universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Models of genetic hitchhiking

angestrebter akademischer Grad

Magistra der Naturwissenschaften
(Mag. rer. nat.)

Verfasserin:	Theresa Dvorak BSc
Studienrichtung:	A405 Mathematik
Betreuer:	Univ.-Prof. Dr. Reinhard Bürger

Wien, im Juli 2012

Abstract

The hitchhiking effect has been introduced into the field of population genetics by Maynard Smith and Haigh [29] in 1974. In this pioneering work, genetic hitchhiking has been defined as the process by which a newly arising favourable mutant on its way to fixation alters the frequency of alleles at linked neutral loci, i.e., the neutral alleles *hitchhike* with the selected substitution. This process causes a loss of preexisting heterozygosity at linked neutral loci. In the model of Maynard Smith and Haigh, the change of frequency of the favourable mutant has been treated deterministically. This results in an underestimate of the effects of hitchhiking. More recent models remedy this weakness. The stochastic model of Kaplan *et al.* [17], proposed in 1989, applies coalescent theory to study the cumulative effect on the number of segregating sites of all selective substitutions that occurred in the history of a sample. As an alternative, with the potential of further generalization, Stephan *et al.* [41] developed a diffusion approximation of the fully stochastic Markov-chain model. The present work presents these three models in order to provide a profound introduction to the effects of genetic hitchhiking. Nowadays, this concept has been extended in various ways and, therefore, the term genetic hitchhiking is used in a more general sense and describes all indirect effects of positively selected substitutions on neighbouring loci.

Zusammenfassung

Im Bereich der mathematischen Populationsgenetik wurde der „hitchhiking effect“ erstmals 1974 von Maynard Smith und Haigh [29] ausführlich beschrieben. In dieser Arbeit wurde „genetic hitchhiking“ als jener Prozess definiert, in dem die Allelhäufigkeiten auf neutralen Loci durch die Verbreitung einer neu aufgetretenen Mutante mit positivem Selektionsvorteil auf einem gekoppeltem Locus verändert werden, das heißt die neutralen Loci *hitchhiken* mit der selektierten Mutante. Dieser Prozess führt zur Reduktion der Heterozygosität auf gekoppelten neutralen Loci. Die deterministische Modellierung des Wachstumsprozesses der selektierten Mutante bei Maynard Smith und Haigh resultierte in einer quantitativen Unterschätzung des „hitchhiking effects“. Neuere Modelle behoben diesen Schwachpunkt. Im Jahre 1989 präsentierten Kaplan *et al.* [17] ein stochastisches Modell, welches auf Methoden der Koaleszenztheorie zurückgreift um den kumulativen Effekt aller selektiven Substitutionen, die seit dem ältesten gemeinsamen Vorfahren einer Stichprobe aufgetreten sind, untersuchen zu können. Drei Jahre später veröffentlichten Stephan *et al.* [41] eine Diffusionsapproximation des stochastischen Markov-Ketten Models. Ziel ihrer Arbeit war es ein generalisiertes Modell zu entwickeln. In der vorliegenden Arbeit werden die eben genannten drei Modelle dargestellt um eine profunde Einführung zum Thema „genetic hitchhiking“ zu bieten. Das zugrundeliegende Konzept wurde in jüngster Vergangenheit unterschiedlich erweitert und adaptiert. Daher wird der Begriff „genetic hitchhiking“ heutzutage in einem allgemeineren Sinn verwendet und inkludiert alle indirekten Effekte eines oder mehrerer selektierter Loci auf den Rest des Genoms.

Danksagung

Vielen lieben Dank an all jene, die mich während meines Studiums begleitet haben. Insbesondere möchte ich jenen Menschen danken, die zur Entstehung dieser Arbeit beigetragen haben:

- Prof. Reinhard Bürger, für die hervorragende Betreuung meiner Arbeit und die vielen Stunden, die er dafür aufgebracht hat.
- Meinen Eltern, Christine und Johann Dvorak, die mich immer dabei unterstützt haben meinen eigenen Weg zu gehen.
- Meinen Schwestern, auf die ich immer zählen kann.
- Meinen Freundinnen und Freunden, für die Zerstreuung und viele aufmunternde Gespräche. Weil der kürzeste Weg nicht immer der Richtige ist.
- Sebastian Novak, für seine fachliche Unterstützung und viele konstruktive Gespräche, für zahlreiche aufbauende und motivierende Worte, für uns und dafür, dass er mich jeden Tag zum Lachen bringt.

Theresa Dvorak, am 19. Juli 2012

Contents

Chapter 1

Introduction

The hitchhiking effect has been introduced into the field of mathematical population genetics by Maynard Smith and Haigh [29] in 1974. It refers to the process by which a neutral allele increases in frequency due to a positively selected mutation on a linked locus ([29], p. 23).

Strongly selected, favourable mutations that survive the initial generations, during which they are rare and prone to extinction by genetic drift, will increase in frequency and go to fixation. Other neutral alleles on linked loci will be inherited together as long as they do not get separated by recombination, i.e., the neutral alleles *hitchhike* with the selected substitution. If the favourable allele goes to fixation, it will fix rapidly, i.e., the strongly selected mutation *sweeps* through the population. This process leads to a reduction or elimination of genetic diversity on loci close to the favourable mutation and is referred to as a *selective sweep*.

At the time when Maynard Smith and Haigh defined the term hitchhiking, the idea of such a process was not entirely new. In 1967, Kojima and Schaffer [23] studied the effect of "hitch-hiking genes" as a special case of the "survival process of linked mutant genes" ([23], pp. 525-527). They used probability generating functions in order to obtain a joint distribution of the number of offspring of a pair of linked mutant genes. The evaluation of the ultimate survival probabilities of the mutants showed that hitchhiking of a weakly selected mutation increases its survival probability. Maynard Smith and Haigh ([29], p. 23) pointed out that, in this previous model, the probability of replication of each genotype has been assumed to be constant regardless of the other genotype frequencies and therefore, the results turned out to be inaccurate by several orders of magnitude.

In 1968, Kimura [19] proposed the neutral theory of molecular evolution, which states that the major part of genetic diversity is caused by the balance of neutral mutations and random genetic drift rather than by selection. In

contrast, experimental data from the early seventies (for example [6, 27]) proved evidence for a lower level of heterozygosity in abundant species than predicted by the neutral theory. Furthermore, the heterozygosity between various species turned out to be surprisingly constant. These facts motivated Maynard Smith and Haigh [29] to study another process that regulates the amount of allelic polymorphisms in addition to random genetic drift. Their deterministic model is a pioneering work about genetic hitchhiking, and will be presented in Chapter 3.

The deterministic model of Maynard Smith and Haigh [29] does not consider the stochastic effects caused by the finite population size, and therefore, provides quantitatively different results than more recent models. The stochastic model of Kaplan *et al.* [17] remedies these weaknesses. Kaplan *et al.* modeled the hitchhiking effect in terms of genealogies which are especially appropriate to study sequence data. Their analysis has been stimulated by a large quantity of DNA sequence data (for example [1, 39]) that proved the reduction of standing variation in regions of restricted recombination. Since the distribution of segregating sites is determined by the population dynamics since the most recent common ancestor of the sample, their stochastic model considers the cumulative effect of recurring selected substitutions that occurred in the history of the sample. The model of Kaplan *et al.* [17] will be treated in Chapter 4. The required theoretical framework of coalescent theory will be presented in Chapter 2.

The third model, presented in Chapter 5, considers both the effect of a single selected substitution and the effect of recurring selected substitutions. It has been proposed by Stephan *et al.* [41] and models the hitchhiking effect by means of diffusion theory. The aim of this approach was to provide a model that can readily be generalized in order to describe polymorphisms on weakly selected loci instead of neutral loci based on the assumption that a significant part of mutations at the molecular level is not selectively neutral.

All three models of hitchhiking explain the loss of genetic variation on neutral loci as an indirect effect of selection on a linked locus. However, not all forms of selection reduce heterozygosity. Kaplan *et al.* [16], for example, showed that in models with balancing selection the expected number of segregating sites exceeds the predicted number of the neutral theory. Thus, balancing selection causes an increase of heterozygosity rather than a reduction. The concept of hitchhiking can be extended to cases where selection eliminates deleterious mutations or fluctuates in space or time and linked loci are not necessarily neutral (for example [3, 10]). Therefore, the term "genetic hitchhiking" is nowadays used in a broader sense and includes all indirect effects of positively selected loci on their chromosomal neighbourhood.

Chapter 2

Theoretical background

This chapter provides the basic theory about the coalescent that will be applied in the stochastic model of genetic hitchhiking described in Chapter 4.

All models assume linked loci, which means that the alleles on the loci will be inherited together as long as they do not get separated by recombination. The probability of separation is given by the recombination rate, which is proportional to the probability of a recombination event between the loci. If there is no recombination, we talk about complete linkage. In this case, the loci are often treated like a single locus. This is very likely if the loci are located next to each other on the chromosome. The greater the distance between the loci, the greater is the probability that they will be separated by recombination. Therefore, the recombination rate can be considered as a measure of the physical distance of the loci. If the loci get separated with probability of $\frac{1}{2}$, the loci are called unlinked.

2.1 Coalescent theory

The "coalescent" was first described by Kingman [22] in 1982, but Hudson [14] and Tajima [42] also discovered it independently. Kingman introduced the coalescent as a stochastic process that describes the genealogical history of a sample of n genes (therefore often called the "n-coalescent" to underscore the dependency on the sample size, or "Kingman's coalescent" named after its discoverer). The ancestral tree realized in this process shows, which of the n lineages coalesce in which generation backward in time until the most recent common ancestor, referred to as the MRCA, is encountered. The model was originally based on a neutral Wright-Fisher model, but it can be

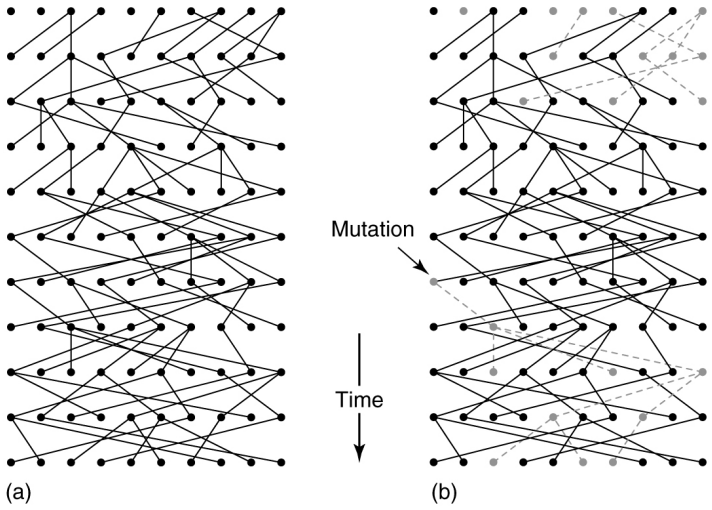

Figure 2.1: A realization of the genealogical relationships in the Wright-Fisher model for a diploid population with $2N = 10$ genes during 10 generations. Neutral mutations can be superimposed after the genealogy has been created.

shown that many different (and even more complex) neutral models give rise to Kingman's coalescent ([33], p. 844).

2.1.1 The coalescent in neutral models

Consider a diploid population of size N that evolves according to the neutral Wright-Fisher model. Generations are discrete and non overlapping and individuals reproduce by random mating. The number of offspring of each individual is binomially distributed with parameters $2N$ and $p = \frac{1}{2N}$ (the probability of being chosen as a parent). Note that for a randomly chosen daughter all parental genes are equally likely to be the parent. The joint offspring distribution of all N individuals is symmetrically multinomial ([33], p. 844).

Figure 2.1 shows a graph of the genealogical relationships forward in time for a sample of individuals according to the Wright-Fisher model. We observe that, going forward in time, a branching of lineages corresponds to a number of offspring greater than or equal two, and lineages end, when no offspring is produced. However, going backward in time, two lineages coalesce whenever they choose the same parent. Accordingly, the number of lineages decreases and eventually reaches one, when the MRCA occurs.

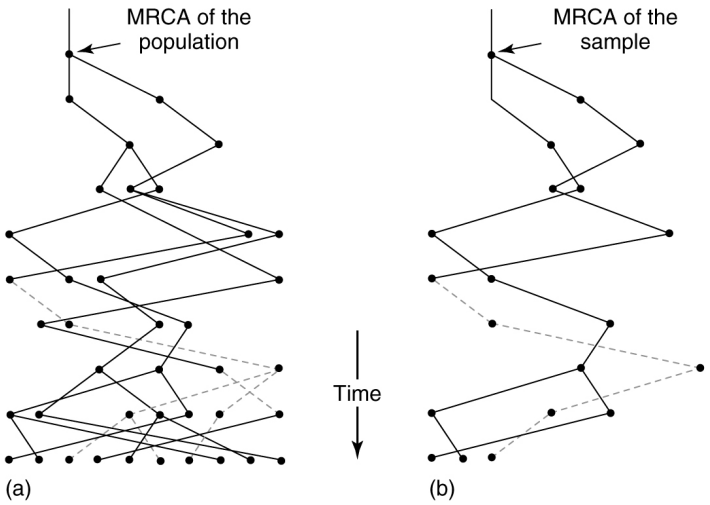

Figure 2.2: Many branches of Figure 2.1 do not have to be considered anymore, when creating the genealogy backward in time. There is a single underlying genealogy of the population to estimate.

Two fundamental insights can be deduced from these considerations. First, since neutral mutations arising during the process do not affect reproductive success, it is possible to separate "state" from "descent" (see Figure 2.1), which means separating the dynamics of neutral allelic variants from the genealogical process. As a consequence, the neutral mutation process can be modelled by "gene dropping", i.e., given a realization of the genealogical relationships, mutations can be superimposed forward in time. This implies that offspring and parent are genetically identical unless there is a mutation. Tracing the lineages of the sample back in time until the MRCA, we find that most of the details of the Wright-Fisher process forward in time are irrelevant, and the genealogy of a sample can be constructed without caring about the whole population (see Figure 2.2). This leads to the second insight that we only have to consider the sample to model their genealogy backward in time. This property allows very efficient computer simulations. Note that, independent of the sample size, there is a unique underlying ancestry that needs to be inferred ([33], pp. 845-846).

The essential information necessary to simulate an ancestral tree is which two lineages coalesce in which generation. In the discrete Wright-Fisher model, all lineages are equally likely to coalesce with probability $\frac{1}{2N}$. Since individuals reproduce independently, the probability that two randomly chosen lineages do not coalesce in the first $t-1$ generations (backward in time)

is $\left(1-\frac{1}{2N}\right)^{t-1}$. Thus, the probability that they coalesce in generation t is

$$\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{t-1},$$

and so, the waiting time for a coalescence event, given two lineages, has a geometric distribution with mean $2N$. The expected coalescence time is $2N$ generations.

For convenience, from now on we will measure time in units of $2N$ generations. Therefore, the probability that two given lineages do not coalesce for at least τ units of time ($2N\tau = t$) is given by

$$\left(1-\frac{1}{2N}\right)^{\lfloor 2N\tau \rfloor}, \tag{2.1}$$

where $\lfloor 2N\tau \rfloor$ denotes the largest integer less than or equal to $2N\tau$. If N goes to infinity, we obtain a continuous-time approximation to the discretet Wright-Fisher model. The probability given in (2.1) tends to

$$\left(1-\frac{1}{2N}\right)^{\lfloor 2N\tau \rfloor} \longrightarrow e^{-\tau}, \tag{2.2}$$

as $N \to \infty$. Hence, in the limit $N \to \infty$, we obtain an exponential distribution with mean one for the coalescence time for two randomly chosen lineages.

In the Wright-Fisher model in discrete time, the probability that three randomly chosen lineages coalesce in one generation, i.e., that three individuals choose the same parent, is given by $\frac{1}{4N^2}$. In general, the probability that more than two lineages coalesce, referred to as multiple mergers, is of order $O\left(\frac{1}{N^2}\right)$.

Considering k lineages and assuming $k \ll 2N$, the probability of no coalescence event during one generation is given by

$$\begin{aligned}\prod_{i=1}^{k-1}\left(1-\frac{i}{2N}\right) &= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\left(1-\frac{3}{2N}\right)\ldots\left(1-\frac{k-1}{2N}\right)\\ &= 1-\binom{k}{2}\frac{1}{2N}+O\left(\frac{1}{N^2}\right),\end{aligned} \tag{2.3}$$

where $\binom{k}{2}\frac{1}{2N}$ is the probability of a single coalescence event, and $O\left(\frac{1}{N^2}\right)$ describes the probability of two or more simultaneous coalescence events, referred to as simultaneous mergers.

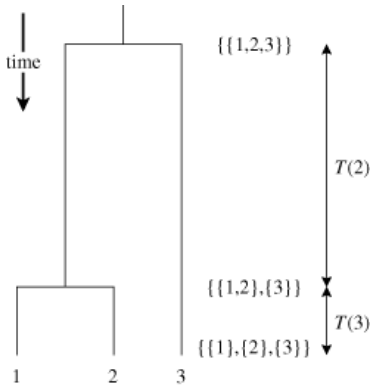

Figure 2.3: Example of an ancestral tree for a sample of size $n = 3$ genes in the coalescent.

Note that multiple mergers and simultaneous mergers occur with probability $O\left(\frac{1}{N^2}\right)$. In the limit $N \to \infty$, these probabilities are negligible and multiple events can be ignored. Therefore, the structure of the genealogy simplifies substantially. These considerations motivated the emergence of Kingman's coalescent [22] as a continuous-time approximation to the Wright-Fisher model for large populations, i.e., $N \to \infty$. Consequently, the coalescent only takes account of single binary mergers.

