Mathematical Population Genetics

Lecture Notes

Joachim Hermisson

January 10, 2018

University of Vienna
Mathematics Department
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria

These lecture notes are based on previous course material from the "Population Genetics Tutorial" by Peter Pfaffelhuber, Pleuni Pennings, and JH (2009), and from the lecture "Mathematical Population Genetics" by Reinhard Bürger and JH (2015).

Literature

- Reinhard Bürger (2000). The Mathematical Theory of Selection, Recombination, and Mutation. Wiley, New York. *The standard textbook for deterministic modelling approaches in population genetics.*
- Warren J. Ewens (2004). Mathematical Population Genetics. Springer, New York. *The standard textbook for stochsastic modelling approaches in population genetics.*
- Sean H. Rice (2004). Evolutionary Theory. Sinauer, Sunderland. *Well-written didactic introduction into the field.*
- John Wakeley (2009). Coalescent Theory. Roberts & Co., Greenwood Village. *Comprehensive introduction into the coalescent.*
- Thomas Nagylaki (1992). Introduction to Theortical Population Genetics. *Classical textbook with many worked examples.*
- Linda J. S. Allen (2011). An Introduction to Stochastic Processes with Applications to Biology. Taylor & Francis, Boca Raton.
- Brian Charlesworth and Deborah Charlesworth (2010). Elements of Evolutionary Genetics. Roberts & Co., Greenwood Village.
- Sarah P. Otto and Troy Day (2007). A Biologist's Guide to Mathematical Modeling. Princeton University Press, Princeton.

What is population genetics? – Basic concepts and definitions

Evolution describes the change in heritable characteristics of biological populations over time. Depending on the type of these characteristics, and depending on the time-scale of interest, we can distinguish different branches of evolutionary research.

- *Phylogenetics* is concerned with the construction of the *tree of life*, following Darwin's insight that all life on earth (and the fossil record) is connected via common ancestors. Changes in traits and characteristics among species, or the emergence of new traits happen over macroevolutionary time-scales, millions or billions of years. Differences among individuals within each species can usually be ignored relative to the differences among species. Each species is therefore generally only represented by a single data point, such as a consensus DNA sequence ("the" human genome).
- *Population genetics* and *quantitative genetics* are interested in the microevolutionary process within a population. Microevolution is concerned with heritable characteristics that differ among individuals in a population and describes how the distribution of these characteristics changes across generations. Going back to Darwin (again) and to Wallace, the elementary forces that drive these changes are well-understood:

mutation, selection, recombination, genetic drift, and gene flow / migration. In contrast to phylogenetics, which basically is a historical science, microevolution has a mechanistic basis that can be used to construct theoretical models and to make predictions for the future.

Genotype and phenotype

Each individual in a population can be characterized by a large number of morphological, physiological, and behavioral traits, which collectively define its *phenotype*. Individual phenotypes may be more or less adapted to the environmental conditions and influence the viability or reproductive success of its carrier. As a consequence, selection operates on phenotypes. Phenotypes themselves are not inherited, but phenotypic characteristics, such as body size, are influenced by heritable and non-heritable factors.

The part of an individual that is directly heritable is its *genotype*. The genotype of each individual is largely encoded in its *genome* and represented by its DNA (DeoxyriboNucleic Acid) sequence. DNA is a polymer consisting of four types of nucleotides, which differ in the *base* that they contain: adenine (A), guanine (G), thymine (T), and cytocine (C). The nucleotides are organized in two polynucleotide chains that form a double-helix with A-T and G-C base pairings. In eukaryotic cells (animals, plants, fungi), the cell nucleus contains several such DNA double strands, called *chromosomes*. In prokaryotic cells (bacteria and archea), DNA typically forms a single ring (bacterial chromosome). Through development, the genotype determines (the heritable part of) the phenotype, but the connection is exceedingly complex for most phenotypic traits. The genotype naturally decomposes into *genes*, functional units of DNA that encode a single protein. Quantitative traits of interest (such as milk yield in cows) are usually affected by a large number of genes.

Due to the complexity of the genotype-phenotype map, all models of (micro)evolution must rely on simplifying assumptions. Models of quantitative genetics rely on phenotypic data, but often do not resolve individual genes. They rather infer heritable and non-heritable parts of phenotypic traits directly from trait measurements across generations. On the other hand, models of population genetics directly follow the frequencies of genotypes and variants of genes in a population. They often do not refer to phenotypes at all, but assume that selection acts directly on the genes, no matter where the selection pressure comes from and how it is transmitted via the genotype-phenotype map.

Genes, loci, and alleles

Population genetics is concerned with the evolutionary dynamics of genotypes. It follows the frequencies of genetic variants or *alleles* that differ between individuals. The complete genotype of each individual is given by its DNA sequence (≈ 3 billion base pairs in the human genome, ≈ 130 million in *Drosophila*). Usually however, one is only interested in changes at certain aspects of the genotype, such as the genomic positions, or *genetic loci*, that affect a phenotypic trait of interest. On the molecular level, a locus is the position of a single base in the DNA. There are four alleles, corresponding to the four different bases, A, T, G, and C. Frequently, however, the term locus is used on a coarser level as the position of a gene (or some other significant stretch of sequence) . It is always assumed that a locus is a "unit of recombination" that is not broken up during reproduction. There can be many different alleles at a single locus (4^n different alleles for a gene that is represented by a DNA sequence of fixed length n), but usually one considers classes of equivalent alleles. Many population genetic models only distinguish two classes: an ancestral *wildtype* and a *mutant* allele.

Genetic loci can have different levels of *ploidy*. Most simple life forms (bacteria, mosses, algae, fungi) have a single copy of each chromosome, they are *haploid*. For haploids, a single-locus genotype is determined by a single allele. Almost all higher plants and animals are *diploid*, i.e., most of their chromosomes (the so-called *autosomes* = non-sex-chromosomes) are represented twice in each adult cell. Some organisms (mostly plants) have even a higher ploidy level (e.g., *tetraploid* with a 4-fold set). Consequently, single-locus genotypes in diploids are given by a pair of alleles (4 alleles in tetraploids, etc).

Mathematical methods

The art of mathematical modeling is to choose the appropriate mathematical methods to address a scientific question. Since population genetics is concerned with the change of allele frequencies as a function of time, natural mathematical methods come from fields that describe such processes. Often, the main decision for a given problem is to decide whether a deterministic or a stochastic framework is appropriate.

- Deterministic models in population genetics use methods from the theory of dynamical systems and of differential equations (both ODE's and PDE's). On the biological side, this is appropriate if stochastic effects due to a finite population size (genetic drift) can be ignored. this is usually the case if selection is the dominant population genetic force and if the total number of individuals that carry a certain allele is not very small. The dynamics can be modeled in discrete time (using discrete dynamical systems) if a generation is a natural time unit in the biological system, like in annual plants. In other cases, a continuous-time dynamics (building of differential equations) is more appropriate and/or more convenient.

- If genetic drift has a strong effect on the evolutionary process, stochastic models are needed. Basically all these models build on Markov processes, typical examples being birth-death processes or branching processes. As in the deterministic case, they can proceed in discrete or in continuous time. Coalescent theory, in particular, is a stochastic process that proceeds in the reversed time direction, from the present to the past. It turns out that this is particularly appropriate if we want to explain observed patterns of diversity in data by past evolutionary processes. If population sizes are large and if selection is not too strong, allele frequencies can be treated as a continuous random variable on the unit interval. This leads to diffusion processes as a model of evolutionary change. Indeed, parts of the theory of diffusions was developed in the early 20th century with applications in population genetics in mind.

1 Selection at a single locus

1.1 Selection at a single haploid locus

Consider a haploid population of size N. We characterize the genotype by the allelic type at a single locus. There are k alleles, denoted $\{A_1, \ldots, A_k\}$. Generations are discrete and we assume that the population is sufficiently large that stochastic effects due to genetic drift can be ignored. Assume that there are initially n_i individuals with allele A_i. The frequency of A_i in the population is thus $p_i = n_i/N$. Reproduction is clonal, offspring inherit the genotype of their (single) parent, without any modification (no mutation). We are interested in the change of allele frequencies due to selection across a single generation.

Fitness

The fundamental property of individuals that leads to selection and drives adaptive evolution is their fitness. In population genetics, we assign fitness values directly to genotypes or alleles, as follows:

- The *viability* $v_i \geq 0$ measures the probability that a newborn A_i individual survives to reproductive age ($v_i = 0$ means that the individual is inviable).
- The *fecundity* $f_i \geq 0$ measures the expected number of offspring of an adult A_i individual ($f_i = 0$ means that the individual is sterile).
- Finally, the *(absolute) fitness* of allele A_i is defined as

  $$w_i = v_i \cdot f_i \,.$$

  $w_i \geq 0$ measures the expected number of offspring of a newborn A_i individual. Ignoring stochastic effects, we thus have $n_i' = w_i n_i$ for the number n_i' of A_i individuals in the next generation.

For the change in a single generation, we obtain

$$N' = \sum_i n_i' = \sum_i w_i n_i = \Big(\sum_i w_i p_i\Big) N =: \bar{w} N \tag{1.1}$$

where $\bar{w} = \sum_i p_i w_i$ is the *mean fitness* in the population. For the change in allele frequencies, the canonical selection equation for a single haploid locus reads

$$p_i' = \frac{w_i}{\bar{w}}\, p_i \quad \text{or:} \qquad \Delta p_i = p_i' - p_i = \frac{w_i - \bar{w}}{\bar{w}}\, p_i \,. \tag{1.2}$$

We see that any fitness differences among alleles that are represented in the population ($w_i \neq w_j$ for $p_i, p_j > 0$) entails evolutionary change due to selection.

For allele frequency changes across multiple generations, we need to account for the fact that absolute fitness values, as defined above, are usually not constant across generations.

Indeed, $w_i = w_i(N, \mathbf{p}, t)$ is usually not only a function of the allelic type A_i, but also of the population size N (or density), the distribution of allele frequencies $\mathbf{p} = (p_1, \ldots, p_k)$, and of generation time t.

- Imagine first that fitness does not depend on the population size (or density). We then have $n_i' = w_i(\mathbf{p}, t) n_i$ and usually obtain unlimited growth or decline of n_i over multiple generations. This is clearly unrealistic.

- Assume next that fitness does depend on density, but not on the allelic state. We then have $\bar{w}(N, \mathbf{p}, t) = w_i(N, \mathbf{p}, t) =: w(N, t)$ and thus

  $$p_i' = p_i \quad ; \quad N' = w(N, t) N \,.$$

  This means we have only changes in the population size (population dynamics), but no changes in the allele frequencies (population genetics) and thus no evolution. Pure population dynamics is a topic of theoretical ecology. With models like logistic growth ($w(N) = r - cN$), population sizes can be regulated and converge to a finite, no-zero value.

- To obtain a reasonable evolutionary model, we need to combine a model of population regulation with a model of evolutionary change. A canonical approach that is implicit to most models in population genetics is to assume that population size regulation is independent of selection. Absolute fitness values then decompose into two parts

  $$w_i(N, \mathbf{p}, t) := w(N, t) \cdot w_i(\mathbf{p}, t) \,.$$

  This leads to

  $$p_i' = \frac{w_i(N, \mathbf{p}, t)}{\bar{w}} \, p_i = \frac{w(N, t) w_i(\mathbf{p}, t)}{w(N, t) \sum_i w_i(\mathbf{p}, t)} \, p_i = \frac{w_i(\mathbf{p}, t)}{\sum_i w_i(\mathbf{p}, t)} \, p_i \tag{1.3}$$

  and the density dependence drops out. Following this idea, population genetic models usually do not work with absolute fitness values, but only the *relative fitness* values. If population size regulation is independent of selection, relative fitnesses are density independent ($w_i = w_i(\mathbf{p}, t)$). We can then ignore changes in the population size in population genetic models and only follow the dynamics of allele frequencies. Note that we use the same symbol w_i for relative fitness. Since all fitness values in the following are *relative* fitness values, this should not lead to any confusion.

- Since any factor that is common to all fitness values w_i drops out of the selection equation, relative fitness values w_i are only defined up to a constant factor. We can use this freedom to normalize the fitness of some reference allele A_1 (often: the ancestral wildtype allele) to $w_1 = 1$.

- Following these leads, the easiest model of selection results if we assume constant *relative* fitness values for all alleles, $w_i = \text{const}$. We say that selection is time homogeneous and frequency-independent. The change in p_i across t generations follows as
$$p_i(t) = \frac{n_i(t)}{N} = \frac{w_i^t\, n_i(0)}{\sum_j w_j^t\, n_j(0)} = \frac{w_i^t\, p_i(0)}{\sum_j w_j^t\, p_j(0)}\,. \tag{1.4}$$
If $w_1 > w_j,\ j \geq 2$, we obtain
$$p_i(t) = \frac{p_i(0)}{\sum_j \left(\frac{w_j}{w_i}\right)^t \cdot p_j(0)} \quad \xrightarrow{t \to \infty} \quad \frac{p_i(0)}{p_1(0) \lim_{t\to\infty} \left(\frac{w_1}{w_i}\right)^t} = \delta_{1,i}\,.$$
We conclude that with time-homogeneous and frequency-independent selection in haploids only the fittest allele survives and fixes in the population. There is no genetic variation maintained.

1.2 Selection at a single diploid locus

Consider a diploid locus with two alleles (wildtype and mutant), A and a. In principle, there can be $2 \times 2 = 4$ genotypes at the locus, but if there is no *position effect* (i.e. it does not matter on which DNA strand an allele is located), there are only three: the two *homozygous* genotypes AA and aa and the *heterozygous* genotype Aa $(= aA)$. Let x, y, and z be the frequencies of genotypes AA, Aa, and aa, respectively. We can express the frequencies $p = x + y/2$ of the A allele and $q = z + y/2$ of the a allele in terms of the genotype frequencies, but note that this is generally not possible vice-versa.

Random mating and Hardy-Weinberg proportions

To describe evolutionary dynamics in diploids, even without selection, we first need a model for the change in genotype frequencies under reproduction. Most diploids reproduce sexually. Under *Mendelian inheritance*, each newborn inherits a single allele from both parents at each autosomal locus. In general, the change of genotype frequencies across generations depends on the mating pattern. For example, males and females often prefer mating partners with similar phenotypic characteristics such as body size (assortative mating). However, the simplest mating scheme that also serves that is used as default in population genetic models just assumes that matings are random. We also assume that sexes are equivalent and there are no differences in genotype frequencies among males and females in the population (this is necessarily true for monoecius species, where all individuals act in male and female roles). We can then summarize the offspring frequencies for each mating type in a table:

The third column of the table gives the probability of the mating pair under random mating and columns 4 to 6 the genotype frequencies in the offspring generation under Mendelian inheritance, conditioned on the mating pair. The total (unconditioned) genotype frequencies in the offspring generation derived by summing over all mating pairs. We observe:

♀	♂	mating prob.	x'	y'	z'
AA	AA	x^2	1	0	0
	Aa	xy	1/2	1/2	0
	aa	xz	0	1	0
Aa	AA	xy	1/2	1/2	0
	Aa	y^2	1/4	1/2	1/4
	aa	yz	0	1/2	1/2
aa	AA	xz	0	1	0
	Aa	yz	0	1/2	1/2
	aa	z^2	0	0	1

$$x' = 1 \cdot x^2 \,+\, 2\,\frac{1}{2}\,xy \,+\, \frac{1}{4}\,y^2 = \Big(x + \frac{y}{2}\Big)^2 = p^2$$

$$\begin{aligned} y' &= 2\,\frac{1}{2}\,xy \,+\, 2\,\frac{1}{2}\,yz \,+\, 2xz \,+\, \frac{1}{2}\,y^2 \\ &= 2\Big(x + \frac{y}{2}\Big)\Big(z + \frac{y}{2}\Big) = 2pq \end{aligned}$$

$$z' = 1 \cdot z^2 \,+\, 2\,\frac{1}{2}\,yz \,+\, \frac{1}{4}\,y^2 = \Big(z + \frac{y}{2}\Big)^2 = q^2$$

- The genotype frequencies after a single generation of random mating are determined by the allele frequencies, $(x', y', z') = (p^2, 2pq, q^2)$: *Hardy-Weinberg proportions.*

- The allele frequencies do not change under random mating

  $$p' = x' + \frac{1}{2}y' = p \quad ; \quad q' = z' + \frac{1}{2}y' = q\,.$$

  There is thus no loss of genetic variation under Mendelian inheritance.

- The so-called *Hardy-Weinberg law* states that, after a single generation of random mating, both the allele frequencies and the genotype frequencies remain invariant: They are in *Hardy-Weinberg equilibrium.*

- It is easy to extend the Hardy-Weinberg law to an arbitrary number of alleles $\{A_1, \ldots, A_k\}$. Let $P_{ij} = P_{ji}$ denote the frequency of the genotype A_iA_j. The allele frequency of A_i is $p_i = P_{ii} + \frac{1}{2}\sum_{j\neq i} P_{ij}$. A straight-forward extension of the 2-allele derivation shows that $p_i' = p_i$, $P_{ii}' = p_i^2$ and, for $j \neq i$, $P_{ij}' = 2p_ip_j$.

The important consequence of the Hardy-Weinberg (HW) law for population genetic models is that it is sufficient to follow k allele frequencies, rather than the $k(k+1)/2$ frequencies of diploid genotypes. However, the law is only valid under a number of assumptions.

- Random mating: with other mating schemes (e.g. assortative mating or selfing), we obtain different equilibrium frequencies *and* generally only gradual (asymptotic) convergence to this equilibrium, rather than convergence in a single generation.

- Discrete Generations: Convergence to HW proportions is only asymptotic if generations are overlapping (individuals do not all reproduce and die at the same time).

- Equivalent sexes: If the initial allele frequencies in males and females differ, HW proportions are only reached in two generations of random mating.

- Autosomal loci: For X-linked loci (that are diploid in females, but haploid in males), HW proportions are only reached asymptotically.

- No selection, mutation, or drift: all evolutionary forces readily lead to deviations from HW proportions. However, as we will see below, we can often still make use of the HW law at certain stages of a diploid *life cycle*.

Viability selection at a single diploid locus

Consider a diploid population with discrete generations and equivalent sexes and a single locus with two alleles, A and a with frequencies p and q, respectively. We also assume that selection acts on the viability, the probability that newborn diploid individuals reach reproductive age. We can then dissect the life-cycle of the population into two phases: a selection phase, during which juveniles grow up and a reproductive phase where adults mate and produce offspring. The key assumption is that selection and reproduction can be separated and occur at different stages.

- Consider the reproductive phase first. If reproduction works via random mating as described above, we can use the results of the HW law: Allele frequencies are conserved during the reproductive step and genotype frequencies will be in HW equilibrium directly after reproduction (for zygotes (= newly fertilized eukaryotic cell) not yet affected by selection).

- We still need a model for the change of allele and genotype frequencies during the reproductive phase. We assign fitness values w_{AA}, w_{Aa}, and w_{aa} to the three genotypes AA, Aa, and aa, respectively. The genotypes frequencies are P_{AA}, P_{Aa}, and P_{aa}, and the allele frequencies are $p = P_{AA} + P_{Aa}/2$ and $q = P_{aa} + P_{Aa}/2$. We can the define *marginal fitness* values for the alleles A and a,

$$w_A = \frac{w_{AA}2P_{AA} + w_{Aa}P_{Aa}}{2P_{AA} + P_{Aa}} \quad , \quad w_a = \frac{w_{aa}2P_{aa} + w_{Aa}P_{Aa}}{2P_{aa} + P_{Aa}} \,.$$

The mean fitness in the population follows as

$$\bar{w} = w_{AA}P_{AA} + w_{Aa}P_{Aa} + w_{aa}P_{aa} = w_A p + w_a q \,.$$

With these definitions, the changes in genotype and allele frequencies over a life cycle can easily be expressed. They are summarized in the following table.

	AA	Aa	aa	A	a
frequency after random mating	$P_{AA} = p^2$	$P_{Aa} = 2pq$	$P_{aa} = q^2$	p	q
frequency after selection	$p^2\frac{w_{AA}}{\bar{w}}$	$2pq\frac{w_{Aa}}{\bar{w}}$	$q^2\frac{w_{aa}}{\bar{w}}$	$p\frac{w_A}{\bar{w}} = p'$	$q\frac{w_a}{\bar{w}} = q'$
next gener. frequency after random mating	$P'_{AA} = p'^2$	$P'_{Aa} = 2p'q'$	$P'_{aa} = q'^2$	p'	q'

Note that the diploid selection equation for the allele frequencies takes the same functional form as in the haploid case if we replace the allelic fitness value by the corresponding marginal fitness. In general, marginal fitnesses and the mean fitness depend on the genotype frequencies and the equations on the level of allele frequencies do not form a closed dynamical system. In the special case of random mating and viability selection, however, we can express genotype frequencies as HW proportions and the dynamical system for the allele frequencies closes. In particular, the marginal fitness values simplify to

$$w_A = w_{AA}p + w_{Aa}q \quad , \quad w_a = w_{aa}q + w_{Aa}p \,.$$

Selection scenarios

We have seen that viability selection on a single diploid locus with random mating leads to a selection equation that is formally equivalent to the haploid case. The difference is that the marginal fitness values for the alleles depend on the allele frequencies, even if the genotypic fitness values are constant. This leads to differences in the evolutionary dynamics. To characterize these differences, we use the following classical parametrization of the genotypic fitness values.

$$w_{aa} = 1 \qquad \text{normalization of the (relative) wildtype fitness} \tag{1.5a}$$

$$w_{AA} = 1 + s \qquad s\text{: selection coefficient for the homozygote mutant} \tag{1.5b}$$

$$w_{Aa} = 1 + hs \qquad h\text{: dominance coefficient for heterozygote fitness} \tag{1.5c}$$

Depending on the value of the dominance coefficient, we distinguish the following biological scenarios for the mutant allele A

$$h \begin{cases} > 1 & \text{overdominant} \\ = 1 & \text{(fully) dominant} \\ \in (\frac{1}{2}, 1) & \text{partially dominant} \\ = \frac{1}{2} & \text{codominant (or no dominance)} \\ \in (0, \frac{1}{2}) & \text{partially recessive} \\ = 0 & \text{(fully) recessive} \\ < 0 & \text{underdominant} \end{cases}$$

For all cases, the marginal allele fitnesses and mean fitness in HW equilibrium follow as

$$w_a = 1 + p \cdot hs \tag{1.6a}$$

$$w_A = 1 + q \cdot hs + p \cdot s \tag{1.6b}$$

$$\bar{w} = 1 + 2pq \cdot hs + p^2 \cdot s \tag{1.6c}$$

and the allele frequency change per generation of the mutant allele is

$$\Delta p = p' - p = \frac{w_A - \bar{w}}{\bar{w}}\, p = pq\, \frac{s(h + (1-2h)p)}{\bar{w}} \,.$$

In contrast to the haploid case, there is usually no explicit solution for the allele frequency $p(t)$ as a function of time. However, it is straightforward to derive the equilibrium frequencies of the dynamical system. We have $\Delta p = 0$ for

$$p = 0 \quad , \quad p = 1 \;\; [\Leftrightarrow q = 0] \qquad\qquad (\textit{monomorphic equilibria})$$

$$h + (1-2h)p = 0 \quad \Rightarrow \quad p = \hat{p} = \frac{h}{2h-1} \qquad (\textit{polymorphic equilibrium})$$

The equilibrium at $\hat{p}$ is in the interior of the frequency space, $0 < \hat{p} < 1$, if and only if either $h > 1$ (A is overdominant) or $h < 0$ (A is underdominant). We can distinguish three parameter ranges, based on the dominance coefficient, that lead to qualitatively different dynamical behavior.

1. In the whole parameter range $0 \leq h \leq 1$, ranging from complete recessiveness to complete dominance of the mutant allele A, we have

   $$h + (1-2h)p > 0 \quad \text{for} \quad 0 < p < 1$$

   and thus $\Delta p > 0$ for a beneficial mutant ($s > 0$), resp. $\Delta p < 0$ for a deleterious mutant ($s < 0$). The dynamical system therefore converges monotonically either to the equilibrium at $p = 1$ or to $p = 0$ for the beneficial or deleterious case, respectively.

2. If an equilibrium $\hat{p}$ at an intermediate frequency exists, we can write

   $$\Delta p = \frac{pqs(2h-1)}{\bar{w}}\,(\hat{p} - p)\,.$$

   With $s > -1$, we also have

   $$\bar{w} - pqs(2h-1) = 1 + 2pqhs + p^2 s - pqs(2h-1) = 1 + ps > 0\,.$$

   We therefore obtain monotone convergence of $p(t)$ toward the polymorphic equilibrium $\hat{p}$ for the overdominant case ($h > 1$) if $s > 0$ and for the underdominant case ($h < 0$) if $s < 0$. In both cases, the heterozygote is the fittest genotype (*heterozygote advantage*). Note that an underdominant allele A corresponds to an overdominant allele a. The term *overdominance* is often used as synonymous to *heterozygote advantage*, implicitly using the allele with the higher fitness as reference.

3. Analogously, we find monotonic divergence from $\hat{p}$ toward either $p = 0$ or $p = 1$ for the underdominant beneficial case ($h < 0$ and $s > 0$) and for the overdominant deleterious case ($h > 1$ and $s < 0$).

We see that heterozygote advantage ("overdominance") is necessary and sufficient for the maintenance of genetic variation under selection at a single diploid locus. With $\bar{w} = 1 + 2p(1-p)hs + p^2 s$ we have

$$\frac{\partial \bar{w}}{\partial p} = 2s\big(h + p(1-2h)\big)$$

and thus

$$\Delta p = \frac{1}{2}\, pq\, \frac{1}{\bar{w}}\, \frac{\partial \bar{w}}{\partial p} = \frac{pq}{2}\, \frac{\partial \ln \bar{w}}{\partial p}\,.$$

We can therefore understand the evolutionary dynamics also as a process in the direction of increasing mean fitness $\bar{w}$. This is an example of *Fisher's fundamental theorem* (see below). Mathematically, this means that $\bar{w}$ is a so-called Lyapunov function of the dynamics.

Multiple alleles

It is easy to extend the 2-alleles case for a single diploid locus to the general case of k alleles, $\{A_1, \ldots, A_k\}$ with frequencies $\{p_1, \ldots, p_k\}$. Let $w_{ij} = w_{ji}$ be the fitness value of genotype A_iA_j, with frequency P_{ij} in the population. After random mating, the population is in HW equilibrium, thus $P_{ii} = p_1^2$ and $P_{ij} = 2p_ip_j$ for $i \neq j$. The marginal allelic fitnesses and the mean fitness are

$$w_i = \sum_j w_{ij}p_j \quad , \quad \bar{w} = \sum_i w_ip_i = \sum_{i,j} w_{ij}p_ip_j$$

and the change in allele frequencies is

$$p_i' = \frac{w_i}{\bar{w}}\, p_i \qquad \text{resp.} \qquad \Delta p_i = p_i' - p_i = \frac{w_i - \bar{w}}{\bar{w}}\, p_i\,. \tag{1.7}$$

Using

$$w_i = \frac{1}{2}\, \frac{\partial \bar{w}}{\partial p_i} \qquad \text{and} \qquad \bar{w} = \frac{1}{2} \sum_i p_i \frac{\partial \bar{w}}{\partial p_i}$$

we can also write

$$\Delta p_i = \frac{p_i}{2\bar{w}} \Big(\frac{\partial \bar{w}}{\partial p_i} - \sum_j p_j \frac{\partial \bar{w}}{\partial p_j} \Big) \tag{1.8}$$

Defining

$$\vec{\nabla}(\ln \bar{w}) = \Big(\frac{\partial \ln \bar{w}}{\partial p_1}, \ldots, \frac{\partial \ln \bar{w}}{\partial p_k} \Big)^{(t)}$$

$(\cdots)^{(t)}$ denoting transposition, and

$$\mathbf{G} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \cdots \\ -p_2p_1 & p_2(1-p_2) & p_2p_3 & \cdots \\ \vdots & & \ddots & \end{pmatrix}$$

we can write the evolution equation in matrix form,

$$\Delta \mathbf{p} = \frac{1}{2}\, \mathbf{G} \cdot \vec{\nabla}(\ln \bar{w})\,. \tag{1.9}$$

Here, the *selection gradient* $\vec{\nabla}(\ln \bar{w})$ is a vector that points into the direction of steepest ascent in mean fitness $\bar{w}$. $\mathbf{G}$ is the covariance matrix of allele frequencies ($\mathbf{G}_{ij} = \mathrm{cov}[r_i, r_j]$, where $r_i = 1$ if a randomly drawn allele is A_i and zero otherwise). The diagonal elements $p_i(1-p_i)$ are the variances of Bernoulli distributions with parameter p_i, they are largest for $p_i = 1/2$. We see that selection does not necessarily drag a population into the direction of steepest fitness increase (direction of $\vec{\nabla}(\ln \bar{w})$), but always needs to "work on" genetic variation that is present in the population. Evolution proceeds in a direction where the fitness gain *times the genetic variation* is largest.

