively, and let D be the migration rate over

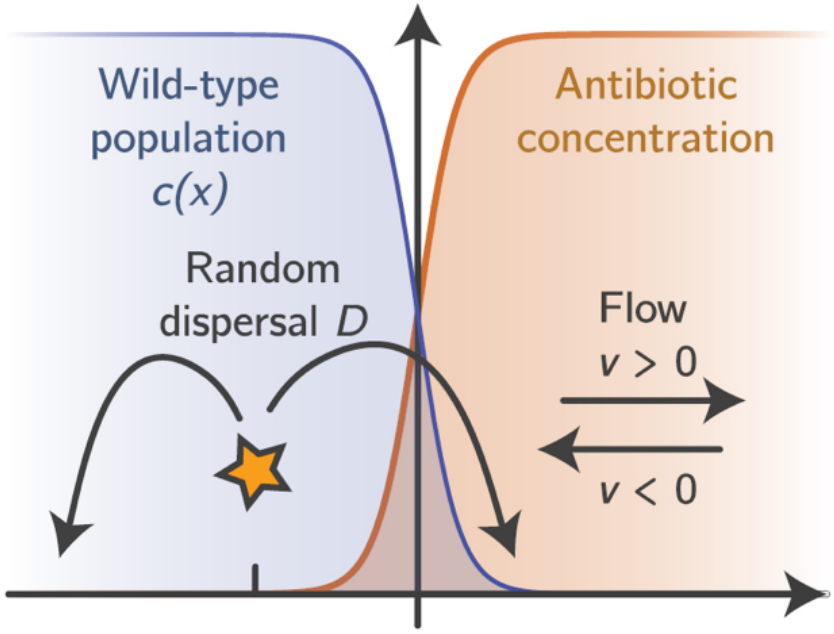


Fig. 1.8 Representation of an antibiotic gradient and the corresponding steady-state wild-type population ($c(x)$). Mutations that confer resistance can occur and generate a mutant population able to live in the antibiotic region. The populations can in principle also be subject to flow.

a distance δx (i.e., one lattice site). Then, the $u(x,t)$ after a short time interval ϵ satisfies the equation

$$u_x(t+\epsilon) = \epsilon a(x)\left\{1-[1-u_x(t)]^2\right\} + \{1-\epsilon\,[a(x)+b(x)]\}\,u_x(t) + \epsilon D\left\{u_{x+\delta x}(t)+u_{x-\delta x}(t)-2u_x(t)\right\}. \tag{1.3}$$

The first term on the right-hand side accounts for the fact that when the initial mutant divides then there are two mutants, and the probability of survival of at least one lineage is 1 minus the square of both lineages disappearing. The second term describes the case of nothing happening in the time interval ϵ.

Letting $\epsilon \to 0$ and performing the Taylor expansion in δx similarly to what we did to derive the FKPP, we obtain

$$\partial_t u = D\partial_x^2 u + s(x)u - a(x)u^2, \tag{1.4}$$

where $s(x) = a(x) - b(x)$.

We assume that the birth rate of wild-type and mutant is identical and constant, $a_{WT}(x) = a_{MT}(x) = a_0$, while the wild-type drug-induced death rate $b_{WT}(x)$ ranges from $-a_0$ to a_0. This implies that effect of the antibiotic is to increase the death rate of the wild type, while the drug-induced death rate of the resistant mutant is zero. The effective growth rate of the mutants is thus determined purely through competition with the wild type, i.e., $s_{\mathrm{MT}}(x) = a_0[1-c(x)]$.

A step-like increase in concentration at $x = 0$ gives rise a net growth rate of a_0 for $x < 0$ and $-a_0$ for $x > 0$, i.e.,

$$s_{WT}(x) = a_0 \left[1 - 2\Theta(\xi)\right]. \tag{1.5}$$

Such a sharp gradient could emerge, for instance, at the boundary of different tissues or organs with different affinities to store antibiotics. Upon rescaling the spatial coordinate by the characteristic length scale $\ell = \sqrt{D/a_0}$, which can be intuitively understood as the typical distance that a mutant individual travels through random dispersal before replicating, the wild-type population density in this case becomes

$$0 = \partial_\xi^2 c + [1 - 2\Theta(\xi)]c - c^2 \tag{1.6}$$

where $\xi = x/\ell$.

For both $\xi > 0$ and $\xi < 0$, this equation can be solved by a mechanical analogy with a particle in the "potential" $U(c) = \pm\frac{c^2}{2} - \frac{c^3}{3}$. For $\xi < 0$, we have the potential $U(c) = c^2/2 - c^3/3$ and the boundary condition $c(-\infty) = 1$ and $c(0) = c_0$, which gives a total energy $E = K + U = 1/6$ because the kinetic energy $K = 0$ at $-\infty$. The population density $c(\xi)$ is then determined through the integral

$$\xi = \int_{c_0}^{c(\xi)} \frac{dc'}{\sqrt{-2U(c') + 2E}} = \int_{c_0}^{c(\xi)} \frac{dc'}{\sqrt{-c^2 + c^3/3 + 1/3}}. \tag{1.7}$$