It is convenient to describe the genealogy of a sample of size n in the coalescent (see Figure 2.3) in terms of equivalence classes of genes and branch lengths, that correspond to the times between transitions of equivalence classes, i.e., the waiting times for the next coalescence event ([33], p. 847). Two genes are considered to be part of the same equivalence class at a certain point in time, if and only if they have a common ancestor at that time. A set of convenient equivalence classes forms an equivalence relation on $\{1, \ldots, n\}$. Assuming $\{\{1\}, \ldots, \{n\}\}$ to be the first equivalence relation and $\{\{1, \ldots, n\}\}$ to be the equivalence relation at the time of the MRCA, the coalescent describes the probabilities of the other equivalence relations to arise during the time in between. The following description of the details are based on [5], unless stated otherwise.

The length of the ancestral tree

Let $T(k)$ denote the coalescence time for k lineages, i.e., the time until the first coalescence event happens, when there are k lineages left in the ancestral tree. $T(k)$ can also be interpreted as the branch length (measured in $2N$ generations), when the ancestral tree consists of exactly k branches. Using

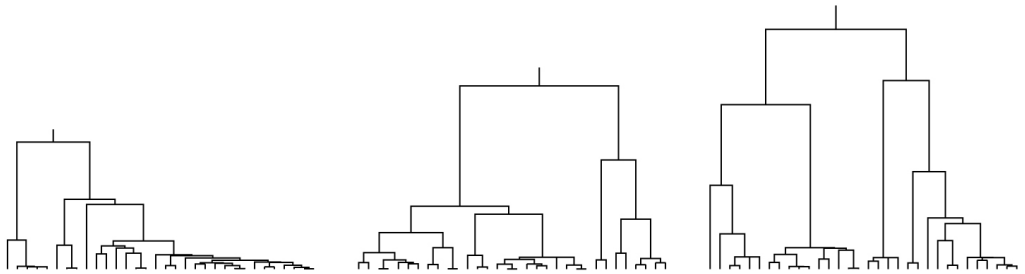

Figure 2.4: Three realizations of the ancestral tree for a sample of size $n = 32$ on the same scale.

the same argument as in (2.3), we obtain

$$P\left[T(k) > \tau\right] = \left[1 - \binom{k}{2}\frac{1}{2N} + O\left(\frac{1}{N^2}\right)\right]^{\lfloor 2N\tau \rfloor} \longrightarrow e^{-\binom{k}{2}\tau}, \tag{2.4}$$

as $N \to \infty$. Thus, in the limit $N \to \infty$, $T(k)$ is exponentially distributed with parameter $\binom{k}{2}$ and expectation

$$\mathrm{E}\left[T(k)\right] = \frac{1}{\binom{k}{2}} = \frac{2}{k\left(k-1\right)}. \tag{2.5}$$

In continuous time, the $T(k)$, $k = 1, 2, \ldots, n-1$, are independent random variables.

Applying (2.5) we are able to calculate the expected time to the MRCA for a sample of size n

$$\begin{aligned}\mathrm{E}\left[T_{MRCA}\right] &= \mathrm{E}\left[\sum_{k=2}^{n} T(k)\right] = \sum_{k=2}^{n} \mathrm{E}\left[T(k)\right] \\ &= \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right).\end{aligned} \tag{2.6}$$

Note that $\mathrm{E}\left[T(2)\right] = 1$, which means that the waiting time for the last coalescence event is greater than half the expected time to the MRCA.

Finally, we obtain the expected total length of the ancestral tree T,

$$\mathrm{E}\left[T\right] = \mathrm{E}\left[\sum_{k=2}^{n} kT(k)\right] = \sum_{k=1}^{n-1} \frac{2}{k} \longrightarrow 2\left(\gamma + \log n\right), \tag{2.7a}$$

as $n \to \infty$, where $\gamma \approx 0.577216$ denotes Euler's constant ([33], p. 850). The variance of T is

$$\mathrm{Var}\,[T] = \sum_{k=1}^{n-1} \frac{4}{k^2} \longrightarrow 4\frac{\pi^2}{6}, \tag{2.7b}$$

as $n \to \infty$ ([16], p. 820).

Note that the total length of the tree grows very slowly with n, which is reasonable, because the lineages coalesce at rate $\binom{k}{2}$ and thus more rapidly in the early states of the coalescent. Ergo, the length of the ancestral tree is mainly determined by the deep branches, when few lineages are left (see Figure 2.4). Donnelly [8] showed that, as a consequence, increasing the sample size often has negligible effect.

Kingman's coalescent

Kingman's coalescent can be described in a brief mathematical manner as a continuous-time Markov process. The state space $\mathcal{E}_n$ of the process is given by the set of equivalence relations on $\{1, \ldots, n\}$ as described above (see p. 7). Suppose $n \ll 2N$ and let φ be a possible equivalence relation and ψ denote another equivalence relation, that is generated by coalescing two equivalence classes of φ. The number of equivalence classes in φ is denoted by k. Measuring time in $2N$ generations ($\tau = \frac{t}{2N}$, $\Delta = \frac{1}{2N}$), the transition probabilities (cf. (2.3)) are given by

$$\begin{aligned}
&P\left[\varphi \mapsto \psi \text{ in } (\tau, \tau + \Delta)\right] = \Delta \\
&P\left[\varphi \mapsto \varphi \text{ in } (\tau, \tau + \Delta)\right] = 1 - \frac{k(k-1)}{2}\Delta + o\left(\Delta\right) \\
&P\left[\geq 2 \text{ coalescence events in } (\tau, \tau + \Delta)\right] = o\left(\Delta\right).
\end{aligned}$$

As $N \to \infty$ and, consequently, $\Delta \to 0$, these probabilities describe a Poisson process with parameter $\frac{k(k-1)}{2}$. Therefore, Kingman's coalescent can be understood as a stochastic process, that results from considering a sequence of Poisson processes with parameters $\frac{k(k-1)}{2}$, where k runs through $n, n-1, \ldots, 2$ and is reduced by one every time a coalescence event occurs.

Another way of describing the coalescent is to introduce an infinitesimal

generator $Q = (q_{\varphi\eta})_{\varphi,\eta\in\mathcal{E}_n}$ given by

$$q_{\varphi\eta} = \begin{cases} \frac{k(k-1)}{2}, & \text{if } \varphi = \eta, \\ 1, & \text{if } \varphi \prec \eta, \\ 0, & \text{otherwise,} \end{cases} \tag{2.8}$$

where $\varphi \prec \eta$ if and only if η is generated by coalescing two equivalence classes of φ ([33], p. 848).

Neutral mutations

In the beginning, we realized that the occurrence of neutral mutations can be modelled separately, after the genealogy of the sample has been created. For this purpose, we assume the selectively neutral infinite sites model. Since its introduction by Kimura [20] in 1969, studies of nucleotide variations have often been based on it. According to the infinite sites model, each small stretch of DNA (respectively each gene) consists of an infinite sequence of nucleotides. Since there are infinitely many, each mutation is supposed to arise at a site, where no mutation has occurred before. In that way, there are at most two different nucleotide types at each site, a wild type and a mutant type. Thus, each mutation causes a new segregating site, referred to as a single nucleotide polymorphism (SNP). Mutations are assumed to be selectively neutral.

Let ν denote the mutation rate per gene per generation. Hence, on average, there are $2N\nu$ segregating sites arising in a diploid population of size N per generation. According to Kimura ([9], p. 238), considering the mean number of generations for which a mutant is neither fixed nor extinct in the population, the total mean number of segregating sites in the population at stationarity is given by $\theta = 4N\nu$, where $\frac{\theta}{2} = 2N\nu$ indicates the mean number of mutations per generation. In the coalescent, time is measured in units of $2N$ generations. Thus, the mean number of mutations per unit of scaled time is calculated as $\frac{\theta}{2}2N$. The population consists of $2N$ genes. Therefore, the mutation rate per gene per unit of scaled time is given by $\frac{\theta}{2}\frac{2N}{2N} = \frac{\theta}{2}$.

Since the mutation rate is small and N large, it is reasonable to model the mutation process according to a Poisson distribution with parameter $\frac{\theta}{2}$. Mutations are equally likely to occur anywhere on the branch and are independent of the underlying genealogy ([33], p. 869). Thus, the number of mutations that occur on a branch of length τ units of scaled time is Poisson

distributed with parameter $\frac{\theta\tau}{2}$. On the other hand, if there are currently k lineages in the ancestral tree (which equates to k genes), mutations occur at rate $\frac{k\theta}{2}$ per unit of time.

The distribution of segregating sites

The number of segregating sites, S, in a sample of n genes is an important summary statistic for genetic polymorphisms. We are going to study the distribution of the number of segregating sites under the coalescent process. Examining the coalescent with neutral mutations, we have to consider two Poisson processes simultaneously. In the previous sections, we showed that, given k lineages, coalescence events occur at rate $\frac{k(k-1)}{2}$ and mutations at rate $\frac{k\theta}{2}$ per unit of time. So, the probability that the next event is a coalescence event is given by

$$P\left[\text{coalescence}|\text{event}\right] = \frac{\frac{k(k-1)}{2}}{\frac{k(k-1)}{2} + \frac{k\theta}{2}} = \frac{k-1}{\theta + k - 1}, \tag{2.9a}$$

and the probability that the next event is a mutation is

$$P\left[\text{mutation}|\text{event}\right] = 1 - \frac{k-1}{\theta + k - 1} = \frac{\theta}{\theta + k - 1}. \tag{2.9b}$$

Let S_k denote the number of segregating sites that appear when k lineages are left, and $S = \sum_{k=2}^{n} S_k$. Using (2.9a) and (2.9b) we find

$$P\left[S_k = j\right] = \frac{k-1}{\theta + k - 1}\left(\frac{\theta}{\theta + k - 1}\right)^j, \tag{2.10}$$

where $j = 0, 1, 2, \ldots$; thus, S_k is distributed geometrically with mean and variance

$$\mathrm{E}\left[S_k\right] = \frac{1}{\frac{k-1}{\theta+k-1}} - 1 = \frac{\theta}{k-1}, \tag{2.11a}$$

$$\mathrm{Var}\left[S_k\right] = \frac{\theta}{k-1} + \left(\frac{\theta}{k-1}\right)^2. \tag{2.11b}$$

Applying (2.11a) and (2.11b), we are able to determine the mean and the

variance of S, which are

$$\begin{aligned} \mathrm{E}[S] &= E\left[\sum_{k=2}^{n} S_k\right] = \sum_{k=2}^{n} E[S_k] \\ &= \theta \sum_{k=1}^{n-1} \frac{1}{k}, \end{aligned} \tag{2.12a}$$

$$\begin{aligned} \mathrm{Var}[S] &= Var\left[\sum_{k=2}^{n} S_k\right] = \sum_{k=2}^{n} Var[S_k] \\ &= \theta \sum_{k=1}^{n-1} \frac{1}{k} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2}. \end{aligned} \tag{2.12b}$$

The distribution of S can even be calculated explicitly by

$$P[S=j] = \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} P[S_i = j]. \tag{2.13}$$

An important application of (2.12a) is the estimation of θ as

$$\theta \approx \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}. \tag{2.14}$$

Comparison of (2.12) with (2.7) shows that the moments of S follow directly from the moments of T,

$$\mathrm{E}[S] = \mu \, \mathrm{E}[T] \tag{2.15a}$$

$$\mathrm{Var}[S] = \mu \, \mathrm{E}[T] + \mu^2 \, \mathrm{Var}[T], \tag{2.15b}$$

where $\mu = \frac{\theta}{2}$. Hence, inserting the limit values of (2.7a) and (2.7b) for $n \to \infty$, the expectation and the variance of S can be approximated by

$$\mathrm{E}[S] \sim 2\mu (\gamma + \log n), \tag{2.16a}$$

$$\mathrm{Var}[S] \sim 2\mu (\gamma + \log n), \tag{2.16b}$$

as $n \to \infty$ ([43], p. 258).

The number of neutral mutations that occur on a branch of length τ units of scaled time is Poisson distributed with parameter $\mu\tau$, where μ is the expected number of selectively neutral mutations per gene per unit of

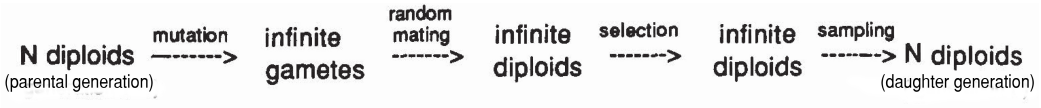

Figure 2.5: The life cycle in models with selection.

scaled time (cf. 2.1.1). Thus, the number of segregating sites conditioned on a genealogy of total length T can be approximated by a Poisson distribution with mean μT. Therefore,

$$P\left[S = j\right] = \int_0^\infty e^{-\mu\tau} \frac{(\mu\tau)^j}{j!} \, \mathrm{d}F(\tau), \tag{2.17}$$

$$F(\tau) = P\left[T \leq \tau\right], \quad \tau \geq 0,$$

where $T = \sum_{k=2}^{n} kT(k)$ denotes the total length of the ancestral tree in units of scaled time ([16], p. 820). The limit values of (2.16a) and (2.16b) confirm the appropriateness of the Poisson distribution. On the other hand, (2.15a) and (2.15b) follow directly from (2.17).

2.1.2 The coalescent in models with selection

In Kingman's coalescent, due to selective neutrality, each parental gene is equally likely to be chosen by a daughter gene. This very property of the standard coalescent looses its validity under selection, because the reproductive success of a gene depends on its allelic state. Consequently, it is no longer possible to separate "state" from "descent". Nevertheless, there are two different approaches to solve this problem. The first approach leads to the so called "ancestral selection graph". The coalescent gets extended by including branching events into the realization process of the genealogy and by removing the branches that carry deleterious alleles thereafter ([24, 32]). The second approach is known as the "conditional structured coalescent", which allows us to study the effects of selection on the underlying genealogy. It was pioneered by Kaplan *et al.* [16] and will be described here.

The coalescent of a single selected locus

As in the last section, we assume a Wright-Fisher model based on the infinite sites model for a diploid population of size N, where time is measured

in generations. The life cycle of the genes is shown in Figure 2.5. We restrict our study to the case where selection is acting on a single nucleotide site. Consequently, the probability of being chosen as a parent depends on the allelic state of the selected site. Mutations hitting this site, referred to as selective mutations, change its allelic state and thus influence the fitness of the gene. In order to create the ancestral tree of a sample of size n, we have to observe both the waiting time for the next coalescence event and the time to the next selective mutation. Obviously, these waiting times are no longer independent random variables (cf. 2.1.1, p. 8) and therefore, we need to study their distribution conditioned on the distribution of the ancestral frequencies. Note that, since there is no recombination, the model assumes all sites to be completely linked. Therefore, the genealogy of the neutral loci is the same as the genealogy of the selected locus.

Consider a single diallelic selected locus A with allelic types A_1 and A_2. We refer to the current generation as generation 0, and to the population t generations back in time as the ancestral generation t. Let $X(t)$ denote the ancestral frequency of A_1 in the ancestral generation t. The fitnesses of the three possible genotypes A_1A_1, A_1A_2 and A_2A_2 are given by ω_{11}, ω_{12} and ω_{22}. The mean fitness of the population in the ancestral generation t is

$$\bar{\omega}(t) = \omega_{11}X(t)^2 + \omega_{12}X(t)\left(1 - X(t)\right) + \omega_{22}\left(1 - X(t)\right)^2.$$

Selective mutations from A_1 to A_2 and vice versa, forward in time, occur at rate u and v, respectively. Assume

$$\omega_{ij} = 1 + O\left(\frac{1}{N}\right) \qquad i, j \in \{1, 2\},$$
$$u = \frac{\beta_1}{2N} + O\left(\frac{1}{N^2}\right), \quad v = \frac{\beta_2}{2N} + O\left(\frac{1}{N^2}\right),$$

where $\beta_1 > 0$ and $\beta_2 > 0$ are appropriate constants.