- In the case of *multiplicative fitness*, $w_{ij} = v_i v_j$, the mean fitness factorizes as $\bar{w} = (\sum_j v_j p_j)^2 =: \bar{v}^2$ and the marginal fitness is $w_i = v_i \sum_j v_j p_j = v_i \bar{v}$. The change in allele frequencies therefore becomes

  $$\Delta p_i = \frac{v_i - \bar{v}}{\bar{v}}$$

  and reproduces the dynamics of the haploid case. The haploid dynamics is therefore a special case of the diploid dynamics. Also, multiplicative fitness is the only case where selection does not lead to deviations from HW equilibrium.

- The equilibrium structure of the diploid dynamics can be complex. There are clearly at least k equilibria (all monomorphic states). In general, we have either $p_i = 0$, or $w_j = \bar{w}$ for all j with $p_j \neq 0$, at any equilibrium. For each choice of a (potentially empty) subset $S \subset \{1, \ldots, k\}$, with $p_i = 0$ for $i \in S$ and $p_j > 0$ for $j \notin S$ (with j_1 the smallest index with $p_j > 0$), we can express the equilibrium condition as a linear equation system

  $$\begin{aligned} p_i &= 0, && i \in S \\ w_j &= w_{j_1}, && j \notin S,\ j > j_1 \\ \sum_{i=1}^{k} p_i &= 1\,. \end{aligned} \tag{1.10}$$

  There are $2^k - 1$ such equation systems for different choices of the subsets S. Each system has either zero, one, or infinitely many solutions. We conclude that in non-degenerate cases with a finite number of equilibria there are at most $2^k - 1$ equilibrium points. In particular, there is at most a single fully polymorphic equilibrium with $p_i > 0$ for all i.

- An easy example of a system with $2^k - 1$ coexisting equilibria is given by the fitness values $w_{ii} = 1$ for all i and $w_{ij} = 0$ for $i \neq j$. Frequently one is interested in the number of *stable* equilibria that can coexist. The maximum number of coexisting stable equilibria for a single diploid locus with k alleles is unknown.

- One can show that dominance of the fitter allele for all pairs of alleles is a necessary condition for a *stable fully polymorphic equilibrium*,

  $$w_{ij} > \frac{w_{ii} + w_{jj}}{2}, \qquad \forall i \neq j$$

[see Nagylaki 1992, p. 62]. However, pairwise overdominance, $w_{ij} > \max[w_{ii}, w_{jj}]$, is *not* necessary, while the stronger condition of *global overdominance*, $w_{ij} > \max_k w_{kk}$, is in general not sufficient. Computer simulations show that overdominance alone is not very efficient in maintaining many alleles segregating at a single locus.

- Since mean fitness is non-decreasing (see below) and thus a Lyapunov function, we can exclude cycling behavior for the dynamics.

Theorem: Mean fitness does not decrease [following Nagylaki 1992, p. 57/58] Mean fitness is a non-decreasing function under selection on a single diploid locus, $\bar{w}' \geq \bar{w}$. For a proof, we use *Jensen's inequality*

$$\sum_i p_i x_i^\alpha \;\leq\; \Big(\sum_i p_i x_i\Big)^\alpha, \qquad \alpha \leq 1; \;\; p_i \text{ probabilities}$$

to obtain a lower bound for the mean fitness $\bar{w}'$ of the offspring generation,

$$\begin{aligned}
\bar{w}' &= \sum_{ij} p_i' p_j' w_{ij} = \frac{1}{\bar{w}^2} \sum_{ij} p_i w_i p_j w_j w_{i,j} && /\, w_i = \sum_k p_k w_{ik} \\
&= \frac{1}{\bar{w}^2} \sum_{i,j,k} p_i p_j p_k w_{ij} w_{ik} \, \frac{w_j + w_j}{2} = \frac{1}{\bar{w}^2} \sum_{i,j,k} p_i p_j p_k w_{ij} w_{ik} \, \frac{w_j + w_k}{2} && /\, (a+b) \geq 2\sqrt{ab} \\
&\geq \frac{1}{\bar{w}^2} \sum_{i,j,k} p_i p_j p_k w_{ij} w_{ik} (w_j w_k)^{1/2} = \frac{1}{\bar{w}^2} \sum_i p_i \Big(\sum_j p_j w_{ij} w_j^{1/2} \Big)^2 && /\ \text{(Jensen)} \\
&\geq \frac{1}{\bar{w}^2} \Big(\sum_i p_i \sum_j p_j w_{ij} w_j^{1/2} \Big)^2 = \frac{1}{\bar{w}^2} \Big(\sum_j p_j w_j^{3/2} \Big)^2 && /\ \text{(Jensen)} \\
&\geq \frac{1}{\bar{w}^2} \Big(\Big(\sum_j p_j w_j \Big)^{3/2} \Big)^2 = \bar{w} && (1.11)
\end{aligned}$$

Continuous time model for selection

Mathematically, our model so far for the evolutionary dynamics has been a discrete dynamical system. A model in discrete time is realistic for some biological species (e.g. annual plants), it has some technical advantages (in particular, it allows for a separation of reproduction and selection) and it is easy to simulate on a computer. However, we have also seen that the number of explicit mathematical results that we can obtain is limited. It is often more convenient (and/or more realistic biologically) to model evolution in continuous time. As we will see, we naturally obtain such a model if we study the discrete time model in a limit of weak selection. To this end, consider again a single locus with k alleles, $\{A_1, \ldots, A_k\}$ with genotypic and marginal fitness values defined as

$$w_{ij} := 1 + \varepsilon\, m_{ij}\,; \qquad w_i = 1 + \varepsilon\, m_i\,; \qquad m_i = \sum_j m_{ij} p_j$$

and corresponding mean fitness

$$\bar{w} = 1 + \varepsilon\,\bar{m}\,; \qquad \bar{m} = \sum_{i,j} m_{ij} p_i p_j$$

The dynamical equation reads

$$\Delta p_i = p_i' - p_i = \frac{\varepsilon\,(m_i - \bar{m})}{1 + \varepsilon\,\bar{m}}\, p_i\,.$$

Assume now that fitness differences per generation are small (weak selection), generation time is short (we thus have many generations per unit time interval). This can be done by scaling both fitness differences and generation time by ε. In the limit of $\varepsilon \to 0$, we then obtain

$$\dot{p}_i = \frac{dp_i(t)}{dt} = \lim_{\varepsilon\to 0} \frac{p_i(t+\varepsilon) - p_i(t)}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{p_i' - p_i}{\varepsilon}$$

and thus

$$\dot{p}(t) = (m_i - \bar{m})\, p_i\,. \tag{1.12}$$

This is the evolution equation in continuous time. Note that m_i and $\bar{m}$ depend on all allele frequencies and $\bar{m}$ is quadratic in the p_j. We thus have a system of coupled non-linear ordinary differential equations.

- The m_{ij} and m_i are also called *Malthusian fitness* parameters (or *log fitness*). They have the interpretation of growth rates per generation,

$$m_i = \lim_{\varepsilon\to 0} \frac{\log w_i}{\varepsilon}\,.$$

- The continuous-time evolution equation can alternatively be derived from a growth model with overlapping generations, where birth and death events occur continuously with given rates [e.g. Rice 2004, p. 15-17]. We note that the ODE (1.12) is only approximate, since a diploid population under selection in continuous time will always deviate from HW equilibrium (unless Malthusian fitness is additive, $m_{ij} = m_i + m_j$). Deviations only vanish in the limit of weak selection. A comprehensive derivation should also account for age structure in a population, with birth and death rates depending on age [see Nagylaki 1992 for a detailed model]. In this case, the dynamics can be much more complicated, but can reproduce (1.12) after the population has reached a *stable age distribution* [see the *Mathematical ecology* lecture for more on age-structured populations].

- The continuous-time single-locus selection dynamics has the same equilibria as the discrete-time model and does therefore not lead to a simplification for this particular problem. However, as we will see below and in the next section, model derivations and extensions to include further evolutionary forces are often easier in continuous time.

Sir Ronald A. Fisher, 1890–1962, is well-known for both his work in statistics and genetics. He is one of the founding fathers of population genetics (together with JBS Haldane and S Wright) that combined Darwinian selection and Mendelian inheritance in the so-called *Modern Synthesis* and led to the breakthrough of Darwinism in the early 20th century. Fisher's 1930 article on *The Genetical Theory of Natural Selection* defined large parts of the field.
In statistics, Fisher's key achievement was his invention of the analysis of variance, or ANOVA. This statistical procedure allows to connect the observed deviations in experimental data to different controlled and uncontrolled underlying factors. It constituted a notable advance over the prevailing procedure of varying only one factor at a time in an experiment. Fisher summed up his statistical work in his book Statistical Methods and Scientific Inference (1956). Fisher became Galton Professor of Eugenics at University College, London in 1933. From 1943 to 1957 he was Balfour Professor of Genetics at Cambridge. He was knighted in 1952 and spent the last years of his life conducting research in Australia (adapted from Encyclopedia Britannica and Wikipedia).

- It is straightforward to show that mean Malthusian fitness is non-decreasing under (1.12),

$$\begin{aligned}\dot{\bar{m}} &= \sum_{ij} m_{ij}(\dot{p}_i p_j + p_i \dot{p}_j) = 2\sum_{ij} m_{ij} p_i p_j (m_i - \bar{m}) \\ &= 2\sum_i p_i m_i (m_i - \bar{m}) = 2\sum_i p_i (m_i - \bar{m})(m_i - \bar{m}) \\ &= 2V_g \qquad (1.13)\end{aligned}$$

  where $V_g > 0$ is the *genetic variance in fitness.* We thus see that the increase in mean fitness is not just non-negative, but is always given by (twice) the current variance in fitness in the population. This is the assertion of *Fisher's fundamental theorem of natural selection* that goes back to R.A. Fisher (1930) and has been discussed in many population genetic textbooks. However, this theorem is only exact for a single locus in continuous time and only holds approximately for discrete time and in more general evolutionary situations.

1.3 Mutation-selection models

The ultimate source of all genetic variation in a population is mutation. So far, we have assumed that genetic variation is just given as an initial condition and have not modeled its creation explicitly. Since selection is usually a much stronger force that mutation and leads to allele frequency changes over shorter time scales, this is often a reasonable approximation. However, for a more complete description of evolution over longer time scales, we need to include mutation into the model.

Only mutation

Usually, mutation occurs during reproduction (or: the production of gametes) due to errors in DNA copying. Each generation, there is a probability that an offspring individual does not inherit the allelic state of (one of) its parent(s), but rather a mutated allele. For a single locus and two alleles, A and a, assume that there is a fixed probability μ that an ancestor carrying the ancestral allele a produces an offspring with A allele. Vice-versa, there is a probability ν that A mutates back to a during reproduction. If the frequency of A alleles is p, the single-generation dynamics reads

$$\Delta p = p' - p = \mu(1-p) - \nu p \tag{1.14}$$

with equilibrium ($\Delta p = 0$)

$$p = \hat{p} = \frac{\mu}{\mu + \nu} \,.$$

For an arbitrary number of alleles $A_1, \ldots, A_k$ and mutation probability from allele A_i to allele A_j denoted as μ_{ij} ($= \mu_{i \to j}$), we can define a *mutation matrix* $\mathbf{U}$ as follows

$$U_{ij} = \mu_{ij} \,, \quad i \neq j \,; \qquad U_{ii} = 1 - \sum_{j \neq i} \mu_{ij} \,. \tag{1.15}$$

With $\mathbf{p} = (p_1, \ldots, p_k)$ the probability vector of allele frequencies, we can then write the evolution equation in matrix form as

$$\mathbf{p}' = \mathbf{p} \cdot \mathbf{U} \quad \text{and} \qquad \mathbf{p}^{(n)} = \mathbf{p} \cdot \mathbf{U}^n \quad \text{for } n \text{ generations} \,. \tag{1.16}$$

We note the following

- The mutation matrix $\mathbf{U}$ is a stochastic matrix that transforms probability vectors into probability vectors. All entries are non-negative and all rows of the matrix sum to 1.

- Assume that $\mathbf{U}$ is *irreducible* and *aperiodic*. This is always the case, in particular, if all diagonal entries are positive (all alleles can be inherited without being mutated) and if we can get from each allele state to any other allele through mutation (or a series of mutations). Then we can use the *Perron-Frobenius theorem* to conclude that $\mathbf{U}$ has a unique largest eigenvalue $\lambda_{\max} = 1$ with corresponding left eigenvector $\mathbf{p}_{\max}$ that holds the equilibrium frequencies of the alleles A_1 to A_k under mutation. [See e.g. the lecture *Mathematical Ecology* for a proof of the Perron-Frobenius theorem.]

- Note that the mutation dynamics is the same for haploids or diploids.

Mutation and selection in discrete time

In discrete time, we can simply include mutation as a separate step into the life cycle. We define the allele frequency change during one generation, starting with newborn zygotes, as $p_i \to p_i^{(s)} \to p_i'$ with

$$p_i^{(s)} = \frac{w_i}{\bar{w}}\, p_i\,, \tag{1.17a}$$

$$p_i' = \Big(1 - \sum_j \mu_{ij}\Big) p_i^{(s)} + \sum_j \mu_{ji} p_j^{(s)}\,. \tag{1.17b}$$

The scheme applies to both haploids and (random mating) diploids, with w_i as marginal fitness for diploids. The first step accounts for viability selection, the second step for mutation during reproduction. It is easy to check that mutation in HW equilibrium changes the allele frequencies, but maintains HW proportions.

In the haploid case, we can still write down the explicit solution of the evolutionary process. Defining the *mutation-selection matrix* $\mathbf{C} = \mathbf{W} \cdot \mathbf{U}$, where $\mathbf{U}$ is the mutation matrix (1.15) and $\mathbf{W} = \mathrm{diag}[w_1, w_2, \ldots]$ is the diagonal matrix holding the fitness values, we obtain

$$\mathbf{p}' = \frac{\mathbf{p} \cdot \mathbf{C}}{\bar{w}}\,, \tag{1.18a}$$

$$\mathbf{p}(t) = \frac{\mathbf{p}(0) \cdot \mathbf{C}^t}{\sum_i \big[\mathbf{p}(0) \cdot \mathbf{C}^t\big]_i}\,. \tag{1.18b}$$

The denominator in both equations just serves for the normalization of the frequency vector. If the matrix $\mathbf{C}$ is primitive (aperiodic and irreducible), there is a unique globally stable and fully polymorphic equilibrium of the dynamics.

Perturbation analysis

For the diploid mutation-selection equation, there is no explicit solution and multiple equilibria can exist (just like for the case without mutation). For two alleles A and a, genotypic fitnesses as in Eq. (1.5), and forward and backward mutation rates μ and ν, respectively, the frequency change of the mutant allele A is

$$p' = f(p) = (1 - \nu)\frac{w_A}{\bar{w}}\, p + \mu\Big(1 - \frac{w_A}{\bar{w}}\, p\Big) \tag{1.19}$$

with

$$w_A = 1 + (1 - p) \cdot hs + p \cdot s \tag{1.20a}$$

$$\bar{w} = 1 + 2p(1 - p) \cdot hs + p^2 \cdot s \tag{1.20b}$$

as in Eq. (1.6). The equilibrium points $\hat{p} = f(\hat{p})$ are zeros of a third-order polynomial. Since the absorptions points $\hat{p} = 0$ and $\hat{p} = 1$ are no longer equilibria, the exact solution is no longer a simple expression. However, we can make use of the fact that mutation rates are usually very small ($\sim 10^{-8}$ per base or $\sim 10^{-5}$ per gene and generation). Because we know the solution in the absence of mutation, we can obtain an approximate solution for the mutation-selection dynamics by means of perturbation analysis. Since this is a technique that can be used more generally, we will present the steps in some detail, with the above problem as application.

Step 1 Identify the small parameters of the problem and write them as a product of a constant times a small perturbation parameter ε,

$$\mu = \tilde{\mu} \cdot \varepsilon \; ; \quad \nu = \tilde{\nu} \cdot \varepsilon \, .$$

This way, the dynamical equation (1.19) becomes a function of ε, $p' = f(\varepsilon, p)$.

Step 2 Write the (unknown) equilibrium solution as a formal power series in ε,

$$\hat{p}(\varepsilon) := \hat{p}_0 + \varepsilon \hat{p}_1 + \varepsilon^2 \hat{p}_2 + \dots$$

and define

$$\begin{aligned} F(\varepsilon) :&= f(\varepsilon, \hat{p}(\varepsilon)) - \hat{p}(\varepsilon) \\ &= (1 - \tilde{\nu}\varepsilon - \tilde{\mu}\varepsilon) \, \frac{1 + hs(1 - \hat{p}(\varepsilon)) + s\hat{p}(\varepsilon)}{1 + 2hs\hat{p}(\varepsilon)(1 - \hat{p}(\varepsilon)) + s(\hat{p}(\varepsilon))^2} \, \hat{p}(\varepsilon) + \tilde{\mu}\varepsilon - \hat{p}(\varepsilon) \, . \end{aligned}$$

We can then write the equilibrium condition as $F(\varepsilon) = 0$.

Step 3 Expand $F(\varepsilon)$ into a Taylor series in ε around 0

$$F(\varepsilon) = F(0) + F^{(1)}(0) \cdot \varepsilon + \frac{1}{2} F^{(2)}(0) \cdot \varepsilon^2 + \frac{1}{6} F^{(3)}(0) \cdot \varepsilon^3 + \dots$$

where $F^{(n)}(0) = \partial^n F(\varepsilon) / \partial \varepsilon^n |_{\varepsilon = 0}$ denotes the nth derivative of $F(\varepsilon)$ at $\varepsilon = 0$. Assuming that the Taylor series converges, $F(\varepsilon) = 0$ implies $F(0) = 0$ and $F^{(n)} = 0$ for all derivatives.

Step 4 Solve the equations to the order desired.

(i) $F(0) = 0$ The 0th order leads back to the selection dynamics without mutation.

$$\frac{1 + hs(1 - \hat{p}_0) + s\hat{p}_0}{1 + 2hs\hat{p}_0(1 - \hat{p}_0) + s\hat{p}_0^2} \, \hat{p}_0 - \hat{p}_0 = 0$$

We can pick any solution of this equation to derive correction terms. Here, we assume that the A mutant is deleterious, $s < 0$, in which case $\hat{p}_0 = 0$ is the stable equilibrium of the selection dynamics in the absence of overdominance.

(ii) $F^{(1)}(0) = 0$

$$F^{(1)}(0) = \hat{p}_0 \cdot \frac{\partial}{\partial \varepsilon} \left[(1 - \tilde{\nu}\varepsilon - \tilde{\mu}\varepsilon) \, \frac{\dots}{\dots} \right] + \hat{p}_1 \, \frac{1 + hs(1 - \hat{p}_0) + s\hat{p}_0}{1 + 2hs\hat{p}_0(1 - \hat{p}_0) + s\hat{p}_0^2} - \hat{p}_1 + \tilde{\mu} = 0$$

Since $\hat{p}_0 = 0$, this results in $\hat{p}_1(1 + hs) - \hat{p} + \tilde{\mu} = 0$ and thus

$$\hat{p}_1 = -\frac{\tilde{\mu}}{hs} \, . \tag{1.21}$$

(iii) $F^{(2)}(0) = 0$

$$\begin{aligned} F^{(2)}(0) &= 2\hat{p}_2(1+hs) + 2\hat{p}_1\Big[(-\tilde{\nu} - \tilde{\mu} - 2hs\hat{p}_1)(1+hs) + s\hat{p}_1(1-h)\Big] - 2\hat{p}_2 \\ &= 2hs\hat{p}_2 - \frac{2\tilde{\mu}}{hs}\Big[(\tilde{\mu} - \tilde{\nu})(1+hs) + \tilde{\mu} - \tilde{\mu}/h\Big] \end{aligned}$$

using $\hat{p}_0 = 0$ and $\hat{p}_1$ from (1.21). Thus

$$\hat{p}_2 = \Big(\frac{\tilde{\mu}}{hs}\Big)^2\Big[1 - \frac{1}{h} + \Big(1 - \frac{\tilde{\nu}}{\tilde{\mu}}\Big)(1+hs)\Big]\,. \tag{1.22}$$

Step 5 Collect all terms in $\hat{p}(\varepsilon)$, using the original biological parameters $\tilde{\mu}\varepsilon = \mu$ and $\tilde{\nu}\varepsilon = \nu$,

$$\hat{p} = \frac{\mu}{|hs|} + \Big(\frac{\mu}{|hs|}\Big)^2\Big[1 - \frac{1}{h} + \Big(1 - \frac{\nu}{\mu}\Big)(1 - |hs|)\Big] + \mathcal{O}[\varepsilon^3]\,. \tag{1.23}$$

We see that frequency of a deleterious mutant in mutation-selection balance is

$$\hat{p} \approx \frac{\mu}{|hs|} \tag{1.24}$$

as long as selection is much stronger than mutation, $|hs| \gg \mu$. Back mutations do not affect this leading-order result.

In our derivation above, we have derived an approximation for the equilibrium frequency in mutation-selection balance, using perturbation theory around the stable equilibrium of the pure selection dynamics. To complete our analysis, we still need to assess the stability of the perturbed equilibrium. For a general discrete dynamical system $p' = f(p)$, an equilibrium point $\hat{p}$ is asymptotically stable if for sufficiently small δ,

$$|\delta'| = |f(\hat{p} + \delta) - f(\hat{p})| < |\delta|$$

If f is continuously differentiable, we have

$$\delta' = \lambda_{\hat{p}} \cdot \delta + \mathcal{O}(\delta^2)\ ; \quad \lambda_{\hat{p}} = \frac{\partial}{\partial p} f(p)|_{\hat{p}}$$

and thus a stable equilibrium at $\hat{p}$ if $|\lambda_{\hat{p}}| < 1$. For $|\lambda_{\hat{p}}| > 0$ the equilibrium is unstable, while for $|\lambda_{\hat{p}}| = 1$ stability depends on higher-order derivatives. At an perturbed equilibrium, we derive

$$\begin{aligned} \lambda_{\hat{p}}(\varepsilon) :&= \frac{\partial}{\partial p} f(\varepsilon, p)|_{\hat{p}(\varepsilon) = \hat{p}_0 + \varepsilon\hat{p}_1 + \dots} \\ &= \lambda_{\hat{p}}(0) + \frac{\partial}{\partial\varepsilon}\lambda_{\hat{p}}(\varepsilon)|_{\varepsilon=0} \cdot \varepsilon + \mathcal{O}[\varepsilon^2] \end{aligned}$$

If the unperturbed equilibrium is stable with $|\lambda_{\hat{p}}(0)| = |\lambda_{\hat{p}_0}| < 1$, this implies stability of the perturbed equilibrium for sufficiently small ε. Similarly. $|\lambda_{\hat{p}}| > 1$ implies that also the perturbed equilibrium is unstable. For the mutation-selection dynamics discussed above, we have $\lambda_{\hat{p}} = \lambda_0 = 1 + hs < 1$ for a deleterious mutant. In general, for $hs \neq 0$, the stability of the perturbed equilibrium is the same as the unperturbed one.

Mutation load

The effect of a deleterious mutation on an individual is measured by the corresponding reduction in fitness, given by the selection coefficient s (or by hs in a diploid heterozygote). Similarly, we can assess the effect of a deleterious mutation on the population level by the reduction in mean fitness in mutation-selection balance. The standard measure is the *mutation load*

$$L_m = \frac{w_{\text{opt}} - \bar{w}}{w_{\text{opt}}}\,, \tag{1.25}$$

where w_{opt} is the fitness of an "optimal" genotype that is free from deleterious mutations. If we normalize the optimal fitness $w_{\text{opt}} = 1$ (as in the mutation model) studied above, we simply have $L_m = 1 - \bar{w}$. With (1.20b) we obtain at the equilibrium $\hat{p} = -\mu/(hs) + \mathcal{O}[\mu^2]$,

$$L_m = -2hs\hat{p} - (s - 2hs)\hat{p}^2 = 2\mu + \mathcal{O}[\mu^2], \tag{1.26}$$

as long as $\mu \ll |hs|$. For arbitrarily many deleterious alleles A_i with mutation rates μ_{i0} from a fittest wildtype a, we obtain more generally

$$L_m = 2\sum_{i\geq 1} \mu_{i0} + \mathcal{O}[\mu^2]\,.$$

We see that the mutation load depends (to leading order) only on the mutation rates, but not on the fitness effects of the deleterious mutations. The reason is that a milder mutation with small $|hs|$ will segregate at a higher frequency $\hat{p} = \mu/|hs|$ in the population. To leading order, the effects of mutation frequency and mutation size on the mean fitness just cancel. This is also called *Haldane's rule* or the *Haldane-Muller principle.* This has relevant consequences for programs of public health that aim for an increase of population-level parameters like the mean fitness. Indeed, according to the Haldane-Muller principle, the mean fitness in a population is neither altered by eugenics (birth control for diseased people, effectively increasing the deleterious fitness effect of a mutation) nor by a partial cure of a genetic disease (reduction of $|hs|$). For population-level fitness, mildly deleterious mutations are as harmful as strongly deleterious ones. Only the reduction of mutation *rates* has a lasting effect on mean fitness.

Mutation and selection in continuous time

There are various ways to write down a mutation-selection equation in continuous time. The most widely used formalism is simply to assume that mutation and selection are independent processes that occur in parallel. This leads to the differential equation

$$\dot{p}_i = (m_i - \bar{m})p_i + \sum_j \big(\mu_{ji}p_j - \mu_{ij}p_i\big) \tag{1.27}$$

extending Eq. (1.12). The m_i are Malthusian fitness values (marginal fitnesses for diploids) and the μ_{ij} have the interpretation of mutation rates per time unit. Like for discrete time,

JBS (John Burdon Sanderson) Haldane, 1892–1964, was a British geneticist, biometrician, physiologist, and popularizer of science who opened new paths of research in population genetics and evolution. Together with R.A. Fisher and Sewall Wright, but in separate mathematical arguments, he related Darwinian evolutionary theory and Gregor Mendel's laws of heredity. Haldane also contributed to the theory of enzyme action and to studies in human physiology. He possessed a combination of analytic powers, literary abilities, a wide range of knowledge, and a force of personality that produced numerous discoveries in several scientific fields and proved stimulating to an entire generation of research workers.
Haldane announced himself a Marxist in the 1930s but later became disillusioned with the official party line and with the rise of the controversial Soviet biologist Trofim D. Lysenko. In 1957 Haldane moved to India, where he took citizenship and headed the government Genetics and Biometry Laboratory in Orissa (adapted from Encyclopedia Britannica).

Herrmann Joseph Muller 1890–1967, Nobel laureate in Medicine (1946) for his discovery of the mutagenic effect of X-rays was very concerned about the reduction of mean fitness in humans by radiation, also due to nuclear fallout caused by nuclear testing. Together with fellow scientists, he was a vocal critic of nuclear weapons testing (from Wikipedia).

the dynamics for haploids is fully solvable also in continuous time. For $m_i = \text{const}$ and $\bar{m} = \sum_i m_i p_i$, we can write

$$\dot{\mathbf{p}} = \mathbf{p} \cdot \mathbf{A} - \mathbf{p}\,\bar{m} \tag{1.28}$$

with matrix $\mathbf{A}$ with entries $a_{ij} = m_i \delta_{ij} + \mu_{ij} - \sum_l \mu_{il} \delta_{ij}$. Like in the discrete case, the mean fitness $\bar{m}$ is a common factor for all allele frequencies. It enforces normalization of the frequency vector $\mathbf{p}$ during the dynamics, but does not affect the relative size of allele frequencies p_i/p_j. We thus have a normalized linear dynamics for $\mathbf{p}$ with solution

$$\mathbf{p}(t) = \frac{\exp[\mathbf{p}(0) \cdot \mathbf{A}\, t]}{\sum_i \big[\exp[\mathbf{p}(0) \cdot \mathbf{A}\, t]\big]_i}\,. \tag{1.29}$$

For a diploid locus with two alleles a and A, mutation rates μ from a to A and ν from A to a, and Malthusian fitness values

$$m_{aa} = 0\,; \quad m_{Aa} = hs\,; \quad m_{AA} = s \tag{1.30}$$

we have $m_A = sp_A + hs(1 - p_A)$ and $\bar{m} = sp_A^2 + 2hsp_A(1 - p_A)$ and Eq. (1.27) results in

$$\dot{p}_A = s\big(p_A + h(1 - 2p_A)\big)p_A(1 - p_A) + \mu(1 - p_A) - \nu p_A\,. \tag{1.31}$$

This is a thrid-order equation unless $h = 1/2$ (case of no dominance), where the diploid dynamics reduces to an effective haploid one ($m_A - \bar{m} = (s/2)(1 - p_A)$, as in the haploid model with $m_A = s/2$ and $\bar{m} = (s/2)p_A$).