Similarly, for $\xi > 0$, we have $U(c) = -c^2/2 - c^3/3$ and $E = 0$, and $c(\xi)$ is determined through the integral

$$\xi = \int_{c_0}^{c(\xi)} \frac{dc'}{\sqrt{c^2 + c^3/3}}. \tag{1.8}$$

Both integrals can be solved exactly, and the derivatives matched at $\xi = 0$. The result is

$$c(\xi) = \begin{cases} \frac{3}{2}\tanh\left[\frac{\xi - \xi_-}{2}\right]^2 - \frac{1}{2}, & \xi < 0, \\ \frac{3}{2}\tanh\left[\frac{\xi + \xi_+}{2}\right]^2 - \frac{3}{2}, & \xi \geq 0, \end{cases} \tag{1.9}$$

where $\xi_\pm = 2\text{arctanh}(\frac{1}{3}\sqrt{6 \pm 3 + \sqrt{6}})$.

The population density transitions from 1 to 0 exponentially fast, and we can approximate $c(\xi) \approx \Theta(-\xi)$ when computing the establishment probability, which then approximately obeys the equation

$$0 = \partial_\xi^2 u + \Theta(\xi)u - u^2. \tag{1.10}$$

Using the same mechanical analogy as above, and again matching derivatives at $\xi = 0$, we find

$$u(\xi) = \begin{cases} \frac{1/\sqrt{3}}{\left(1 - \xi/\sqrt{6\sqrt{3}}\right)^2}, & \xi < 0, \\ \frac{3}{2}\tanh\left[\frac{\xi + \xi_u}{2}\right]^2 - \frac{1}{2}, & \xi \geq 0, \end{cases} \tag{1.11}$$

where $\xi_u = 2\text{arctanh}(\frac{1}{3}\sqrt{3 + 2\sqrt{6}})$.

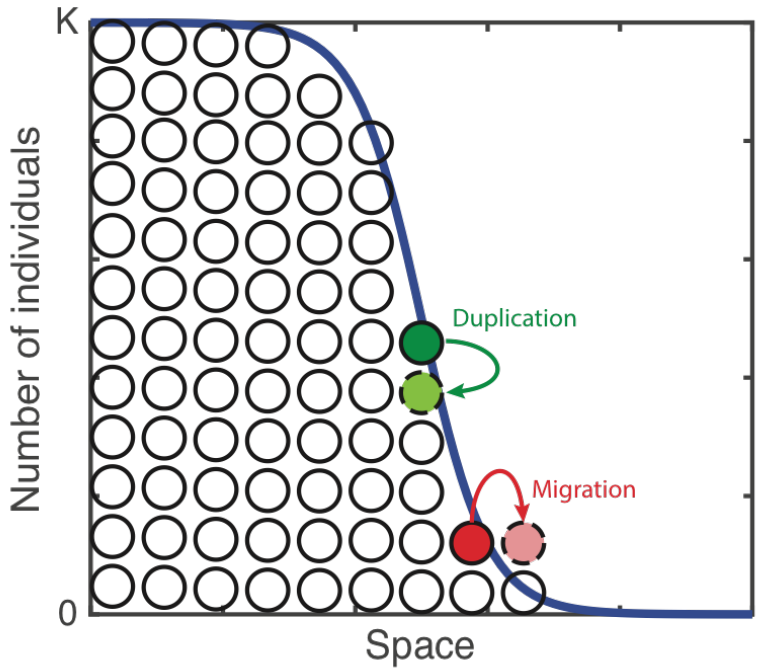


Fig. 1.9 Representation of a stochastic simulation that mimics the growth of a population in 1 dimensional space.

1.5.4 Stochasticity in the FKPP equation

The sections above removed any stochasticity from the FKPP equation. However, genetic drift equally acts in spatial models as it did in well-mixed conditions. In particular, since in the FKPP equation the dynamics is dominated by the individuals at the front of the wave, the corresponding effective population size N_e can be very small and genetic fluctuations very high. Indeed, in these conditions, stochastic effects can be so strong that selection becomes very inefficient and deleterious mutations may, by chance, accumulate in the population, with a phenomenon called expansion load. To better understand this, let's look at the stochastic nature of the FKPP equation by imagining to simulate it.

The FKPP can be used to model the expansion of a population in virgin territory (see Fig. 1.9). At each point in time and space, the population density is defined by $q(x,t)$. At each step in our simulation, one individual is selected and can replicate. A generation will occur once every individual at position x is given a chance to replicate, therefore 1 generation consists of N steps. Or, in other words, 1 simulation step happens at a rate $1/N$. Because of logistic growth, replication occurs if we pick a random individual *and* a random free position (see Fig. 1.9). This means

that at time $t + dt$

$$q(t+dt) = \begin{cases} q+1 & \text{with prob. } \frac{dt}{N} q(N-q) \\ q & \text{with prob. } 1 - \frac{dt}{N} q(N-q) \end{cases}$$