Let $p_{ij}(t)$ denote the probability that a randomly chosen gene in the ancestral generation t is of allelic type A_i and its parent from the ancestral generation $t+1$ of allelic type A_j. Then,

$$\begin{aligned} p_{11}(t) &= \frac{1-u}{\bar{\omega}(t+1)}\left(X(t+1)^2\omega_{11} + X(t+1)\left(1 - X(t+1)\right)\omega_{12}\right) \\ &= X(t+1) + O\left(\frac{1}{N}\right), \end{aligned} \tag{2.18a}$$

$$p_{21}(t) = \frac{u}{\bar{\omega}(t+1)}\left(X(t+1)^2\omega_{11} + X(t+1)\left(1-X(t+1)\right)\omega_{12}\right)$$
$$= \frac{\beta_1}{2N}X(t+1) + O\left(\frac{1}{N^2}\right), \tag{2.18b}$$
$$p_{22}(t) = \frac{1-v}{\bar{\omega}(t+1)}\left(\left(1-X(t+1)\right)^2\omega_{22} + X(t+1)\left(1-X(t+1)\right)\omega_{12}\right)$$
$$= 1 - X(t+1) + O\left(\frac{1}{N}\right), \tag{2.18c}$$
$$p_{12}(t) = \frac{v}{\bar{\omega}(t+1)}\left(\left(1-X(t+1)\right)^2\omega_{22} + X(t+1)\left(1-X(t+1)\right)\omega_{12}\right)$$
$$= \frac{\beta_2}{2N}\left(1-X(t+1)\right) + O\left(\frac{1}{N^2}\right). \tag{2.18d}$$

Let $p_i(t)$ indicate the probability that a randomly chosen gene from the ancestral generation t is of allelic type A_i, regardless of the allelic identity of its parent. Using (2.18) we obtain

$$p_1(t) = p_{11}(t) + p_{12}(t)$$
$$= X(t+1) + O\left(\frac{1}{N}\right) \tag{2.19a}$$
$$p_2(t) = p_{21}(t) + p_{22}(t)$$
$$= 1 - X(t+1) + O\left(\frac{1}{N}\right). \tag{2.19b}$$

Thus, the probability that a randomly chosen A_1 allele from the ancestral generation t has a parental A_2 allele in the ancestral generation $t+1$ is

$$\frac{p_{12}(t)}{p_1(t)} = \frac{\beta_2}{2N}\frac{\left(1-X(t+1)\right)}{X(t+1)} + O\left(\frac{1}{N^2}\right). \tag{2.20}$$

The marginal fitness of A_1 in the ancestral generation $t+1$ is $\omega_1(t+1) = X(t+1)\omega_{11} + \left(1-X(t+1)\right)\omega_{12}$. Fix a particular parental gene carrying A_1 in the ancestral generation $t+1$. Assuming that no mutation event occurs, the frequency of its daughter genes in the ancestral generation t is given by

$$\frac{1}{2N}\frac{\omega_1(t+1)}{\bar{\omega}(t+1)} = \frac{p_{11}(t)}{2NX(t+1)},$$

where $\frac{1}{2N}$ is the frequency of a single gene (cf. (2.18a)). Thus, the probability that a randomly chosen A_1 allele has this particular A_1 allele as a parent is

$$\frac{p_{11}(t)}{2NX(t+1)p_1(t)}.$$

Since the number of A_1 alleles in the ancestral generation $t+1$ is given by $2NX(t+1)$, the probability that two randomly chosen A_1 alleles from the ancestral generation t have the same A_1 parental gene equals

$$2NX(t+1)\left(\frac{p_{11}(t)}{2NX(t+1)p_1(t)}\right)^2 = \frac{1}{2NX(t+1)} + O\left(\frac{1}{N^2}\right). \tag{2.21}$$

The Q process

The different changes, both coalescence events and selective mutations, during the coalescent can be described by a two dimensional jump process. We define $Q(t) = (i, j)$ if the ancestral generation t consists of i alleles of type A_1 and j alleles of type A_2. On the one hand, $Q(t)$ counts the number of the two different alleles, on the other hand, $|Q(t)| = i + j$ denotes the total number of ancestral genes in the ancestral generation t. Let T_k denote the waiting time (measured in generations backward in time) for the k^{th} event, either coalescence or mutation. The set $\{Z_k\}_{k\in\mathbb{N}}$ describes the successive random states of the process starting in generation 0, i.e., at the time of sampling, and going backward in time. The corresponding T_k for each state Z_k denotes the number of generations the process stays in this state, i.e., the time until the next jump (see Figure 2.6). Hence, the representation of the Q process is given by

$$Q(t) = \begin{cases} Q(0) & \text{if } t < T_1, \\ Z_k & \text{if } \sum_{i=1}^{k} T_i \le t < \sum_{i=1}^{k+1} T_i, \end{cases} \tag{2.22}$$

where $k \ge 1$. Naturally, $|Q(t)|$ decreases with time and the process eventually reaches $(0, 1)$ or $(1, 0)$, when the MRCA occurs.

Given $Q(t) = (i, j)$, the probability that one of the i genes carrying A_1 from the ancestral generation t has a parental gene carrying A_2 is the same as the probability that the Q process changes its state to $(i-1, j+1)$ in generation $t+1$. Thus, using (2.20), the probability that the Q process changes its state to $(i-1, j+1)$ due to a selective mutation (from A_2 to A_1 going forward in time) within one generation is

$$i\frac{\beta_2}{2N}\frac{(1 - X(t+1))}{X(t+1)} + O\left(\frac{1}{N^2}\right). \tag{2.23}$$

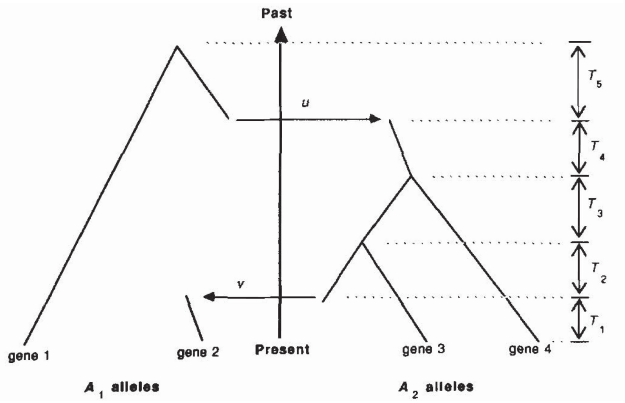

Figure 2.6: A realization of an ancestral tree in the coalescent in models with selection for a sample of size $n = 4$. The Q process starts in generation 0 with $(2, 2)$ and changes its state in the ancestral generation T_1 to $(1, 3)$ due to a selective mutation. In the ancestral generation T_1+T_2 the first two lineages coalesce and the Q process moves to the state $(1, 2)$. Another coalescence event follows that changes the state to $(1, 1)$. In the ancestral generation $T_1 + T_2 + T_3 + T_4$ another mutation occurs and the process changes to $(2, 0)$. Finally, with the last coalescence event in the ancestral generation $T_1 + T_2 + T_3 + T_4 + T_5$ the MRCA arises and the Q process reaches the state $(1, 0)$.

Similarly, the probability that the Q process changes its state to $(i+1, j-1)$ due to a selective mutation from the ancestral generation t to the ancestral generation $t + 1$ is

$$j\frac{\beta_1}{2N}\frac{X(t+1)}{(1-X(t+1))} + O\left(\frac{1}{N^2}\right). \tag{2.24}$$

The probability that two or more mutation events happen in one generation is of order $O\left(\frac{1}{N^2}\right)$.

Next consider coalescence events. The probability that two of the i genes carrying A_1 from the ancestral generation t have the same parental gene carrying A_1 is the same as the probability that the Q process changes its state to $(i-1, j)$ going from t to $t+1$. Thus, using (2.21), the probability that the Q process changes its state to $(i-1, j)$ due to a coalescence event of two genes carrying A_1 within one generation is

$$\binom{i}{2}\frac{1}{2NX(t+1)} + O\left(\frac{1}{N^2}\right). \tag{2.25}$$

Similarly, the probability that the Q process changes its state to $(i, j-1)$

due to a coalescent event of two A_2 carrying genes from t to $t+1$ is

$$\binom{j}{2}\frac{1}{2N\left(1-X(t+1)\right)}+O\left(\frac{1}{N^2}\right). \tag{2.26}$$

The probability that two or more coalescence events happen in one generation is of order $O\left(\frac{1}{N^2}\right)$. Obviously, all these probabilities depend on the ancestral frequency process, $\{X(t), t>0\}$, and thus, the dynamics of the Q process have to be studied conditioned on it.

The distribution of $Q(t+1)$ conditioned on $Q(t)=(i,j)$ and $X(t+1)$ follows from (2.23)-(2.26) and is given by

$$P[Q(t+1)=Q(t)]=1-\frac{1}{2N}h_{ij}\left(X(t+1)\right)+O\left(\frac{1}{N^2}\right), \tag{2.27a}$$

where

$$h_{ij}(x)=\frac{\binom{i}{2}}{x}+\frac{\binom{j}{2}}{1-x}+\frac{j\beta_1 x}{1-x}+\frac{i\beta_2(1-x)}{x}. \tag{2.27b}$$

The jump probabilities, given that an event happens, are

$$\begin{aligned}
&P[Q(t+1)=(i-1,j)|Q(\tau)=(i,j)]=q_{i-1,j}\left(X(t+1)\right)+O\left(\frac{1}{N}\right),\\
&P[Q(t+1)=(i,j-1)|Q(\tau)=(i,j)]=q_{i,j-1}\left(X(t+1)\right)+O\left(\frac{1}{N}\right),\\
&P[Q(t+1)=(i+1,j-1)|Q(\tau)=(i,j)]=q_{i+1,j-1}\left(X(t+1)\right)+O\left(\frac{1}{N}\right),\\
&P[Q(t+1)=(i-1,j+1)|Q(\tau)=(i,j)]=q_{i-1,j+1}\left(X(t+1)\right)+O\left(\frac{1}{N}\right),
\end{aligned} \tag{2.27c}$$

where

$$\begin{aligned}
&q_{i-1,j}(x)=\binom{i}{2}\frac{1}{xh_{ij}(x)}, && q_{i,j-1}(x)=\binom{j}{2}\frac{1}{(1-x)h_{ij}(x)},\\
&q_{i+1,j-1}(x)=\frac{j\beta_1 x}{(1-x)h_{ij}(x)}, && q_{i-1,j+1}(x)=\frac{i\beta_2(1-x)}{xh_{ij}(x)}.
\end{aligned} \tag{2.27d}$$

It follows from (2.27c) that if N is large, there are only four possible states the Q process can jump to. These are $(i-1,j)$, $(i,j-1)$, $(i+1,j-1)$ and $(i-1,j+1)$, where the first two describe coalescence events and the second two selective mutations.

It is possible to derive the joint distribution of the random variables $\{T_k\}$ and $\{Z_k\}$ conditioned on the ancestral frequency process $\{X(t), t > 0\}$ by repeatedly using the formulas in (2.27):

$$P\left[T_k = t_k, Z_k = z_k | Q(0) = z_0\right] = \\ = \frac{1}{(2N)^m} E\left[\prod_{k=1}^{m} q_{z_k}(X(s_k))\; h_{z_{k-1}}(X(s_k)) \prod_{l=s_{k-1}+1}^{s_k-1} \left(1 - \frac{h_{z_{k-1}}(X(l))}{2N}\right)\right] \\ + O\left(\frac{1}{N^{m+1}}\right). \tag{2.28}$$

Here $s_k = \sum_{i=1}^{k} t_i$, $s_0 = 0$, $t_k > 0$ and $m \geq 1$ denotes the total number of events until the MRCA. We note that depending on the value of z_{k-1}, z_k can assume one of the four possible values in (2.27c) and, if $s_{k-1} + 1 > s_k - 1$ the second product in the expectation is set equal to 1.

Consider time measured in units of $2N$ generations, i.e., $2N\tau = t$ and $2N\Delta = 1$. Therefore, $X(\tau)$ is the frequency of A_1 at time τ, which equals the frequency of A_1 in the ancestral generation $2N\tau$. It can be shown (see [16], p. 823) that in the limit $N \to \infty$, the distribution of the Q process conditioned on the ancestral frequency process in scaled time, $\{X(\tau), \tau > 0\}$, can be approximated by a time inhomogeneous Markov jump process. Its representations follows from (2.27), considering that $h_{ij}\left(X(\tau + \Delta)\right) = h_{ij}\left(X(\tau)\right) + O\left(\Delta\right)$, and is given by

$$P[Q(\tau + \Delta) = Q(\tau)] = 1 - h_{ij}\left(X(\tau)\right)\Delta + O\left(\Delta^2\right) \tag{2.29a}$$

as $\Delta \to 0$, where

$$h_{ij}(x) = \left(\frac{\binom{i}{2}}{x} + \frac{i\beta_2(1-x)}{x}\right)\chi(x > 0) + \left(\frac{\binom{j}{2}}{1-x} + \frac{j\beta_1 x}{1-x}\right)\chi(x < 1). \tag{2.29b}$$

We set $\chi(x > 0)$ equal to 1 if $x > 0$ and equal to 0 otherwise and $\chi(x < 1)$ is defined in a similar way.

The conditional jump probabilities are

$$P[Q(\tau + \Delta) = (i-1, j)] = q_{i-1,j}\left(X(\tau)\right) + O\left(\Delta\right), \\ P[Q(\tau + \Delta) = (i, j-1)] = q_{i,j-1}\left(X(\tau)\right) + O\left(\Delta\right), \\ P[Q(\tau + \Delta) = (i+1, j-1)] = q_{i+1,j-1}\left(X(\tau)\right) + O\left(\Delta\right),$$

$$P[Q(\tau+\Delta)=(i-1,j+1)] = q_{i-1,j+1}\left(X(\tau)\right) + O\left(\Delta\right), \tag{2.29c}$$

where

$$\begin{aligned} q_{i-1,j}(x) &= \binom{i}{2}\frac{\chi(x>0)}{xh_{ij}(x)}, \quad q_{i,j-1}(x) = \binom{j}{2}\frac{\chi(x<1)}{(1-x)h_{ij}(x)}, \\ q_{i+1,j-1}(x) &= \frac{j\beta_1 x\ \chi(x<1)}{(1-x)h_{ij}(x)}, \quad q_{i-1,j+1}(x) = \frac{i\beta_2(1-x)\ \chi(x>0)}{xh_{ij}(x)}. \end{aligned} \tag{2.29d}$$

The length of the ancestral tree

Let $T(i,j)$ denote the total length of the ancestral tree (measured in units of scaled time) of a sample that consists of i alleles of type A_1 and j alleles of type A_2, $i+j \geq 2$. If the ancestral frequency process can be approximated by a diffusion process, then the moments of $T(i,j)$ can be calculated by solving a system of second order differential equations (see [16], p. 826).

The distribution of segregating sites

In general, when selection is acting on multiple sites, the number of segregating sites can be written as $S = S_{neu} + S_{sel}$, where S_{neu} denotes the number of neutral segregating sites and S_{sel} the number of segregating sites where selection is acting on. If there is just one selected segregating site, as assumed above, S_{sel} is no more than one and therefore negligible compared to S_{neu}. In this case, the properties of S can be derived from those of S_{neu} using equations (2.15) and (2.17) (with $T = T(i,j)$). Other cases, where S_{sel} is not negligible, have to be studied separately ([16], pp. 819-820).

2.1.3 The coalescent in models with selection and recombination

If an ancestral gene of a sample was created due to recombination, then the lineage of that gene, going backward in time, splits into two. Thus, the number of lineages is increasing instead of decreasing. Therefore, including recombination into the coalescent framework, the genealogical history of a sample can no longer be represented by a single ancestral tree, but by the so called "ancestral recombination graph" [13]. As a consequence, the resulting stochastic process is in general quite complicated. Nevertheless, the coalescent in selectively neutral models with recombination has been studied by a

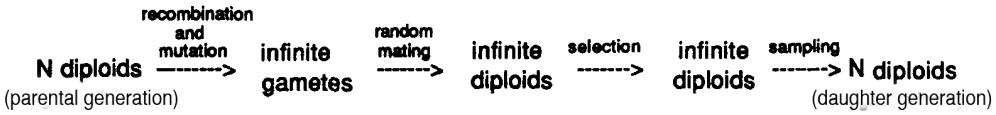

Figure 2.7: The life cycle in models with selection and recombination.

number of authors and the distributional properties of S at equilibrium are well characterized ([14, 38, 44]). The goal of this section is to study a special case, where both recombination and selection are acting at the same time.

The coalescent of a single neutral locus linked to a selected locus

The theory of the coalescent of a single selected locus (see 2.1.2) can easily be extended to the case where, in addition, recombination is acting between the selected locus and a single neutral locus [15]. The assumptions for this model are essentially the same as in 2.1.2. The population of size N evolves according to a Wright-Fisher model based on the infinite sites model. Therefore, the life cycle of the genes changes slightly (see Figure 2.7). Consider time measured in generations. If the two loci are completely linked, then the genealogy of the neutral site is exactly the same as the genealogy of the selected site. If recombination is acting between the two loci, this complete linkage gets destroyed. The effects of recombination on the coalescent of the neutral locus will be discussed here.

Consider two linked loci. Assume that selection is acting on the A locus and its different allelic states are denoted by A_1 and A_2, respectively. The second locus, referred to as the B locus, is supposed to be selectively neutral. Let $X(t)$ denote the ancestral frequency of A_1 in the ancestral generation t. The fitness of the different diploid genotypes of the selected locus A_1A_1, A_1A_2 and A_2A_2 are given by ω_{11}, ω_{12} and ω_{22}. In addition to the selective mutations that occur at rate u and v, respectively, recombination is acting at rate r given by

$$r = \frac{R}{2N} + O\left(\frac{1}{N^2}\right),$$

where $R > 0$ is an appropriate constant.

Let $p_{ij}(t)$ denote the probability that a randomly chosen gene at locus B in the ancestral generation t is linked to A_i and its parent from the ancestral

generation $t+1$ is linked to A_j. Then,

$$\begin{aligned} p_{11}(t) &= \frac{1}{\bar{\omega}(t+1)}\left(X(t+1)^2\omega_{11} + X(t+1)\left(1-X(t+1)\right)\omega_{12}\right) \\ &= X(t+1) + O\left(\frac{1}{N}\right), \end{aligned} \tag{2.30a}$$

$$\begin{aligned} p_{21}(t) &= \frac{1}{\bar{\omega}(t+1)}\left(uX(t+1)^2\omega_{11} + (u+r)X(t+1)\left(1-X(t+1)\right)\omega_{12}\right) \\ &= \frac{\beta_1 + R(1-X(t+1))}{2N}X(t+1) + O\left(\frac{1}{N^2}\right), \end{aligned} \tag{2.30b}$$

$$\begin{aligned} p_{22}(t) &= \frac{1}{\bar{\omega}(t+1)}\left(\left(1-X(t+1)\right)^2\omega_{22} + X(t+1)\left(1-X(t+1)\right)\omega_{12}\right) \\ &= 1 - X(t+1) + O\left(\frac{1}{N}\right), \end{aligned} \tag{2.30c}$$

$$\begin{aligned} p_{12}(t) &= \frac{1}{\bar{\omega}(t+1)}\left(v\left(1-X(t+1)\right)^2\omega_{22} + (v+r)X(t+1)\left(1-X(t+1)\right)\omega_{12}\right) \\ &= \frac{\beta_2 + RX(t+1))}{2N}\left(1-X(t+1)\right) + O\left(\frac{1}{N^2}\right). \end{aligned} \tag{2.30d}$$

The Q process

In accordance with the last section we introduce a two dimensional jump process describing the different states of the sample in the coalescent with selection and recombination. The process, called the Q process, is defined in the same way as before (cf. 2.1.2), except that $Q(t) = (i,j)$ means that i ancestral genes are linked to A_1 and j ancestral genes are linked to A_2 in the ancestral generation t. The representation of the Q process is given by

$$Q(t) = \begin{cases} Q(0) & \text{if } t < T_1, \\ Z_k & \text{if } \sum_{i=1}^{k} T_i \leq t < \sum_{i=1}^{k+1} T_i, \end{cases} \tag{2.31}$$

where T_k denotes the waiting time for the k^{th} event (measured in generations), Z_k describes the corresponding random state of the process, and $k \geq 1$ (see Figure 2.8). Obviously, $|Q(t)|$ never increases and the process eventually reaches $(0,1)$ or $(1,0)$, which means that the MRCA is either linked to A_1 or A_2.