2 Recombination

In diploid organisms recombination happens during *meiosis* (the production of gametes). Recombination mixes paternal and maternal material before it is transferred to the next generation. Each gamete that is produced by an individual therefore contains material from the maternal and the paternal side. To see what this means, take a look at your two chromosomes number 1, one of which came from your father and one from your mother. The one that stems from your father is in fact a mosaic of pieces from his mother and his father, your two paternal grandparents. In humans these mosaics are such that a chromosome is made of a couple of chunks or recombination blocks. There is generally more than one such block, but rarely more than ten per generation. Chromosomes that do not recombine are not mosaics. The Y-chromosome does not recombine at all, males inherit it completely from their father and paternal grandfather, etc. Mitochondrial DNA also does not normally recombine, both females and males inherit mitochondria from their mother, maternal grandmother, etc. The X-chromosome only recombines when it is in a female.

There are various mechanisms for recombination. The most well-known one is *crossing over*, where matching regions in homologous chromosomes (which pair during meiosis) experience a *double strand break* and subsequently are reconnected to the other chromosome (see Fig. 2.1). There are other recombination mechanisms like *gene conversion*, where a stretch of DNA is copied from one chromosome to the matching region of its homologous partner. Exchange of genetic material can also happen in haploid individuals. In this case two different individuals exchange pieces of their genome.

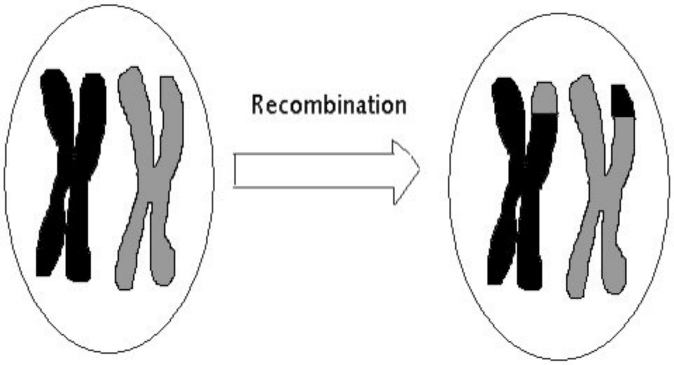

Figure 2.1: Single recombination event by *crossing over* of chromosomes during meiosis. The figure shows a pair of homologous chromosomes after the initial duplication, before and after recombination. Black and gray parts derive from different parents. Subsequently, the duplicated pair will segregate into four gametes, two recombined and two not recombined.

2.1 Linkage and linkage disequilibrium

Linkage

Mendel's second law (of *independent assortment*) states that genes are inherited independently of each other. It means that the probability of inheriting a gene at some locus $\mathcal{A}$ from one grandmother is independent of whether or not a gene at a different locus $\mathcal{B}$ has been inherited from the same grandmother. This "law" is generally only true for gene loci that are located on different chromosomes: they are *unlinked.* On the other hand, if genes are on the same chromosome, they are said to be *physically linked.* Linked genes are not inherited independently of each other. In particular, if gene loci are very close to each other, recombination between them is rare and they are typically inherited together. Mathematically, this is expressed by the *recombination fraction* $r = r_{AB}$ between loci $\mathcal{A}$ and $\mathcal{B}$, which defines the probability that genes inherited from different grandparents at these loci end up on the same parental gamete (sperm, egg, pollen) that contributes to the offspring genotype,

$$\begin{pmatrix} a_1b_1 \\ a_2b_2 \end{pmatrix} \longrightarrow \left\{ \begin{array}{ll} \left.\begin{array}{c} a_1b_1 \\ a_2b_2 \end{array}\right\} & \text{freq. } \frac{1}{2}(1-r) \text{ each} \\ & \\ \left.\begin{array}{c} a_1b_2 \\ a_2b_1 \end{array}\right\} & \text{freq. } \frac{1}{2}\cdot r \text{ each} \end{array} \right. . \tag{2.1}$$

Here, a_1 and b_1 (res. a_2 and b_2) do not denote an allelic state, but only the origin of the gene either from grandparent 1 or 2.

- r is often also called a *recombination rate*, but it is really a probability in discrete generation models. We generally have $r = 1/2$ as upper limit for unlinked loci on different chromosomes and $0 \le r < 1/2$ for linked loci.

- We can define a molecular recombination probability ρ as the probability for recombination between neighboring base pairs along a chromosome. Typical values are $\rho \approx 10^{-8}$ per generation. However, ρ generally depends strongly on the genomic position x. The estimation of recombination maps $\rho(x)$ from data is an important task of genomics.

- For a given recombination map, we can define a *recombination distance* d along a chromosome in units of *Morgans* (named after Thomas Morgan). A distance of $d = 1M$ indicates that there is on average one recombination breakpoint per generation within the stretch (e.g., due to crossing over). Typical lengths of chromosome regions measure in *centi-Morgans* (cM).

- The recombination fraction r between loci on the same chromosome is the probability of an odd number of recombination breakpoints between these loci. Ignoring interference of recombination events in neighboring regions, r relates to the recombination distance d via *Haldane's mapping function*

$$r = \frac{1}{2}\big(1 - \exp[-2d]\big)\,. \tag{2.2}$$

Linkage disequilibrium

Assume now that there are k alleles $\{A_1, \ldots, A_k\}$ at locus $\mathcal{A}$ and l alleles $\{B_1, \ldots, B_l\}$ at locus $\mathcal{B}$. There are then $k \times l$ gametes (or haplotypes) A_iB_j with frequency denoted as $P_{A_iB_j}$. The allele frequencies derive as

$$P_{A_i} = \sum_{j=1}^{l} P_{A_iB_j}\ ; \quad P_{B_j} = \sum_{i=1}^{k} P_{A_iB_j}\,. \tag{2.3}$$

As a measure of non-random association of alleles A_i and B_j at different loci on the same gamete (or haplotype), we define the *linkage disequilibrium* (LD)

$$D_{A_iB_j} = P_{A_iB_j} - P_{A_i}P_{B_j}\,. \tag{2.4}$$

If the linkage disequilibrium is zero, $D_{A_iB_j} = 0$, we say that alleles A_i and B_j are in *linkage equilibrium* (LE).

- Mathematically, D is simply the covariance of two indicator random variables that take value 1 if a randomly picked haplotype shows the corresponding allele at locus $\mathcal{A}$ resp. $\mathcal{B}$, and zero otherwise. Linkage disequilibria depend strongly on the allele frequencies and (since $P_{A_iB_j} \leq \max[P_{A_i}, P_{B_j}]$) we see that

  $$D_{A_iB_j} \leq \max[P_{A_i}(1 - P_{B_j}), P_{B_j}(1 - P_{A_i})]\,.$$

  In order to make disequilibria between different pairs of alleles better comparable, one therefore often uses the normalized measure

  $$r^2_{A_iB_j} = \frac{D^2_{A_iB_j}}{P_{A_i}(1 - P_{A_i})P_{B_j}(1 - P_{B_j})}\,, \tag{2.5}$$

  which corresponds to the (squared) correlation coefficient of the indicator variables.

- In addition to two-locus disequilibria, we can also define higher-order linkage disequilibria between alleles at three or more loci (e.g. as higher-order cross-locus cumulants, see chapter 5 of the book by R. Bürger).

- Note that *linkage* and *linkage disequilibrium* are concepts on different levels. While linkage is a property of loci and manifests in each individual, linkage disequilibrium is a population property and related to allele/haplotype frequencies. Unlinked loci can certainly have non-zero linkage disequilibria among their alleles, while alleles at linked loci (even with $r = 0$) can be in linkage equilibrium.

2.2 Two-locus model

Only recombination

Consider the two-locus model as described above. Without mutation or selection (or drift), the single-locus allele frequencies in the population stay constant, $P'_{A_i} = P_{A_i}$. However, recombination will change the haplotype frequencies. Assuming HW proportions in the germ cells prior to meiosis (and recombination), we obtain

$$P'_{A_iB_j} = (1-r)P_{A_iB_j} + r \cdot P_{A_i}P_{B_j} = P_{A_iB_j} - r \cdot D_{A_iB_j} \,. \tag{2.6}$$

Indeed, a fraction of $(1-r)$ of all gametes that contribute to the new generation has not undergone any recombination. In this part of the population, haplotype frequencies maintain their value from the previous generation. Conversely, a fraction of r of new gametes are recombination products. In HW equilibrium, the probability for them to result in a A_iB_j haplotype is $P_{A_i}P_{B_j}$. For the change in linkage disequilibrium, we obtain

$$D'_{A_iB_j} = P'_{A_iB_j} - P'_{A_i}P'_{B_j} = (1-r)P_{A_iB_j} + r \cdot P_{A_i}P_{B_j} - P_{A_i}P_{B_j} = (1-r) \cdot D_{A_iB_j} \,. \tag{2.7}$$

- We thus see that for $r > 0$ all linkage disequilibria decay to zero at geometric rate $(1-r)$. The population approaches linkage equilibrium among all alleles, $P_{A_iB_j} = P_{A_i}P_{B_j}$.

- Note that, in contrast to HW equilibrium, linkage equilibrium among alleles at different loci is *not* reached in a single generation, but only asymptotically – even for unlinked loci with $r = 1/2$.

Recombination and selection in discrete time

Consider a model with two loci under selection and focus on the case of two alleles at each locus. We can write the fitness schemes for haploid or diploid individuals as follows

$$\begin{array}{ccc} & B & b \\ A & w_{AB} & w_{Ab} \\ a & w_{aB} & w_{ab} \end{array} \quad ; \quad \begin{array}{cccc} & BB & Bb & bb \\ AA & w_{ABAB} & w_{ABAb} & w_{AbAb} \\ Aa & w_{ABaB} & w_{ABab} & w_{Abab} \\ aa & w_{aBaB} & w_{aBab} & w_{abab} \end{array}$$

The diploid scheme assumes that the fitness of a genotype depends only on the number and type of alleles in the genotype, but not on the association of the allele to a particular haplotype (no *position effect*). I.e., the fitness of the diploid genotype (Ab, aB) is the same as the one of (AB, ab). Assuming HW proportions for diploids in zygote state, marginal fitness values for the 2-locus haplotypes follow in the usual way, $w_{AB} = w_{ABAB}\,P_{AB} + w_{ABAb}\,P_{Ab} + w_{ABaB}\,P_{aB} + w_{ABab}\,P_{ab}$, etc. The mean fitness for both haploids and diploids is

$$\bar{w} = w_{AB}\,P_{AB} + w_{Ab}\,P_{Ab} + w_{aB}\,P_{aB} + w_{ab}\,P_{ab} \,.$$

It is convenient to write the linkage disequilibrium as

$$\begin{aligned} D_{AB} &= P_{AB} - P_A P_B \\ &= P_{AB}(P_{AB} + P_{Ab} + P_{aB} + P_{ab}) - (P_{Ab} + P_{AB})(P_{aB} + P_{AB}) \\ &= P_{AB}P_{ab} - P_{Ab}P_{aB} \,. \end{aligned} \tag{2.8}$$

It is easy to verify that

$$D := D_{AB} = D_{ab} = -D_{Ab} = -D_{aB} \,.$$

Discrete time dynamics

Like for the mutation-selection model, we can construct a recombination-selection model by including both events as separate steps into a life cycle. This is best done on the level of haplotype frequencies. Indeed, with random mating, whole genotype frequencies decompose into haplotype frequencies also in a diploid population. On the other hand, haplotype frequencies do not factor into allele frequencies as long as $D \neq 0$. Starting with zygotes, we first have selection, followed by recombination during reproduction. This results in

$$P'_{AB} = \hat{P}_{AB} - r\hat{D} \,, \tag{2.9a}$$

$$\begin{aligned} D' &= (\hat{P}_{AB} - r\hat{D})(\hat{P}_{ab} - r\hat{D}) - (\hat{P}_{Ab} + r\hat{D})(\hat{P}_{aB} + r\hat{D}) \\ &= \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB} - r\hat{D} \,, \end{aligned} \tag{2.9b}$$

and similar expressions for the other haplotype frequencies. $\hat{P}_{..}$ and $\hat{D}$ are the values for the frequencies and for LD after selection. We have

$$\hat{P}_{AB} = \frac{w_{AB}}{\bar{w}} P_{AB} \,.$$

For $\hat{D}$, we need to distinguish the haploid and diploid dynamics. For haploids that recombine after random union of gametes after the selective phase, we obtain

$$\hat{D} = \hat{P}_{AB}\hat{P}_{ab} - \hat{P}_{Ab}\hat{P}_{aB} = \frac{w_{AB}w_{ab}}{\bar{w}^2} P_{AB}P_{ab} - \frac{w_{Ab}w_{aB}}{\bar{w}^2} P_{Ab}P_{aB}$$

and thus

$$D' = (1-r)\left(\frac{w_{AB}w_{ab}}{\bar{w}^2} P_{AB}P_{ab} - \frac{w_{Ab}w_{aB}}{\bar{w}^2} P_{Ab}P_{aB}\right) \,. \tag{2.10}$$

For diploids, recombination occurs in the diploid phase after selection and we get

$$\hat{D} = \widehat{P_{AB}P_{ab}} - \widehat{P_{aB}P_{Ab}} = \frac{w_{ABab}}{\bar{w}}\big(P_{AB}P_{ab} - P_{Ab}P_{aB}\big) = \frac{w_{ABab}}{\bar{w}} D$$

resulting in

$$D' = \frac{w_{AB}w_{ab}}{\bar{w}^2} P_{AB}P_{ab} - \frac{w_{aB}w_{Ab}}{\bar{w}^2} P_{Ab}P_{aB} - r\frac{w_{ABab}}{\bar{w}} D \tag{2.11}$$

Assume that, initially, $D = P_{AB}P_{ab} - P_{Ab}P_{aB} = 0$. Eqs. (2.10) and (2.11) show that selection will create positive or negative LD, depending on the fitness values for haplotypes and on the so-called level of *epistasis*. In both cases

$$w_{AB}w_{ab} - w_{Ab}w_{aB} \begin{cases} > 0 & \text{positive epistasis, creates positive LD} \quad D' > 0 \\ = 0 & \text{no epistasis, maintains LE} \quad D' = D = 0 \\ < 0 & \text{negative epistasis, creates negative LD} \quad D' < 0\,. \end{cases} \qquad (2.12)$$

For haploids, we can normalize the fitness of the *wildtype* (ab) to 1 and set

$$w_{ab} = 1\;; \quad w_{Ab} = v_A\;; \quad w_{aB} = v_B\;; \quad w_{AB} = v_A v_B + \varepsilon$$

where v_A and v_B are the single-mutant fitness values and the *epistasis parameter* measures the deviation of the double mutant fitness from the multiplicative effects of the single mutants. Obviously, $\varepsilon > 0$ ($\varepsilon < 0$) implies positive (negative) epistasis and leads to positive (negative) LD if evolution starts in LE. For the diploid case, the haplotype fitnesses are marginal fitnesses and depend on the haplotype frequencies. We can still verify that epistasis vanishes for all frequencies if the genotype fitnesses are multiplicative across loci ($w_{ABAb} = v_{AA}v_{Bb}$, etc). In contrast to the haploid case, there is more than one epistasis parameter needed to parametrize deviations from multiplicative fitnesses in the full fitness scheme. Even without epistasis, the dynamics in discrete time is complex and can only be solved in special cases (see the books by Bürger, chapter 2 and by Nagylaki, chapter 8).

Continuous time dynamics

As in the case of mutation and selection we assume that selection and recombination occur in parallel and independently of each other in continuous time. This is a good approximation, in particular, if selection and recombination are both weak. For simplicity, we focus on the haploid case. We assign Malthusian fitness values to the four hapotypes, m_{ab}, m_{Ab}, m_{aB}, and m_{AB}. The dynamical equations for the haplotype frequencies read

$$\dot{P}_{AB} = P_{AB}(m_{AB} - \bar{m}) - rD \qquad (2.13a)$$
$$\dot{P}_{Ab} = P_{Ab}(m_{Ab} - \bar{m}) + rD \qquad (2.13b)$$
$$\dot{P}_{aB} = P_{aB}(m_{aB} - \bar{m}) + rD \qquad (2.13c)$$
$$\dot{P}_{ab} = P_{ab}(m_{ab} - \bar{m}) - rD \qquad (2.13d)$$

where $\bar{m}$ is the mean Malthusian fitness. We focus on the case of no epistasis. On the logarithmic scale of Malthusian fitnesses, this corresponds to additive contributions across loci. Normalizing the wildtype fitness to zero, we have

$$m_{ab} = 0\;; \quad m_{Ab} = m_A\;; \quad m_{aB} = m_B\;; \quad m_{AB} = m_A + m_B\,.$$

The dynamics of the mean fitness then becomes independent of the recombination rate,

$$\begin{aligned}\dot{\bar{m}} &= \dot{P}_{Ab}m_A + \dot{P}_{aB}m_B + \dot{P}_{AB}(m_A + m_B)\\ &= P_{Ab}m_A(m_A - \bar{m}) + P_{aB}m_B(m_B - \bar{m}) + P_{AB}(m_A + m_B)(m_A + m_B - \bar{m})\\ &= P_{Ab}(m_A - \bar{m})^2 + P_{aB}(m_B - \bar{m})^2 + P_{AB}(m_A + m_B - \bar{m})^2 + P_{ab}(0 - \bar{m})^2 \,. \qquad (2.14)\end{aligned}$$

We see that mean fitness is non-decreasing (a Lyapunov function), with $\dot{\bar{m}} = 0$ if and only if the allele frequencies are at an equilibrium point. We conclude that $P_{AB}(m_{AB} - \bar{m}) = 0$ and thus with Eq. (2.13a) also $D = 0$ at each equilibrium. The dynamics of the disequilibrium is

$$\begin{aligned}\dot{D} &= \dot{P}_{AB}P_{ab} + P_{AB}\dot{P}_{ab} - \dot{P}_{Ab}P_{aB} - P_{Ab}\dot{P}_{aB}\\ &= P_{AB}P_{ab}(m_{AB} + m_{ab} - 2\bar{m}) - P_{Ab}P_{aB}(m_{Ab} + m_{aB} - 2\bar{m}) - rD\\ &= (m_A + m_B - 2\bar{m} - r)D \,. \qquad (2.15)\end{aligned}$$

Like in discrete time, the dynamics with non-epistatic fitness thus maintains LE $D = 0$. For the search of equilibrium points, we can thus restrict the dynamics to the LE manifold, where we obtain

$$\begin{aligned}\dot{p}_A = \dot{P}_{AB} + \dot{P}_{Ab} &= P_{AB}\big(m_A(1 - p_A) + m_B(1 - p_B)\big) + P_{Ab}\big(m_A(1 - p_A) - m_B p_B\big)\\ &= p_A(1 - p_A)m_A + (P_{AB} - p_A p_B)m_B\\ &= p_A(1 - p_A)m_A \,, \qquad (2.16)\end{aligned}$$

and equivalently for p_B. We see that the dynamics on the LE manifold simply reduces to the single locus dynamics.

- The result shows under which conditions the use of simple single locus models is meaningful in complex biological scenarios: If fitness epistasis can be ignored and if loci are in LE, the multi-locus dynamics reduces to the single-locus dynamics. Furthermore, even if starting conditions are not in LE, but $D = 0$ for all equilibria, we can use the single-locus formalism to describe the equilibrium structure and the long-term dynamics.

- Unless epistasis is very strong and/or linkage very tight, the mutilocus dynamics usually converge to a parameter range very close to the LE manifold. One can then solve the problem under the assumption of LE first and treat linkage disequilibria as a perturbation. This is the idea of the *quasi linkage equilibrium* approximation, see e.g. the book by Bürger.

3 Genetic Drift

In the first part of the lecture, we have described the evolutionary dynamics using a *deterministic* framework that does not allow for stochastic fluctuations of any kind. In a deterministic model, the dynamics of allele (or genotype) frequencies is governed by the expected values: mutation and recombination rates determine the expected number of mutants or recombinants, and fitness defines the expected number of surviving offspring individuals. In reality, however, the number of offspring of a given individual (and the number of mutants and recombinants) follows a distribution. Altogether, there are three possible reasons why an individual may have many or few offspring:

- *Good or bad genes*: the heritable genotype determines the distribution for the number of surviving offspring. Fitness, in particular, is the expected value of this distribution and determines the allele frequency change due to natural selection.
- *Good or bad environment*: the offspring distribution and the fitness value may also depend on non-heritable ecological factors, such as temperature or humidity. These factors can be included into a deterministic model with space- or time-dependent fitness values.
- *Good or bad luck*: the actual number of offspring, given the distribution, will depend on random factors that are not controlled by either the genes nor the external environment. This gives rise to the stochastic component in the change of allele frequencies: *random genetic drift*.

For a general evolutionary system, we can define classes of individuals according to genotypes and environmental parameters. Because of the law of large numbers, genetic drift can be ignored if and only if the number of individuals in each class tends to infinity (or if the variance of the offspring distribution is zero). Note that effects of genetic drift may be relevant even in infinite populations if the number of individuals in a focal allelic class is finite.

3.1 The Wright-Fisher model

The Wright-Fisher model (named after Sewall Wright and Ronald A. Fisher) is maybe the simplest population genetic model for genetic drift. We will introduce the model for a single locus in a haploid population of constant size N. Further assumptions are no mutation and no selection (neutral evolution) and discrete generations. The life cycle is as follows

1. Each individual in the parent generation produces an equal and very large number of gametes (or seeds). In the limit of seed number $\to \infty$, we obtain a so-called *infinite gamete pool*.
2. We sample N individuals from this gamete pool to form the offspring generation.

Sewall Wright, 1889–1988, was an American geneticist. Wright's earliest studies included investigation of the effects of inbreeding and crossbreeding among guinea pigs, animals that he later used in studying the effects of gene action on coat and eye color, among other inherited characters. His papers on inbreeding, mating systems, and genetic drift make him a principal founder of theoretical population genetics, along with R.A. Fisher and JBS Haldane. Wright's most eminent contribution to population genetics is his concept of *genetic drift* and his development of mathematical theory combining drift with the other evolutionary forces. He was also the inventor/discoverer of key concepts like the *fitness landscape* and the *inbreeding coefficient* and originated a theory to guide the use of inbreeding and crossbreeding in the improvement of livestock (adapted from Encyclopedia Britannica and Wikipedia).

Obviously, this just corresponds to *multinomial sampling with replacement* directly from the parent generation according to the rule:

- Each individual from the offspring generation picks a parent at random from the previous generation and inherits the genotype of the parent.

Remarks

- Mathematically, the probability for $k_1, \ldots, k_N$ offspring for individual number $1, \ldots, N$ in the parent generation is given by the multinomial distribution with
$$\Pr\Big[k_1, \ldots, k_N \big| \textstyle\sum_i k_i = N\Big] = \frac{N!}{\prod_i k_i! N^N}. \tag{3.1}$$
- The number of offspring of a given parent individual is binomially distributed with parameters $n = N$ (number of trials) and $p = 1/N$ (success probability):
$$\Pr\big[k_1\big] = \binom{N}{k_1} \left(\frac{1}{N}\right)^{k_1} \left(1 - \frac{1}{N}\right)^{N-k_1}.$$
- Under the assumption of *random mating* (or *panmixia*), a diploid population of size N can be described by the haploid model with size $2N$, if we follow the lines of descent of all gene copies separately. Technically, we need to allow for selfing with probability $1/N$.
- The Wright-Fisher model can easily be extended to non-constant population size $N = N(t)$, simply by taking smaller or larger samples to generate the offspring generation.
- As long as the population is unstructured and evolution is neutral, the offspring distribution is invariant with respect to exchange of individuals in each generation. We can use this symmetry to disentangle the genealogies, as shown in Figure (3.3).
- Inclusion of mutation, selection, and migration (population structure) is straightforward, as shown in later sections.

Figure 3.1: The first generation in a Wright-Fisher Model of 5 diploid or 10 haploid individuals. Each of the haploids is represented by a circle.

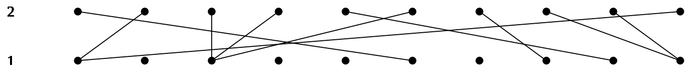

Figure 3.2: The second generation (first offspring generation) in a Wright-Fisher Model.

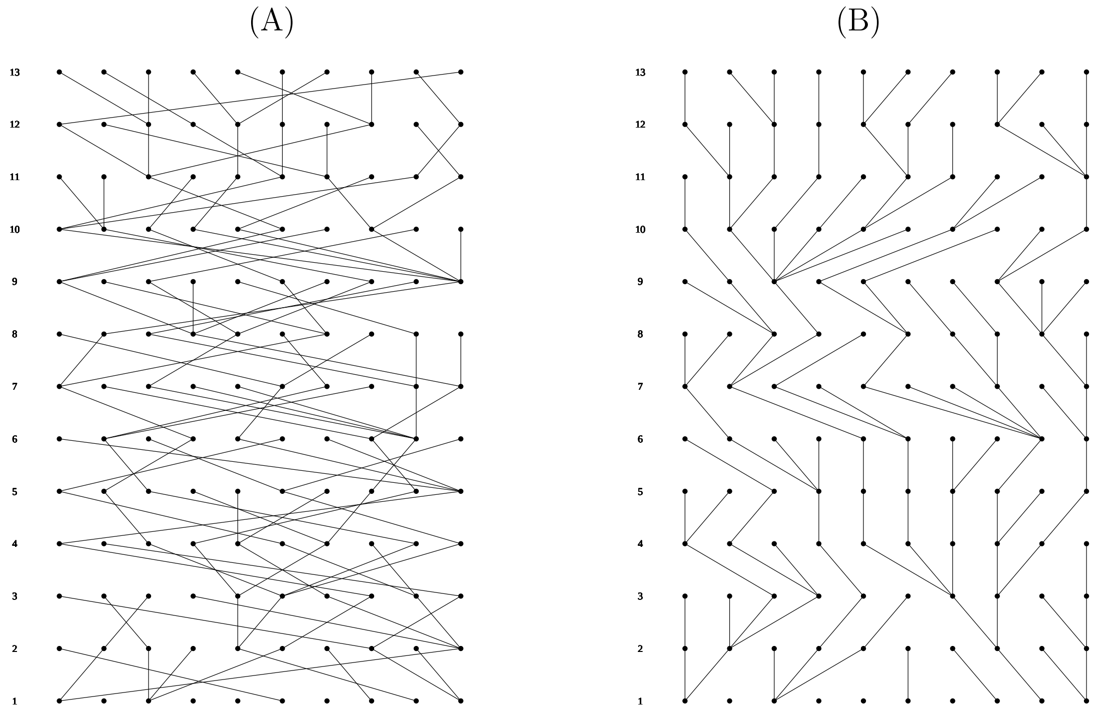

Figure 3.3: The *tangled* and *untangled version* of the Wright-Fisher Model after several generations. Both pictures show the same process, except that the individuals in the untangled version have been shuffled to avoid line crossings. The genealogical relationships are still the same, but the children of one parent are now put next to each other and close to the parent. (Web resource: www.coalescent.dk, Wright-Fisher simulator.)

3.2 Consequences of genetic drift

Genetic drift is the process of random changes in allele frequencies in populations. We will now study the effects of genetic drift quantitatively using the Wright-Fisher model. To this end, consider a single locus with two neutral alleles a and A in a diploid population of size N. We thus have a haploid population size (= number of gene copies) of $2N$. We denote the number of A alleles in the population at generation t as n_t and its frequency as $p_t = n_t/2N$. The transition probability from state $n_t = i$ to state $n_{t+1} = j$, $0 \leq i, j, \leq 2N$ is given by

$$P_{ij} := \Pr[n_{t+1} = j | n_t = i] = \binom{2N}{j} \cdot \Big(\frac{i}{2N}\Big)^j \cdot \Big(1 - \frac{i}{2N}\Big)^{2N-j} . \tag{3.2}$$

This defines the transition matrix $\mathbf{P}$ with elements P_{ij}, $0 \leq i, j \leq 2N$, of a time-homogeneous Markov chain. If $\mathbf{x}_t$ is the probability vector (of length $2N+1$) on the state space at generation t, we have $\mathbf{x}_{t+1} = \mathbf{x}_t\mathbf{P}$. Some elementary properties of this process are:

1. For the expected number of A alleles, we have $\mathrm{E}[n_1|n_0] = 2N \cdot \frac{n_0}{2N} = i = n_0$, and thus $\mathrm{E}[n_1] = \mathrm{E}[n_0]$ and
$$\mathrm{E}[p_t] = \mathrm{E}[p_0] .$$
The expected allele frequency is constant. The stochastic process defined by the neutral Wright-Fisher model is thus a *martingale*. This holds true, in more general, for any neutral model of *pure random drift* (no mutation and selection) in an unstructured population. We can also express this in terms of the expected change in allele frequencies as $\mathrm{E}[\delta p|p = p_0] = \mathrm{E}[p_1 - p_0] = 0$.