Therefore, at position x,

$$\langle q(x, t+dt) \rangle = q(x,t) + \frac{dt}{N} q(x,t)(N - q(x,t))$$

and variance

$$Var(q(x, t+dt) = \frac{dt}{N} q(x,t)(N - q(x,t))$$

Defining the white noise R_t such that $\langle R_t \rangle = 0$ and $\langle R_t^2 \rangle = 1$, we can write for position x

$$q(x, t+dt) = q(x,t) + \frac{dt}{N} q(x,t)(N - q(x,t)) + R_t \sqrt{\frac{dt}{N} q(x,t)(N - q(x,t))}$$

In the limit of $dt \to 0$, at position x

$$\partial_t q = \frac{q(N-q)}{N} + \eta_t \sqrt{\frac{q(N-q)}{N}}$$

where $\langle \eta_t \eta_{t'} \rangle = \delta(t - t')$. Adding the diffusion term and defining $c = q/N$ as the population density, we get the stochastic FKPP equation:

$$\partial_t c = \partial_x^2 c + c(1-c) + \eta_t \sqrt{\frac{c(1-c)}{N}}$$

.

From this equation, you can see that at the very front of the wave, where we have only one individual, $c \approx 1/N$, and therefore the noise term and the growth term are both $\mathcal{O}(1/N)$ and fluctuations cannot be neglected.

One of the consequences of these large fluctuations is that, similarly to the well-mixed case, weak selection effects can behave almost neutrally so that weakly beneficial mutations may easily get lost and weakly deleterious mutations may instead accumulate at the front of the wave.

1.6 Ecological scales: multispecies evolution

Credit: Models of Life by Kim Sneppen

If we look at evolution over much larger timescales, we often observe signs of cooperative behavior, where many species have often been replaced almost "simultaneously". To model macro-evolutionary patterns we start with units on the size of the main

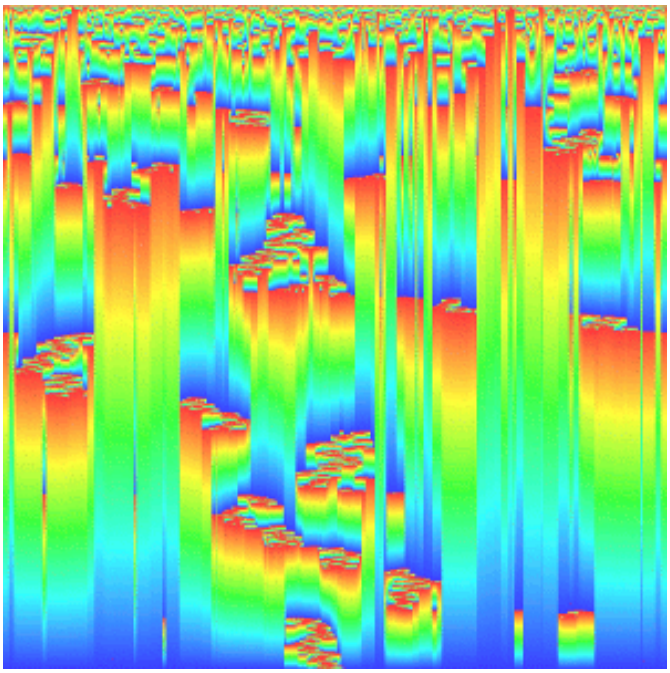


Fig. 1.10 Sample of Bak–Sneppen model evolution: the x-axis represent the position of the different species in the ecosystem and on the y-axis (from top to the bottom) we see the history of the population. Each discontinuity represents an evolution. The color codes the age of the species (red=origin, blue=extinction). (From Wikipedia) Note the spatial clustering that naturally emerges along the simulation.

players at this scale: species. A species consists of many individuals, and the dynamics of a species represent the coarse-grained view of the dynamics of these entire populations. Whereas population dynamics may be influenced by fitness, species dynamics are governed by stabilities. We characterize a species by a number B_i that specifies its stability on very long timescales. An ecosystem of species then consistst of N numbers B_i, each representing a different species, with links that could be predation, collaboration or niche maintanance.

A simple model that describes the dynamics of these species is the Bak-Sneppen model. Let's assume that the numbers B_i with $i = 1, 2, ..., N$ are placed on a line, mimicking a one dimensional model ecosystem. At each time step, one changes the least stable of these species. The fitness of a given species if a function of the species it interacts with, and accordingly, the neighboring species will also change their stabilities. We will use the following update rule: at each step, the smallest $\{B_i\}_{i=1,N}$ are located. Its value and those of its nearest neighbors are replaces by new random numbers in $[0, 1]$.

This is the simplest model that exhibits a phenomenon called self-organized criticality. As the system evolves, the smallest of B_i is eliminated and after a transient period, a statistically stationary

distribution of B is obtained. For a system $N \to \infty$, this distribution is a step-function, where the selected minimum B_{min} is always below a critical B_c, and the distribution is constant above B_c. The value of B_c depends on dimensionality and the details of the update move: for $d = 1$ and two nearest neighbors update, $B_c = 0.667$.

Importantly, over the course of evolution, active sites tend to be localized in the same region of the model ecosystem (Fig. 1.10. Thus evolution is reinforced locally, bridging punctuated equilibrium in single-species evolution to larger quantum evolution and origination of new taxonomic groups.