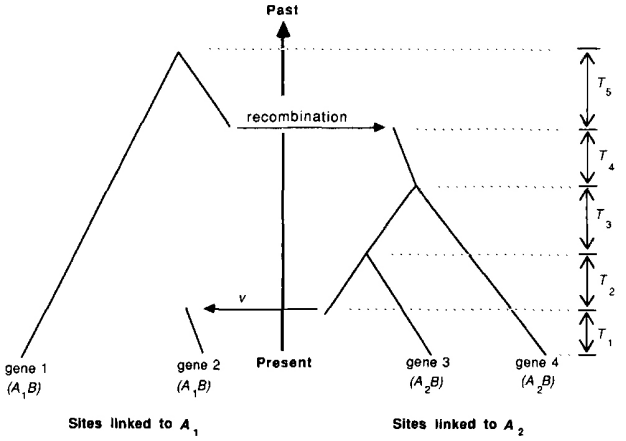

Figure 2.8: A realization of an ancestral tree in the coalescent in models with selection and recombination for a sample of size $n = 4$. The Q process starts in generation 0 with $(2, 2)$ and changes its state in the ancestral generation T_1 to $(1, 3)$ due to a selective mutation. In the ancestral generation T_1+T_2 the first two lineages coalesce and the Q process moves to the state $(1, 2)$. Another coalescence event follows that changes the state to $(1, 1)$. In the ancestral generation $T_1+T_2+T_3+T_4$ a recombination event occurs and the process changes to $(2, 0)$. Finally, with the last coalescence event in the ancestral generation $T_1+T_2+T_3+T_4+T_5$ the MRCA arises and the Q process reaches the state $(1, 0)$.

It follows from (2.30) that all formulas of 2.1.2 apply if we replace β_1 and β_2 by

$$\beta_1(x) = \beta_1 + R(1-x) \qquad \text{and} \qquad \beta_2(x) = \beta_2 + Rx, \tag{2.32}$$

respectively. Thus, the probability of an event (cf. (2.27b)) and the conditional jump probabilities (cf. (2.27d)) change to

$$h_{ij}(x) = \frac{\binom{i}{2}}{x} + \frac{\binom{j}{2}}{1-x} + \frac{j(\beta_1 + R(1-x))x}{1-x} + \frac{i(\beta_2 + Rx)(1-x)}{x}, \tag{2.33a}$$

$$\begin{aligned} q_{i-1,j}(x) &= \binom{i}{2}\frac{1}{xh_{ij}(x)}, & q_{i,j-1}(x) &= \binom{j}{2}\frac{1}{(1-x)h_{ij}(x)}, \\ q_{i+1,j-1}(x) &= \frac{j(\beta_1 + R(1-x))x}{(1-x)h_{ij}(x)}, & q_{i-1,j+1}(x) &= \frac{i(\beta_2 + Rx)(1-x)}{xh_{ij}(x)}. \end{aligned} \tag{2.33b}$$

In the same way, we can derive the limit approximation of the coalescent of a linked single neutral locus, which is a time inhomogeneous Markov jump process (cf. (2.29)).

Chapter 3

A deterministic model

Maynard Smith and Haigh [29] introduced the term "hitchhiking effect" into the field of population genetics. In 1974, they presented a deterministic model, which describes how a single selective substitution can affect the heterozygosity on a linked neutral locus. It is a pioneering work about genetic hitchhiking and successive studies almost always compare their results to it. In the following treatment, some results will be generalized in order to reach a better comparability with other models.

3.1 The model

Consider two linked loci. Selection is acting on the first locus, referred to as the A locus. Assume that a favourable mutation arises in a single copy with selective advantage $s > 0$. This selectively favourable mutation is denoted by A_1, while the deleterious wild type is called A_2. The other locus is selectively neutral and referred to as the B locus. Its two different allelic types are denoted by B_1 and B_2. The favourable mutant A_1 arises either on a gene carrying B_1 or B_2, also referred to as its background. The frequency of A_1 is given by p and we assume that it can be modelled deterministically. This implies that the mutation is eventually going to fixation. Obviously, the frequency of A_2 is given by $1 - p$.

Let Q_{A_1} denote the fraction of B_1 alleles among genes carrying A_1, and let Q_{A_2} denote the fraction of B_1 alleles among genes carrying A_2. Then, the total frequency of B_1 in the whole population is given by

$$Q_{A_1} p + Q_{A_2}(1 - p), \tag{3.1}$$

and the total frequency of B_2 alleles is given by $(1 - Q_{A_1})\, p + (1 - Q_{A_2})\, (1 - p)$.

genotypes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
fitnesses	$\omega_{11} = 1 + s$	$\omega_{12} = 1 + s$	$\omega_{21} = 1$	$\omega_{22} = 1$
frequencies	P_{11}	P_{12}	P_{21}	P_{22}

Table 3.1: The fitnesses and frequencies of the four possible genotypes in the haploid case.

The frequencies of the four possible genotypes A_1B_1, A_1B_2, A_2B_1 and A_2B_2 are denoted by P_{11}, P_{12}, P_{21} and P_{22}, respectively. We suppose that the fractions Q_{A_1} and Q_{A_2} of B_1 alleles on background A_1 and A_2 will change depending on the frequency p of A_1. Therefore, we calculate these variables by

$$p = P_{11} + P_{12}, \tag{3.2}$$

$$Q_{A_1} = \frac{P_{11}}{P_{11} + P_{12}} = \frac{P_{11}}{p}, \tag{3.3}$$

$$Q_{A_2} = \frac{P_{21}}{P_{21} + P_{22}} = \frac{P_{21}}{1 - p}. \tag{3.4}$$

Note that the frequencies of the different genotypes can also be expressed in terms of p, Q_{A_1} and Q_{A_2}; indeed $P_{11} = p\ Q_{A_1}$, $P_{12} = p\ (1 - Q_{A_1})$, $P_{21} = (1 - p)\ Q_{A_2}$ and $P_{22} = (1 - p)\ (1 - Q_{A_2})$. For further investigations, we distinguish between the haploid and the diploid case.

3.1.1 The haploid case

Since the selective advantage of A_1 is given by s and the B locus is supposed to be neutral, the fitnesses of the four different genotypes are given by $\omega_{11} = \omega_{12} = 1+s$ for those containing A_1, and $\omega_{21} = \omega_{22} = 1$ for those containing A_2 (see Table 3.1). Let r denote the recombination rate, which is the probability of a recombination event between the two loci per generation.

We start in generation zero with a single selectively favourable mutant A_1. Maynard Smith and Haigh restricted their study to the case, where the mutant arises on a gene carrying B_2. However, this restriction is not necessary and we will consider the general case, where the mutant arises either on a gene carrying B_1 or B_2 (cf. [37]).

For convenience, we assume that in the life cycle random mating follows selection. Let P'_{ij} denote the frequencies of the four different genotypes A_iB_j

in the next generation, $i, j \in \{1, 2\}$. These frequencies can be calculated by

$$P'_{ij} = \sum_{k,l=1}^{2} \sum_{s,t=1}^{2} \frac{P_{kl}\omega_{kl}}{\bar{\omega}} \frac{P_{st}\omega_{st}}{\bar{\omega}} P\left[A_k B_l, A_s B_t \to A_i B_j\right],$$

where

$$\bar{\omega} = \sum_{i,j=1}^{2} \omega_{ij} P_{ij} = \omega_{11} p + \omega_{22}(1-p) = 1 + sp \tag{3.5}$$

denotes the mean fitness of the population and $P\left[A_k B_l,\ A_s B_t \to A_i B_j\right]$ is the probability that parental genotypes $A_k B_l$ and $A_s B_t$ produce an $A_i B_j$ offspring. For example, the frequency of $A_1 A_2$ in the next generation can be calculated by

$$\begin{aligned} P'_{11} &= \frac{1}{\bar{\omega}^2}\big[P_{11}^2\omega_{11}^2 + P_{11}\omega_{11}P_{12}\omega_{12} + \\ &\quad + P_{11}\omega_{11}P_{21}\omega_{21} + P_{11}\omega_{11}P_{22}\omega_{22}(1-r) + P_{12}\omega_{12}P_{21}\omega_{21}r\big] \\ &= (1+s)^2 p^2 Q_{A_1}^2 + (1+s)^2 p^2 \left(1 - Q_{A_1}\right)\left[\frac{1}{2}Q_{A_1} + \frac{1-r}{2}\left(1 - Q_{A_1}\right)\right]. \end{aligned}$$

Let p', Q'_{A_1} and Q'_{A_2} indicate the frequency of A_1 and the fractions of B_1 on its respective backgrounds in the next generation. Then, using (3.2)-(3.4), we obtain

$$p' = P'_{11} + P'_{12} = \frac{p\omega_{11}}{\bar{\omega}} = \frac{(1+s)p}{(1+s)p + (1-p)}, \tag{3.6}$$

$$\begin{aligned} Q'_{A_1} &= \frac{P'_{11}}{P'_{11} + P'_{12}} \\ &= Q_{A_1} + \frac{\omega_{22} r}{\bar{\omega}}(1-p)\left(Q_{A_2} - Q_{A_1}\right) \\ &= Q_{A_1} + \frac{r}{1+sp}(1-p)\left(Q_{A_2} - Q_{A_1}\right), \end{aligned} \tag{3.7}$$

$$\begin{aligned} Q'_{A_2} &= \frac{P'_{21}}{P'_{21} + P'_{22}} \\ &= Q_{A_2} + \frac{\omega_{11} r}{\bar{\omega}} p\left(Q_{A_1} - Q_{A_2}\right) \\ &= Q_{A_2} + \frac{r}{1+sp} p(1+s)\left(Q_{A_1} - Q_{A_2}\right). \end{aligned} \tag{3.8}$$

3.1.2 The diploid case

In the diploid case, we assign fitnesses to the three possible genotypes on the A locus. These fitnesses and the frequencies of the four possible gametes

genotypes	A_1A_1	A_1A_2	A_2A_2	
fitnesses	$w_{A_1A_1} = 1 + s$	$w_{A_1A_2} = 1 + hs$	$w_{A_2A_2} = 1$	
gametes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
fitnesses	w_{11}	w_{12}	w_{21}	w_{22}
frequencies	P_{11} $= p\ Q_{A_1}$	P_{12} $= p\ (1 - Q_{A_1})$	P_{21} $= (1-p)\ Q_{A_2}$	P_{22} $= (1-p)\ (1 - Q_{A_2})$

Table 3.2: The fitnesses of the three possible genotypes and the fitnesses and frequencies of the four possible gametes in the diploid case.

are shown in Table 3.2. The fitness w_{ij} of a single gamete A_iB_j equals the marginal fitness w_i of A_i, $i, j \in \{1, 2\}$, and can be calculated by

$$\begin{aligned} w_1 = w_{11} = w_{12} &= w_{A_1A_1}(P_{11} + P_{12}) + w_{A_1A_2}(P_{22} + P_{21}) \\ &= (1 + s)p + (1 + hs)(1 - p) \\ &= 1 + s\,(h + p(1 - h))\,, \qquad (3.9) \\ w_2 = w_{22} = w_{21} &= w_{A_1A_2}(P_{11} + P_{12}) + w_{A_2A_2}(P_{22} + P_{21}) \\ &= (1 + hs)p + (1 - p) \\ &= 1 + shp. \qquad (3.10) \end{aligned}$$

The mean fitness $\bar{w}$ in the diploid population is given by

$$\begin{aligned} \bar{w} &= (1 + s)p^2 + 2(1 + hs)p(1 - p) + (1 - p)^2 \\ &= 1 + sp[2h + p(1 - 2h)]. \qquad (3.11) \end{aligned}$$

The frequencies P'_{ij} of the four different genotypes A_iB_j in the next generation, $i, j \in \{1, 2\}$ can be calculated by

$$\begin{aligned} \bar{w}P'_{11} &= w_1P_{11} - w_{A_1A_2}rD, \qquad (3.12a) \\ \bar{w}P'_{12} &= w_1P_{12} + w_{A_1A_2}rD, \qquad (3.12b) \\ \bar{w}P'_{21} &= w_2P_{21} + w_{A_1A_2}rD, \qquad (3.12c) \\ \bar{w}P'_{22} &= w_2P_{22} - w_{A_1A_2}rD, \qquad (3.12d) \end{aligned}$$

where D denotes the linkage disequilibrium given by

$$D = P_{11}P_{22} - P_{12}P_{21}. \qquad (3.12e)$$

Analogously to the haploid model (cf. (3.6), (3.7) and (3.8)) we obtain

$$p' = P'_{11} + P'_{12} = \frac{pw_1}{\bar{w}}$$

$$= \frac{p\left[(1+s)p + (1+hs)(1-p)\right]}{(1+s)p^2 + 2(1+hs)p(1-p) + (1-p)^2}$$
$$= \frac{p\left[1+s\left(h+p(1-h)\right)\right]}{1+sp[2h+p(1-2h)]}, \tag{3.13}$$

$$Q'_{A_1} = \frac{P'_{11}}{P'_{11}+P'_{12}} = Q_{A_1} + \frac{w_{A_1A_2}r}{w_1}(1-p)(Q_{A_2}-Q_{A_1})$$
$$= Q_{A_1} + \frac{r(1+hs)}{(1+s)p+(1+hs)(1-p)}(1-p)\left(Q_{A_2}-Q_{A_1}\right)$$
$$= Q_{A_1} + \frac{r(1+hs)}{1+s\left(h+p(1-h)\right)}(1-p)\left(Q_{A_2}-Q_{A_1}\right), \tag{3.14}$$

$$Q'_{A_2} = \frac{P'_{21}}{P'_{21}+P'_{22}} = Q_{A_2} + \frac{w_{A_1A_2}r}{w_2}p\left(Q_{A_1}-Q_{A_2}\right)$$
$$= Q_{A_2} + \frac{r(1+hs)}{(1+hs)p+(1-p)}p\left(Q_{A_1}-Q_{A_2}\right)$$
$$= Q_{A_2} + \frac{r(1+hs)}{1+shp}p\left(Q_{A_1}-Q_{A_2}\right). \tag{3.15}$$

3.1.3 Weaknesses of the deterministic model

When the mutant A_1 arises in the population, there is a good chance that it will be eliminated even though it is selectively favourable. The deterministic approach assumes that the favourable allele always gets fixed in the population and does not consider the cases where A_1 gets lost by random drift. Therefore, the deterministic model is not realistic as long as the frequency of A_1 is low. If the favourable allele gets fixed, its frequency initially has to increase much faster than in the deterministic model.

On the other hand, if the frequency of A_1 is near one, there is a good chance that the mutant fixes much faster than in the deterministic model due to chance fluctuations. Therefore, the number of recombination events until fixation of A_1 will be smaller and the effect on the neutral locus will be greater. The deterministic model underestimates the decrease in frequency of B_1 and thus it is not a good approximation when the frequency of A_1 is very high.

3.2 The effect of a single substitution

In order to quantify the effect of a single selected substitution we will examine its impact on different population variables and statistics. First, we will determine the change in frequencies on the neutral B locus, where we distinguish between the haploid and the diploid case. Second, we will consider heterozygosity on the neutral locus. Finally, we assume a selectively maintained polymorphism on the B locus with low selection compared to the A locus, and consider the probability of complete fixation.

3.2.1 The effect on neutral allele frequencies in the haploid case

The formulas in (3.6), (3.7) and (3.8) can be used to derive general equations for the fraction of B_1 alleles on background A_1 or A_2 in the n^{th} generation, denoted by $Q_{A_1}^{(n)}$ and $Q_{A_2}^{(n)}$. Consequently, we write $p^{(0)}$, $Q_{A_1}^{(0)}$ and $Q_{A_2}^{(0)}$ for the initial frequency and fractions. Iterating (3.7) and using (3.6) leads to

$$Q_{A_1}^{(n+1)} = Q_{A_1}^{(0)} + r\left(Q_{A_2}^{(0)} - Q_{A_1}^{(0)}\right)\left(1 - p^{(0)}\right)\sum_{k=0}^{n}\frac{(1-r)^k}{(1-p^{(0)}) + (1+s)^{k+1}p^{(0)}}. \tag{3.16}$$

Consider a population of size N. Thus, the initial frequency of A_1 is given by $p^{(0)} = \frac{1}{N}$. If N is sufficiently large, then $p^{(0)} \approx 0$, i.e., initially, the population consists nearly exclusively of genes carrying A_2. Therefore, according to (3.1), the total initial frequencies of B_1 and B_2 in the whole population are determined by $Q_{A_2}^{(0)}$ and $1 - Q_{A_2}^{(0)}$, the initial fractions of B_1 and B_2 on background A_2.