2. For the variance among replicate offspring populations from a founder population with frequency $p_0 = n_0/2N$ of the A allele, we obtain: $\mathrm{Var}[n_1|n_0] = 2Np_0(1 - p_0)$ and thus
$$V := \mathrm{Var}[p_1|p_0] = \frac{p_0(1 - p_0)}{2N} .$$
The variance is largest for $p_0 = 1/2$. In terms of allele frequency changes, we can also write $\mathrm{Var}[\delta p|p = p_0] = \mathrm{Var}[p_1 - p_0] = \mathrm{Var}[p_1|p_0] = V$.

3. There are two absorbing states of the process: Fixation of the A allele at $p_t = 1$, corresponding to a probability vector $\mathbf{x}^{(1)} = (0, 0, \ldots, 1)$, and loss of the allele at $p_t = 0$, corresponding to $\mathbf{x}^{(0)} = (1, 0, \ldots, 0)$. Both $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$ are left eigenvectors of the transition matrix with eigenvalue 1.

4. The absorption probabilities in $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$ for an initial frequency $p_0 = n_0/2N$ are given by the corresponding right eigenvectors, $\mathbf{y}^{(0)}$ and $\mathbf{y}^{(1)}$, with normalization $\mathbf{x}^{(i)} \cdot \mathbf{y}^{(j)} = \delta_{ij}$: If we define as π_i the fixation probability (absorption in $\mathbf{x}^{(1)}$) for a process that starts in state $p_0 = i/2N$, then $\mathbf{y}^{(1)} = (0, \pi_1, \pi_2, \ldots, \pi_{2N-1}, 1)$. Indeed, we have the single-step iteration
$$\pi_i = \sum_{j=0}^{2N} P_{ij}\pi_j$$
which is just the eigenvalue equation for $\mathbf{P}$ with eigenvalue $\lambda = 1$.

5. For a neutral process with two absorbing states, we can immediately determine the fixation probability from the martingale property of the process. Assume that we start in state $p_0 = i/2N$. Since any process will eventually be absorbed in either $\mathbf{x}^{(0)}$ or in $\mathbf{x}^{(1)}$, we have

$$\lim_{t\to\infty} \mathrm{E}[p_t] = \frac{i}{2N} = \pi_i \cdot 1 + (1-\pi_i)\cdot 0 \quad \Rightarrow \quad \pi_i = \frac{i}{2N}\,.$$

In particular, the fixation probability of a single new mutation in a population is $\pi_1 = 1/2N$.

Random genetic drift has consequences for the variance of allele frequencies among and within populations. For the variance among colonies that derive from the same ancestral founder population, we have already derived above that $V = p_0(1-p_0)/2N$ after a single generation. After a long time, we get

$$V_\infty = \lim_{t\to\infty}\Big(\mathrm{E}[(p_t)^2] - \big(\mathrm{E}[p_t]\big)^2\Big) = p_0 - p_0^2 = p_0(1-p_0)\,.$$

The variance among populations thus increases with drift to a finite limit. To measure variance within a population, we define the homozygosity F_t and the heterozygosity H_t as follows

$$F_t = p_t^2 + (1-p_t)^2 \quad ; \quad H_t = 2p_t(1-p_t) = (1-F_t)\,.$$

The homozygosity (heterozygosity) is the probability that two randomly drawn individuals carry the same (a different) allelic state, where the same individual may be drawn twice (i.e. with replacement). We can generalize this definition for a model with k different alleles with frequencies $p_t^{(1)},\ldots,p_t^{(k)}$ and $\sum_i p_t^{(i)} = 1$,

$$F_t = \sum_{i=1}^{k}\big(p_t^{(i)}\big)^2 = 1 - H_t\,.$$

We obtain the single-step iteration

$$F_t = \frac{1}{2N} + \Big(1 - \frac{1}{2N}\Big)F_{t-1}$$

Indeed, if we take two random alleles (with replacement) from the population in generation t, the probability that we have picked the same allele twice is $1/2N$. If this is not the case, we choose parents for both alleles in the previous generation $t-1$. By definition, the probability that these parents carry the same state is F_{t-1}. From this we get for the heterozygosity

$$H_t = \Big(1-\frac{1}{2N}\Big)H_{t-1} = \Big(1-\frac{1}{2N}\Big)^t H_0 \approx H_0\exp[-t/2N]\,.$$

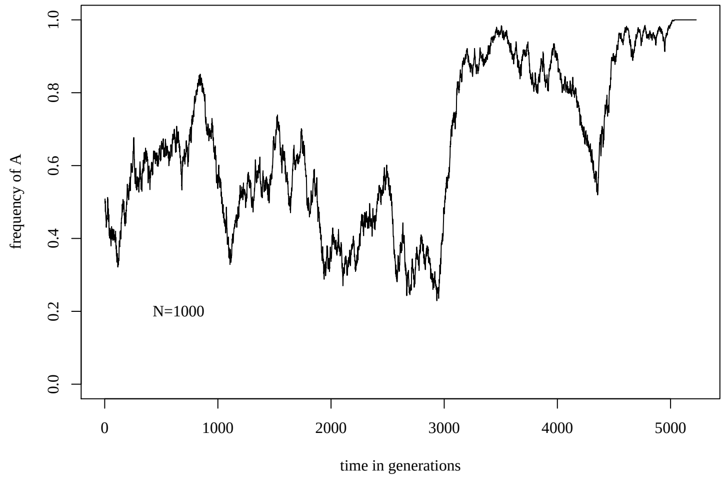

Figure 3.4: Frequency curve of one allele in a Wright-Fisher Model. Population size is $2N = 2000$ and time is given in generations. The initial frequency is 0.5.

We see that drift reduces variability within a population and $H_t \to 0$ as $t \to \infty$. The characteristic time for approaching a monomorphic state is given by the (haploid) population size. We can derive the half time for H_t as follows

$$\frac{H_t}{H_0} = \Big(1 - \frac{1}{2N}\Big)^{t_{1/2}} \approx \exp[-t_{1/2}/2N] := \frac{1}{2}$$

and thus

$$t_{1/2} = 2N \log[2] \approx 1.39N\,.$$

The half time scales with the population size. Note that the time scale to approach a monomorphic state does not depend on the number of alleles that are initially present.

Note finally that heterozygosity and homozygosity (as defined here) should not be confused with the frequency of heterozygotes and homozygotes in a population. Both quantities only coincide under the assumption of random mating. For this reason, some authors (e.g. Charlesworth and Charlesworth 2010) prefer the terms *gene diversity* for H_t and *identity by descent* for F_t.

Exercises

1. We have defined homozygosity and heterozygosity by drawing individuals with replacement. How do the formulas look like if we define these quantities without replacement (which is sometimes also done in the literature)?

2. Consider the neutral Wright-Fisher model with a variable population size. What is then the fixation probability of a new mutant that arises in generation 1?

4 Neutral theory

In a pure drift model, genetic variation within a population can only be eliminated, but never created. To obtain even the most basic model for evolution, we need to include mutation as the ultimate source for new variation. Just these two evolutionary forces, mutation and drift, are the only ingredients of the so-called *neutral theory*, developed by Motoo Kimura in the 50s and 60s. Kimura famously pointed out that models without selection already explain much of the observed patterns of polymorphism within species and divergence between species. Importantly, Kimura did not claim that selection is not important for evolution. It is obvious that purifying selection is responsible for the maintenance of functional important parts of the genome (e.g. in coding regions). However, Kimura claimed that most differences that we see within and among populations are not influenced by selection. Today, selection is thought to play an important role also for these questions. However, the neutral theory is the standard null-model of population genetics. This means, if we want to make the case for selection, we usually do so by rejecting the neutral hypothesis. This makes understanding of neutral evolution key to all of population genetics.

> Motoo Kimura, 1924–1994, published several important, highly mathematical papers on random genetic drift that impressed the few population geneticists who were able to understand them (most notably, Wright). In one paper, he extended Fisher's theory of natural selection to take into account factors such as dominance, epistasis and fluctuations in the natural environment. He set out to develop ways to use the new data pouring in from molecular biology to solve problems of population genetics. Using data on the variation among hemoglobins and cytochromes-c in a wide range of species, he calculated the evolutionary rates of these proteins. Extrapolating these rates to the entire genome, he concluded that there could not be strong enough selection pressures to drive such rapid evolution. He therefore decided that most evolution at the molecular level was the result of neutral processes like mutation and drift. Kimura spent the rest of his life advancing this idea, which came to be known as the "neutral theory of molecular evolution" (adapted from `http://hrst.mit.edu/groups/evolution`.)

4.1 Mutation schemes

There are three widely used schemes to introduce (point) mutations to a model of molecular evolution:

1. With a finite number of alleles, we can define transition probabilities from any allelic state to any other state. For example, there may be k different alleles A_i, $i = 1, \ldots, k$ at a single locus and a mutation probability from A_i to A_j given by μ_{ij}. Then $\mu_i = \sum_{j \neq i} \mu_{ij}$ is the total mutation rate per generation in state A_i. Mutation according to this scheme is most easily included into the Wright-Fisher model as an additional

step on the level of the infinite gamete pool,

$$\mathbf{p}_t \rightarrow \mathbf{p}'_{t+1} = \mathbf{p}_t \cdot \mathbf{U}$$

where $\mathbf{p}_t$ is the (row) vector of allele frequencies and the mutation matrix $\mathbf{U}$ has elements μ_{ij} for $i \neq j$ and $\mu_{ii} \equiv 1 - \mu_i$. We then obtain the frequencies in the next generations $\mathbf{p}_{t+1}$ from $\mathbf{p}'_{t+1}$ by multinomial sampling as in the model without mutation.

2. If we take a whole gene as our locus, we get a very large number of possible alleles if we distinguish different amino acid sequences. In particular, back mutation to an ancestral allelic state becomes very unlikely. In this case, it makes sense to assume an effectively infinite number of alleles in an evolutionary model,

$$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow \ldots$$

Usually, a uniform mutation rate u from one allelic state to the next is assumed. Formally, the *infinite alleles model* corresponds to a Markov chain with an infinite state space.

3. In the infinite alleles model, we assume that the latest mutation erases all the memory of the previous state. Only the latest state is visible. However, for a stretch of DNA, point mutation rates at a single site (or nucleotide position) are very small. We can thus assume that subsequent point mutations will always happen at different sites and remain visible. This leads to the so-called *infinite sites model* for mutation that is widely applied in molecular evolution. In particular, under the assumptions of the infinite sites model (no “double hits”), we can count the number of mutations that have occurred in a sequenced region – given that we have information about the ancestral sequence.

4.2 Predictions from neutral theory

We can easily derive several elementary consequences of neutral theory, given one of the mutation schemes above.

- Under the infinite sites model, new mutations enter a population at a constant rate $2Nu$, where u is the mutation rate per generation and per individual for the locus (stretch of DNA sequence) under consideration. Since any new mutation has a fixation probability of $1/(2N)$, we obtain a neutral substitution rate of

$$k = 2Nu \cdot \frac{1}{2N} = u\,.$$

Importantly, the rate of neutral evolution is independent of the population size and also holds if $N = N(t)$ changes across generations. As long as the mutation rate u can be assumed to be constant, neutral substitutions occur constant in time. They define a so-called *molecular clock*, which can be used for molecular dating of phylogenetic events.

- For the evolution of the homozygosity F_t or heterozygosity H_t under mutation and drift, we obtain for the infinite alleles model or the infinite sites model

  $$F_t = 1 - H_t = (1-u)^2\left(1 - \left(1 - \frac{1}{2N}\right)H_{t-1}\right).$$

  In the long term, the population will approach a state where both forces, mutation and drift balance. We thus reach an equilibrium, $H_t = H_{t-1} = H$, with

  $$H = \frac{1-(1-u)^2}{1-(1-u)^2(1-1/2N)} = \frac{\Theta(1-u/2)}{\Theta(1-u/2)+(1-u)^2} \approx \frac{\Theta}{\Theta+1}$$

  where $\Theta = 4Nu$ is the population mutation parameter. In the case with a finite number of alleles, we need to account for cases where one allelic state can be produced by multiple mutations (i.e., F_t measures the identity in state rather than just the identity by descent). For two alleles with symmetric mutation at rate u in both directions,

  $$1 - H_t = (1-2u)\left(1 - \left(1 - \frac{1}{2N}\right)H_{t-1}\right) + 2u\left(1 - \frac{1}{2N}\right)H_{t-1}$$

  and thus

  $$H = \frac{\Theta}{2\Theta+1-4u} \approx \frac{\Theta}{2\Theta+1}.$$

- For the special case of the expected *nucleotide diversity*, denoted as $\mathrm{E}[\pi]$, where the focus is on a single nucleotide site, we usually have $\Theta \ll 1$. We can then further approximate

  $$\mathrm{E}[\pi] = H_{\text{nucleotide}} \approx \Theta\,,$$

  independently of the mutational scheme that is used.

5 The coalescent

Until now, in our outline of the Wright-Fisher model, we have shown how to predict the state of the population in the next generation $(t + 1)$ given that we know the state in the current generation (t). This is the classical approach in population genetics and follows the evolutionary process forward in time. This view is most useful if we want to predict the evolutionary outcome under various scenarios of mutation, selection, population size and structure, etc. that enter as parameters into the model. However, these model parameters are not easily available in natural populations. Usually, we rather start out with data from a present-day population. In molecular population genetics, this will be mostly sequence polymorphism data from a population sample. The key question then becomes: What are the evolutionary forces that have shaped the observed patterns in our data? Since these forces must have acted in the history of the population, this naturally leads to a genealogical view of evolution backward in time. This view in captured by the so-called coalescent process (or simply *the coalescent*), which has caused a small revolution in molecular population genetics since its introduction in the 1980's. There are three main reasons for this:

- The coalescent is a valuable mathematical tool to derive analytical results that can be directly linked to observable data.
- The coalescent leads to very efficient simulation procedures.
- Most importantly, the coalescent allows for an intuitive understanding of patterns in DNA polymorphism data and of how these patterns result from evolutionary processes.

For all these reasons, we will introduce this modern backward view of evolution in parallel to the classical forward picture.

The coalescent process describes the genealogy of a population sample. The key event of this process is therefore that, going backward in time, two or more individuals share a common ancestor. We can ask, for example: what is the probability that two individuals from the population today (t) have the same ancestor in the previous generation $(t - 1)$? For the neutral Wright-Fisher model, this can easily be calculated because all individuals pick a parent at random. If the population size is $2N$ the probability that two individuals choose the same parent is

$$p_{c,1} = \Pr[\text{common parent one generation ago}] = \frac{1}{2N}. \tag{5.1}$$

Given the first individual picks its parent, the probability that the second one picks the same one by chance is 1 out of $2N$ possible ones. This can be iterated into the past. Given that the two individuals did not find a common ancestor one generation ago maybe they found one two generations ago and so on. We say that the lines of descent from the two individuals *coalescence* in the generation where they find a common ancestor for the first time. The probability for coalescence of two lineages exactly t generations ago is therefore

$$p_{c,t} = \Pr\left[\begin{array}{l}\text{two lineages coalesce}\\ t\text{ generations ago}\end{array}\right] = \frac{1}{2N}\Big(1 - \frac{1}{2N}\Big)^{t-1}.$$

Mathematically, we can describe the *coalescence time* as a random variable that is geometrically distributed with success probability $\frac{1}{2N}$. Figure 5.1 shows an example for the common ancestry like it can be generated by a simulation animator, such as the Wright-Fisher animator on www.coalescent.dk. In this case the history of just two individuals is highlighted. Going back in time there is always a chance that they choose the same parent. In this case they do so after 11 generations. In all the generations further back in time they will automatically also have the same ancestor. The common ancestor in the 11th generation in the past is therefore called the *most recent common ancestor* (MRCA).

The coalescence perspective is not restricted to a sample of size two but can be applied to any number of individuals. For a sample of size n from the Wright-Fisher model of size $2N$, the probability of coalescence in a single generation is

$$\begin{aligned} p_{c,1}^{(n)} &= 1 - \Big(1 - \frac{1}{2N}\Big)\cdot\Big(1 - \frac{2}{2N}\Big)\cdots\Big(1 - \frac{n-1}{2N}\Big) = 1 - \prod_{i=1}^{n-1}\Big(1 - \frac{i}{2N}\Big) \\ &= \frac{1}{2N}\sum_{i=1}^{n-1} i + \mathcal{O}\Big[\Big(\frac{n}{N}\Big)^2\Big] = \frac{1}{2N}\binom{n}{2} + \mathcal{O}\Big[\Big(\frac{n}{N}\Big)^2\Big]. \end{aligned} \tag{5.2}$$

We can interprete this result as follows. In a sample of size n, there are $\binom{n}{2}$ possible coalescence events between pairs of individuals. If we assume that $n \ll N$, multiple coalescence events in a single generation can be ignored and the leading order term in $p_{c,1}^{n}$ just accounts for the probability of a single pairwise coalescence event in the sample in the previous generation. Multiple coalescence events and coalescence events of more than two lineages simultaneously (so-called "multiple mergers") only contribute to the error term $\sim \mathcal{O}[N^-2]$, which can be ignored for small samples in a large population. In this approximation, the coalescence probability after t generation in a sample of size n becomes

$$p_{c,t}^{(n)} \approx \frac{1}{2N}\binom{n}{2}\cdot\left(1 - \frac{1}{2N}\binom{n}{2}\right)^{t-1}. \tag{5.3}$$

We can then construct the genealogical history of the sample in a two-step procedure:

1. First, fix the topology of the coalescent tree. I.e., decide (at random), which pairs of genealogical lineages from individuals in a sample coalesce first, second, etc., until the MRCA of the entire sample is found.

2. Second, specify the times in the past when these coalescence events have happened. I.e., draw a so-called coalescent time for each coalescent event. This is independent of the topology.

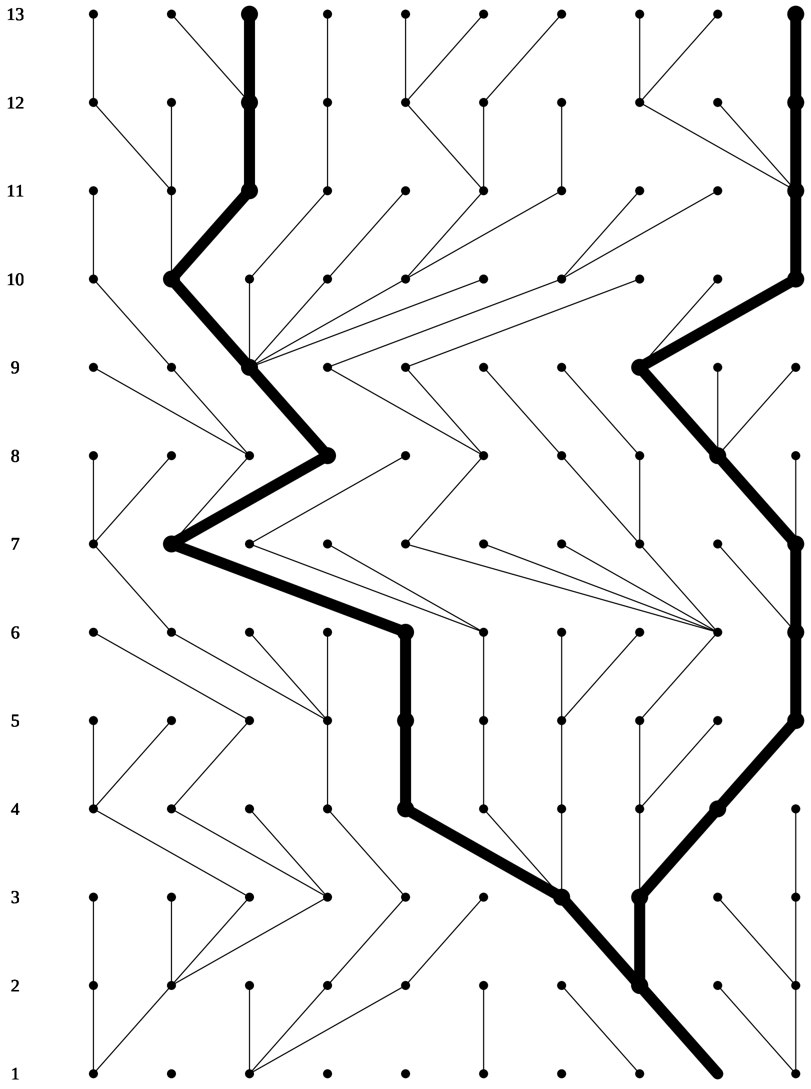

Figure 5.1: The coalescent of two lines in the Wright-Fisher Model

5.1 Topologies

With only pairwise coalescence events, the topology of a coalescence tree is easy to model. Consider a sample of size n and represent the ancestry of this sample as coalescing lineages back in time. Since each coalescence event reduces the number of ancestral lines by one, it takes $n-1$ such events to reach the MRCA as the root of the tree. We say that the tree is *in state* k at some time t in the past if there are k ancestral lines at this time. Looking further back in time, all $k(k-1)/2$ pairs of lines that can be chosen from these k lines are equally likely to be involved in the next coalescence event. If we start the coalescent process with labeled individuals (representing the n tips of the tree in our sample), we thus have

$$\prod_{k=2}^{n} \binom{k}{2} = \frac{n!(n-1)!}{2^{n-1}} \tag{5.4}$$

different *labeled and time-ordered histories*, where we do not only distinguish who coalesces with whom, but also different time orders in the coalescence events. In many cases, however, we are not interested in the genealogy of a specific sample, but in the statistical properties of (e.g. neutral) coalescent trees, such as the number of subtrees of a certain size – irrespectively of any labels at the tips of the tree. In this case, it is sometimes easier to construct the coalescent tree forward in time: For a tree currently in state k, we simply pick one of the lines at random and split it to obtain state $k+1$. For example, we can prove the following

Theorem 1 *Take a random coalescent tree of size n and consider the k branches that exist at state k of the tree. Let λ_i, $i = 1, \ldots, k$ be the number of offspring of the ith branch. Such a branch is also called a branch of* size λ_i. *Then, the offspring $(\lambda_1, \ldots, \lambda_k)$ of all k branches is uniformly distributed over all k-dim vectors with entries $\lambda_i \in \mathbb{N}_+$ and $\sum_i \lambda_i = n$.*

- Note that there are

  $$\binom{n-1}{k-1} \tag{5.5}$$

  such vectors. To see this, imagine that we distribute n (identical) balls over k (labeled) groups. We can put all n balls next to each other in a single line and then place vertical lines between the balls to delimit the groups. Then, $k-1$ demarcation lines are needed, which can go in any of the $n-1$ spaces between the balls.

Proof To prove the theorem, consider first a specific history, forward in time, starting at state k: Imagine that the first $\lambda_1 - 1$ split events all occur in descendants of the first branch, followed by $\lambda_2 - 1$ split events in offspring of the second branch, and so on until the k-th branch, which needs $\lambda_k - 1$ split events in its offspring. The probability of this particular history is

$$\left(\frac{1}{k}\cdot\frac{2}{k+1}\cdots\frac{\lambda_1-1}{k+\lambda_1-2}\right)\cdot\left(\frac{1}{k+\lambda_1-1}\cdots\frac{\lambda_2-1}{k+\lambda_1+\lambda_2-3}\right)\cdots$$
$$\cdots\left(\frac{1}{k+\sum_{i=1}^{k-1}\lambda_i-k+1}\cdots\frac{\lambda_k-1}{k+\sum_{i=1}^{k}\lambda_i-k-1}\right)=\frac{(k-1)!\prod_{i=1}^{k}(\lambda_i-1)!}{(n-1)!}\,. \tag{5.6}$$

As long as there are λ_i-1 splitting events in the descendants of the ith branch ($i=1,\dots,k$), we will always obtain the same distribution $(\lambda_1,\dots,\lambda_k)$, irrespective of the order of these splitting events. If we can calculate the probability of each of these alternative histories in a stepwise procedure like in (5.6), it is easy to see that the only difference to (5.6) is a permutation of the numbers in the numerator. We conclude that the probability of all alternative histories to obtain a specific offspring distribution $(\lambda_1,\dots,\lambda_k)$ is identical. The number of alternative histories for a given distribution is given by the multinomial coefficient $\binom{n-k}{\lambda_1-1,\dots,\lambda_k-1}$, and thus

$$\Pr[(\lambda_1,\dots,\lambda_k)]=\frac{(k-1)!(n-k)!}{(n-1)!}=\binom{n-1}{k-1}^{-1}\,. \tag{5.7}$$

- The splitting scheme is also known as the *Polya urn scheme* in the mathematical literature. This scheme starts with an urn containing k balls with k different colors. Then, each round, take out one ball, put it back in and add another ball of the same color.

- For $k=2$, the result says that if we pick one of the branches after the first split, the size of this branch will be uniformly distributed on $1,2,\dots,n-1$. In the limit $n\to\infty$, we obtain a coalescent tree of the "whole population". Then, the proportion X of lines that derive from the left branch after the first split is uniformly distributed on the interval $(0,1)$. Consider now the coalescent tree of a random sample of size m. The MRCA of the sample tree will be the same one as for the population tree unless either all m lines or no lines at all trace back to the left branch after the first split of the population tree. This occurs with probability

  $$\int_0^1\left(x^m+(1-x)^m\right)dx=\frac{2}{m+1}\,.$$

  The probability that the population MRCA coincides with the sample MRCA is thus

  $$1-\frac{2}{m+1}=\frac{m-1}{m+1}\,. \tag{5.8}$$

  In more general, if we pick a subsample of size m of a sample of size n, the probability that both samples go back to the same MRCA is

  $$\frac{(m-1)(n+1)}{(m+1)(n-1)}\,. \tag{5.9}$$

For a proof, consider the n-tree after the first split and calculate the probability that all m lines of the subsample go back to the left branch,

$$p_l = \frac{1}{n-1}\sum_{k=m}^{n-1} \frac{k}{n}\frac{k-1}{n-1}\cdots\frac{k-m+1}{n-m+1} = \frac{m!(n-m)!}{(n-1)n!}\sum_{k=m}^{n-1}\binom{k}{m} = \frac{n-m}{(n-1)(m+1)}$$

using the summation formula Eq. (10.3). The result (5.9) is then obtained as $1 - 2p_l$, since the m lines can either go back to the left or the right branch with equal probability.

- In general, the uniform distribution over the branch sizes leads to a much higher variance in branch size than expected under a binomial or multinomial distribution: neutral coalescent trees can be both balanced or unbalanced.

Number of possible rooted and unrooted trees

In the examples above, we did not distinguish trees according to their branch length, but we have still accounted for the order of coalescence events. However, we can also count coalescence trees without any reference to time order.

For a sample of size n, we have $n-1$ coalescence events until we reach the MRCA (the *root*). This creates $2n-1$ so-called *vertices* in the tree: n are external (the *leaves*) and $n-1$ are internal. Every vertex has a branch directly leading to the next coalescence event. Only the root, which is also a vertex in the tree, does not have a branch. This makes $2n-2$ branches in a rooted tree with n leaves. As two branches lead to the root, the number of branches in an unrooted tree with n leaves is $2n-3$.

Let $\mathcal{B}_n$ be the number of topologies of unrooted trees with n leaves. We can derive this number recursively. Assume we have a tree with $n-1$ leaves, representing the first $n-1$ sampled sequences. We can ask in how many ways the nth sequence can be added to this tree. There are $2n-5$ branches in a tree with $n-1$ leaves. Since any branch can have the split leading to the nth leave, we obtain

$$\mathcal{B}_n = (2n-5)\mathcal{B}_{n-1}.$$

It is easy to see that there is only a single unrooted tree with three leaves. Thus

$$\mathcal{B}_n = 1\cdot 3\cdot 5\cdots(2n-7)\cdot(2n-5) = (2n-5)!!\,. \qquad (5.10)$$

5.2 Coalescence times

For the branch lengths of the coalescent tree, we need to know the coalescence times. For a sample of size n, we need $n-1$ times until we reach the MRCA. As stated above, these times are independent of the topology. Mathematically, we obtain these times most conveniently by an approximation of the geometrical distribution by the exponential distribution for large N:

- If X is geometrically distributed with small success probability p and t is large then

  $$\Pr[X \geq t] = (1-p)^t \approx e^{-pt}.$$

  This is the distribution function of an exponential distribution with parameter p.

Let t_n be the time until the first coalescence occurs in a smaple of size n. This time is geometrically distributed according to

$$\Pr[t_n > t] = \left[1 - \frac{\binom{n}{2}}{2N}\right]^t = \left[1 - \frac{n(n-1)}{4N}\right]^t. \tag{5.11}$$

The mean waiting time until the first coalescence event is $\mathrm{E}[t_n] = 4N/n(n-1)$ and thus proportional to the population size. It is standard to integrate this dependence into a "coalescent time scale"

$$\tau := \frac{t}{2N}.$$

We can then take the limit $N \to \infty$ to obtain a stochastic process with a continuous time parameter τ. Coalescence times $T_n := t_n/2N$ in this limiting process are distributed like

$$\Pr[T_n > \tau] = \lim_{N\to\infty} \left[1 - \frac{\binom{n}{2}}{2N}\right]^{2N\tau} = \exp\left[-\tau\binom{n}{2}\right]. \tag{5.12}$$

In a sample of size n, the time to the first coalescence is thus exponentially distributed with parameter $\lambda = n(n-1)/2$. The fact that in the coalescent the times are exponentially distributed enables us to derive several important quantities.

- The time to the MRCA,

  $$T_{\mathrm{MRCA}}(n) = \sum_{k=2}^{n} T_k,$$

  is the sum of $n-1$ mutually independent exponentially distributed random variables. Its expectation and variance derive to

  $$\mathrm{E}[T_{\mathrm{MRCA}}(n)] = \sum_{k=2}^{n} \mathrm{E}[T_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{n}\Big) \tag{5.13}$$

  and

  $$\mathrm{Var}[T_{\mathrm{MRCA}}(n)] = \sum_{k=2}^{n} \mathrm{Var}[T_k] = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} = 8\sum_{k=2}^{n}\frac{1}{k^2} - 4\Big(1 - \frac{1}{n}\Big)^2. \tag{5.14}$$

  We have $\mathrm{E}[T_{\mathrm{MRCA}}(n)] \to 2$ for large sample sizes $n \to \infty$. Note that $\mathrm{E}[T_{\mathrm{MRCA}}(2)] = 1$, so that in expectation more than half of the total time to the MRCA is needed for the last two ancestral lines to coalesce. Similarly, $\mathrm{Var}[T_{\mathrm{MRCA}}(n)] \to 4\pi^2/3 - 12 \approx 1.16$ for $n \to \infty$ is dominated by $\mathrm{Var}[T_2] = 1$.

- Due to the independence of the coalescence times, the full distribution of $T_{\mathrm{MRCA}}(n)$ can be derived as an $(n-2)$-fold convolution,

$$f_{T_{\mathrm{MRCA}}(n)}(\tau) = \sum_{k=2}^{n} \binom{k}{2} \exp\left[-\binom{k}{2}\tau\right] \prod_{j=2, j\neq k}^{n} \frac{\binom{j}{2}}{\binom{j}{2}-\binom{k}{2}}. \tag{5.15}$$

- For the total tree length,

$$L(n) = \sum_{k=2}^{n} kT_k,$$

  we obtain the expected value

$$\mathrm{E}[L(n)] = \sum_{k=2}^{n} k\,\mathrm{E}[T_k] = 2\sum_{k=2}^{n} \frac{1}{k-1} = 2\sum_{k=1}^{n-1}\frac{1}{k}. \tag{5.16}$$

  and the variance

$$\mathrm{Var}[L(n)] = \sum_{k=2}^{n} k^2\,\mathrm{Var}[T_k] = 4\sum_{k=1}^{n-1}\frac{1}{k^2}. \tag{5.17}$$

  Increasing the sample size will mostly add short twigs to a coalescent tree. As a consequence, also the total branch length

$$\mathrm{E}[L(n)] \approx 2(\log(n-1)+\gamma) \quad ; \quad \gamma = 0.577216\ldots.$$

  increases only very slowly with the sample size (γ is the Euler constant). The variance even approaches a finite limit $2\pi^2/3 \approx 6.58$ for $n \to \infty$.

- Again, also the entire distribution can be derived and takes a relatively easy form,

$$f_{L(n)}(\tau) = \frac{n-1}{2}\exp[-\tau/2]\Big(1-\exp[-\tau/2]\Big)^{n-2} \tag{5.18}$$

- As we have seen above, the probability that the coalescent of a sample of size n contains the MRCA of the whole population is $(n-1)/(n+1)$ (for large, finite N). An important practical consequence of these findings is that, under neutrality, relatively small sample sizes (typically 10-20) will usually be enough to gain all statistical power that is available from a single locus.

5.3 Polymorphism patterns

In order to generate DNA diversity patterns using the coalescent, we need to add mutations to the process. This can be done according to any of the mutation schemes introduced in section (4.1). Most frequently used are the infinite sites and the infinite alleles model, which we will discuss in the following.

The key insight for the description of neutral DNA diversity using the coalescent is that neutral mutations do not interfere with the genealogy: *state* (the genotype) and *descent* (the genealogical relationships) are decoupled for neutral evolution. This is easy to see from the time-forward dynamics, since parents carrying different variants of a neutral allele are still equivalent concerning the distribution of their offspring in all future generations. If we want to create a random neutral polymorphism pattern using the coalescent process, we can therefore pick a genealogy first (as described in the previous section) and decide on the state later on. This is done by so-called *mutation dropping*, where mutations are added to all branches of the tree.

Let us first discuss the infinite sites mutation scheme. I.e. each mutation hits a new site (and thus leads to a new allele) and all mutations on a genealogy remain visible. If a mutation occurs on a branch of size i in the genealogy of n individuals, it will give rise to a polymorphism with frequency i/n of the derived (mutant) allele. Note that we do not need to know the precise time for the origin of the mutations in the genealogy, all that is needed is the total number of mutations that fall on each branch. On genealogical time scales (as opposed to phylogenetic time scales), we can usually assume that the mutation rate u (per haploid individual and generation) is constant.

For a branch of length l, we therefore directly get the number of neutral mutations on this branch by drawing from a Poisson distributed with parameter $2Nlu$. The factor $2N$ accounts for the fact that branch length l is measured on the coalescent time scale (in units of $1/2N$). In particular, the total number of mutations in an entire coalescent tree of length L is Poisson distributed with parameter $2NLu$. Let S be the number of segregating (polymorphic) sites in a sample. Since each polymorphic site corresponds to exactly one mutation on the tree under the infinite sites model, we have

$$\Pr[S=k] = \int_0^\infty \Pr[S=k|\ell]\cdot f_{L(n)}(\ell)d\ell = \int_0^\infty e^{-2N\ell u}\frac{(2N\ell u)^k}{k!}\cdot f_{L(n)}(\ell)d\ell\,.$$

For the expectation that means

$$\begin{aligned}\mathrm{E}[S] = \sum_{k=0}^{\infty} k\Pr[S=k] &= \int_0^\infty \frac{\ell\theta}{2}e^{-\ell\theta/2}\Big(\sum_{k=1}^{\infty}\frac{(\ell\theta/2)^{k-1}}{(k-1)!}\Big)\cdot f_{L(n)}(\ell)d\ell \\ &= \frac{\theta}{2}\int_0^\infty \ell\, f_{L(n)}(\ell)d\ell = \frac{\theta}{2}\,\mathrm{E}[L(n)] = \theta\sum_{i=1}^{n-1}\frac{1}{i} = a_n\theta\end{aligned} \tag{5.19}$$

with

$$a_n = \sum_{i=1}^{n-1}\frac{1}{i}\,, \tag{5.20}$$

and where

$$\theta = 4Nu$$

is the standard population mutation parameter. Note that the distribution of S does not depend on the coalescent topologies, but only on the distribution of the coalescence times.

The mismatch distribution

For a Poisson distributed random variable, the time interval between consecutive events is exponentially distributed. There is therefore an alternative way to derive the equilibrium heterozygosity H (or the number of polymorphic sites in a sample of size 2) using the coalescent. If we follow the genealogy of two copies of a homologous site back in time, two things can happen first: (1) either one of the two mutates or (2) they coalesce. If they coalesce first they are identical by descent, if one of the two mutates, they are not identical. For both processes, the time back to the first event is exponentially distributed. Since mutation (in either lineage) occurs at rate $2u$ and coalescence occurs at rate $1/2N$, we directly obtain using Eq. (10.17),

$$H(u,N) = \frac{2u}{2u + (1/2N)} = \frac{\theta}{\theta + 1}\,. \tag{5.21}$$

We can easily extend this result and ask for the probability that we find precisely k differences among the two sequences. Under the assumptions of the infinite-sites model, and using that we can re-start the Poisson process after every event,

$$\Pr[\pi = k] = \Big(\frac{\theta}{\theta+1}\Big)^k \frac{1}{\theta+1}\,, \tag{5.22}$$

which is a modified geometrical distribution. Note that this is not the distribution of pairwise mismatches in a larger sample, which will be correlated due to a shared history. However, under the standard neutral model, we should see this distribution if we sequence from independent loci along the genome (e.g. counting mismatches among the two copies carried by a diploid individual).

The site frequency spectrum

The total number S of polymorphic sites is the simplest so-called *summary statistic* of polymorphism data. There are many more. As a next step, we can ask for the number S_i of mutations of a given size i. To derive the expected value $\mathrm{E}[S_i]$, we proceed in two steps. First, we ask for the probability that a branch at state k of the coalescent process is of size i,

$$P[i|k] := \Pr[\text{Probability for branch at state } k \text{ to be of size } i]\,.$$

From Theorem 1 we directly obtain $P[i|k]$ as

$$P[i|k] = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \tag{5.23}$$

Here, the numerator counts all different possibilities to distribute $n - i$ descendants over $k - 1$ branches – i.e. all remaining descendants after we have assigned i descendants to the focal branch. Note that $P[i|k]$ does not depend at all on the coalescent times, but only on aspects of the topology. In the second step, we ask for the expected number $\mathrm{E}[S^{(k)}]$ of mutations on a branch at state k. Noting that the length of such a branch is T_k, this is easily derived (analogous to Eq. 5.19),

$$\mathrm{E}[S^{(k)}] = \frac{\theta}{2}\,\mathrm{E}[T_k] = \frac{\theta}{k(k-1)}\,. \tag{5.24}$$

In contrast to $P[i|k]$, this expression does not depend on the topologies, but only on the coalescent times. Using the independence of coalescent times and topologies, we now obtain the expected number of mutations of size i as

$$\begin{aligned}\mathrm{E}[S_i] &= \sum_{k=2}^{n} kP[i|k]\cdot\mathrm{E}[S^{(k)}]\\ &= \sum_{k=2}^{n}\frac{\theta}{k-1}\frac{(n-i-1)!(k-1)!(n-k)!}{(k-2)!(n-i-k+1)!(n-1)!}\\ &= \frac{\theta}{i\binom{n-1}{i}}\sum_{k=2}^{n}\binom{n-k}{i-1}\\ &= \frac{\theta}{i\binom{n-1}{i}}\sum_{k=2}^{n}\left(\binom{n-k+1}{i}-\binom{n-k}{i}\right)\\ &= \frac{\theta}{i\binom{n-1}{i}}\cdot\binom{n-1}{i} = \frac{\theta}{i}\,,\end{aligned} \tag{5.25}$$

where we have used Eq. (10.2). The expected number of mutations of size i is thus θ/i. Together, these numbers define the (expected) *site frequency spectrum* of sample taken from a standard neutral population.

- The frequencies of the expected normalized site frequency spectrum are $p_i = 1/(a_n i)$. They are independent of θ. The characteristic $(1/i)$-shape is a prime indicator of "neutrality".

- We can easily obtain an empirical site frequency spectrum from any polymorphism data. This empirical spectrum can then be compared to the spectrum predicted under neutrality. Note that we need data from many independent (unlinked) loci to observe the *expected* spectrum. For any single locus, the spectrum can differ considerably, because we only have a single coalescent history.

- To determine the size of a given polymorphism in the sample, we need to know the ancestral state at the locus. In practice, this is inferred from a so-called outgroup (usually a single consensus sequence from a closely related sister species). If the ancestral state cannot be determined, we can work with the so-called *folded site frequency spectrum*, with mutation classes $\tilde{S}_i = S_i + S_{n-i}$ for $i < n/2$ and $\tilde{S}_i = S_i$ for $i = n/2$.

Infinite alleles and haplotype statistics

So far, we have considered polymorphism patterns under the assumption of the infinite sites model, where all mutations that occur during the genealogy of a sample remain visible as a polymorphic site. Depending on the type of the mutation, however, this may not always be true. For example, the infinite sites model does not easily generalize to insertion/deletion mutations. Alternatively, we may focus on the entire haplotype in a chromosomal window and just ask for the distribution of different types (ignoring any information about the mutational distances between these types). Questions like these can be addressed within the framework of the infinite alleles model.

Just like in the case of the infinite sites model, we can construct the genealogical tree first and add mutations later on. However, for the infinite alleles model, only the latest mutations (the ones closest to the leaves of the tree) will be observed. As a consequence, major parts of the genealogy do not influence the pattern. We can account for this by adding mutations already as we build the genealogy. Once we encounter the first mutation in the ancestry of an individual, we know the state of this this ancestor and of all its descendants. So, before we construct the genealogy further back in time, we can stop (or *kill*) this branch. This leads to the so-called *coalescent with killings*, where we have two kinds of events:

1. As before, coalescence events occur at rate $k(k-1)/2$ on the coalescence time scale for a tree in state k (i.e. with k ancestral lines).

2. In addition, we directly account for mutation events, which occur at rate $k\theta/2$ in state k. Each mutation "kills" the corresponding branch.

Let K_n be the number of different haplotypes that we observe in a sample of size n. We are interested in the probability that K_n takes a certain value k. By following the coalescent with killings back in time to the first event (either coalescence or mutation), we can relate the values for the distribution of K_n to the corresponding values for K_{n-1},

$$P[K_n = k] = \frac{\theta}{\theta + n - 1} \cdot P[K_{n-1} = k-1] + \frac{n-1}{\theta + n - 1} \cdot P[K_{n-1} = k]\,. \tag{5.26}$$

As initial condition, we have $P[K_1 = 1] = 1$. To solve this recursion, observe that the denominator in both terms in (5.26) is the same. Note also, that for $K_n = k$, we need to choose "mutation" $k-1$ times before we reach a sample of size 1. Each time, we pick up a factor of θ, like in the first term of (5.26). We thus can write

$$P[K_n = k] = \frac{\theta^{k-1}}{(\theta+1)(\theta+2)\cdots(\theta+n-1)} \cdot S(n,k) = \frac{\theta^k}{\theta_{(n)}} \cdot S(n,k) \tag{5.27}$$

where

$$\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$$

and the $S(n,k)$ are the so-called *Stirling numbers of the first kind*, which follow the recursion relation

$$S(n,k) = S(n-1,k-1) + (n-1)\cdot S(n-1,k)\,. \tag{5.28}$$

In analogy to the allele frequency spectrum, we can also ask for the frequency distribution of haplotypes. Let A_j be the number of haploytpes that appear j times in a sample of size n. For $K_n = k$, we thus have

$$\sum_{j=1}^{n} A_j = k \quad \text{and} \quad \sum_{j=1}^{n} jA_j = n\,,$$

and let $\boldsymbol{a} = (a_1, \ldots, a_n)$ be a realization of $(A_1, \ldots, A_n)$. We can prove the following

Theorem 2 *The combined distribution of the number and frequencies of haplotypes under the standard neutral model is given by the so-called Ewens' sampling formula,*

$$P_n[\boldsymbol{a}] = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!}\,. \tag{5.29}$$

- One interpretation of this result is to view the A_j as independently Poisson distributed random variables with parameter (= expected value) θ/j, and then consider the marginal distribution under the condition $\sum_{j=1}^{n} jA_j = n$. Note that the distribution is strongly influenced by θ. For large $\theta > 1$, the distribution is dominated by singleton haplotypes $\sim \theta^{a_1}$, for small θ, a large number of singletons (large a_1) is unlikely.

Proof To prove the theorem, we extend the recursion method that we have used in our proof of the distribution of K_n. Note first that for $n = 1$, we have $a_1 = 1$ with probability 1, in accordance with (5.29). Define $\boldsymbol{e}_i = (0, \ldots, 1, 0, \ldots)$ as the ith unit vector (with entry 1 in the ith position). Now, start with a sample of size n and go back to the first event. With probability $\theta/(\theta + n - 1)$, this is a mutation, which creates a new haplotype. This relates the partition $\boldsymbol{a}$ of the n haplotypes in the sample to the partition $\boldsymbol{a} - \boldsymbol{e}_1$ of the remaining $n-1$ types. If the first event is coalescence (which it will be with probability $(n-1)/(\theta + n - 1)$), this decreases the frequency of one of the haplotypes (with at least two copies) by one. In terms of the allelic partitions, this turns $\boldsymbol{a}$ into $\boldsymbol{a} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}$ for some $j \in \{1, \ldots, n-1\}$. Conditioned on this latter partition for the tree at state $n-1$, the probability that (forward in time) the next split event will be in one of the haplotype classes with j copies is thus $(a_j + 1)j/(n-1)$. This results in the recursion

$$P_n[\boldsymbol{a}] \;=\; \frac{\theta}{\theta + n - 1} P_{n-1}[\boldsymbol{a} - \boldsymbol{e}_1] + \frac{n-1}{\theta + n - 1} \sum_{j=1}^{n-1} \frac{(a_j + 1)j}{n-1} P_{n-1}[\boldsymbol{a} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}]. \tag{5.30}$$

It remains to be shown that (5.29) fulfills this recursion. For this, note that (5.29) implies that

$$P_{n-1}[\boldsymbol{a}-\boldsymbol{e}_1] = \frac{(\theta+n-1)a_1}{n\theta}P_n[\boldsymbol{a}] \tag{5.31}$$

$$P_{n-1}[\boldsymbol{a}+\boldsymbol{e}_j-\boldsymbol{e}_{j+1}] = \frac{(\theta+n-1)a_{j+1}(j+1)}{(a_j+1)nj}P_n[\boldsymbol{a}] \tag{5.32}$$

Inserting this into (5.30) yields

$$1 = \frac{a_1}{n} + \sum_{j=2}^{n-1}\frac{a_{j+1}(j+1)}{n} = \frac{1}{n}\sum_{j=1}^{n} a_j\,,$$

which holds true, since $\sum_j a_j = n$.

- The underlying combinatorial problem is also known as the *Hoppe urn scheme* in the mathematical literature. This scheme starts with k colored balls like the *Polya urn*, but adds a special back ball with weight θ. Each time a colored ball is drawn, the ball is returned with another ball of the same color. Each time the black ball is drawn, it is put back with another ball of a new color.

- Note that the marginal distribution for the allelic partition given the number of haplotypes can be written as

$$P_n[\boldsymbol{a}|K_n = k] = \frac{P_n[\boldsymbol{a}]}{P[K_n = k]} = \frac{n!}{S(n,k)}\prod_{j=1}^{n}\frac{1}{a_j!j^{a_j}}\,, \tag{5.33}$$

  which is, in particular, independent of θ. In statistical terms this means that all information about θ is already contained in the number of haplotypes found in a sample: K_n is a *sufficient statistic*. Knowledge about their distribution does not add any further information.

5.4 Coalescent and statistics

Coalescent trees show the genealogical relationships between two or more sequences that are drawn from a population. This should not be confounded with a phylogenetic tree that shows the relation of two or more species. Indeed, both “trees” have entirely different roles for the theory of evolution. In phylogenetics, one is usually interested in the one “true tree” and the parameters of this tree (such as split times) are estimated from data. In contrast, there is no single “true tree” for a set of individuals from a population. Indeed, the genealogy will usually be different for different loci. For example, at a mitochondrial locus your ancestor is certainly your mother and her mother. However, if you are a male, the ancestor for the loci on your Y-chromosome is your father and his father. So the genealogical tree will look different for a mitochondrial locus than for a Y-chromosomal locus. But even for a single locus, we are usually not able to reconstruct a single "true coalescence tree" and this is not the goal in coalescent studies. Instead, coalescent histories are used as a statistical tool for inferences about an underlying model.

The general idea is as follow. We define an evolutionary model that depends on a number of biological parameters (such as mutation rates, population sizes, selection coefficients). Under this model, we obtain a distribution of coalescent histories and (consequently) a distribution of polymorphism patterns that is predicted under this model. We can then compare measured data with the predicted distribution to make statistical inferences. Usually, there is a twofold goal:

1. to reject (or not) the underlying model. This is true, in particular, for the neutral model as the standard null model of population genetics.

2. to estimate model parameters. Note that the parameters of the coalescent tree (coalescent times, topology) are generally not model parameters. They are "integrated out" in the statistical treatment.

In some easy cases (notably the neutral model), key aspects of the distribution of polymorphism patterns can be obtained analytically using coalescent theory. In many other cases, this is no longer possible. However, even in these cases, the coalescent offers a highly efficient simulation framework that is routinely used in statistical simulation packages.

Estimators for the mutation parameter θ

All population genetic models, whether forward or backward in time, depend on a set of biological parameters that must be estimated from data. In the standard neutral model, there are two such parameters: the mutation rate u and the population size N. However, since both parameters only occur in the combination $\theta = 4Nu$, the population mutation parameter is effectively the only parameter of the model. From our derivation of the expected site frequency spectrum, we easily obtain several estimators for θ. In principle, we can use the total number of mutations of any size class to define an unbiased estimator $\hat{\theta}_i$,

$$\mathrm{E}[S_i] = \frac{\theta}{i} \quad \longrightarrow \quad \hat{\theta}_i := i \cdot S_i \,. \tag{5.34}$$

In practice, widely used estimators are linear combinations across mutations of different size classes. They can be distinguished according to the relative weight that is put on a certain class. The most important ones are the following:

1. *Watterson's estimator*,

$$\hat{\theta}_{\mathrm{W}} := \frac{S}{a_n} = \frac{1}{a_n} \sum_{i=1}^{n-1} S_i = \frac{1}{a_n} \sum_{1 \leq i \leq n/2} \tilde{S}_i \,, \tag{5.35}$$

uses the total number of segregating sites and puts an equal weight on each mutation class. The last equation expresses $\hat{\theta}_{\mathrm{W}}$ in terms of frequencies of the folded spectrum.

Remember that the distribution of S – and thus of $\hat{\theta}_{\mathrm{W}}$ – is independent of the coalescent topologies, but only depends on the coalescent times.

2. Let π_{ij} be the number of differences among two sequences i and j from our sample. We have $\mathrm{E}[\pi_{ij}] = \mathrm{E}[S(n=2)] = \theta$. If the sample size is just two, this corresponds to Watterson's estimator. In a larger sample, we can still take the pairwise difference as our basis and average over all $n(n-1)/2$ pairs. This leads to the *diversity-based estimator* (sometimes also called *Tajima's estimator*),

$$\hat{\theta}_\pi := \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij} \,. \tag{5.36}$$

We can also express $\hat{\theta}_\pi$ in terms of the (folded) frequency spectrum as follows,

$$\hat{\theta}_\pi = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i(n-i)S_i = \binom{n}{2}^{-1} \sum_{1\le i\le n/2} i(n-i)\tilde{S}_i \,. \tag{5.37}$$

Whereas Watterson's estimator weights all frequency classes equally, $\hat{\theta}_\pi$ puts the highest weight on classes with an intermediate frequency. In contrast to $\hat{\theta}_{\mathrm{W}}$, it also depends on the distribution of tree topologies. The estimator is often also just written as $\hat{\pi}$.

3. *Fay and Wu's estimator*,

$$\hat{\theta}_{\mathrm{H}} := \binom{n}{2}^{-1} \sum_{i=1}^{n-1} i^2 S_i \,, \tag{5.38}$$

puts a hight weight on mutation classes of the unfolded spectrum with a high frequency of the derived allele. In contrast to the other estimators, it is not a summary statistic of the folded spectrum and thus requires knowledge of the ancestral state.

4. Finally, the *singleton estimator* $\hat{\theta}_{\mathrm{s}}$ uses the singletons of the folded spectrum,

$$\hat{\theta}_{\mathrm{s}} := \frac{n-1}{n}\Big(S_1 + S_{n-1}\Big) = \frac{n-1}{n}\tilde{S}_1 \,. \tag{5.39}$$

It has all its weight at both ends of the unfolded spectrum.

Test statistics for neutrality tests

Estimators of any model parameter, such as θ, will only produce meaningful results if the assumptions of the underlying model hold. In our case, we have assumed standard neutral evolution. In addition to the absence of selection, this includes the assumptions of a constant population size and no population structure. But how can we know whether these assumptions do hold (at least approximately) for a given data set? This question asks for a test of the model assumptions. As it turns out, the availability of various different estimators of the same quantity θ is helpful for the construction of such a test.

The key idea is to consider the difference among two different estimators, such as $\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}$. Under standard neutrality, this quantity should be close to zero, whereas significant deviations indicate that the model should be rejected. The most widely used test statistic that is constructed in such a way is *Tajima's D*,

$$D_\mathrm{T} := \frac{\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_\mathrm{W}]}} \,. \tag{5.40}$$

The denominator of D_T is used for normalization and makes the distribution of the statistic (almost) independent of θ and of the sample size. Tajima has shown that D_T is approximately β-distributed. Today, however, the exact distribution under the standard neutral null model is usually obtained (resp. approximated to arbitrary precision) by computer simulations. For a given significance level α, one can then specify the critical upper and lower bounds for D_T, beyond which the null model should be rejected. Test statistics that are constructed in a similar way are *Fu and Li's D*,

$$D_\mathrm{FL} := \frac{\hat{\theta}_\mathrm{W} - \hat{\theta}_\mathrm{s}}{\sqrt{\mathrm{Var}[\hat{\theta}_\mathrm{W} - \hat{\theta}_\mathrm{s}]}} \tag{5.41}$$

and *Fay and Wu's H*,

$$H_\mathrm{FW} := \frac{\hat{\theta}_\pi - \hat{\theta}_\mathrm{H}}{\sqrt{\mathrm{Var}[\hat{\theta}_\pi - \hat{\theta}_\mathrm{H}]}} \,. \tag{5.42}$$

To understand, which kind of deviations from the standard neutral model are picked up by the three summary statistics, it is instructive to consider the contribution of the site frequency classes S_i to the numerator of each statistic. For example, D_T will be negative if we have an excess of very low or very high frequency alleles, whereas it will be positive if many sites segregate at intermediate frequencies.

6 Effective population size

In the previous chapter, we have constructed the coalescent for an idealized Wright-Fisher population. Our assumptions have included the following:

1. neutral evolution with identical offspring distribution for all individuals,
2. a constant population size,
3. no population structure: i.e. offspring choose their parents with equal probability for all individuals from the parent generation,
4. offspring choose their parents independently of each other: as a consequence, the distribution of offspring for each parent is binomial (and approximately Poisson),
5. generations are discrete, individuals are haploid, and there are no separate sexes ...

One may wonder whether such a simplified theory can tell us much about nature. In statistical terms: if we construct a null model under a large number of assumptions, rejecting this null model does not provide us with a lot of information. Indeed, any of the assumptions could have been violated – for most biological populations we even know in advance that several assumption won't hold.

Luckily, the situation is not so bleak as it may look and we can often still use the theory that we have developed. As it turns out, many biological factors can be taken care of by an appropriate adjustment of the model parameters. This leads to the concept of the effective population size.

6.1 The concept

The number of individuals in a natural population is referred to as the *census population size* or *per-capita population size*. We have seen that finite population size often leads to genetic drift and, *prima facie*, it seems natural to identify the number of individuals (or individual gene copies) in a Wright-Fisher model with the census population size of a natural population. However, as it turns out, this is usually not appropriate. The point of the Wright-Fisher model (and similar models, like the Moran model) is to capture genetic drift. It should therefore be chosen in such a way that the strength of drift in the natural system is equal to the strength of drift in the model. The idea is to choose the size of an ideal Wright-Fisher population in such a way, that this correspondence holds. The size that is needed is called the *effective* population size. The remaining question is which measure for genetic drift we should use. Unfortunately, there is more than one measure that is commonly used. This leads to some ambiguity in the definition of the effective population size. In general, we use the following philosophy:

> Let $\bullet$ be some measurable quantity that relates to the strength of genetic drift in a population. This can be e.g. the rate of loss of heterozygosity or the

> probability of identity by descent. Assume that this quantity has been measured in a natural population. Then the effective size N_e of this population is the size of an ideal (neutral panmictic constant-size equilibrium) Wright-Fisher population that gives rise to the same value of the measured quantity $\bullet$. To be specific, we call N_e the $\bullet$-effective population size.

With an appropriate choice of this measure we can then use a model based on the ideal population to make predictions about the natural one. Although a large number of different concepts for an effective population size exist, there are two that are most widely used.