Since $r > 0$, according to (3.16), the fraction Q_{A_1} of B_1 on background A_1 decreases or increases monotonically depending on the sign of $\left(Q_{A_2}^{(0)} - Q_{A_1}^{(0)}\right)$. If A_1 arises on background B_1, we set $Q_{A_1}^{(0)} = 1$ and $1 - Q_{A_1}^{(0)} = 0$. Thus, $\left(Q_{A_2}^{(0)} - Q_{A_1}^{(0)}\right) < 0$ and Q_{A_1} decreases, which implies that $1 - Q_{A_1}$ increases and the mutant also appears on background B_2.

Otherwise, if the mutant arises on background B_2, we set $1 - Q_{A_1}^{(0)} = 1$ and $Q_{A_1}^{(0)} = 0$. Thus, $\left(Q_{A_2}^{(0)} - Q_{A_1}^{(0)}\right) > 0$ and Q_{A_1} increases, i.e., the A_1 allele arises on background B_1. How long it takes until the favourable allele on the selected locus appears on both backgrounds depends on the recombination rate r.

When the favourable mutation has fixed in the population, the distribution on the neutral locus is determined by the fractions of B_1 and B_2 among genes carrying A_1. Since the frequency p of A_1 after fixation is equal to one, these fractions are equal to the total frequencies of B_1 and B_2 in the whole population (cf. (3.1)). Therefore, we derive the frequencies of B_1 and B_2 after fixation, i.e., the equilibrium frequencies given by $Q_{A_1}^{\infty}$ and $1 - Q_{A_1}^{\infty}$, respectively. Using (3.16), we obtain for $n \to \infty$

$$Q_{A_1}^{\infty} = Q_{A_1}^{(0)} + \left(Q_{A_2}^{(0)} - Q_{A_1}^{(0)}\right) \Omega, \tag{3.17}$$

where

$$\Omega = r\left(1 - p^{(0)}\right) \sum_{k=0}^{\infty} \frac{(1-r)^k}{(1-p^{(0)}) + (1+s)^{k+1} p^{(0)}}. \tag{3.18}$$

Note that the sum in (3.18) always exists.

Iterating (3.8) and using (3.6) leads to

$$Q_{A_2}^{(n+1)} = Q_{A_2}^{(0)} + (1+s) r p^{(0)} \left(Q_{A_1}^{(0)} - Q_{A_2}^{(0)}\right) \sum_{k=0}^{n} \frac{(1+s)^k (1-r)^k}{(1-p^{(0)}) + (1+s)^{k+1} p^{(0)}}. \tag{3.19}$$

The fraction Q_{A_2} of B_1 on background A_2 decreases or increases monotonically depending on the sign of $\left(Q_{A_1}^{(0)} - Q_{A_2}^{(0)}\right)$. Thus, the fraction of B_1 on background A_2 increases if it decreases on background A_1 and vice versa. This dynamics during the selective sweep can be specified if we take a look at the differences of the fractions Q_{A_1} and Q_{A_2}. We use (3.7) and (3.8) to calculate this difference in the first generation, which is $Q_{A_1}^{(1)} - Q_{A_2}^{(1)} = (1-r)\left(Q_{A_1}^{(0)} - Q_{A_2}^{(0)}\right)$. Iterating this recursion leads to

$$Q_{A_1}^{(n)} - Q_{A_2}^{(n)} = (1-r)^n \left(Q_{A_1}^{(0)} - Q_{A_2}^{(0)}\right). \tag{3.20}$$

Obviously, the difference between the two fractions decreases geometrically, which implies $Q_{A_1}^{(n)} \sim Q_{A_2}^{(n)}$ for $n \to \infty$. Thus, during the selective sweep, the distribution of B_1 and B_2 on background A_1 tends to be asymptotically the same as the distribution on background A_2. The velocity of this process depends on the recombination rate r.

Approximation

The expression in (3.17) does not simplify, but we can approximate $Q_{A_1}^{\infty}$, the frequency of B_1 after fixation, in the case where selection is weak and

recombination is low compared to selection, i.e., for $0 < r \ll s \ll 1$, $p_0 \ll 1$. The recursion equations for the frequency of A_1 and the fractions of B_1 on its respective backgrounds, given by (3.6), (3.7) and (3.8), can be approximated by the differential equations

$$\dot{p} = \frac{\mathrm{d}}{\mathrm{d}t}p = \frac{1}{\bar{\omega}}sp(1-p), \tag{3.21}$$

$$\dot{Q}_{A_1} = \frac{\mathrm{d}}{\mathrm{d}t}Q_{A_1} = \frac{r}{\bar{\omega}}(1-p)(Q_{A_2} - Q_{A_1}), \tag{3.22}$$

$$\dot{Q}_{A_2} = \frac{\mathrm{d}}{\mathrm{d}t}Q_{A_2} = \frac{r}{\bar{\omega}}p(1+s)(Q_{A_1} - Q_{A_2}). \tag{3.23}$$

The difference between $\dot{Q}_{A_1}$ and $\dot{Q}_{A_2}$ is given by $\dot{Q}_{A_2} - \dot{Q}_{A_1} = -r(Q_{A_2} - Q_{A_1})$, and integration leads to $Q_{A_2} - Q_{A_1} = (Q^{(0)}_{A_2} - Q^{(0)}_{A_1})e^{-rt}$.

Assume $Q^{(0)}_{A_1} = 0$, i.e., the favourable allele arises on background B_2. Dividing (3.22) by (3.21) and using the formula for $Q_{A_2} - Q_{A_1}$ provides

$$\frac{\mathrm{d}Q_{A_1}}{\mathrm{d}p} = \frac{r}{sp}Q^{(0)}_{A_2}e^{-rt}. \tag{3.24}$$

Assume r to be sufficiently small such that $e^{-rt} \approx 1$ during the selective sweep, i.e., the time when the frequency of A_1 increases from $p^{(0)}$ to 1. Then, we obtain by integration

$$\frac{Q^{\infty}_{A_1}}{Q^{(0)}_{A_2}} = \frac{r}{s}\ln\frac{1}{p^{(0)}}. \tag{3.25}$$

The frequency after fixation of B_1 is given by $Q^{\infty}_{A_1}$ and $Q^{(0)}_{A_2}$ denotes the initial frequency of B_1. Thus, the ratio $Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$ provides the factor by which the frequency of B_1 is reduced due to the hitchhiking effect. Some values of (3.25) and (3.17) are shown in Table 3.3. According to Maynard Smith and Haigh ([29], p. 26), the approximation is good when $Q^{\infty}_{A_1}/Q^{(0)}_{A_2} < 0.1$, but when $Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$ reaches 0.5, it substantially overestimates the exact value of (3.17).

Analogous results are obtained for $Q^{(0)}_{A_1} = 1$, i.e., when the favourable allele arises on background B_1.

3.2.2 The effect on neutral allele frequencies in the diploid case

In the diploid case, we do not derive a closed form for $Q^{(n)}_{A_1} - Q^{(n)}_{A_2}$ and or $Q^{\infty}_{A_1}$. In most cases, however, it is sufficient to use an approximation.

N	s	r	$\frac{r}{s}\log N$	$Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$	N	s	r	$\frac{r}{s}\log N$	$Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$
10^2	0.01	$3\cdot10^{-5}$	0.0138	0.0138	10^2	0.01	$3\cdot10^{-4}$	0.1382	0.1283
10^2	0.1	$3\cdot10^{-4}$	0.0138	0.0142	10^2	0.1	$3\cdot10^{-3}$	0.1382	0.1325
10^6	0.01	10^{-5}	0.0138	0.0138	10^6	0.01	10^{-4}	0.1382	0.1295
10^6	0.1	10^{-4}	0.0138	0.0143	10^6	0.1	10^{-3}	0.1382	0.1344
10^2	0.01	$1.5\cdot10^{-4}$	0.0691	0.0667	10^2	0.01	$1.5\cdot10^{-3}$	0.6908	0.4822
10^2	0.1	$1.5\cdot10^{-3}$	0.0691	0.0689	10^2	0.1	$1.5\cdot10^{-2}$	0.6908	0.4951
10^6	0.01	$5\cdot10^{-5}$	0.0691	0.0670	10^6	0.01	$5\cdot10^{-4}$	0.6908	0.4984
10^6	0.1	$5\cdot10^{-4}$	0.0691	0.0697	10^6	0.1	$5\cdot10^{-3}$	0.6908	0.5130

Table 3.3: The table compares the approximated values $\frac{r}{s}\log\frac{1}{p^{(0)}}$ of (3.25) to the exact values $Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$ of (3.17) for a population of size N. The ratio $Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$ of the frequency after fixation and the initial frequency of B_1 measures the proportional reduction of B_1 in the case where the favourable allele arises on background B_2. ([29], Table 1)

Approximation

To obtain an approximation for $Q^{\infty}_{A_1}$, the equilibrium frequency of B_1, we assume, as in the haploid case, $0 < r \ll s \ll 1$, $p_0 \ll 1$. The frequency of A_1 and the fractions of B_1 on its respective backgrounds, given by (3.13), (3.14) and (3.15), can be approximated by the differential equations

$$\begin{aligned}\dot{p} = \frac{\mathrm{d}}{\mathrm{d}t}p &= \frac{(h+p(1-2h)}{(1+s)p^2+2(1+hs)p(1-p)+(1-p)^2}sp(1-p)\\ &= \frac{(h+p(1-2h)}{1+sp[2h+p(1-2h)]}sp(1-p), \qquad (3.26)\\ \dot{Q}_{A_1} = \frac{\mathrm{d}}{\mathrm{d}t}Q_{A_1} &= \frac{r(1+hs)}{(1+s)p+(1+hs)(1-p)}(1-p)\left(Q_{A_2}-Q_{A_1}\right)\\ &= \frac{r(1+hs)}{1+s\left(h+p(1-h)\right)}(1-p)\left(Q_{A_2}-Q_{A_1}\right), \qquad (3.27)\\ \dot{Q}_{A_2} = \frac{\mathrm{d}}{\mathrm{d}t}Q_{A_2} &= \frac{r(1+hs)}{(1+hs)p+(1-p)}p\left(Q_{A_1}-Q_{A_2}\right)\\ &= \frac{r(1+hs)}{1+shp}p\left(Q_{A_1}-Q_{A_2}\right). \qquad (3.28)\end{aligned}$$

Assume $Q^{(0)}_{A_1} = 0$, i.e., the favourable allele arises on background B_2, and r to be sufficiently small such that $Q_{A_1} \ll Q_{A_2}$ and Q_{A_2} is approximately constant. Dividing (3.27) by (3.26) and using $Q_{A_2} - Q_{A_1} \approx Q^{(0)}_{A_2}$ provides

$$\frac{1}{Q^{(0)}_{A_2}}\frac{\mathrm{d}Q_{A_1}}{\mathrm{d}p} \cong \frac{r(1+hs)}{s}\frac{1+sp(2h+p(1-2h))}{p(h+p(1-2h)(1+s(h+p(1-h)))}. \qquad (3.29)$$

Integration of (3.29) over $[p^{(0)}, 1]$ yields qualitatively different results depending on h.

For example, in the recessive case ($h = 0$) we obtain

$$\frac{Q^{\infty}_{A_1}}{Q^{(0)}_{A_2}} \approx \frac{r}{s}\frac{1}{p^{(0)}}, \tag{3.30}$$

and in the additive case ($h = \frac{1}{2}$) the reduction of B_1 relative to its initial frequency is given by

$$\frac{Q^{\infty}_{A_1}}{Q^{(0)}_{A_2}} \approx \frac{2r}{s}\ln\frac{1}{p^{(0)}}. \tag{3.31}$$

In the dominant case ($h = 1$), we have to integrate over $[p^{(0)}, 1 - p^{(0)}]$, since the integral diverges at $p = 1$, then

$$\frac{Q^{\infty}_{A_1}}{Q^{(0)}_{A_2}} \approx \frac{2r}{s}\ln\frac{1}{p^{(0)}}. \tag{3.32}$$

The initial frequency of A_1 in the diploid case is $p^{(0)} = \frac{1}{2N}$. Note that the approximated values for $Q^{\infty}_{A_1}/Q^{(0)}_{A_2}$ in the haploid case (3.25) and in the diploid case with additive selection (3.31) are effectively the same. The reduction in frequency of the neutral allele, which has not been the initial background of the favourable mutant, is much greater if the favourable allele is fully or partly dominant. Since r increases with physical distance to the selected locus and $r \ll s$, the effect is very local even in the dominant case.

3.2.3 The effect on heterozygosity

Assume a diploid population of size N. Heterozygosity is the probability that two randomly chosen genes are heterozygous. It is one of the most important measures of genetic diversity in a population. Let H_0 denote the initial heterozygosity and H_n the heterozygosity after n generations. In the neutral case, i.e., assuming random drift but no selection and no mutation, Kimura [18] showed that

$$H_n = H_0\ e^{-\frac{n}{2N}}. \tag{3.33}$$

We expect that genetic diversity on a neutral locus is also reduced by selective sweeps. Therefore, we will estimate the decay of heterozygosity in our model in order to be able to compare it to Kimura's result. Consider a diallelic neutral locus B. The frequencies of B_1 and B_2 are given by Q_{A_2} and $1 - Q_{A_2}$. Then, the heterozygosity at the B locus is given by $H =$

$2Q_{A_2}(1-Q_{A_2})$. The initial frequencies of B_1 and B_2 are given by $Q^{(0)}_{A_2}$ and $1-Q^{(0)}_{A_2}$. Hence, the initial heterozygosity is given by

$$H_0 = 2\, Q^{(0)}_{A_2}\left(1-Q^{(0)}_{A_2}\right). \tag{3.34}$$

The frequencies of B_1 and B_2 after fixation of the favourable allele are given by $Q^{\infty}_{A_1}$ and $1-Q^{\infty}_{A_1}$ and the mean heterozygosity after fixation is

$$H_\infty = 2\, Q^{\infty}_{A_1}\left(1-Q^{\infty}_{A_1}\right). \tag{3.35}$$

Thus, the reduction rate of heterozygosity due to the hitchhiking effect is determined by

$$\frac{H_\infty}{H_0} = \frac{Q^{\infty}_{A_1}\left(1-Q^{\infty}_{A_1}\right)}{Q^{(0)}_{A_2}\left(1-Q^{(0)}_{A_2}\right)}. \tag{3.36}$$

To obtain an unconditioned expression for H_∞ we use (3.17) for the equilibrium frequencies of B_1 and B_2. The probability that the mutant A_1 arises on background B_1 is given by $Q^{(0)}_{A_2}$. In this case, we set $Q^{(0)}_{A_1}=1$. The probability that the mutant A_1 arises on background B_2 is given by $1-Q^{(0)}_{A_2}$. In this case, we set $Q^{(0)}_{A_1}=0$. Therefore, combining the two cases, the heterozygosity after fixation can be calculated by

$$\begin{aligned} H_\infty &= Q^{(0)}_{A_2}\big[1-(1-Q^{(0)}_{A_2})\Omega\big]\Big[1-\big[1-(1-Q^{(0)}_{A_2})\Omega\big]\Big] \\ &\quad + \left(1-Q^{(0)}_{A_2}\right)Q^{(0)}_{A_2}\Omega\big[1-Q^{(0)}_{A_2}\Omega\big] \\ &= Q^{(0)}_{A_2}\left(1-Q^{(0)}_{A_2}\right)\Omega\,(2-\Omega) \\ &= H_0\Omega\,(2-\Omega)\,. \end{aligned} \tag{3.37}$$

with Ω as in (3.18).

Approximation

Maynard Smith and Haigh [29] proceeded with an approximation of the reduction rate $\frac{H_\infty}{H_0}$ of heterozygosity on the neutral B locus conditioned on $Q^{(0)}_{A_1}=0$, i.e., the favourable allele arises on background B_2. Assume additive fitness and $Q^{\infty}_{A_1} \ll Q^{(0)}_{A_2}$, such that the haploid case and the diploid case are effectively the same. Then, we use (3.31) with $p^{(0)}=\frac{1}{2N}$ to approximate the reduction rate of heterozygosity by

$$\frac{H_\infty}{H_0} \approx \frac{2r}{s(1-Q^{(0)}_{A_2})}\log\frac{1}{p^{(0)}}. \tag{3.38}$$

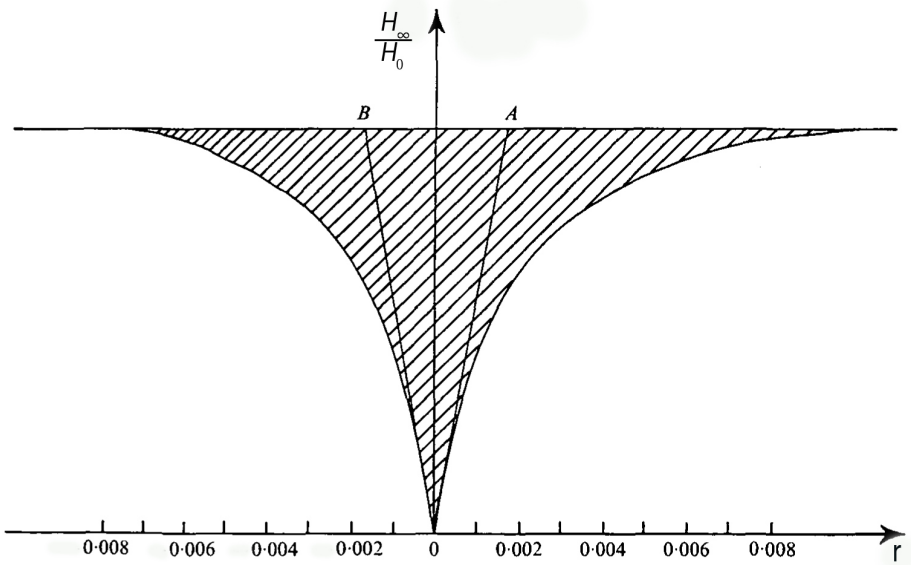

Figure 3.1: The typical pattern caused by the hitchhiking effect for $Q_{A_2}^{(0)} = 0.5$, $N = 10^6$, $s = 0.01$. The exact values of the reduction rate H_∞/H_0 of heterozygosity according to (3.36) are plotted as a function of the recombination rate r. The approximation in (3.38) is represented by the lines OA and OB.

The approximation in (3.38) is a homogeneous linear function of the recombination rate r with slope

$$\beta = \frac{2}{s\left(1 - Q_{A_2}^{(0)}\right)} \log \frac{1}{p^{(0)}}. \tag{3.39}$$

Figure 3.1 shows the exact values of the reduction rate $\frac{H_\infty}{H_0}$ of heterozygosity according to (3.17) as a function of r. The approximation in (3.38) is represented by the lines $0A$ and $0B$. The smaller the recombination rate r, the smaller the value of $\frac{H_\infty}{H_0}$, the greater is the reduction of heterozygosity on the neutral locus.

The probability of a recombination event between the selected locus and a neutral locus increases with the distance between them. Thus, the probability of escaping the hitchhiking effect also increases with the distance. Figure 3.1 shows that neutral loci near the selected locus have a greater chance to suffer a greater loss of heterozygosity and next to the selected locus heterozygosity is annihilated. The total loss of heterozygosity is given by the area of the hatched region.

3.2.4 The equivalent region made homozygous

Maynard Smith and Haigh ([29], pp. 28-29) argue that the region where heterozygosity is reduced can be translated into an equivalent region, where the loci reach complete homozygosity during the fixation process of A_1.

The total loss of heterozygosity is given by the area of the hatched region. In order to translate this region where heterozygosity gets lost into an equivalent region where heterozygosity gets completely lost, we model a coextensive rectangle, shown as the light grey shaded area in Figure 3.2. The side lengths of this rectangle are 1 and L, where L denotes the physical distance on the chromosome within which the loci will be homozygous after fixation of the favourable allele (measured in percent of the recombination rate). We use the approximation of the reduction rate of heterozygosity of (3.2.3) to calculate the size of L.

The area of the hatched region in Figure 3.2 is approximated by the area of the triangle $A0B$ (shown as the dark shaded region). Since the slope of $0A$ is given by β and the height of the triangle is equal to one, the area of the dark shaded region is equal to $\frac{1}{\beta}$ (see Figure 3.2). We define α as the ratio of the hatched and the dark shaded area, i.e., $\alpha = \frac{A_{A0B}}{A_{\text{hatched area}}}$. Using (3.39), L can be calculated by

$$L \approx \alpha \, \frac{1}{\beta} \, 100 \approx \frac{50\alpha s \left(1 - Q_{A_2}^{(0)}\right)}{\log \frac{1}{p^{(0)}}}. \tag{3.40}$$

According to [29], simulations for different values of $Q_{A_2}^{(0)}$, s and $p^{(0)}$ show that it is reasonable to choose $\alpha = 2$.

The expectation of L can be calculated by integrating over all possible initial frequencies x of B_2 $\left(x = 1 - Q_{A_2}^{(0)}\right)$, which is the probability that A_1 arises on background B_2. Assume a uniform distribution of x over $[0, 1]$ and choose $\alpha = 2$. The corresponding value of L is given by

$$L = \frac{100sx}{\log \frac{1}{p^{(0)}}}.$$

and the expectation of the physical distance within which the loci reach homozygosity during the fixation of A_1 ([29], p. 29) is

$$\mathrm{E}(L) = 2 \int_0^1 \frac{100sx}{\log \frac{1}{p^{(0)}}} \, x \, \mathrm{d}x$$

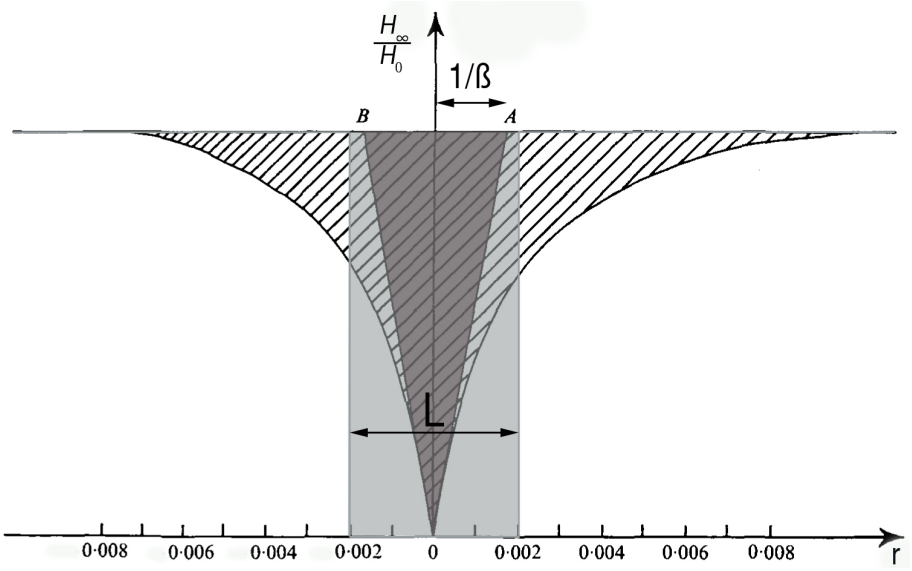

Figure 3.2: The approximated physical distance $1/\beta$ (see (3.39)), within which the chromosome will be completely homozygous after the selective sweep, is given by the area of the dark shaded region. The real value of this physical distance L (see (3.40)) equals the area of the light grey shaded region, which is equivalent to the area of the hatched region.

$$= \frac{200s}{3 \log \frac{1}{p^{(0)}}}. \tag{3.41}$$

Note that we will achieve the maximum effect if selection is strong, i.e., the selective advantage s is very large. The expectation of L is also influenced by the population size N. Increasing N, increases $\frac{1}{p^{(0)}}$, which increases $\log \frac{1}{p^{(0)}}$ and, therefore, decreases $\mathrm{E}(L)$, but the impact of N is rather small compared to the influence of s.

3.2.5 The probability of complete fixation

Now we assume that selection is also acting on the B locus, i.e., it is not neutral anymore. However, the selective forces that maintain the polymorphism on the B locus are small compared to those acting on the A locus. In this way, during a selective sweep the frequencies of the B locus will be affected like the frequencies of a neutral locus, but in contrast to neutral loci, a long-lasting effect can only be achieved if one of the alleles on the B locus gets completely fixed. Otherwise, the selective forces acting on the B locus will turn the frequencies to an equilibrium value as soon as the fixation of

genotypes	A_1A_1	A_1A_2	A_2A_2	
fitness	$1+s$	$1+\frac{s}{2}$	1	
gametes	A_1B_1	A_1B_2	A_2B_1	A_2B_2
frequencies	0	p	$(1-p)Q_{A_2}$	$(1-p)(1-Q_{A_2})$

Table 3.4: The fitnesses of the three possible genotypes and the frequencies of the four possible gametes in the diploid case, when the favourable allele A_1 has arises on background B_2 exclusively.

the favourable allele on the A locus has been completed.

We restrict our study to the case where the favourable mutant A_1 arises on background B_2. The neutral allele B_2 will get a lift in frequency and we ask for the probability that B_2 will get fixed in the population, i.e., the probability that B_1 extincts.

Consider a diploid population of size N and assume additive fitness. Table 3.4 shows the fitnesses determined by the A locus and the initial frequencies of the four possible genotypes. Note that the favourable allele A_1 has not yet arisen on the other background B_1.

Assume that recombination is sufficiently small such that we can ignore terms of order $O(rs)$. Then, the probability that an A_1B_1 zygote arises in the next generation is given by $rp(1-p)Q_{A_2}$. Let x denote the probability that the new zygote will get fixed in the population, i.e., the probability that it will not extinct. The magnitude of x depends on the offspring distribution we choose and particularly on the selective advantage $\bar{s}$ of the new zygote A_1B_1. If we choose a Poisson offspring distribution and $\bar{s}$ small, the probability of fixation is given by $x \approx 2\bar{s}$. This approximation often overestimates the real value and thus, to provide flexibility, we choose $x = k\bar{s}$ for an appropriate constant k. Using (3.9) and (3.11) with $h = \frac{1}{2}$, the selective advantage $\bar{s}$ of the zygote A_1B_1 can be calculated by

$$\begin{aligned}\bar{s} &= \frac{w_{11}-\bar{w}}{\bar{w}} \\ &= \frac{(1+s)p + (1+\frac{1}{2}s)(1-p)}{(1+s)p^2 + 2(1+\frac{1}{2}s)p(1-p) + (1-p)^2} - 1 \\ &\approx \frac{s(1-p)}{2(1+sp)}. \end{aligned} \tag{3.42}$$

Thus, the probability g that a randomly chosen gamete in the next generation

is of allelic type A_1B_1 and will be fixed in the population is given by

$$\begin{aligned} g &= rp(1-p)Q_{A_2}kx \\ &\approx \frac{krsQ_{A_2}p(1-p)^2}{2(1+sp)}. \end{aligned} \tag{3.43}$$

Since we consider $2N$ gametes in the whole population, the probability that in one generation there is no gamete with allelic type A_1B_1, which eventually gets fixed in the population, arising is $(1-g)^{2N} \approx e^{-2Ng}$. Hence, the probability P_0 that A_1B_1 never arises or gets fixed, i.e., the probability that B_1 extincts and B_2 gets fixed, is

$$P_0 \approx e^{-2N\sum g}, \tag{3.44}$$

where $\sum g$ denotes the summation of the values of (3.43) over all generations during which the selective substitution takes place. The values of g depend on the frequency p, which changes with time during the process.

To approximate P_0, we assume r to be constant and let T denote the point in time when A_1 has fixed. Then, (3.44) can be expressed by

$$\log P_0 \approx -sNkrQ_{A_2} \int_0^T \frac{p(t)(1-p(t))^2}{1+sp(t)}\ \mathrm{d}t. \tag{3.45}$$

Adapting (3.21) to the diploid model by replacing s by $\frac{s}{2}$, we derive $\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{sp(1-p)}{2+sp}$. Then, substitution leads to

$$\log P_0 \approx -NkrQ_{A_2} \int_{p^{(0)}}^1 (1-p)\frac{2+sp}{1+sp}\ \mathrm{d}p. \tag{3.46}$$

The integral in (3.46) is approximately one if s and $p^{(0)}$ are small. Thus, we find

$$P_0 \cong e^{-kNrQ_{A_2}} \tag{3.47}$$

for the probability of complete fixation of B_2.

The value of k will usually be chosen between one and two and $Q_{A_2} \in [0,1]$. Thus, there will be reasonable high chance of fixation if we choose $r \leq \frac{1}{N}$.

Note that (3.47) does not depend on the selective advantage s of the favourable mutant. Maynard Smith and Haigh ([29], p. 33) argued that if s is large the mutant will fix very fast; this implies that the time within which A_1 could arise on the other background is very short and thus, the probability that A_1B_1 arises is low and the chance of fixation of B_2 quite high. On the other hand, if A_1B_1 arises, there is a good chance it will be fixed, i.e., B_2 will extinct. These forces annul each other.

Since the derivation of (3.47) includes many approximations and assumptions, its adequacy has been approved by simulations (see [29], pp. 33-34).

3.3 Conclusion

Let ν denote the neutral mutation rate per gene per generation. In the neutral theory, the heterozygosity H at stationarity is given by

$$H = \frac{4N\nu}{1 + 4N\nu}. \tag{3.48}$$

It is sensitive to changes of the population size N, i.e., for values of H between 0.1 and 0.5 the population size has to take values between 0.028ν and 0.25ν ([29], p. 34).

In contrast, species of a large population size do not show the expected high level of heterozygosity. Furthermore, the heterozygosity between various species turns out to be surprisingly constant (see, for example, [27]). Maynard Smith and Haigh [29] tried to explain the observed uniformity of H between species by introducing another process that regulates the amount of allelic polymorphisms apart from random genetic drift.

In fact, the deterministic model of Maynard Smith and Haigh [29] partially confirmed their contention that the effect of genetic hitchhiking reduces the expected heterozygosity on linked neutral loci. At least for populations of size equal or greater than 10^6, the hitchhiking effect seems to be more important for the regulation of heterozygosity than random genetic drift. A comparison of the expected half-life of a neutral polymorphism towards fixation by genetic drift, which is $2N \ln 2$ generations ([18], cf. (3.33)), i.e., of order $O(N)$, versus $\frac{3m \ln 2\left(\ln(1/p^{(0)})\right.}{200s \ln(1+d)}$ by hitchhiking ([29], Equation (32)), where m denotes the total length of the genome and $1 + d$ is the mean number of offspring of the fittest genotype, shows that altering the population size from 10^4 to 10^8 alters the expected half-life by a factor 10^4 due to drift and in contrast only by a factor about 4 due to hitchhiking.

Hitchhiking causes a loss of heterozygosity at neighbouring linked loci (3.2.3). The reduction rate depends mainly on the factor $\frac{r}{s}$. According to (3.41), the greatest effect will be achieved if selection is strong, i.e., if the selective advantage s of the favourable mutation is large.

In Section 3.2.5, we showed that the hitchhiking effect on selectively maintained polymorphisms is very local. Indeed, we have to choose $r \leq \frac{1}{N}$ to obtain a reasonably high chance of fixation on a linked locus. Therefore, only loci within a physical distance of order $\frac{1}{N}$ to the selected locus will reach homozygosity.

The deterministic model presented by Maynard Smith and Haigh in 1974 [29] may exhibit some weaknesses (cf. 3.1.3), but considering that they could not rely on the large amount of sequence data which is available today, it has been a great achievement to provide a model that offers an explanation for the unexpected uniformity of heterozygosity that we observe between different species and the surprisingly low level of heterozygosity in abundant populations.

More recent models remedy the weaknesses of the model of Maynard Smith and Haigh [29] regarding the omission of stochastic effects caused by the finite population size.

Chapter 4

A stochastic model

In the previous chapter we described the effect of a strongly selected mutation on linked neutral loci while it is on its way to fixation. The weaknesses of this first model are based on the omission of stochastic effects caused by the finite population size. Furthermore, the distribution of segregating sites is determined by the population dynamics since the MRCA of the sample and therefore, the effects of all selective sweeps that occurred in the past have to be considered. In 1989, Kaplan *et al.* [17] presented a stochastic model that considers the effect of recurring selected substitutions at different times in the past on the number of segregating sites using coalescent theory. First, we start by describing the effect of a single linked ancestral substitution and second, the effect of recurring selected substitutions on neutral linked loci and the number of segregating sites.

4.1 The effect of a single linked ancestral substitution

Consider a sample of n genes from a randomly mating diploid population of size N. The region of investigation consists of L completely linked neutral sites and a single selected site. Since complete linkage is assumed, it is appropriate to consider the L neutral sites as a single locus. Therefore, our model simplifies to a two-locus model. We call the selected locus A, and A_1 denotes the favourable mutation and A_2 the deleterious wild type. The neutral locus is called B. Since we want to apply the coalescent framework developed in 2.1, we refer to the current generation as generation 0 and to the population t generations back in time as the ancestral generation t. According to 2.1.3, the frequency of A_1 in the ancestral generation t is given by $X(t)$,

genotypes	A_1A_1	A_1A_2	A_2A_2
fitnesses	$1+2s$	$1+s$	1

Table 4.1: The fitnesses of the three possible genotypes.

$t > 0$. The fitnesses of the three possible genotypes A_1A_1, A_1A_2 and A_2A_2 are $w_{11} = 1 + 2s$, $w_{12} = 1 + s$ and $w_{22} = 1$ (see Table 4.1), i.e., selection is additive (cf. 3.1.2). Assume that the strongly selected mutation A_1 arises in a single copy in the $2N\tau_a$ ancestral generation and fixes in the $2N\tau_f$ ancestral generation ($\tau_f < \tau_a$). As customary in the coalescent framework, from now on time will be measured in units of $2N$ generations, i.e., $2N\tau = t$. We refer to the time interval $[\tau_f, \tau_a]$ in which the selective sweep occurs, as the selective phase.

4.1.1 The model

Define $\alpha = 2Ns$, where s denotes the selective advantage of the favourable allele A_1. Assume N to be sufficiently large and consider strong selection, which is achieved if α is about $10^3 \leq \alpha \leq 2N \cdot 10^{-2}$. In this case, it is well known that the dynamics of the frequency process X can be modelled deterministically as long as it does not reach a value close to the boundary values 0 and 1 ([26], [34]). Near the boundaries, i.e., if the frequency of A_1 is either very low or very high, the deterministic approach is not appropriate for the reasons discussed in 3.1.3. Therefore, the process will be modelled stochastically near the boundaries and deterministically otherwise. The deterministic part can be approximated by the solution of a differential equation which is defined in (4.3).

Consider $\varepsilon > 0$ small. Therefore, the representation of the X process is given by

$$X(\tau) = \begin{cases} X(\tau) & \text{if } 0 \leq \tau \leq \tau_{1-\varepsilon} \\ x(\tau) & \text{if } \tau_{1-\varepsilon} \leq \tau \leq \tau_\varepsilon \\ X(\tau) & \text{if } \tau > \tau_\varepsilon, \end{cases} \tag{4.1}$$

where

$$\tau_{1-\varepsilon} = \inf\{\tau : X(\tau) = 1 - \varepsilon\}, \qquad \tau_\varepsilon = x^{-1}(\varepsilon), \tag{4.2}$$

and $x(\tau)$ is determined by the differential equation

$$\frac{\mathrm{d}x(\tau)}{\mathrm{d}\tau} = -\alpha x(\tau)\left(1 - x(\tau)\right), \tag{4.3}$$

$$x\left(\tau_{1-\varepsilon}\right) = 1-\varepsilon, \qquad \tau \geq \tau_{1-\varepsilon},$$

which is a classical continuous time selection model, that describes the change in frequency of an allele on a selected diallelic locus in the case of no dominance in a diploid population.