The coalescent (or inbreeding) effective population size

One of the most basic consequences of a finite population size - and thus of genetic drift - is that there is a finite probability for two randomly picked individuals in the offspring generation to have a common ancestor in the parent generation. This is the single-generation *probability of identity by descent*, which translates into the single-generation *coalescence probability* of two lines $p_{c,1}$ in the context of the coalescent. In the absence of population structure and demography, we can iterate the single-generation step across multiple generations: conditioned on non-coalescence in a single generation, the two parents are again random picks from a population of the same size. We thus obtain $p_{c,t} = p_{c,t}(1 - p_{c,1})^{(t-1)}$ as a simple function of $p_{c,1}$. For the ideal Wright-Fisher model with $2N$ (haploid) individuals, we have $p_{c,1} = 1/2N$. Knowing $p_{c,1}$ in a natural population, we can thus define the coalescent effective population size

$$N_e^{(c)} = \frac{1}{2p_{c,1}}. \tag{6.1}$$

The degree of inbreeding is one of the factors that influences $N_e^{(c)}$. For historic reasons, $N_e^{(c)}$ is therefore often referred to as *inbreeding effective population size.* The more relevant connection is the one to coalescent times. Assume that $N_e^{(c)}$ is constant over generations. Then all coalescent times are directly proportional to this size. One also says that $N_e^{(c)}$ fixes the *coalescent time scale.*

The variance effective population size

Another key aspect about genetic drift is that it leads to random variations in the allele frequencies among generations. Assume that p is the frequency of an allele A in an ideal Wright-Fisher population of size $2N$. In Section 3, we have seen that the number of A alleles in the next generation, $2Np'$, is binomially distributed with parameters $2N$ and p, and therefore

$$\mathrm{Var}_{\mathrm{WF}}[p'] = \frac{1}{(2N)^2}\mathrm{Var}[2Np'] = \frac{p(1-p)}{2N}.$$

For a natural population where the variance in allele frequencies among generations is known, we can therefore define the variance effective population size as follows

$$N_e^{(v)} = \frac{p(1-p)}{2\mathrm{Var}[p']}. \tag{6.2}$$

As we will see below, the inbreeding and variance effective sizes are often identical or at least very similar. However, there are exceptions and then the correct choice of an effective size depends on the context and the questions asked. Finally, there are also scenarios (e.g. changes in population size over large time scales) where no type of effective size is satisfactory. We then need to abandon the most simple ideal models and take these complications explicitly into account.

Estimating the effective population size

For the Wright-Fisher model, we have seen in Section 5 that the expected number of segregating sites S in a sample is proportional to the mutation rate and the total expected length of the coalescent tree, $\mathrm{E}[S] = u\,\mathrm{E}[L]$. The expected tree-length $\mathrm{E}[L]$, in turn, is a simple function of the coalescent times, and thus of the coalescent effective population size $N_e^{(c)}$. Under the assumption of (1) the infinite sites model (no double hits), (2) a constant $N_e^{(c)}$ over the generations (constant coalescent probability), and (3) a homogeneous population (equal coalescent probability for all pairs) we can therefore estimate the effective population size from polymorphism data if we have independent knowledge about the mutation rate u (e.g. from divergence data). In particular, for a sample of size 2, we have $\mathrm{E}[S_2] = 4N_e^{(c)}\,u$ and thus

$$N_e^{(c)} = \frac{\mathrm{E}[S_2]}{4u}.$$

In a sample of size n, we can estimate the expected number of pairwise differences to be $\widehat{\mathrm{E}}[S_2] = \hat{\theta}_\pi$ (see (5.36)) and obtain the estimator of $N_e^{(c)}$ from polymorphism data as

$$\hat{N}_e^{(c)} = \frac{\hat{\theta}_\pi}{4u}.$$

A similar estimate can be obtained from Watterson's estimator $\hat{\theta}_{\mathrm{W}}$, see Eq. (5.35). While the assumption of the infinite sites model is often justified (as long as $4N_e^{(c)}\,u_n \ll 1$, with u_n the per-nucleotide mutation rate), the assumption of constant and homogeneous coalescent rates is more problematic. We will come back to this point in the next section when we discuss variable population sizes and population structure.

6.2 Factors affecting N_e

Let us now discuss the main factors that influence the effective population size. For simplicity, we will focus on $N_e^{(c)}$. We will always assume that there is only a single deviation from the ideal Wright-Fisher population.

Offspring variance

One assumption of the ideal model is that the offspring distribution for each individual is binomial (approximately Poisson). In natural populations, this will usually not be the case. Note that the average number of offspring must always be 1, as long as we keep the (census) population size constant. The offspring variance σ^2, however, can take any value in a wide range. Let x_i be the number of offspring of individual i with $\sum_i x_i = 2N$. Then the probability that individual i is the parent of two randomly drawn individuals from the offspring generation is $x_i(x_i-1)/(2N(2N-1))$. Thus, the expected probability for identity by descent of two random offspring individuals is

$$p_{c,1} = \mathrm{E}\Big[\sum_{i=1}^{2N} \frac{x_i(x_i-1)}{2N(2N-1)}\Big] = \sum_{i=1}^{2N} \mathrm{E}\Big[\frac{x_i(x_i-1)}{2N(2N-1)}\Big]. \tag{6.3}$$

With $\mathrm{E}[x_i] = 1$ and $\mathrm{E}[x_i^2] = \sigma^2 + 1$ and the definition (6.1) we arrive at

$$N_e^{(c)} = \frac{1}{2p_{c,1}} = \frac{N-1/2}{\sigma^2} \approx \frac{N}{\sigma^2}. \tag{6.4}$$

By a slightly more complicated derivation (not shown), we can establish that the variance effective population size $N_e^{(v)}$ takes the same value in this case.

Separate sexes

A large variance in the offspring number leads to the consequence that in any single generation some individuals contribute much more to the offspring generation than others. So far, we have assumed that the offspring distribution for all individuals is identical. In particular, the expected contribution of each individual to the offspring generation was equal (= 1). Even without selection, this is not necessarily the case. An important example is a population with separate sexes and unequal sex ratios in the breeding population. Consider the following example:

Imagine a zoo population of primates with 20 males and 20 females. Due to dominance hierarchy only one of the males actually breeds. What is the inbreeding population size that informs us, for example, about loss of heterozygosity in this population? 40? or 21??

Let, in general, N_f be the number of breeding females and N_m the number of breeding males. Then half of the genes in the offspring generation will derive from the N_f parent females and half from the N_m parent males. Now draw two genes at random from two individuals of the offspring generation. The chance that they are both inherited from males is $\frac{1}{4}$. In this case, the probability that they are copies from the same paternal gene is $\frac{1}{2N_m}$. Similarly, the probability that two random genes are descendents from the same maternal gene is $\frac{1}{4}\frac{1}{2N_f}$. We thus obtain the probability of finding a common ancestor one generation ago

$$p_{c,1} = \frac{1}{4}\frac{1}{2N_m} + \frac{1}{4}\frac{1}{2N_f} = \frac{1}{8}\Big(\frac{1}{N_m} + \frac{1}{N_f}\Big)$$

and an effective population size of

$$N_e^{(c)} = \frac{1}{2p_{c,1}} = \frac{4}{\frac{1}{N_m} + \frac{1}{N_f}} = \frac{4N_f N_m}{N_f + N_m}.$$

In our example with 20 breeding females and 1 breeding male we obtain

$$N_e^{(c)} = \frac{4 \cdot 20 \cdot 1}{20 + 1} = \frac{80}{21} \approx 3.8.$$

The coalescent (or inbreeding) effective population size is thus much smaller than the census size of 40 due to the fact that all offspring have the same father. Genetic variation will rapidly disappear from such a population. In contrast, for an equal sex ratio of $N_f = N_m = \frac{N}{2}$ we find $N_e^{(c)} = N$.

Sex chromosomes and organelles

Take two random Y-chromosome alleles from a population. What is the probability that they have the same ancestor one generation ago? This is the probability that they have the same father, because Y-chromosomes come only from the father. So this probability is $\frac{1}{N_m}$ where N_m is the number of males in the population, so $N_e^{(c)} = N_m$. Similarly, for mitochondrial genes $N_e^{(c)} = N_f$ where N_f is the number of females in the population. In birds the W-chromosome is the female determining chromosome. WZ individuals are female and ZZ individuals are male. So for the W-chromosome $N_e = N_f$. For the X-chromosome in mammals and the Z-chromosome in birds it is a bit more complicated. Take two random X-chromosome alleles, what is the probability that they are from the same ancestor? This is

$$\frac{1}{2}\Big(\frac{N_f}{N_m} + \frac{N_m}{N_f}\Big) \cdot \frac{1}{2N_f + N_m}.$$

Exercise 6.1. Explain the last formula. (Hint: You need to treat males and females in both the offspring and parent generations seperately.) What is $N_e^{(c)}$ for the X-chromosome if the sex ratio is 1:1? □

Fluctuating Population Sizes

Consider the evolution of a population with periodically varying size over a period of T_p generations. Imagine that we have already calculated the effective population size for each individual generation N_0 to N_{T_p-1}. The N_i take breeding structure etc. into account. The expected reduction of heterozygosity over T_p generations then is

$$\begin{aligned} H_{T_p} &= \Big(1 - \frac{1}{2N_0}\Big) \cdots \Big(1 - \frac{1}{2N_{T_p-1}}\Big) H_0 \\ &= (1 - \bar{p}_{c,1})^{T_p} H_0 \end{aligned}$$

where $\bar{p}_{c,1}$ is the relevant average single-generation coalescence probability that describes the loss of heterozygosity. We then have

$$1-\bar{p}_{c,1} = \Big[\Big(1-\frac{1}{2N_0}\Big)\cdots\Big(1-\frac{1}{2N_{T_p-1}}\Big)\Big]^{1/T_p} \approx \Big[\exp\Big(-\frac{1}{2N_0}\Big)\cdots\exp\Big(-\frac{1}{2N_{T_p-1}}\Big)\Big]^{1/T_p}$$
$$= \exp\Big(-\tfrac{1}{2T_p}\Big(\frac{1}{N_0}+\ldots+\frac{1}{N_{T_p-1}}\Big)\Big) \approx 1-\frac{1}{2T_p}\Big(\frac{1}{N_0}+\ldots+\frac{1}{N_{T_p-1}}\Big)$$

and get an average (coalescent) effective population size of

$$\bar{N}_e^{(c)} = \frac{1}{2}\frac{1}{\bar{p}_{c,1}} \approx \frac{T_p}{\frac{1}{N_0}+\ldots+\frac{1}{N_{T_p-1}}}.$$

- The (average) inbreeding-effective population size is thus given by the harmonic mean of the population sizes over time. Other than the usual arithmetic mean, the harmonic mean is most strongly influenced by single small values. E.g., if the N_i are given by 100, 4, 100, 100, the arithmetic mean is 76, but we obtain a harmonic mean of just $\bar{N}_e^{(c)} = 14$.

- When can we use such an *average* effective size as a homogeneous time-scale for coalescent calculations? Since population sizes affect coalescent probabilities, the condition is that the population should run through many population-size cycles during a typical coalescent time. I.e., we should have

  $$T_p \ll \binom{n}{2}^{-1}\bar{N}_e^{(c)} = \mathrm{E}[T_n] \qquad (6.5)$$

  (in per-generation scaling). In practice, this is fulfilled, for example, for species with several generations per year and seasonal variation in population size.

- We also note that periodicity in population sizes is not required for the definition of an average effective size. It is generally sufficient that the population sizes experienced during time periods of $E(T_n)$ are representative of a long-term distribution of population sizes.

From the cases studies in this section, we see that most populations are genetically much smaller than they appear from their census size, increasing the effects of drift. There are rare exceptions, as shown in the next section.

Two toy models

Let us deal with two examples of populations that are unrealistic but helpful to understand the concept of effective sizes. The first example that is given represents a zoo population. In order to keep polymorphism as high as possible, care is taken that every parent has exactly one offspring. The ancestry is given by Figure 6.1(A).

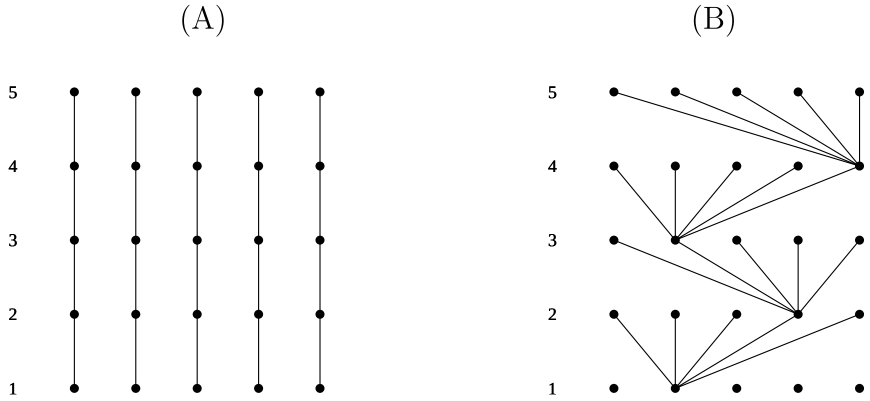

Figure 6.1: (A) The ancestry of a population that is kept as polymorphic as possible. (B) The ancestry of a population where each generation only has one parent

The second example is similar to the example of unequal sex ratios where the population had 20 males, but only one of them had offspring. However, in this case the individuals are haploids, and only one ancestor is the parent of the whole next generation. A scenario like this is shown in Figure 6.1(B).

Exercise 6.2. Figures 6.1(A) and (B) clearly do not come from ideal Wright-Fisher populations. So $N_c \neq N_e$ and we can try to calculate the effective population sizes for them. Given the census size in the two cases is N_c what are the

- variance effective size,
- coalescent effective size (for a single generation).

□

Exercise 6.3. We have introduced the neutral fixation probability and the equilibrium heterozygosity for an ideal Wright-Fisher population. How do we need to adjust the results if we think of a natural population and the various concepts of effective population sizes?

7 Demography: population growth and decline

So far, we have been concerned with biological factors that only change the rescaling of the coalescent process and thus can be captured by an appropriate effective population size N_e. However, some factors lead to more severe deviations from the standard neutral theory. This is obvious for selection, but also occurs for neutral evolution if we allow for major changes in the population size over time, or for population sub-structure. In this section, we will focus on the changes of the neutral coalescent due to large-scale changes in population size.

7.1 Effects of demography

We have seen in the previous section that short-term fluctuations in the population size can be subsumed in an effective population size N_e that is the harmonic mean of the population sizes over the period of the fluctuation. This holds as long as this period is short relative to the typical coalescence time. This is no longer true, however, if population sizes change over longer time scales. The latter will be the case, in particular, if there is no equilibrium distribution of population sizes across generations at all, such as in continued population growth or decline. The simplest model is that of an exponentially growing (or declining) population.

$$N(\tau) = N_0 e^{-\lambda\tau}, \tag{7.1}$$

for some parameter λ which quantifies the speed of growth. Note that we measure time τ in the backward direction. Exponential growth ($\lambda > 0$) or decline ($\lambda < 0$) is a natural consequence if populations regulation does not depend on the population density (through competition), but only on external factors. For the coalescent, this means that two individuals at time τ choose the same ancestor, i.e. coalesce in a single generation, with probability $1/2N(\tau)$. With a growing population, $N(\tau)$ declines as we go back in time and the frequency of coalescent event thus increases relative to the standard neutral model. Graphically this is shown in Figure 7.1: A typical coalescent tree in an expanding population has reduced branch lengths near the root of the tree.

If N_0 is the current (effective) population size, and if the coalescent time scale $\tau = t/2N_0$ is defined relative to this size, then the distribution of coalescence times in a population with variable size is

$$\Pr[T_n > \tau] = \exp\Big[-\binom{n}{2}\int_0^\tau \frac{N_0}{N(t)}dt\Big] = \exp\Big[-\binom{n}{2}\tau'\Big], \tag{7.2}$$

where we have introduced a new time scale

$$\tau' = \int_0^\tau \frac{N_0}{N(t)}dt\,.$$

For an exponentially growing population, for example, we obtain

$$\tau' = \frac{\exp[\lambda\tau]-1}{\lambda} \quad ; \quad \tau = \frac{\log[\lambda\tau'+1]}{\lambda} \tag{7.3}$$

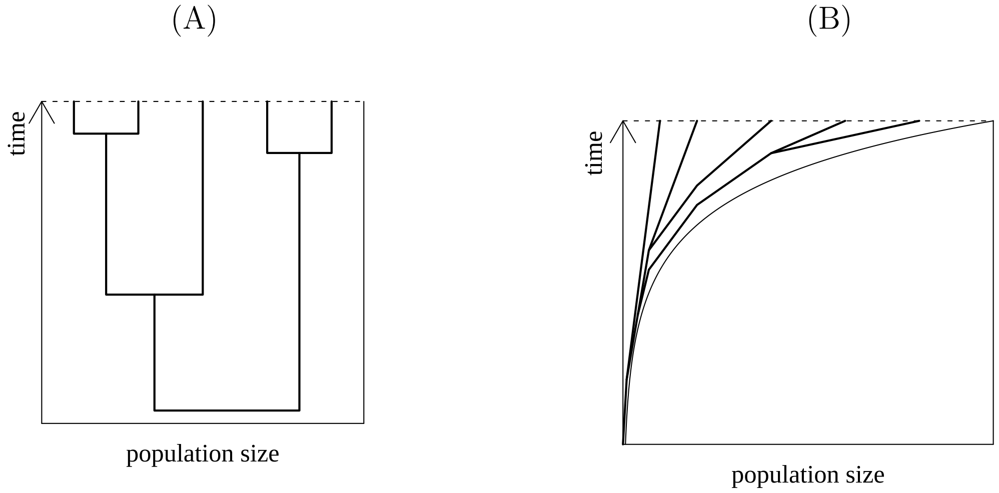

Figure 7.1: (A) The coalescent for a population of constant size and (B) for an expanding population. Real time runs from bottom to top and coalescent time τ from top to bottom.

The point of this transformation is that the coalescence times are exponentially distributed on the new τ' time scale. To obtain a coalescence time on the τ scale, we can thus pick an exponentially distributed random variable on the τ' scale in a first step. Using the reverse transformation, we can then derive the coalescence time on the original scale. Note that in contrast to the cases with a single effective population size N_e, the time rescaling is not by a constant factor, but varies with time. Still, the coalescent topologies do not change at all under such a transformation. Mutations are added to the trees after the transformation (on the τ scale). We can now ask for the effect of such a time-dependent rescaling on the expected summary statistics for the polymorphism pattern.

1. For a growing population, the (reverse) transformation $\tau' \to \tau$ will reduce "older" branches near the root of the tree by a larger factor than "young" branches at the leaves. As a consequence, most mutations will fall on branches near the leaves, where they affect only a single individual in the sample (mutation of size 1). We thus obtain an excess of low-frequency polymorphisms in the site-frequency spectrum relative to the standard neutral model. Looking at the test statistics, we see that Tajima's D_{T} will be negative, while Fay and Wu's H_{FW} will be positive.

2. For a shrinking population, the transformation reduces primarily the branches at the leaves. We typically obtain trees with a deep split, where most of the time during the genealogy is spent for the last two lines to coalesce. Since mutations on these branches produce mutations of any size (from 1 to $n-1$) with an equal probability, this results in a flat site-frequency spectrum, with a reduced number of singletons relative to the standard neutral model. In contrast to population growth, such a pattern is usually characterized by a positive value of D_{T}.

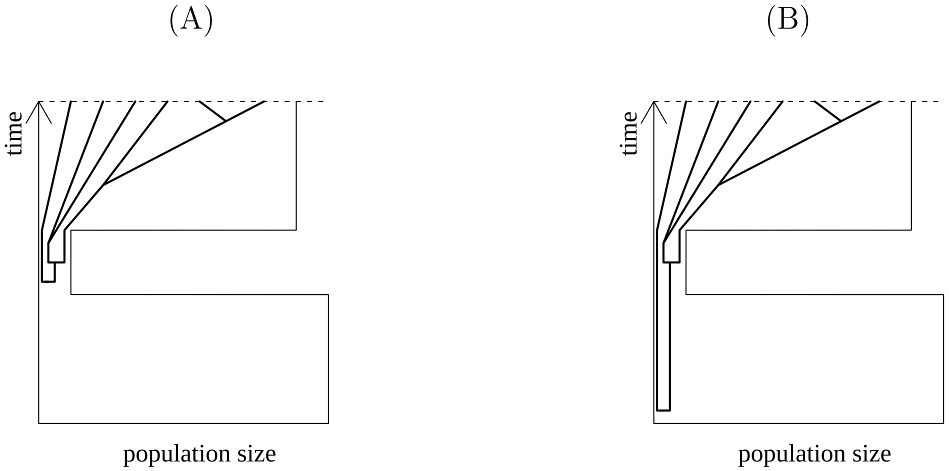

Figure 7.2: Two cases in a bottleneck mode. (A) Only one ancestral line survives the bottleneck. (B) Two or more lines survive which leads to different patterns in observed data.

3. A very complex demographic scenario are so-called bottlenecks, where the population recovers after an intermediate phase with a reduced population size. The consequences of such a demographic history depends on the parameters of the bottleneck in a subtle way. This is seen in two examples in Figure 7.2. On the one hand, a very strong and/or long reduction of the population size can lead full coalescence of the genealogy with a very high probability during that phase. In this case, a bottleneck will look like an expanding population. On the other hand, for a less severe or very short reduction, it is most likely that two or more lines do not coalesce during the bottleneck. As these lines enter (back in time) the ancestral phase with large population size, the time to full coalescence at the MRCA may be very long. The pattern in this case rather mimics one of a declining population. For intermediate bottleneck strengths, both of these genealogical scenarios may occur with a high probability. We then get a mix of patterns and a large increase in the *variance* of most summary statistics if we analyze patterns from different loci along a chromosome.

8 Selection and drift

In the first part of the lecture, we have described the consequences of selection in scenarios, where the effect of genetic drift can be ignored. This can be done in a deterministic framework. In contrast, in the second part, we have concentrated on models of neutral evolution, where drift and mutation are the only evolutionary forces. It is natural to ask, whether we can also combine all three forces, mutation, selection, and drift in a single model to describe their combined effect. There are various approaches how this could be done.

1. The introduction of selection into the Wright-Fisher model is straightforward, as shown below. The problem is that almost no analytical results can be obtained from this framework.

2. Some more progress can be made for the so-called *Moran model.* Several key properties can also be studied using branching processes, an even easier mathematical framework. We will discuss both these models in the first part of this section below.

3. It is also possible to include selection into the coalescent framework. This theory goes under the name of the *ancestral selection graph.* The problem that needs to be addressed is that genotype and genealogy (state and descent) are no longer independent. Since this leads to some major complications, we do not follow this approach here.

4. A more traditional route to proceed is to approximate the Wright-Fisher model by a diffusion process. This approach leads to several valuable results and will be the main focus of the second part of this chapter.

8.1 Selection and the Wright-Fisher model

The Wright-Fisher model assumes a population with discrete generations and the following life-cycle:

$$\begin{array}{ccccccc} p & & p & \textit{random union} & \tilde{p} & \textit{sampling} & p' \\ \text{Adults} & \Longrightarrow & \text{Gametes} & \Longrightarrow & \text{Zygotes} & \Longrightarrow & \text{Adults} \\ (N) & \textit{meiosis} & (\infty) & \textit{mutation} & (\infty) & \textit{drift} & (N) \\ & & & \textit{selection} & & & \end{array}$$

Adults from the parent generation produce a very large – effectively infinite – gamete pool (sperm, eggs, pollen). Subsequently, these gametes randomly combine to diploid zygotes. Zygotes are affected by selection and mutation, still in the infinite-population-size part of the life cycle. Only afterwards, N individuals are sampled from this infinite pool to form the adult population of the next generation. The advantage of this life cycle is that random mating (or random union of gametes), selection, and mutation all occur in an effectively infinite population. It we are interested in the evolutionary change in the frequency of some allele A, we can thus simply use the formulas from deterministic population genetics. Only in a second step, the alleles undergo the sampling step to include drift into the model.

For example, consider a single locus with two alleles a and A and diploid fitness values as in the following table,

Genotype	aa	aA	AA
Fitness	1	$1+hs$	$1+s$

.

Further assume that mutation from a to A occurs at rate μ and back mutation from A to a at rate ν. Let p be the frequency of allele A in the parent generation. Since each adult contributes equally to the gamete pool, this step from a finite to an infinite population size does not change the allele frequency. As in Eq. (1.19), we obtain for the action of selection and mutation

$$\tilde{p} = (1-\nu)\frac{w_A}{\bar{w}} \cdot p + \mu\Big(1 - \frac{w_A}{\bar{w}} \cdot p\Big) \tag{8.1}$$

where

$$\begin{aligned} w_A &= 1 + sp + hs(1-p) \quad \text{and} \\ \bar{w} &= 1 + sp^2 + 2hsp(1-p) \end{aligned}$$

are the marginal fitness of the A allele and the mean fitness, respectively. As for the neutral Wright-Fisher model, drift is modeled by binomial sampling,

$$\Pr[p' = j/2N|\tilde{p}] = \binom{2N}{j}\tilde{p}^j(1-\tilde{p})^{2N-j}\,. \tag{8.2}$$

Although the Wright-Fisher model assumes a quite particular life cycle, it is thought to work as a valid approximation for a broad range of biological scenarios.

- In many species, it may be more natural to assume that mutation happens already during gamete production, prior to selection. This can be easily included by an appropriate modification of the deterministic formula. However, for weak mutation and selection, which we will assume below, this difference vanishes.

- Selection is included into the Wright-Fisher model as viability selection on the zygotes. Since it directly follows random union of the gametes, we can (conveniently) assume Hardy-Weinberg proportions prior to selection. We note that with multiplicative fitness, the selection scheme is equivalent to selection on haploids (cf. page 13). We could then also assume that selection affects the viability of gametes, or their rate of production (i.e. adult fertility).

- Finally, we note that the sampling step in the Wright-Fisher model is done on the level of genes (i.e. haploids) rather than on the level of diploid individuals. As long as diploids are in Hardy-Weinberg equilibrium, this does not make a difference. However, unless fitness is multiplicative, zygotes will deviate from Hardy-Weinberg proportions after selection. Technically, we then would need to assume that zygotes already segregate directly after selection. Adults thus represent a haploid life stage, which is indeed the case for some species (such as many algae). However, for weak selection, only small deviations from Hardy-Weinberg occur, which can usually be ignored. For this reason, the model is used in a much broader context.

Mathematically, Eq. (8.2) defines the transition matrix of a Markov chain with a finite state space. Unfortunately, however, this matrix is so complicated that analytical results cannot be obtained without further assumptions. Below, we will see that in the limit of weak mutation and selection the Wright-Fisher model gives rise to a diffusion process. Alternatively, we can try to construct a model with a simpler transition matrix. This leads to the so-called Moran model.

8.2 The Moran model

The idea of the Moran model is that it breaks down the per-generation change in allele frequencies of the Wright-Fisher model into individual birth-death events. Consider a haploid population of size N and two alleles, a (wild type) and A (mutant) at a single locus. In the neutral version of the model, two individuals from the population are chosen at random (with replacement, i.e. allowing for the same individual to be chosen twice): one for reproduction and the other one for death. The reproducing individual produces exactly one offspring, such that total population size is maintained at all times. Biologically, we can imagine that a random death creates the space for a random offspring (seed) to grow and take its place. We can either assume that individual birth-death intervals occur in discrete time intervals or we assume that they occur in continuous time at a fixed rate. In contrast to the Wright-Fisher model, the survival time of individuals is not fixed, but either geometrically distributed (in the discrete time version) or exponentially distributed (in the continuous time version of the model). N time units in the Moran model (with birth-death rate normalized to 1 in the continuous version) correspond to a single Wright-Fisher generation, but generations are overlapping in the Moran case.

Mathematically, the Moran model corresponds to a simple Markov chain with a finite state space. Define a random vector $\mathbf{x}(t) = (x_0(t), \ldots, x_N(t))$, where $x_k(t)$ is the probability that there are k mutants (A alleles) in the population at time t. The dynamics of the discrete-time Moran model is then given as $\mathbf{x}(t+1) = \mathbf{x}(t)\mathbf{P}$ with transition matrix $\mathbf{P}$ with entries

$$\begin{aligned} p_{k,k+1} &= \frac{k(N-k)}{N^2} && (8.3)\\ p_{k,k-1} &= \frac{k(N-k)}{N^2} = p_{k,k+1} && (8.4)\\ p_{k,k} &= 1 - p_{k,k+1} - p_{k,k-1} && (8.5) \end{aligned}$$

All other entries are zero. The matrix $\mathbf{P}$ is thus tri-diagonal. We also observe the following:

1. For the Moran model in continuous time, the above Markov chain still describes the transitions between states at consecutive birth-death events (the so-called jump-chain of the process). Note that the timing of the events is independent of the state of the process. For a given time interval Δt, we can thus first determine the number of events that have happened during this time (which is Poisson distributed) and then follow the jump chain to derive the probabilities $\mathbf{x}(t_0 + \Delta t)$ from a given initial state $\mathbf{x}(t_0)$.

2. Without new mutations, the process as two absorbing states: fixation of the mutant at $k = N$ or fixation of the wildtype at $k = 0$. In the long term, evolution will always reach one of these states.