It is convenient to approximate the frequency process X by a branching process in its early stages. Hence, the probability that the newly arising mutant A_1 at the selected locus goes extinct is equal to $1-2s$. Therefore, choose ε such that the probability of extinction at time τ_ε, where $x(\tau_\varepsilon) = \varepsilon$ and thus, there are $2N\varepsilon$ copies of A_1 in the population, is approximately zero, i.e., $(1-2s)^{2N\varepsilon} \approx e^{-4Ns\varepsilon} = e^{-2\alpha\varepsilon} \approx 0$. This is valid for, e.g., $2\alpha\varepsilon = 10$, hence $\varepsilon = \frac{5}{\alpha}$. Thus, if α is very large, ε can be very small.

4.1.2 The conditional fixation time

Define $\eta = \tau_a - \tau_f$, the duration of the selective sweep, and let $\mathrm{E}^*(\eta)$ denote the expected time to fixation of the favourable mutation, conditioned on its fixation. The conditioned fixation time can be calculated using the diffusion estimate ([9], eq.(5.52)), which is

$$\mathrm{E}^*(\eta) = \int_{\frac{1}{2N}}^{1} \frac{2\left(e^{2\alpha x}-1\right)\left(e^{2\alpha(1-x)}-1\right)}{2\alpha x(1-x)(e^{2\alpha}-1)}\,\mathrm{d}x. \tag{4.4}$$

This estimate requires $s = \frac{\alpha}{2N}$ to be small, which might be a reason to suspect its adequacy. Therefore, Kaplan *et al.* [17] run a couple of simulations and compared their estimated value $\mathrm{E}^*(\eta)$ with the results of (4.4). Both values agreed in all cases.

The simulations for the approximated conditional expectation $\mathrm{E}^*(\eta)$ have been accomplished in the following way. In the beginning, i.e., for $X(t) \in [0,\varepsilon]$, the process has been simulated by a branching process with a Poisson offspring distribution with mean $1+s$. In the intermediate deterministic phase, the time the process stays in $[\varepsilon, 1-\varepsilon]$ has been calculated by

$$\int_{\varepsilon}^{1-\varepsilon} \frac{\mathrm{d}x}{\alpha x(1-x)} = \frac{-2\ln\varepsilon}{\alpha}. \tag{4.5}$$

Finally, the last phase of the process, i.e., $X(\tau) \in [1-\varepsilon, 1]$, has been simulated by a branching process with a Poisson offspring distribution with mean

$1 - s$. Thus, $E^*(\eta)$ is the sum of the three sojourn times in the different phases of the X process.

It follows from these calculations that, if α is large, the value of $\mathrm{E}^*(\eta)$ is negligible, i.e., for strong selection the selective sweep occurs virtually instantaneous. For example, increasing α from 10^3 to 10^6 leads to a decrease of $\mathrm{E}^*(\eta)$ from $1.6 \cdot 10^{-2}$ to $3.0 \cdot 10^{-5}$.

4.1.3 The coalescent of a single isolated neutral locus

For what follows, we need to distinguish between a single neutral locus without linkage and a single neutral locus linked to a selected one. We will refer to a locus that is not linked to another locus as single isolated neutral locus. In this case, the properties of T, the total length of the ancestral tree of a neutral locus, and S, the number of segregating sites have been discussed in 2.1.1. Since the neutral locus consists of L nucleotide sites the formulas in (2.15) and (2.17) have to be adapted accordingly.

Let υ denote the expected number of selectively neutral mutations per nucleotide site per gene per unit of scaled time. Assume υ to be sufficiently small, such that each mutation causes a new segregating site like in the infinite sites model. The length of the ancestral tree T (of a single nucleotide site) in formula (2.17) has to be substituted by the sum over the ancestral tree sizes of each of the L neutral sites. Since they are completely linked, all trees are equal and the sum is given by LT. In order to apply (2.15), we write υL instead of μ, the expected number of selectively neutral mutations per gene per unit of scaled time. Thus, the distribution of S is given by

$$P\left[S = j\right] = \int_0^\infty e^{-\upsilon\tau} \frac{(\upsilon\tau)^j}{j!} \, \mathrm{d}F(\tau), \tag{4.6}$$

where

$$F(\tau) = P\left[LT \leq \tau\right], \quad \tau \geq 0.$$

Since the $\{T(k)\}$ are independent random variables and approximately exponentially distributed with parameter $\frac{k(k-1)}{2}$, we use

$$T = \sum_{k=2}^{n} kT(k) \qquad \text{and} \qquad T_{MRCA} = \sum_{k=2}^{n} T(k),$$

to calculate T and the time to the MRCA. The expectation and the variance of S follow directly from the distribution of T and can be calculated by

$$\mathrm{E}[S] = L\upsilon\,\mathrm{E}[T] \tag{4.7a}$$
$$\mathrm{Var}[S] = L\upsilon\,\mathrm{E}[T] + (L\upsilon)^2\,\mathrm{Var}[T], \tag{4.7b}$$

4.1.4 The coalescent of a single linked neutral locus

In the case where a single neutral locus is linked to a selected locus, we will refer to it as linked single neutral locus. The coalescent of a single linked neutral locus has been described in 2.1.3. We defined $Q(t) = (i, j)$ if the ancestral generation t consists of i ancestral genes that are linked to A_1 and j ancestral genes that are linked to A_2. Since A_1 has already been fixed at the time of sampling, it is always the case that $Q(0) = (n, 0)$. The favourable allele arises due to a single mutation at the selected locus and is supposed to get fixed in the population. Hence, we do not consider further mutations to occur at the selected locus and set β_1 and β_2 equal to zero. Note that the Q process does not take account of neutral mutations.

Consider time to be measured in units of $2N$ generations. Let R denote the expected number of crossovers between the neutral locus and the selected locus per unit of scaled time. Thus, in the limit $N \to \infty$, the distribution of the Q process conditioned on the ancestral frequency process in scaled time, $\{X(\tau), \tau > 0\}$, can be approximated by a time inhomogeneous Markov jump process (cf. (2.29) and (2.33)) given by

$$P[Q(\tau + \Delta) = Q(\tau)] = 1 - h_{ij}(X(\tau))\,\Delta + O\left(\Delta^2\right) \tag{4.8a}$$

as $\Delta \to 0$, where

$$h_{ij}(x) = \left(\frac{\binom{i}{2}}{x} + iR(1-x)\right)\chi(x > 0) + \left(\frac{\binom{j}{2}}{1-x} + jRx\right)\chi(x < 1). \tag{4.8b}$$

We set $\chi(x > 0)$ equal to 1 if $x > 0$ and equal to 0 otherwise and $\chi(x < 1)$ is defined in a similar way. The conditional jump probabilities are

$$\begin{aligned}
&P[Q(\tau+\Delta) = (i-1, j)|Q(\tau) = (i, j)] = q_{i-1,j}(X(\tau)) + O(\Delta),\\
&P[Q(\tau+\Delta) = (i, j-1)|Q(\tau) = (i, j)] = q_{i,j-1}(X(\tau)) + O(\Delta),\\
&P[Q(\tau+\Delta) = (i+1, j-1)|Q(\tau) = (i, j)] = q_{i+1,j-1}(X(\tau)) + O(\Delta),\\
&P[Q(\tau+\Delta) = (i-1, j+1)|Q(\tau) = (i, j)] = q_{i-1,j+1}(X(\tau)) + O(\Delta),
\end{aligned} \tag{4.8c}$$

where

$$q_{i-1,j}(x) = \binom{i}{2}\frac{\chi(x>0)}{xh_{ij}(x)}, \quad q_{i,j-1}(x) = \binom{j}{2}\frac{\chi(x<1)}{(1-x)h_{ij}(x)},$$
$$q_{i+1,j-1}(x) = \frac{jRx\ \chi(x<1)}{h_{ij}(x)}, \quad q_{i-1,j+1}(x) = \frac{iR(1-x)\ \chi(x>0)}{h_{ij}(x)}. \tag{4.8d}$$

Since the selective mutation A_1 arises at the time τ_a, we write

$$Q(\tau_a+) = (0,j) \qquad \text{if} \qquad Q(\tau_a) = (1,j-1), \qquad j \geq 1, \tag{4.9}$$

where τ_a+ denotes any time after exiting the selective phase (backward in time).

4.1.5 The size of the ancestral tree

The Q process occurs in three different phases. First, consider the dynamics of the Q process starting in generation 0 and going backward in time. Since the favourable allele A_1 fixes in the population at time τ_f, there is no selection causing polymorphism in the time interval from zero to τ_f. Therefore, in this first phase the neutral locus is not influenced by selection on the linked locus and the coalescent behaves like the coalescent of an isolated neutral locus, which we described in 4.1.3. If the number of coalescence events in this first phase is $n-k$, then the selective phase is entered with k ancestral genes, i.e., $Q(\tau_f) = (k,0)$. In this second phase, the Q process behaves like the coalescent of a single linked neutral locus for a sample of size k, which we described in 4.1.4. The dynamics of the coalescent get accelerated due to the selected substitution, depending on the magnitude of α and the recombination rate R. Assume that we exit the selective phase with j ancestral genes, i.e., $Q(\tau_a+) = (0,j)$ where $1 \leq j \leq k$. Then, in the third phase the Q process behaves again like the coalescent of a single isolated neutral locus for a sample of size j until the MRCA occurs. Thus the dynamics of the Q process are only of interest during the selective phase.

If α is sufficiently large, the selective phase occurs virtually instantaneous and with high probability the time of the selective sweep η has negligible effect on the total length of the ancestral tree (cf. 4.1.2). Furthermore, since the number of neutral mutations is proportional to the lengths of the ancestral tree, the number of neutral mutations that occur during the selective phase is also negligible with regard to the number of segregating sites. Thus, the only important quantities in the Q process that we have to consider are

the probabilities of exiting the selective phase with j ancestral genes, if we entered it with k ancestral genes. We denote these probabilities by

$$P_{kj} = P\left[Q(\tau_a+) = (0,j) | Q(\tau_f) = (k,0)\right] \qquad 1 \leq j \leq k \leq n. \tag{4.10}$$

Note that these probabilities are not conditioned on the X process and, therefore, they have to be calculated by averaging over all possible realizations of it.

Let Γ denote the total length of the ancestral tree of a single linked neutral locus for a sample of size n, assuming that only one selected substitution occurs until the MRCA. As mentioned above, except for the selective phase the coalescent of a single linked neutral locus behaves like the coalescent of a single isolated neutral locus (cf. 4.1.3). Therefore, the distribution of the parts of the ancestral tree that lie outside the selective phase can be calculated in the same way as in 4.1.3. In accordance with 4.1.3, the size of these parts will be denoted by T.

In order to calculate the expectation of Γ, we have to distinguish between the different possible realizations of the ancestral tree. On the one hand, going backward in time, the MRCA of the sample could occur before entering the selective phase, i.e., $T_{MRCA} \leq \tau_f$. In this case, we need to consider all realizations of T for a sample of n that fulfill the condition that the MRCA arises before τ_f. On the other hand, assuming that the MRCA occurs during or after the selective phase, we have to split Γ into two parts, first the size of the part of the tree before the selective phase and second, the size of the tree after the selective phase. The first part is equal to the length of an ancestral tree of an isolated neutral locus for a sample of size n for the first τ_f units of time, given that $n-k$ coalescence events occur until τ_f. The second part is equal to size of T for a sample of size j, given that we exit the selective phase with j ancestral genes. In order to calculate the expectation of Γ we have to sum up over all possible values of k and j.
Let $T(\tau)$ denote the total length of the ancestral tree for the first τ units of time. Then, the expectation of Γ is given by

$$\begin{aligned} \mathrm{E}[\Gamma] = {} & \mathrm{E}[T | T_{MRCA} \leq \tau_f,\ Q(0) = (n,0)] \\ & + \sum_{k=2}^{n} \sum_{j=1}^{k} \left(J_{nk}(\tau_f) + L_{nk}(\tau_f) P_{kj} \sum_{i=1}^{j-1} \frac{2}{i} \right), \end{aligned} \tag{4.11}$$

where

$$J_{nk}(\tau) = \mathrm{E}[T(\tau)\chi(Q(\tau) = (k,0)) | Q(0) = (n,0)], \tag{4.12}$$

$$L_{nk}(\tau) = P[Q(\tau) = (k,0)|Q(0) = (n,0)], \qquad \tau \geq 0, \qquad 2 \leq k \leq n. \tag{4.13}$$

The above formulas can be explained as follows. In the case where the MRCA arises during or after the selective phase, the expected lengths of the ancestral tree for the first τ_f units of time for a fixed sample size n and under the condition that $n-k$ coalescence events occur during this time, is given by $J_{nk}(\tau_f)$. The probability of entering the selective phase with k ancestral genes and exiting it with j ancestral genes is given by $L_{nk}(\tau_f)P_{kj}$. Note that $\sum_{i=1}^{j-1} \frac{2}{i}$ is the expected size of the ancestral tree for a sample of size j (cf. (2.7a)). Since η is assumed to be negligible, we treat $\tau_f \approx \tau_a$ as a constant.

It can be shown (see [12]) that the $\{J_{nk}(\tau), \tau \geq 0\}$ and $\{L_{nk}(\tau), \tau \geq 0\}$ are the solutions of the system of differential equations

$$\begin{aligned}
\frac{\mathrm{d}J_{nk}(\tau)}{\mathrm{d}\tau} &= -\binom{k}{2} J_{nk}(\tau) + \binom{k+1}{2} J_{nk+1}(\tau) + kL_{nk}(\tau), \\
\frac{\mathrm{d}L_{nk}(\tau)}{\mathrm{d}\tau} &= -\binom{k}{2} L_{nk}(\tau) + \binom{k+1}{2} L_{nk+1}(\tau), \qquad \tau \geq 0, \qquad 2 \leq k \leq n-1, \\
J_{nn}(\tau) &= n\tau e^{-\binom{n}{2}\tau}, \qquad L_{nn}(\tau) = e^{-\binom{n}{2}}, \qquad \tau \geq 0,
\end{aligned} \tag{4.14}$$

with initial values

$$J_{nk}(0) = 0 \qquad \text{for } 2 \leq k \leq n, \qquad L_{nk}(0) = 0 \qquad \text{for } 1 \leq k \leq n-1, \qquad L_{nn}(0) = 1.$$

The system has been solved explicitly in [12]. The complexity of the solution, however, suggests a numerical treatment for moderate sample sizes.

Note that $J_{n2}(\tau)$ denotes the expected length of an ancestral tree for the first τ units of time, given a sample of size n and $n-2$ coalescence events until τ. Since at time τ there are only two lineages left, the MRCA arises with the next coalescence event. For the expectation of T in the first part of (4.11) we assumed that the MRCA occurs before the selective phase, i.e., $T_{MRCA} \leq \tau_f$. Since we consider η to be of negligible effect on T, we could also include the case where we enter the selective phase with 2 lineages, i.e., $Q(\tau_f) = 2$, and the last coalescence event occurs during the selective phase. Therefore, if we assume that at least one coalescence event occurs during the

selective phase, i.e., $P_{kk} \neq 0$, we obtain

$$\mathrm{E}[T|T_{MRCA} \leq \tau_f,\ Q(0) = (n,0)] = \int_0^{\tau_f} J_{n2}(u)\ \mathrm{d}u. \tag{4.15}$$

The above assumption that at least one coalescence event occurs during the selective phase is reasonable, because otherwise the coalescent of the linked locus would not change substantially compared to the coalescent of an isolated locus.

The interpretation of P_{kj}

To emphasize the dependence of the probabilities P_{kj} on the recombination rate R we write $P_{kj}(R)$ instead of P_{kj}. At the end of the selective phase, all genes of the sample carry the favourable allele A_1. There are two possibilities how a gene reaches the allelic state A_1 at the selected locus. First, the gene is a descendant of an ancestor carrying A_1. Second, the allelic state of the selected locus changes due to recombination.
If R goes to zero, there is only a vanishing chance that a gene receives an A_1 allele due to recombination. Thus, all genes have to inherit the A_1 allele from an ancestor carrying A_1, i.e., due to coalescence events. Since the allele A_1 arises in a single copy only and there is no recombination, coalescence events have to take place if the selective phase is entered by at least two lineages, i.e., if $k \geq 2$. Therefore,

$$P_{kk}(R) \to 0 \qquad \text{if } R \to 0. \tag{4.16}$$

Thus, the hitchhiking effect is greater if the recombination rate R is small.
On the contrary, if R goes to infinity, the linkage between the neutral locus and the selected locus will be mostly eliminated and thus, the Q process behaves like in the coalescent of a single isolated neutral locus for a sample of size k (cf. (2.4)). Hence,

$$P_{kk}(R) \to E\left[e^{-\binom{k}{2}\eta}\right] \qquad \text{if } R \to \infty. \tag{4.17}$$

Thus, if the recombination rate R is high the effect of hitchhiking depends on η and therefore on $\alpha = 2Ns$ (cf. 4.1.2). In this case, if α is large, we expect the fixation time η to be very small and thus, $P_{kk}(R) \approx \mathrm{E}\left[e^{-\binom{k}{2}\eta}\right] \approx 1$, i.e., there is no hitchhiking effect.
Therefore, the probabilities P_{kk} can be understood as the probabilities of escaping the hitchhiking effect.