3. For the neutral process (without mutation and selection) we have $p_{k,k-1} = p_{k,k+1}$ and the expected number of mutant alleles $E[k]$ is conserved under the process (the process is a martingale). If a process is started at k_0, and if $u(k_0)$ is the fixation probability of this process at $k = N$, the martingale property implies $k_0 = E[k(0)] = \lim_{t\to\infty}[E[k(t)]] = u(k_0) \cdot N + (1 - u(k_0)) \cdot 0$ and thus

$$u(k_0) = k_0/N.$$

4. The Moran process is a special case of a birth-death process, which is a standard model of a Markov chain in the mathematical literature. It is also equivalent to a one-dimensional random walk with absorbing boundaries.

Inclusion of mutation and selection into the Moran model is straightforward. For mutation, we simply assume that the newborn individual is a new mutant with a specified mutation rate (and potentially different back mutation rate). Selection can be introduced either via the reproduction rates or the death rates. Let r be the fitness of the mutant allele A, where the fitness of the a allele is normalized to 1. If fitness differences are due to differences in rates of reproduction (production of seeds that could fill empty spots), we have

$$r = \frac{\Pr[\text{repro. mutant}]}{\Pr[\text{repro. wt}]}$$

as the probability that a given mutant is selected for reproduction, relative to a wildtype. If there are k mutants in the population of size N, the total probabilty that a mutant is chosen for reproduction is

$$\frac{r \cdot k}{r \cdot k + N - k} \tag{8.6}$$

and the corresponding probability for a wildtype is

$$\frac{N-k}{r \cdot k + N - k}. \tag{8.7}$$

With death rates as in the neutral case, the transition probabilities of the Moran model with selection (without mutation) are

$$p_{k,k+1} = \frac{r \cdot k(N-k)}{(r \cdot k + N - k)N} \tag{8.8}$$

$$p_{k,k-1} = \frac{k(N-k)}{(r \cdot k + N - k)N}\,. \tag{8.9}$$

The Moran process including selection is no longer a martingale. However, it is still possible to derive fixation probabilities for a general birth-death process with a tri-diagonal transition matrix. Let u_i be the fixation probability at N for a process started at $k(0) = i$. Using the Markov property, we have:

$$u_0 = 0 \tag{8.10a}$$

$$u_i = p_{i,i+1}u_{i+1} + p_{i,i-1}u_{i-1} + p_{i,i}u_i \tag{8.10b}$$

$$u_N = 1\,, \tag{8.10c}$$

which can be written in matrix form as

$$\mathbf{u} = \mathbf{P}\mathbf{u}\,. \tag{8.11}$$

The vector $\mathbf{u}$ of fixation probabilities is thus a (right) eigenvector of the transition matrix $\mathbf{P}$ corresponding to eigenvalue 1. To derive the entries of $\mathbf{u}$, define $y_i := u_i - u_{i-1}$. We have

$$\sum_{i=1}^{N} y_i = u_N - u_0 = 1\,.$$

With $p_{i,i} = 1 - p_{i,i+1} - p_{i,i-1}$ from (8.10b) we further obtain

$$y_{i+1} = \frac{p_{i,i-1}}{p_{i,i+1}} y_i =: \gamma_i y_i$$

and using $y_1 = u_1$

$$y_k = u_1 \prod_{i=1}^{k-1} \gamma_i\,,$$

thus

$$1 = \sum_{k=1}^{N} y_k = u_1 \Big(1 + \sum_{k=2}^{N} \prod_{i=1}^{k-1} \gamma_i\Big)$$

and

$$u_1 = \frac{1}{1+\sum_{k=2}^{N}\prod_{i=1}^{k-1}\gamma_i} = \frac{1}{1+\sum_{k=1}^{N-1}\prod_{i=1}^{k}\gamma_i}\,.$$

The fixation probability for a process started in $k(0) = j$ results in

$$u_j = \sum_{k=1}^{j} y_j = u_1\Big(1+\sum_{k=1}^{j-1}\prod_{i=1}^{k}\gamma_i\Big) = \frac{1+\sum_{k=1}^{j-1}\prod_{i=1}^{k}\gamma_i}{1+\sum_{k=1}^{N-1}\prod_{i=1}^{k}\gamma_i}\,. \tag{8.12}$$

- In the neutral case, $p_{i,i-1} = p_{i,i+1}$ and $\gamma_i = 1$. We thus have

  $$u_j = \frac{j}{N}$$

  as expected.

- With selection

  $$\gamma_i = \frac{p_{i,i-1}}{p_{i,i+1}} = \frac{1}{r}$$

  and

  $$u_i = \frac{\sum_{j=0}^{i-1} r^{-j}}{\sum_{j=0}^{N-1} r^{-j}} = \frac{(1-r^{-i})/(1-r^{-1})}{(1-r^{-N})/(1-r^{-1})} = \frac{1-1/r^i}{1-1/r^N}. \tag{8.13}$$

  The fixation probability of a single A mutant, in particular, is:

  $$p_{\text{fix}}^{(A)} = u_1 = \frac{1-1/r}{1-1/r^N} \approx (r-1)/r \tag{8.14}$$

  für $r^N \gg 1$.

- Vice-versa, the fixation probability at $k = 0$ of a single wildtype in a mutant population ($k_0 = N-1$) is

  $$p_{\text{fix}}^{(a)} = 1-u_{N-1} = \frac{1-r}{1-r^N} \approx (r-1)/r^N\,. \tag{8.15}$$

  We always have

  $$\frac{p_{\text{fix}}^{(a)}}{p_{\text{fix}}^{(A)}} = r^{1-N}\,.$$

Fixation probabilities are small even for strongly advantageous mutants. For mutants with deleterious effects, they are exponentially small, but always positive. Fixation probabilities for a population of size $N = 100$ are given in the following table. $\#_{p_{\text{fix}}=\frac{1}{2}}$ gives the number of consecutive new mutants that is needed to switch a population from wildtype to mutant state with probability $\frac{1}{2}$. (I.e. $(1 - p_{\text{fix}}^{(A)})^{\#_{p_{\text{fix}}=\frac{1}{2}}} = \frac{1}{2}$).

fitness		fixation prob. $p_{\text{fix}}^{(A)}$	$\#_{p_{\text{fix}}=\frac{1}{2}} = \frac{-\log(2)}{\log(1-p_{\text{fix}}^{(A)})}$
$r = 2$	$(+100\%)$	0.5	1
$r = 1.1$	$(+10\%)$	0.09	7
$r = 1.01$	$(+1\%)$	0.016	44
$r = 1$	(neutral)	$0.01(= 1/N)$	69
$r = 0.99$	(-1%)	0.0058	119
$r = 0.9$	(-10%)	0.000003	234861

8.3 Branching processes

Branching processes constitute one of the most explicit class of models for evolutionary or ecological dynamics. The focus is on growth rates and extinction probabilities of populations. Like the Wright-Fisher model, branching process models are individual based. In a general branching process, individuals can have different types, where a type can represent genetic or environmental factors, such as the individual's genotype or its place of birth in a spatially structured population. Individuals can reproduce (i.e. branch) or die with type-dependent probabilities, and they may also change their type (e.g. due to mutation or migration). However, in the most basic framework, the Galton-Watson process, there is only a single type of individuals.

The Galton-Watson Process (GWP)

The GWP is the prototype of a branching process in discrete time. It assumes that there is a single type of individuals. During its lifetime (one generation), each individual produces offspring according to a distribution ρ_{GW} and then dies. The GWP is characterized by the following conditions on the offspring distribution:

- ρ_{GW} is identical for all individuals,
- ρ_{GW} is independent of the number of individuals and of the reproductive success of the other individuals,
- ρ_{GW} is constant over the generations.

Now let X_n be the number of individuals in generation n and let $Z_{i,n}$ be the number of offspring of the ith individual in generation n. According to the definition of the process, all $Z_{i,n}, i \leq X_n$ are independently and identically distributed (i.i.d.) according to ρ_{GW}. Let us define

$$\rho_{\text{GW}}: \quad Z_{i,n} = k \quad \text{with probability} \quad p_k. \tag{8.16}$$

with expected value

$$E[Z_{i,n}] = \sum_{k=0}^{\infty} k p_k := \mu. \tag{8.17}$$

The offspring of all individuals in generation $n-1$ build the nth generation of the process, i.e.

$$X_n = \sum_{i=1}^{X_{n-1}} Z_{i,n-1}. \tag{8.18}$$

The expected number of individuals

The standard way to derive (or prove) properties for the branching process is by induction over the generations, where the induction step uses the conditioned expectation

$$E[X] = E_Y[E[X|Y]].$$

We are interested in the expected size of the population in the nth generation, $E[X_n]$. We can use that $E[X_n|X_{n-1}] = E[\sum_{i=1}^{X_{n-1}} Z_{i,n-1}] = X_{n-1}\mu$ and $E[X_0] = X_0$ for the starting generation. We then find by induction

$$E[X_n] = E[E[X_n|X_{n-1}]] = \mu E[X_{n-1}] = \mu^n X_0. \tag{8.19}$$

The extinction probability

A key question for a branching process is the probability for the extinction of a population. Let π_n be the probability that the population is extinct in generation n, i.e. $\pi_n = Pr[X_n = 0]$. In particular,

$$\pi_\infty := \lim_{n\to\infty} \pi_n = \lim_{n\to\infty} Pr[X_n = 0]$$

is the probability that the population dies out at all. Equivalently, $1-\pi_\infty$ is the probability that it survives for ever. Note that the sequence π_n is monotonic with 1 as upper bound and therefore always converges. Since the descendants of different individuals in the founder generation develop completely independent, we always have $\pi_n[X_0 = m] = (\pi_n[X_0 = 1])^m$. We can thus focus on the case $X_0 = 1$, which we will do in the following. In this case, we have $\pi_1 = p_0$, and for the π_n the following recursion holds:

$$\pi_{n+1} = \phi(\pi_n); \quad \text{with} \quad \phi(t) = \sum_{k=0}^{\infty} p_k t^k. \tag{8.20}$$

To see this, note that $X_1 = k$ with probability p_k and the probability that these individuals have no descendants in generation $n+1$ is π_n^k. Note also that $\phi(t)$ is the generating function of the probability distribution $\rho_{\mathrm{GW}} = p_k$. Since π_n converges as $n \to \infty$, π_∞ must be a fixed point of ϕ, i.e.

$$\pi_\infty = \phi(\pi_\infty) = \sum_{k=0}^{\infty} p_k \pi_\infty^k. \tag{8.21}$$

We can now show the following:

Proposition 1

1. For $p_0 = 0$ we obtain $\pi_\infty = 0$,

2. for $p_0 > 0;\ p_0 + p_1 = 1$ we have $\pi_\infty = 1$,

3. for $p_0 > 0;\ p_0 + p_1 < 1$, π_∞ is given by the smallest positive solution of the fixed point equation $\phi(t) = t$. This fixed point is $\pi_\infty < 1$ if and only if $\mu > 1$.

Proof

1. This is obvious.

2. To see this, note that we have $\mu < 1$ in this case. Since $E[X_n] = \mu^n$ we have

$$\Pr[X_n \geq 1] \leq \mu^n \Rightarrow \Pr[X_n = 0] \geq 1 - \mu^n,$$

which converges to 1 in the limit $n \to \infty$.

3. We use $0 < \phi(0) = p_0 < 1$ and $\phi(1) = 1$. Furthermore, $\phi'(t) = \sum_{k=1}^{\infty} k p_k t^{k-1}$ and $\phi''(t) = \sum_{k=2}^{\infty} k(k-1) p_k t^{k-2}$ are positive for $t \in (0, 1)$. In particular, we have $\phi'(1) = \mu$. Therefore $\phi(t)$ has a fixed point $0 < t^* < 1$ iff $\mu > 1$. Convergence to the smallest fixed point is best seen graphically (”cob-web” picture).

It becomes evident that the expected growth rate μ is the crucial parameter also for the extinction problem. Depending on its value, we distinguish *subcritical* ($\mu < 1$), *critical* ($\mu = 1$) and *supercritical* ($\mu > 1$) branching processes. While subcritical and critical branching processes always die out, there is a “merciless alternative” of either extinction or divergence to infinity for supercritical processes. This is expressed by the following proposition:

Proposition 2

1. Excluding the trivial case with $p_1 = 1$, a branching process does not return to the same non-zero number of individuals infinitely many times,
$\Pr[X_n = k \text{ for infinitely many } n] = 0$ for $k \geq 1$.

2. In the limit $n \to \infty$, a branching process either goes extinct (with probability π_∞) or diverges ($X_n \to \infty$ with probability $1 - \pi_\infty$), but $\lim_{n\to\infty} \Pr[X_n = k] = 0$ for each finite $k \geq 1$.

Proof

1. We distinguish two cases. First, if $p_0 = 0$, the sequence X_n is non-decreasing, $X_{n+1} \geq X_n$. Since, in particular, $\Pr[X_{n+1} = k|X_n = k] = p_1^k < 1$, $X_n = k$ cannot hold for infinitely many n. Second, if $p_0 > 0$, there is a fixed non-zero probability that a population with size k dies out in the next generation, $\Pr[X_{n+1} = 0|X_n = k] = p_0^k > 0$. If we assume $X_n = k$ for infinitely many n, extinction is certain, contradicting the assumption.

2. $\lim_{n\to\infty} \Pr[X_n = k] = 0$ for each finite $k \geq 1$ is a direct consequence of the first part of proposition 2 and the other assertions follow with proposition 1.

Application: Fixation probabilities

Branching processes are often used in ecology in a population dynamical context. In evolutionary genetics, however, the most important application (which already goes back to Haldane and Fisher) is the derivation of the fixation probability of a beneficial mutant allele. To map this problem to a branching process, we need to assume that mutants reproduce independently and with a distribution that is constant across generations. Both these conditions are (approximately) fulfilled as long as the mutant is rare. In particular, (i) we can ignore mutant homozygotes and (ii) the mean fitness is approximately equal to the ancestral wildtype fitness (i.e. the mutant evolves in a constant “wildtype environment”). Both conditions are necessarily violated once the mutant frequency becomes large. However, in a large population the number of mutants at this point is already sufficiently large that extinction is very unlikely. Excluding the case of overdominance, the mutant will then reach fixation and we can thus identify the fixation probability with the probability of non-extinction. According to proposition 1, we obtain the fixation probability p_{fix} as solution of

$$1 - p_{\text{fix}} = \phi(1 - p_{\text{fix}}) \approx \phi(1) - p_{\text{fix}}\phi'(1) + \frac{1}{2}p_{\text{fix}}^2\phi''(1) + \mathcal{O}\Big(p_{\text{fix}}^3\Big).$$

We have

$$\phi(1) = 1 \quad , \quad \phi'(1) = \mu \quad , \quad \phi''(1) = \sum_{k=2}^{\infty} k(k-1)p_k = \sigma^2 + \mu(\mu - 1).$$

where μ and σ^2 are the mean and the variance of the mutant offspring number. Assuming that p_{fix} is small, we can ignore all higher order terms in the Taylor expansion and find

$$p_{\text{fix}} \approx \frac{2(\mu - 1)}{\sigma^2 + \mu(\mu - 1)}.$$

In terms of the selection coefficient with mutant fitness $\mu = 1 + s$ this takes the form

$$p_{\text{fix}} \approx \frac{2s}{\sigma^2} + \mathcal{O}\Big(s^2\Big). \tag{8.22}$$

For the Wright-Fisher model, the offspring distribution is approximately Poisson if N is large and the number of mutants is small. All mutant alleles appear in heterozygotes. We thus have $\sigma^2 = \mu = w_A/\bar{w} \approx 1 + hs$ and obtain

$$p_{\text{fix}} \approx \frac{2hs}{(1+hs)^2} \approx 2hs$$

to leading order in s, a result first derived by JBS Haldane. Note that similar derivations using branching processes are not possible for neutral or deleterious alleles. The crucial condition is that the fate of the mutant is decided while it is rare.

Continuous-time branching process

We have derived the fixation probability for the Moran model explicitly. However, as for the Wright-Fisher model, we can also approximate the Moran model by a branching-process model that is even easier. We then obtain a branching process in continuous time, where each individual independently gives birth at rate r (creating one new individual) or can die at a rate normalized to 1. As before, we derive the probability that a single individual "establishes" (i.e. it has offspring at any time in the future) by jumping to the next event for this individual and by using the fact (Markov property) that any offspring individuals have the same establishment probability (corresponding to a fixation probability) p_{fix}. With probability $r/(r+1)$, the next event will be birth rather than death. Clearly, establishment is only possible in this case. In the case of birth, the focal individual will establish unless both of its offspring do not establish (which has probability $(1-p_{\text{fix}})^2$). We thus obtain the recursion

$$p_{\text{fix}} = \frac{r}{r+1}(1-(1-p_{\text{fix}})^2)$$

which is readily solved by

$$p_{\text{fix}} = 2 - \frac{r+1}{r} = \frac{r-1}{r}\,, \tag{8.23}$$

which can be compared with the exact result Eq. (8.14). We note that the approach only works for $r > 1$ (a beneficial type, corresponding to a supercritical branching process). For $r = 1+s$ we have $p_{\text{fix}} \approx s$ whereas we have $p_{\text{fix}} \approx 2s$ for a haploid Wright-Fisher model. This difference arises because death is stochastic in the Moran model, while it is deterministic in the Wright-Fisher model (where everybody dies after one generation). This leads to a larger variance in allele frequencies due to drift in the Moran model, corresponding to a smaller effective population size.

9 Diffusion models

9.1 Diffusions

Diffusions are a special class of Markov processes in continuous time with a continuous state space. Below, we will give a general (but heuristic) exposition of diffusion theory. Alongside, we apply each step to the Wright-Fisher diffusion as our special case of interest. We define:

- Let $X(t)$, $t \in \mathbb{R}$ be a continuously distributed random variable, which takes values $x \in I \subset \mathbb{R}$. For the Wright-Fisher case, $X(t)$ will measure the allele frequency with values in $I = [0, 1]$.
- Let $f(x, t)$ be the probability density of $X(t)$ at time t, and let $p(x, t|x_0, t_0)$ be the transition probability density from state x_0 at time t_0 to state x at time $t \geq t_0$.
- Finally, we assume that the Markov property holds, i.e. for $t_0 \leq t_1 \leq \cdots \leq t_n$,

$$p(x, t|x_n, t_n, \ldots, x_0, t_0) = p(x, t|x_n, t_n)\,. \tag{9.1}$$

With these conditions, one can easily show that the transition probabilities fulfill the so-called *Chapman-Kolmogorov equation*,

$$p(x_2, t_2|x_0, t_0) = \int_I p(x_2, t_2|x_1, t_1)p(x_1, t_1|x_0, t_0)dx_1\,. \tag{9.2}$$

To define a particular process, we need to specify the transition probabilities. Note that the Chapman-Kolmogorov equation implies that it will be sufficient to know the transition probabilities for arbitrarily small time intervals δt, since we can concatenate these intervals for transitions across larger time spans. This is very convenient for applications in the natural sciences since *laws of nature* usually predict instantaneous changes of a system as a response to external forces. In our context, or example, these forces are mutation, selection, and genetic drift, which change the allele frequency from one generation to the next.

For a *diffusion*, the infinitesimal transition probabilities take a special form: If x and y are the states of the process at times t and $t + \delta t$, respectively, the moments of $\delta x = y - x$ fulfill the following conditions in the limit $\delta t \to 0$,

$$(i) \quad \lim_{\delta t \to 0} \frac{1}{\delta t}\,\mathrm{E}[\delta x|x, t] = \lim_{\delta t \to 0} \frac{1}{\delta t} \int_I (y - x)p(y, t + \delta t|x, t)dy = A(x, t)\,, \tag{9.3}$$

$$(ii) \quad \lim_{\delta t \to 0} \frac{1}{\delta t}\,\mathrm{E}[(\delta x)^2|x, t] = \lim_{\delta t \to 0} \frac{1}{\delta t} \int_I (y - x)^2 p(y, t + \delta t|x, t)dy = D(x, t)\,, \tag{9.4}$$

$$(iii) \quad \lim_{\delta t \to 0} \frac{1}{\delta t}\,\mathrm{E}[(\delta x)^n|x, t] = \lim_{\delta t \to 0} \frac{1}{\delta t} \int_I (y - x)^n p(y, t + \delta t|x, t)dy = 0\,;\ n > 2\,. \tag{9.5}$$

As an important consequence of these conditions, one can show that $\forall\epsilon > 0$

$$\lim_{\delta t\to 0} \frac{1}{\delta t} \int_{|x-y|>\epsilon} p(y, t+\delta t|x,t)dy = 0\,. \tag{9.6}$$

Diffusion processes therefore produce continuous trajectories $x(t)$, even if these can be quite irregular.

- The existence of a Markovian transition density with these limit properties needs a proof that is beyond the scope of this lecture. We note that the choice of conditions on the short-term moments is not at all arbitrary. For example, it is known that if the third moment has a non-zero limit for $\delta t \to 0$, all higher-order moments must be non-zero as well.

- $A(x,t)$ is called the *drift term* of the diffusion – which should not be confused with the notion of genetic drift. $D(x,t)$ is the *diffusion term.*

- The prototype of a diffusion process is *standard Brownian motion* with $A(x,t) = 0$ and $D(x,t) = 1$, with the real numbers as state space.

9.2 The Wright-Fisher diffusion

Wright-Fisher model with weak mutation and selection

We want to derive a diffusion model as an approximation to the evolutionary dynamics defined by the Wright-Fisher model. To achieve this, we need to transform the discrete state space of the Wright-Fisher model to the continuous state space of the diffusion, where allele frequencies take values on the unit interval. Since the distance between neighboring states in the Wright-Fisher model is $1/N$, the idea is to scale of the population size to infinity in such a way that the key properties of the model remain approximately constant. It turns out that this can be done if the (original) population size N is large, and if the mutation rates μ, ν and the selection coefficient s are small. Specifically, in this approximation we only maintain the leading order terms proportional to μ, ν, s, or N^{-1}, but ignore higher order terms, such as s^2, s/N, νs, μ/N, N^{-2}, etc. Formally, this can be achieved via the definition of scaled mutation and selection parameters,

$$\alpha := 2Ns \quad ; \quad \beta_1 := 2N\nu \quad ; \quad \beta_2 := 2N\mu\,. \tag{9.7}$$

With these scaled variables, we can simply develop all expressions as functions of N^{-1} according to a Taylor expansion and truncate after the first (linear) order. This is also called the *weak selection approximation.*

We are interested in the change in the frequency of allele A across a single generation, $\delta x := x' - x$. Due to drift, δx is a random variable. In the weak selection approximation, it is most convenient to express the distribution of δx in terms of its moments. For the first moment, we can use that $\mathrm{E}[x'] = \tilde{x}$ with binomial sampling, and thus

$$\begin{aligned}\mathrm{E}[\delta x] &= \mathrm{E}[x'] - x = \tilde{x} - x \\ &= \frac{1}{2N}\Big(\alpha x(1-x)(x+h(1-2x)) + \beta_2(1-x) - \beta_1 x\Big) + \mathcal{O}\Big[N^{-2}\Big]\,.\end{aligned} \tag{9.8}$$

Note that this equation corresponds, to leading order, with the continuous-time dynamics for mutation and selection, Eq. (1.31). For the variance, we obtain

$$\mathrm{Var}[\delta x] = \mathrm{Var}[x'] = \frac{\tilde{x}(1-\tilde{x})}{2N} = \frac{x(1-x)}{2N} + \mathcal{O}\Big[N^{-2}\Big]\,, \tag{9.9}$$

which thus does not depend on mutation and selection to leading order. Note also that $\mathrm{E}[(\delta x)^2] = \mathrm{Var}[\delta x]$ to the order considered. For all higher order moments, we obtain

$$\mathrm{E}[(\delta x)^n] = \mathcal{O}\Big[N^{-(n-1)}\Big] \quad , n \geq 3\,. \tag{9.10}$$

They are thus ignored to the order of the approximation. These relations show that the first two moments of the distribution of δx are sufficient to describe short-term changes in the allele frequencies.

Diffusion limit

With these preparations, it is now straightforward to define the diffusion approximation for the Wright-Fisher model. Consider the weak selection approximation of the model in Eqs. (9.8 – 9.10). Let δt be the time interval between two generations. Now define a new time scale t with a time unit of $2N$ generations. On this new time scale, we thus have

$$\delta t = \frac{1}{2N}\,. \tag{9.11}$$

We note that this is exactly the same time rescaling that we have done in the coalescent process – just that for the diffusion time is always measured in the forward direction. Following standard notation, we maintain t as the new time variable for (forward) diffusion, whereas we have used τ for the scaled (backward) coalescence time. With this new time scale, let now $N \to \infty$, keeping the values of the rescaled variables for selection α and mutation μ, ν constant. We then obtain a transition to a process in continuous time with a continuous state space. Comparing Eqs. (9.8 – 9.10) in this limit with (9.3 – 9.5) shows that this process neatly fulfills the diffusion conditions with time-independent drift and diffusion terms,

$$A(x,t) = \alpha x(1-x)\big(x + h(1-2x)\big) - \beta_1 x + \beta_2(1-x) =: M(x)\,, \tag{9.12}$$

$$D(x,t) = x(1-x) =: D(x)\,, \tag{9.13}$$

where we now use x as a continuous variable for the allele frequency.

- We stress once again that our treatment is heuristic. A rigorous proof for the convergence of the discrete process to the diffusion needs more advanced concepts from probability theory.

- Note that mutation and selection both enter into the "drift" term $M(x)$, while genetic drift (trough the sampling variance) gives rise to the diffusion term $D(x)$.

- The diffusion scaling $(\delta t)^{-1} = N \to \infty$ should not be considered as an assumption of an infinite population size, but rather as a continuum approximation of a discrete process. In particular, the diffusion limit does preserve the effects of genetic drift at the "same strength" as in the original model. Similarly, the scaling $N \to \infty$ with constant $\alpha = 2Ns$ requires that $s \to 0$ (and similarly for the mutation rates μ and ν). Again, this should not be considered as an assumption of infinitely weak selection (and mutation). The diffusion model preserves also these effects and should rather be seen as an approximation to the model with the original (biological) parameters s, μ, and ν.

- Note that for a good match of diffusion results to the original Wright-Fisher model, we still need moderately large population sizes and small selection coefficients. This is clear from the weak selection approximation of the Wright-Fisher model that is used for the diffusion. In practice, mutation rates and population sizes are rarely ever an issue, while selection coefficients should fulfill $s < 0.1$ (such that $s^2 \ll s$) for satisfactory results. However, it should also be noted that Wright-Fisher model itself is only a crude approximation of any natural population. So, even if there are deviations among the models, this does not necessarily imply that the diffusion model is worse in approximating reality.

9.3 The Kolmogorov forward and backward equations

A key advantage of the diffusion process is that its transition probability can be shown to fulfill two partial differential equations. The solution of the model can thus be reduced to the solution of these equations.

Theorem 3: Kolmogorov forward equation *Let $p(z,t|x,t_0)$ be the transition probability density of a diffusion process on I with drift term $A(x,t)$ and diffusion term $D(x,t)$. Then $p(z,t|x,t_0)$ fulfills the partial differential equation*

$$\frac{\partial}{\partial t} p(z,t|x,t_0) = -\frac{\partial}{\partial z}\Big[A(z,t)p(z,t|x,t_0)\Big] + \frac{1}{2}\frac{\partial^2}{\partial z^2}\Big[D(z,t)p(z,t|x,t_0)\Big]\,. \tag{9.14}$$

Proof Let $R(y)$ be a test function on $C^3[I]$ with $R(y)$ and its derivative vanishing at the boundaries on I. Then

$$\begin{aligned}\int_I R(y)\frac{\partial p(y,t|x,t_0)}{\partial t}dy &= \lim_{\tau\to 0}\frac{1}{\tau}\int_I R(y)\Big(p(y,t+\tau|x,t_0)-p(y,t|x,t_0)\Big)dy\\ &= \lim_{\tau\to 0}\frac{1}{\tau}\int_I R(y)\Big(\int_I p(y,t+\tau|z,t)p(z,t|x,t_0)dz-p(y,t|x,t_0)\Big)dy\\ &= \lim_{\tau\to 0}\frac{1}{\tau}\Big\{\int_I\int_I\Big(R(z)+(y-z)\frac{\partial R(z)}{\partial z}+\frac{1}{2}(y-z)^2\frac{\partial^2 R(z)}{\partial z^2}+\dots\Big)\times\\ &\qquad\times p(y,t+\tau|z,t)p(z,t|x,t_0)dz\,dy-\int_I R(y)p(y,t|x,t_0)dy\Big\}\end{aligned}$$

where we first use the Chapman-Kolmogorov equation and then develop $R(y)$ into a Taylor series. We can now apply the diffusion conditions (9.3 – 9.5) and evaluate the y integral in the limit $\tau\to 0$ to obtain

$$\begin{aligned}\int_I R(y)\frac{\partial p(y,t|x,t_0)}{\partial t}dy &= \int_I p(z,t|x,t_0)\Big(A(z,t)\frac{\partial}{\partial z}R(z)+\frac{1}{2}D(z,t)\frac{\partial^2}{\partial z^2}R(z)\Big)dz\,.\\ &= \int_I\Big\{R(z)-\frac{\partial}{\partial z}\Big[A(z,t)p(z,t|x,t_0)\Big]+\frac{1}{2}\frac{\partial^2}{\partial z^2}\Big[D(z,t)p(z,t|x,t_0)\Big]\Big\}dz\end{aligned}$$

after two times partial integration, using the boundary conditions for $R(y)$. Now the theorem follows since $R(y)$ is an arbitrary test function.

- The PDE is alternatively also called the *Fokker-Planck equation.*
- An analogous equation holds for the probability density,

$$\frac{\partial}{\partial t}f(z,t)=-\frac{\partial}{\partial z}\Big[A(z,t)f(z,t)\Big]+\frac{1}{2}\frac{\partial^2}{\partial z^2}\Big[D(z,t)f(z,t)\Big]\,,\qquad(9.15)$$

  which is obtained from (9.14) by multiplication with $f(x,t_0)$ and integration over x. This is possible since the forward equation is independent of the initial condition x.
- An explicit solution of the forward equation is only possible in some special cases. For standard Brownian motion with $A(x,t)=0$ and $D(x,t)=1$ and initial condition $z(t=0)=z_0$, in particular,

$$f(z,t)=\frac{1}{\sqrt{2\pi Dt}}\exp\Big(-\frac{(z-z_0)^2}{2Dt}\Big)\,.\qquad(9.16)$$

- The Wright-Fisher diffusion can be fully solved only for vanishing $M(z)=0$, i.e., without mutation and selection. The solution (due to Kimura) involves a series of Gegenbauer polynomials and is quite complex. We therefore do not expand on it here, but rather focus on the long-term behavior below, where more general (and transparent) results can be obtained.

Theorem 4: Kolmogorov backward equation *Let $p(z,t|x,t_0)$ be the transition probability density of a diffusion process on I with drift term $A(x,t)$ and diffusion term $D(x,t)$. Then $p(z,t|x,t_0)$ fulfills the partial differential equation*

$$\frac{\partial}{\partial t_0} p(z,t|x,t_0) = -A(x,t_0)\frac{\partial}{\partial x} p(z,t|x,t_0) - \frac{1}{2} D(x,t_0)\frac{\partial^2}{\partial x^2} p(z,t|x,t_0)\,. \tag{9.17}$$

Proof Similar to the forward equation, we derive

$$\begin{aligned}
\frac{\partial p(z,t|x,t_0)}{\partial t_0} &= \lim_{\tau\to 0}\frac{1}{\tau}\Big(p(z,t|x,t_0+\tau) - p(z,t|x,t_0)\Big) \\
&= \lim_{\tau\to 0}\frac{1}{\tau}\Big(p(z,t|x,t_0+\tau) - \int_I p(z,t|y,t_0+\tau)p(y,t_0+\tau|x,t_0)dy\Big) \\
&= \lim_{\tau\to 0}\frac{1}{\tau}\Big\{p(z,t|x,t_0+\tau) - \int_I p(y,t_0+\tau|x,t_0)\times \\
&\qquad \times\Big(p(z,t|x,t_0+\tau) + (y-x)\frac{\partial}{\partial x}p(z,t|x,t_0+\tau) \\
&\qquad\qquad + \frac{1}{2}(y-x)^2\frac{\partial^2}{\partial x^2}p(z,t|x,t_0+\tau) + \dots\Big)dy\Big\} \\
&= -A(x,t_0)\frac{\partial}{\partial x}p(z,t|x,t_0) - \frac{1}{2}D(x,t)\frac{\partial^2}{\partial x^2}p(z,t|x,t_0)\,.
\end{aligned}$$

- For a time-homogeneous process, such as the standard Wright-Fisher diffusion, the transition probability only depends on the time interval $t-t_0$. We can then also write the transition probability as $p(y,t-t_0|x)$, or after a change of variables simply as $p(y,t|x)$, where t now denotes the time interval. The Kolmogorov backward equation then reads
$$\frac{\partial}{\partial t}p(z,t|x) = M(x)\frac{\partial}{\partial x}p(z,t|x) + \frac{1}{2}D(x)\frac{\partial^2}{\partial x^2}p(z,t|x)\,. \tag{9.18}$$

The stationary distribution

In the presence of mutation in both directions, the Wright-Fisher distribution does not have any absorbing states. Instead, the allele frequency distribution will converge to a stationary distribution in the long-time limit. We can obtain this stationary distribution from the forward equation. We proceed in two steps. Integrating (9.14) with respect to z, we obtain

$$F(x,t) := \frac{\partial}{\partial t}\int_0^x f(z,t)dz = -M(x)f(x,t) + \frac{1}{2}\frac{\partial}{\partial z}\Big[D(z)f(z,t)\Big]_{z=x} + C_1\,, \tag{9.19}$$

where C_1 is an integration constant. $M(x)$ and $D(x)$ are the drift and diffusion constants of the Wright-Fisher diffusion as defined in Eq. (9.12) and (9.13). $F(x,t)$ has the interpretation of a *probability flux*. It measures the influx of total probability into the interval $[0,x]$ at time t. We can now use a symmetry argument to determine the constant C_1. To this end, assume that we exchange the roles of the to alleles a and A in the model, i.e. we choose x to be the frequency of a instead of A. This entails the following change of the model parameters,

$$\alpha \to \tilde{\alpha} = -\alpha \quad ; \quad \beta_1 \to \tilde{\beta}_1 = \beta_2 \quad ; \quad \beta_2 \to \tilde{\beta}_2 = \beta_1 \quad ; \quad h \to \tilde{h} = 1 - h\,. \tag{9.20}$$

We also have

$$f(x,t) \to \tilde{f}(x,t) = f(1-x,t) \quad \text{and} \quad F(x,t) \to \tilde{F}(x,t) = -F(1-x,t) \tag{9.21}$$

since the direction of the probability flux is reversed. We further easily derive

$$\tilde{M}(1-x) = -M(x) \quad \text{and} \quad \tilde{D}(1-x) = D(x), \tag{9.22}$$

where $\tilde{M}$ and $\tilde{D}$ are the drift and diffusion constants of the process with exchanged alleles, i.e. with parameters $\tilde{\alpha}$, $\tilde{\beta}_{1,2}$, and $\tilde{h}$. We then obtain

$$\begin{aligned} F(x,t) = -\tilde{F}(1-x,t) &= \tilde{M}(1-x)\tilde{f}(1-x,t) - \frac{1}{2}\frac{\partial}{\partial z}\Big[\tilde{D}(z)\tilde{f}(z,t)\Big]_{z=1-x} - C_1 \\ &= -M(x)f(x,t) + \frac{1}{2}\frac{\partial}{\partial z'}\Big[D(z')f(z',t)\Big]_{z'=x} - C_1 \end{aligned} \tag{9.23}$$

with the substitution $z' = 1 - z$. Comparison with (9.19) shows that $C_1 = 0$.

For the stationary distribution, we have $f(x,t) \to f(x)$ and $F(x,t) \to F(x) = 0$. Thus,

$$0 = -M(x)f(x) + \frac{1}{2}\frac{\partial}{\partial z}\Big[D(z)f(z)\Big]_{z=x} \tag{9.24}$$

with the boundary condition $\int_0^1 f(x)dx = 1$. This first order ordinary differential equation is easily solved to give

$$f(x) = \frac{C}{D(x)}\exp\left[2\int^x \frac{M(y}{D(y)}dy\right] \tag{9.25}$$

where the constant C takes care of the normalization. For Wright-Fisher, we find explicitly

$$\begin{aligned} f(x) &= \frac{C}{x(1-x)} \cdot \exp\left[2\alpha h x - \alpha(2h-1)x^2\right] \cdot \exp\left[2\beta_2 \ln(1-x) - \beta_1 \ln(x)\right] \\ &= C\,\frac{\exp\left[2\alpha h x - \alpha(2h-1)x^2\right]}{x^{1-2\beta_2}(1-x)^{1-2\beta_1}}\,. \end{aligned} \tag{9.26}$$

The latter is also called *Wright's formula.*

- The shape of the equilibrium distribution depends strongly on the strength of mutation. For $\beta_{1,2} > 0.5$, we are in a mutation-dominated regime and $f(x)$ drops to zero near the boundaries at $x = 0$ and $x = 1$. In contrast, for $\beta_{1,2} < 0.5$, the stationary distribution has singularities at the boundaries. This reflects the fact that the A allele is frequently either temporally absent or fixed for extended time periods, until mutation introduces a new A or a allele.

- Selection contributes an exponential factor to the stationary distribution. For strong negative selection on A with $\alpha \ll 0$, the frequency x of A will almost always be small and we can approximate

  $$f(x) = Cx^{2\beta_2 - 1} \exp[-2|\alpha h|x] \,. \tag{9.27}$$

  From this approximate distribution, we obtain a mean mutant frequency in terms of (incomplete) Gamma functions as

  $$\mathrm{E}[x] = \frac{\Gamma(1+2\beta_2) + \Gamma(1+2\beta_2, 2|h\alpha|)}{2|h\alpha|(\Gamma(2\beta_2) + \Gamma(2\beta_2, 2|h\alpha|))} \approx \frac{\Gamma(1+2\beta_2)}{2|h\alpha|\Gamma(2\beta_2)} = \frac{\beta_2}{|h\alpha|} = \frac{\mu}{|hs|} \tag{9.28}$$

  for $|h\alpha| >> \beta_2$, in accordance with the result (1.24) for mutation-selection balance from the deterministic theory. However, the stationary distribution is only peaked around this value if $2\beta_2 \gg 1$.

Fixation probabilities

If we consider the Wright-Fisher diffusion without mutation, the process has two absorbing states in $x = 0$ and $x = 1$. We can then ask: Given the initial frequency x_0, what is the probability for absorption in $x = 1$, rather than in $x = 0$? This probability is also called the *fixation probability* of the allele A. To answer this question, we define the probability for absorption in $x = 1$ by time t as follows,

$$P_1(x_0, t) := \lim_{\epsilon \to 0} F_\epsilon(x_0, t) \quad ; \quad F_\epsilon(x_0, t) = \int_{1-\epsilon}^{1} p(y, t|x_0) dy \,. \tag{9.29}$$

We now use the backward equation on $p(y, t|x_0)$. After exchanging integral and derivatives, and performing the ϵ limit, we see that also $P_1(x_0, t)$ fulfills the backward equation, i.e.,

$$\frac{\partial}{\partial t} P_1(x_0, t) = M(x_0) \frac{\partial}{\partial x_0} P_1(x_0, t) + \frac{D(x_0)}{2} \frac{\partial^2}{\partial x_0^2} P_1(x_0, t) \,. \tag{9.30}$$

Now, let $t \to \infty$, where $P_1(x_0, t) \to P_1(x_0)$ and $\partial P_1(x_0, t)/\partial t \to 0$. Then

$$M(x_0) \frac{\partial}{\partial x_0} P_1(x_0) + \frac{D(x_0)}{2} \frac{\partial^2}{\partial x_0^2} P_1(x_0) = 0 \tag{9.31}$$

with boundary conditions $P_1(0) = 0$ and $P_1(1) = 1$. This ODE is solved by integrating twice, using the boundary conditions to determine the integration constants. We obtain

$$P_1(x_0) = \frac{\int_0^{x_0} \exp\Big[-2 \int^y \frac{M(z)}{D(z)} dz\Big] dy}{\int_0^1 \exp\Big[-2 \int^y \frac{M(z)}{D(z)} dz\Big] dy} \,. \tag{9.32}$$

Using the drift and diffusion terms of the Wright-Fisher diffusion,

$$P_1(x_0) = \frac{\int_0^{x_0} \exp[-\alpha y^2 - 2h\alpha y(1-y)]dy}{\int_0^1 \exp[-\alpha y^2 - 2h\alpha y(1-y)]dy} \tag{9.33}$$

For no dominance with $h = 1/2$, the remaining integral can be evaluated to give

$$P_1(x_0) = \frac{1 - \exp[-\alpha x_0]}{1 - \exp[-\alpha]} \,. \tag{9.34}$$

In particular, if $\alpha \gg 1$, and if we start with a single mutant $x_0 = 1/2N$, this is

$$p_{\text{fix}} = P_1(1/2N) \approx 1 - \exp[-\alpha/2N] \approx \frac{\alpha}{2N} = s \; (= 2hs) \,. \tag{9.35}$$

- Even selection coefficients of strongly beneficial alleles are usually at most $s = 0.1$, and most beneficial alleles rather have $s < 0.01$. We thus see that the overwhelming majority of beneficial alleles that arise as new mutations in a population do not reach fixation, but are lost again due to genetic drift. Note that (to leading order) this does not depend on the population size.

- For a deleterious mutation with $s < 0$ we find that the fixation probability of a new mutant is exponentially small, $p_{\text{fix}} \sim \exp[-|\alpha|]$. Note that this depends strongly on the population size, in contrast to the beneficial case.

Absorption times

Another classical question for a diffusion process with absorbing states is the one for the absorption time. In the population genetic context, in particular, we can ask: given that an allele segregates in the population at an initial frequency of x_0, how long does it take until it reaches one of the absorbing boundaries at $x = 0$ or $x = 1$? The density of the absorption time can be expressed as

$$\phi(x_0, t) = \frac{\partial\big(P_0(x_0,t) + P_1(x_0,t)\big)}{\partial t} \,, \tag{9.36}$$

where $P_0(x_0,t)$ and $P_1(x_0,t)$ is the probability for fixation by time t in $x = 0$ or $x = 1$, respectively. We already know from Eq. (9.30) that $P_1(x_0,t)$ fulfills the backward equation and the same holds true for $P_0(x_0,t)$ by an analogous argument. Exchanging derivatives, we see that also $\phi(x_0,t)$ fulfills the backward equations, and hence (following Ewens 2004, p. 141),

$$\begin{aligned} -1 &= -\int_0^\infty \phi(x_0,t)\,dt \\ &= -t\phi(x_0,t)\Big|_0^\infty + \int_0^\infty t\frac{\partial \phi}{\partial t}\,dt \\ &= 0 + \int_0^\infty t\Big\{M(x_0)\frac{\partial \phi}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 \phi}{\partial x_0^2}\Big\}dt \,, \end{aligned} \tag{9.37}$$

assuming that $t\phi(x_0,t) \to 0$ for $t \to \infty$. We then obtain for the mean fixation time $\bar{t}(x_0) = \int t\phi(x_0,t)dt$ (exchanging integration and differentiation)

$$-1 = M(x_0)\frac{\partial \bar{t}(x_0)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 \bar{t}(x_0)}{\partial x_0^2}\,. \tag{9.38}$$

Using the boundary conditions $\bar{t}(0) = \bar{t}(1) = 0$, this ODE can again be solved by integration. We obtain

$$\bar{t}(x_0) = \int_0^1 \bar{t}(x|x_0)dx \tag{9.39}$$

where

$$\bar{t}(x|x_0) = \bar{t}_<(x|x_0) = \frac{2P_0(x_0)\,\int_0^x \exp\Big[-2\int^y \frac{M(z)}{D(z)}dz\Big]dy}{D(x)\exp\Big[-2\int^x \frac{M(z)}{D(z)}dz\Big]}\,,\quad 0 \le x \le x_0 \tag{9.40}$$

$$\bar{t}(x|x_0) = \bar{t}_>(x|x_0) = \frac{2P_1(x_0)\,\int_x^1 \exp\Big[-2\int^y \frac{M(z)}{D(z)}dz\Big]dy}{D(x)\exp\Big[-2\int^x \frac{M(z)}{D(z)}dz\Big]}\,,\quad x_0 \le x \le 1 \tag{9.41}$$

with $P_1(x_0) = 1 - P_0(x_0)$ is the fixation probability as given in (9.32). Indeed, the function $\bar{t}(x|x_0)$ has the more direct interpretation as *sojourn time density*,

$$\int_{x_1}^{x_2} \bar{t}(x|x_0)dx \tag{9.42}$$

is the mean time the diffusion process started in x_0 spends in the interval $[x_1,x_2]$ before absorption. For the Wright-Fisher diffusion without dominance ($h = 1/2$), we obtain

$$\bar{t}(x|x_0) = \frac{2P_0(x_0)\,(\exp[\alpha x]-1)}{\alpha x(1-x)}\,,\quad 0 \le x \le x_0\,, \tag{9.43}$$

$$\bar{t}(x|x_0) = \frac{2P_1(x_0)\,(1-\exp[-\alpha(1-x)])}{\alpha x(1-x)}\,,\quad x_0 \le x \le 1\,. \tag{9.44}$$

An explicit evaluation of the integral for the mean absorption time is only possible for the neutral case $\alpha \to 0$, where

$$\bar{t}(x|x_0) = \frac{2(1-x_0)}{1-x}\,,\quad 0 \le x \le x_0\,, \tag{9.45}$$

$$\bar{t}(x|x_0) = \frac{2x_0}{x}\,,\quad x_0 \le x \le 1\,, \tag{9.46}$$

and thus

$$\bar{t}(x_0) = -2\big(x_0 \log x_0 + (1-x_0)\log[1-x_0]\big) \tag{9.47}$$

in units of $2N$ generations. We can note the following

- The maximal time to absorption occurs for $x = 1/2$ and derives to
$$\bar{t}(1/2) = 2\log[2] \approx 1.4$$
corresponding to $2.8\,N$ generations.

- On the other hand, for a new mutation that enters the population at frequency $x_0 = 1/2N$ we obtain
$$\bar{t}(1/2N) = [2\log[2N] - (2N-1)\log[1 - 1/2N]]/2N \approx [2\log[2N] + 2]/2N\,,$$
or $2\log[2N] + 2$ on a generation scale, which only scales logarithmically with the population size. Note also that we have $\bar{t}(x|(2N)^{-1}) = 1/(Nx)$. Without recurrent mutation on a single site, we thus obtain a $1/x$ scape of the neutral site-frequency spectrum, consistent with the coalescent results for the infinite-sites model.

Conditioned fixation time Frequently, we are interested specifically in alleles that reach one absorbing state, say $x = 1$, and the expected time until they reach this state. This gives rise to the *conditioned fixation time* $\bar{t}_1(x_0)$. We can derive an expresion for this time using the backward equation following the recipe for the unconditioned fixation time above. Defining
$$\phi_1(x_0, t) = \frac{\partial}{\partial t}\,P_1(x_0, t)\,, \tag{9.48}$$
the density of the conditioned absorption time is $\phi_1(x_0, t)/P_1(x_0)$. We see that $\phi_1(x_0, t)$ fulfills the backward equation just like $\phi(x_0, t)$. Repeating the calculation from above, we obtain
$$\begin{aligned} -P_1(x_0) &= -\int_0^\infty \phi_1(x_0, t)\,dt \\ &= -t\phi_1(x_0, t)\Big|_0^\infty + \int_0^\infty t\frac{\partial \phi_1}{\partial t}\,dt \\ &= 0 + \int_0^\infty t\Big\{M(x_0)\frac{\partial \phi_1}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 \phi_1}{\partial x_0^2}\Big\}dt\,. \end{aligned} \tag{9.49}$$
We have
$$\bar{t}_1(x_0) = \frac{T_1(x_0)}{P_1(x_0)} \quad \text{where} \quad T_1(x_0) = \int_0^\infty t\phi_1(x_0, t)dt$$
and $T_1(x_0)$ fulfills the ODE
$$-P_1(x_0) = M(x_0)\frac{\partial T_1(x_0)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 T_1(x_0)}{\partial x_0^2}\,. \tag{9.50}$$
We also know that $T_1(x_0) \le \bar{t}(x_0)$ for all x_0 and so we must have the boundary conditions $T_1(0) = T_1(1) = 0$. A solution to this ODE is given by
$$T_1(x_0) = \int_0^1 \bar{t}(x|x_0)P_1(x)dx$$

with $\bar{t}(x|x_0)$ given in (9.40) and (9.41) above. The sojourn time density at frequency x before fixation at $x = 1$ follow as

$$\bar{t}_1(x|x_0) = \frac{P_1(x)}{P_1(x_0)}\,\bar{t}(x|x_0)\,.$$

10 Appendix: Mathematical basics

10.1 Calculation rules for binomials

Our derivations frequently use calculation rules for binomial coefficients, which are summarized here. We generally define $\binom{n}{k} = 0$ for $k < 0$ and for $k > n$. We find

1. $$k\binom{n}{k} = \frac{k \cdot n!}{k!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1}, \tag{10.1}$$

2. $$\binom{n}{k-1} + \binom{n}{k} = \frac{n!k + n!(n-k+1)}{k!(n-k+1)!} = \binom{n+1}{k}, \tag{10.2}$$

3. $$\sum_{k=0}^{n}\binom{m+k}{m} = \binom{n+m+1}{m+1}, \tag{10.3}$$

   which is proved by induction since, for $n = 0$, $\binom{m}{m} = \binom{m+1}{m+1} = 1$, and

$$\binom{n-1+m+1}{m+1} + \binom{m+n}{m} = \frac{n(n+m)! + (m+1)(n+m)!}{(m+1)!n!} = \binom{n+m+1}{m+1}.$$

10.2 Probability distributions

The derivations in the main text repeatedly use some elementary properties of probability distributions, which are summarized in this appendix. Generally, two types of random variables are used in the models, depending on the biological question that is addressed. First, a random variable X for the number of events (e.g. number of mutations on a coalescent tree). In discrete time this number generally follows a binomial distribution and in continuous time a Poisson distribution – or a derivative of these distributions. Second, a random variable T asks for the distribution of the waiting time to a specific event. The relevant distributions are the geometrical distribution in discrete time and the exponential distribution in continuous time.

Binomial distribution Let X by a random variable that tells us about the number of successes in an n-fold repeated Bernoulli experiment (with two possible outcomes: "success" or "no success") with success probability p. Then X is binomially distributed with

$$\Pr[X = k] = \binom{n}{k}p^k(1-p)^{n-k}. \tag{10.4}$$

For the expectation of X we obtain, using Eq. (10.1) and the normalization of the binomial distribution,

$$\mathrm{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=0}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} = np\,. \qquad (10.5)$$

Similarly, for the variance

$$\begin{aligned}\mathrm{E}[X^2 - X] &= \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=0}^{n} \binom{n-2}{k-2} p^{k-2} (1-p)^{(n-2)-(k-2)} = n(n-1)p^2\,,\end{aligned}$$

and so

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \mathrm{E}[X^2 - X] + \mathrm{E}[X] - \mathrm{E}[X]^2 = n(n-1)p^2 + np - n^2p^2 \\ &= np - np^2 = np(1-p). \qquad (10.6)\end{aligned}$$

Geometrical distribution Assume that the we perform a sequence of Bernoulli experiments with success probability p. Then the waiting time T to the first success is geometrically distributed with

$$\Pr[T = t] = (1-p)^{t-1} p, \qquad \Pr[T > t] = (1-p)^t\,. \qquad (10.7)$$

The expectation and variance are

$$\begin{aligned}\mathrm{E}[T] &= \sum_{t=1}^{\infty} t\, p(1-p)^{t-1} = -p \frac{\partial}{\partial p} \sum_{t=1}^{\infty} (1-p)^t = -p \frac{\partial}{\partial p} \left(\frac{1}{p} - 1 \right) = \frac{1}{p}\,, \qquad (10.8) \\ \mathrm{E}[T(T-1)] &= \sum_{t=1}^{\infty} t(t-1)\, p(1-p)^{t-1} = p(1-p) \frac{\partial^2}{\partial p^2} \sum_{t=1}^{\infty} (1-p)^t = p(1-p) \frac{2}{p^3}\,, \\ \mathrm{Var}[T] &= \mathrm{E}[T(T-1)] + \mathrm{E}[T] - (\mathrm{E}[T])^2 = \frac{1-p}{p^2}\,. \qquad (10.9)\end{aligned}$$

In many cases, we also use a different time unit, such as n Bernoulli experiments, which are performed during a binomial trial representing a "generation". On this larger time scale, we obtain $\mathrm{E}[T] = 1/(np)$ and $\mathrm{Var}[T] = (1-p)/(np)^2$.

Poisson distribution If a random variable X is binomially distributed with parameters n and p such that n is big and p is small, we can approximate its distribution by letting $n \to \infty$ and $p \to 0$, such that $np = \lambda = \text{const}$. This gives

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n \cdots (n-k+1)}{k!\,n^k} \frac{\lambda^k \left(1 - \frac{\lambda}{n}\right)^n}{(1-p)^k} \xrightarrow[p \to 0]{n \to \infty} e^{-\lambda} \frac{\lambda^k}{k!}. \qquad (10.10)$$

These are the weights of a Poisson distribution with parameter λ. For the expectation and variance of X, we compute

$$\begin{aligned}
\mathrm{E}[X] &= e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda, && (10.11)\\
\mathrm{E}[X(X-1)] &= e^{-\lambda}\sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!} = e^{-\lambda}\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2,\\
\mathrm{Var}[X] &= \mathrm{E}[X(X-1)] + \mathrm{E}[X] - (\mathrm{E}[X])^2 = \lambda. && (10.12)
\end{aligned}$$

More generally, we can imagine the switch from the binomial to a Poisson distribution as a transition from discrete time to continuous time as follows: if X is the number of events during some unit time interval (a generation), we can either assume that all these events occur simultaneously and X is drawn from a binomial distribution. Or we can assume that the events occur in continuous time, uniformly distributed over our unit time interval. In the latter case, we can dissect generation time into smaller intervals of length δt and perform a binomial sampling step for each sub-interval. The success probability for each time step will be $\delta t \cdot p$ and the total number of Bernoulli trials (over all time steps) $n/\delta t$. Letting $\delta t \to 0$ results in a Poisson distribution for X in continuous time with the same expected number of successes during a generation as the discrete time model. Note that we have assumed that all binomial sampling steps in the sub-intervals are independent. As a consequence, it is possible that we have more than n successes during one generation, which is not possible with the original binomial model. This is also reflected by the slightly larger variance of the Poisson model. However, with small p the difference becomes very small.

Exponential distribution Just like for the binomial model in discrete time, we can also ask for the waiting time until the first success for the Poisson model in continuous time. Starting with the discrete model, we use the per-generation scaling, where one time unit corresponds to n Bernoulli trials. Using the same procedure as before, we dissect this generation into time intervals of length δt and perform n Bernoulli trials during this shorter time span, but with a reduced success probability of $\delta t \cdot p$. In the limit $\delta t \to 0$, we get

$$\lim_{\delta t\to 0}\Pr[T > t] = \lim_{\delta t\to 0}(1 - p\cdot\delta t)^{nt/\delta t} = \exp[-pn] = \exp[-\lambda]\,. \qquad (10.13)$$

with $\lambda = np$. This is the distribution function of the exponential distribution with density $f[T] = \lambda\exp[-\lambda]$ and

$$\mathrm{E}[T] = \frac{1}{\lambda}\,, \qquad \mathrm{Var}[T] = \frac{1}{\lambda^2}\,, \qquad (10.14)$$

which should be compared to the corresponding quantities of the geometrical distribution in the generation (n Bernoulli time steps) scaling.

10.3 Multiple Poisson processes

A Poisson process is a process that records events that occur at a constant rate λ per unite time. The number of events in a time interval τ is Poisson distributed with parameter $\lambda\tau$ and the waiting time between events is exponentially distributed with parameter λ. Indeed, the waiting time to the next event starting from any moment is always exponentially distributed with the same parameter λ. For the process, it makes no difference whether the last event had just occurred a moment ago or already a long time ago. The Poisson process is therefore also called a *memoryless* process.

In many applications, we do not only have a single Poisson process, but multiple Poisson processes running in parallel. There is a number of useful results know for this case. Let $X_i(t)$ be independently Poisson distributed,

$$\Pr[X_i(t) = k] = \frac{(\lambda_i t)^k}{k!} \exp[-\lambda_i t] \,. \tag{10.15}$$

Then the time to the first event $T = \min_i[T_i]$ of all these processes is exponentially distributed with parameter $\lambda = \sum_i \lambda_i$, since

$$\Pr[T > t] = \prod_i \Pr[T_i > t] = \prod_i \exp[-\lambda_i t] = \exp\big[-\textstyle\sum_i \lambda_i t\big] \,. \tag{10.16}$$

Again, starting from any moment of time, the time to the next event will always follow this same distribution. The probability that this event is of a given type (say of type 1 counted by the first Poisson process) is

$$\Pr[T = T_i] = \int_0^\infty \lambda_1 \exp[-\lambda_1 t] \prod_{i \geq 2} \Pr[T_i > t] \, dt = \int_0^\infty \frac{\lambda_1}{\exp\big[\sum_i \lambda_i t\big]} \, dt = \frac{\lambda_1}{\sum_i \lambda_i} \,. \tag{10.17}$$

Finally, if we wait repeatedly for the next event of whichever Poisson process for a given time span t, we will account for all events during this time. The number of all events, $X(t) = \sum_i X_i(t)$, will again be Poisson distributed with parameter λt.