The determination of P_{kj}

Let $P_{k,i,j}(\tau)$ denote the probability that $Q(\tau) = (i,j)$ for $\tau \in [\tau_f, \tau_a]$, given there are k lineages left at time τ_f, i.e.,

$$\begin{aligned} P_{k,i,j}(\tau) =& P[Q(\tau) = (i,j) | Q(\tau_f) = (k,0)] \\ & \tau_f \leq \tau \leq \tau_a,\ 1 \leq i+j \leq k,\ k \geq 2. \end{aligned} \tag{4.18}$$

These probabilities describe the distribution of Q during the selective phase conditioned on the number of ancestral genes k with which the Q process enters the selective phase at time τ_f. Above we defined

$$P_{kj} = P\left[Q(\tau_a+) = (0,j) | Q(\tau_f) = (k,0)\right] \qquad 1 \leq j \leq k \leq n,$$

and

$$Q(\tau_a+) = (0,j) \qquad \text{if} \qquad Q(\tau_a) = (1, j-1), \qquad j \geq 1. \tag{4.19}$$

Thus, P_{kj} can be written as

$$\begin{aligned} P_{kj} &= P\big[Q(\tau_a) = (0,j) | Q(\tau_f) = (k,0)\big] + P\big[Q(\tau_a) = (1,j-1) | Q(\tau_f) = (k,0)\big] \\ &= P_{k,0,j}(\tau_a) + P_{k,1,j-1}(\tau_a), \end{aligned}$$

which is the probability that the favourable allele A_1 arises on a gene in the population which has not been sampled plus the probability that it arises on a gene in the sample.

It follows from the Markov structure of the Q process that, if $\varepsilon > 0$ is sufficiently small, it is very unlikely that it changes its state between τ_f and $\tau_{1-\varepsilon}$. Therefore,

$$P[Q(\tau_{1-\varepsilon}) = (k,0) | Q(\tau_f) = (k,0)] \approx 1,$$

and the $P_{k,i,j}$ can be conditioned on the number of ancestral genes k at time $\tau_{1-\varepsilon}$ instead, i.e.,

$$P_{k,i,j}(\tau) \approx P[Q(\tau) = (i,j) | Q(\tau_{1-\varepsilon}) = (k,0)] \qquad \text{for } \tau \geq \tau_{1-\varepsilon}.$$

On the time interval $[\tau_{1-\varepsilon}, \tau_\varepsilon]$ the X process has been modelled deterministically (see 4.1) and the Q process is a time inhomogeneous Markov process given by (4.8c). Therefore, we can use the associated Kolmogorov forward

differential equations to calculate the $P_{k,i,j}$ on the interval $[\tau_{1-\varepsilon}, \tau_\varepsilon]$. Using (4.8c) and the solution of (4.3), we derive

$$\begin{aligned}
P_{k,i,j}(\tau+\Delta) = {} & P\left[Q(\tau+\Delta)=Q(\tau)\right] P_{k,i,j}(\tau) \\
& + h_{ij}\left(x(\tau)\right)\Delta\left\{P\left[Q(\tau+\Delta)=(i,j)|Q(\tau)=(i+1,j)\right] P_{k,i+1,j}(\tau)\right. \\
& + P\left[Q(\tau+\Delta)=(i,j)|Q(\tau)=(i,j+1)\right] P_{k,i,j+1}(\tau) \\
& + P\left[Q(\tau+\Delta)=(i,j)|Q(\tau)=(i-1,j+1)\right] P_{k,i-1,j+1}(\tau) \\
& \left. + P\left[Q(\tau+\Delta)=(i,j)|Q(\tau)=(i+1,j-1)\right] P_{k,i+1,j-1}(\tau)\right\},
\end{aligned}$$

which is equal to

$$\begin{aligned}
\frac{P_{k,i,j}(\tau+\Delta)-P_{k,i,j}(\tau)}{\Delta} = {} & -h_{ij}\left(x(\tau)\right) P_{k,i,j}(\tau) \\
& + h_{ij}\left(x(\tau)\right)\left\{\binom{i+1}{2}\frac{1}{x(\tau)h_{ij}\left(x(\tau)\right)} P_{k,i+1,j}(\tau)\right. \\
& + \binom{j+1}{2}\frac{1}{(1-x(\tau))h_{ij}\left(x(\tau)\right)} P_{k,i,j+1}(\tau) \\
& + \frac{(j+1)Rx(\tau)}{h_{ij}\left(x(\tau)\right)} P_{k,i-1,j+1}(\tau) \\
& \left. + \frac{(i+1)R(1-x(\tau))}{h_{ij}\left(x(\tau)\right)} P_{k,i+1,j-1}(\tau)\right\} + O\left(\Delta\right).
\end{aligned}$$

As $\Delta \to 0$, we obtain the system of differential equations

$$\begin{aligned}
\frac{\mathrm{d}P_{k,i,j}(\tau)}{\mathrm{d}\tau} = {} & -h_{ij}\left(x(\tau)\right) P_{k,i,j}(\tau) + \frac{i(i+1)}{2x(\tau)} P_{k,i+1,j}(\tau) \\
& + \frac{j(j+1)}{2(1-x(\tau))} P_{k,i,j+1}(\tau) + (i+1)R(1-x(\tau)) P_{k,i+1,j-1}(\tau) \\
& + (j+1)Rx(\tau) P_{k,i-1,j+1}(\tau), \qquad (4.20)
\end{aligned}$$

with initial conditions $P_{k,k,0}(\tau_{1-\varepsilon}) = 1$ and $P_{k,i,j}(\tau_{1-\varepsilon}) = 0$ for any other configuration of the indices i, j, and $P_{k,i,j}(\tau) = 0$ if $i + j > k$.

Finally, since we do not know the value of τ_a, but we can calculate the value of τ_ε for a given $\varepsilon > 0$ using (4.2), we relate the $\{P_{k,i,j}(\tau_\varepsilon)\}$ to the $\{P_{k,i,j}(\tau)\}$. Assume that the $\{P_{k,i,j}(\tau_\varepsilon), 2 \le i \le k\}$ are negligible, i.e.,

$$P[Q(\tau_\varepsilon)=(0,j)|Q(\tau_{1-\varepsilon})=(k,0)] + P[Q(\tau_\varepsilon)=(1,j-1)|Q(\tau_{1-\varepsilon})=(k,0)] \approx 1,$$

such that no more than one gene of the sample carries A_1 at time τ_ε. Then, the P_{kj} can be defined as

$$P_{kj} = P_{k,0,j}(\tau_\varepsilon) + P_{k,1,j-1}(\tau_\varepsilon). \qquad (4.21)$$

In other cases, where recombination events could still occur in $[\tau_\varepsilon, \tau_a]$, i.e., the $\{P_{k,i,j}(\tau_\varepsilon), 2 \leq i \leq k\}$ are not negligible, the value of P_{kj} as defined in (4.21) underestimates the true value ([17], p. 892).

A sample of size $n = 2$

Consider a sample of size $n = 2$. Then, the expected length of the ancestral tree Γ (cf. (4.11)) is

$$\begin{aligned} E[\Gamma] =& E[T|T_{MRCA} \leq \tau_f,\ Q(0) = (2,0)] \\ &+ \sum_{j=1}^{2} \left(J_{22}(\tau_f) + L_{22}(\tau_f) P_{2j} \sum_{i=1}^{j-1} \frac{2}{i} \right) \\ =& \, 2\left(1 + (\tau_f + P_{22} - 1)e^{-\tau_f}\right), \end{aligned} \tag{4.22}$$

where we have used

$$J_{22}(\tau_f) = 2\tau_f e^{-\tau_f}, \qquad L_{22}(\tau_f) = e^{-\tau_f},$$

and

$$\begin{aligned} E[T|T_{MRCA} \leq \tau_f,\ Q(0) = (2,0)] &= \int_0^{\tau_f} J_{22}(u)\ \mathrm{d}u \\ &= \int_0^{\tau_f} 2\tau e^{-\tau}\ \mathrm{d}\tau \\ &= 2\left(1 - (1 + \tau_f)e^{-\tau_f}\right). \end{aligned}$$

The system of differential equations in (4.20) simplifies to

$$\begin{aligned} \frac{\mathrm{d}P_{2,2,0}(\tau)}{\mathrm{d}\tau} &= -\left(\frac{1}{x(\tau)} + 2R(1 - x(\tau))\right) P_{2,2,0}(\tau) + Rx(\tau)P_{2,1,1}(\tau), \\ \frac{\mathrm{d}P_{2,1,1}(\tau)}{\mathrm{d}\tau} &= -RP_{2,1,1}(\tau) + 2R(1 - x(\tau))P_{2,2,0}(\tau) + 2Rx(\tau)P_{2,0,2}(\tau), \\ \frac{\mathrm{d}P_{2,0,2}(\tau)}{\mathrm{d}\tau} &= -\left(\frac{1}{1 - x(\tau)} + 2Rx(\tau)\right) P_{2,0,2}(\tau) + R(1 - x(\tau))P_{2,1,1}(\tau), \end{aligned}$$

with initial conditions $P_{2,2,0}(\tau_{1-\varepsilon}) = 1$ and $P_{2,1,1}(\tau_{1-\varepsilon}) = P_{2,0,2}(\tau_{1-\varepsilon}) = 0$. This system can be solved numerically.

We use (4.21) to calculate $P_{22}(R)$, assuming $\varepsilon = \frac{10^2}{2N}$ and $2N = 10^8$. Figure 4.1 shows $P_{22}(R)$ as a function of R for different values of α. The higher the recombination rate R, the higher the probability P_{22}, i.e., the lower the

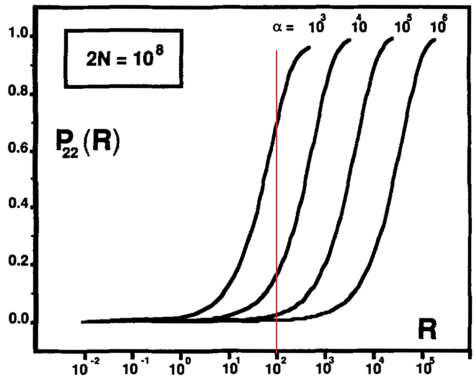

Figure 4.1: The probability of escaping the hitchhiking effect $P_{22}(R)$ according to (4.21) for a sample of size $n = 2$ as a function of R for different values of α.

hitchhiking effect. Thus, $P_{22}(R)$ is an increasing function of R. If we fix a value of R, e.g., $R = 10^2$, we observe that the values of P_{22} decrease if α increases. Hence, there is a greater hitchhiking effect, if α is large, i.e., if selection is strong. Another interpretation would be that the region of the genome that is affected by the selective sweep increases with α. According to Kaplan *et al.* ([17], p. 895), the curves in Figure 4.1 are quite insensitive to the population size as long as $2N \geq 100\alpha$.

The expected length of the ancestral tree Γ is shown in Figure 4.2 as a function of τ_a, the time when the favourable allele A_1 arises in the population. The left panel shows $E[\Gamma]$ for different values of the recombination rate R and fixed $\alpha = 10^4$. On the right, $E[\Gamma]$ is shown for different values of α and fixed $R = 10$. Naturally, $E[\Gamma]$ is an increasing function of τ_a, which means that, the longer the appearance of the favourable allele dates back in time, the greater is the total size of the tree. We observe that the expectation is an increasing function of R, indicating that the effect of hitchhiking on the length of the ancestral tree is low if recombination is strong. On the other hand, $E[\Gamma]$ decreases with α, i.e., if selection is strong the coalescent gets accelerated and hitchhiking reduces the size of the ancestral tree.

The effect on heterozygosity

The quantity P_{22} can be used to calculate the expected heterozygosity at the neutral locus after fixation, H_∞ (cf. 3.37). Let H_0 denote the probability that the ancestral genes of the two sampled genes are different just before the favourable mutation A_1 arises in the population. If no coalescence event occurs during the selective phase, the two genes are still different, i.e., het-

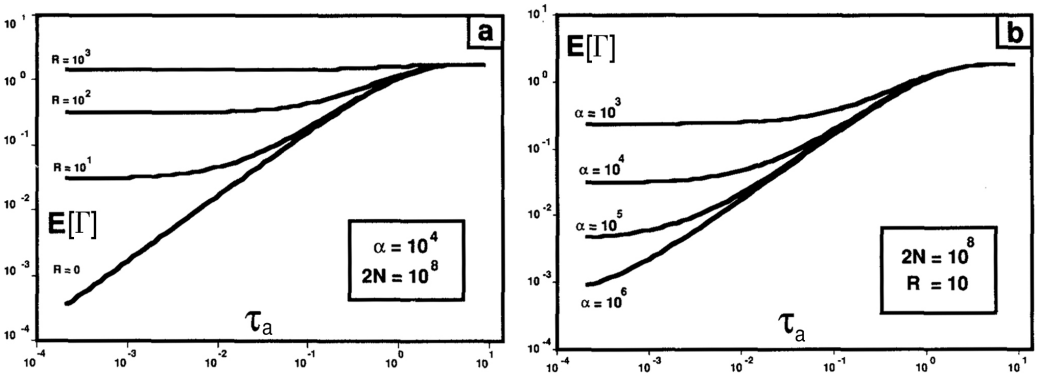

Figure 4.2: The expected total length of the ancestral tree $E[\Gamma]$ according to (4.22) as a function of τ_a, the time when the selective mutation arises in the population.

erozygous, after fixation of A_1. Since these two events are independent, we obtain

$$H_\infty = H_0 P_{22}. \tag{4.23}$$

Thus, if recombination is low and selection strong, the expected heterozygosity at the neutral locus will be reduced due to the selective sweep.

4.2 The effect of recurring selected substitutions

Let λ denote the expected number of selected substitutions per nucleotide site per generation. Thus, $\Lambda = 2N\lambda$ is the expected number of selected substitutions per nucleotide site per unit of time. Assume that selected mutations arise according to a time-homogeneous Poisson process with mean Λ and are located randomly on the genome. Let m denote the physical distance of a selected substitution to the neutral region measured in base pairs (bp). If the expected number of crossovers per nucleotide site per genome per generation is given by c, then $C = 2Nc$ indicates the expected number of crossovers per nucleotide site per genome per unit of time. So, the recombinational distance of a selected substitution can be define as Cm, which is the expected number of crossovers between the neutral locus and the selected locus in m base pairs distance per genome per unit of time.

Since $P_{nn}(Cm) \approx E\left[e^{-\binom{n}{2}\eta}\right] \approx 1$ as m goes to infinity (cf. (4.17)), the effect of selected substitutions that occur far from the neutral locus will have negligible effect on the coalescence process. Thus, we define a limit value M

of the recombinational distance, i.e., substitutions with a recombinational distance larger than M are assumed to have no effect on the coalescent (for a specification of M we refer to [17], p. 894). In order to relate the expected number of selected substitutions to the expected number of crossovers we introduce

$$\Lambda_r = \frac{\Lambda}{C}.$$

Assume Λ_r to be sufficiently small, such that at each time of the coalescent there is at most one selected substitution which effects the coalescent on its way to fixation. Thus, the theory of a single ancestral selected substitution, which we developed in the last section (4.1), can be applied.

4.2.1 The Q process

Consider a sample of size n. Analogously to the last section, the Q process is defined by $Q(\tau) = k$ if the number of ancestral genes at time τ is equal to k, $1 \leq k \leq n$. The process changes its state either due to a coalescence event or due to a selected substitution during which at least one coalescence event occurs. Hence, the distribution of the Q process is given by

$$\begin{aligned}
&P[Q(\tau+\Delta) = k | Q(\tau) = k] \\
&\qquad = 1 - \left(\binom{k}{2} + 2\Lambda_r I_{kk}(M) \right) \Delta + O\left(\Delta^2\right), \\
&P[Q(\tau+\Delta) = k-1 | Q(\tau) = k] \\
&\qquad = \left(\binom{k}{2} + 2\Lambda_r I_{kk-1}(M) \right) \Delta + O\left(\Delta^2\right), \\
&P[Q(\tau+\Delta) = j | Q(\tau) = k] \\
&\qquad = 2\Lambda_r I_{kj}(M)\Delta + O\left(\Delta^2\right), \qquad 1 \leq j \leq k-2, \qquad (4.24)
\end{aligned}$$

where

$$\begin{aligned}
I_{kk}(M) &= \int_0^M (1 - P_{kk}(u)) \ \mathrm{d}u, \\
I_{kj}(M) &= \int_0^M P_{kj}(u) \ \mathrm{d}u, \qquad 1 \leq j \leq k-1,
\end{aligned}$$

i.e., the Q process is a Markov process. We already know that $P_{kk}(Cm)$ denotes the probability of escaping the hitchhiking effect. Thus, $(1 - P_{kk}(Cm))$

is the probability, that the coalescent will be affected due to a selective substitution with recombinational distance Cm. Therefore, $I_{kk}(M)$ is the probability that the process is influenced by a selective sweep that occurs within a recombinational distance M (which equals a physical distance of $\frac{M}{C}$ bp) from the neutral region. $I_{kj}(M)$ is the probability that an selective sweep occurs within a distance of $\frac{M}{C}$ bp from the neutral region during which $k-j$ coalescence events occur.

Let $\Gamma(k)$ denote the waiting time for the next event when there are k ancestral genes left. These waiting times are exponentially distributed with parameter $\binom{k}{2} + 2\Lambda_r I_{kk}(M)$ and expectation

$$\mathrm{E}\left[\Gamma(k)\right] = \frac{1}{\binom{k}{2} + 2\Lambda_r I_{kk}(M)}. \tag{4.25}$$

The jump probabilities of the Q process are

$$q_{k,k-1} = \frac{\binom{k}{2} + 2\Lambda_r I_{kk-1}(M)}{\binom{k}{2} + 2\Lambda_r I_{kk}(M)} \tag{4.26}$$

$$q_{k,j} = \frac{2\Lambda_r I_{kj}(M)}{\binom{k}{2} + 2\Lambda_r I_{kk}(M)}, \qquad 1 \leq j \leq k-2. \tag{4.27}$$

The expected total length of the ancestral tree, Γ, can be calculated by

$$\mathrm{E}\left[\Gamma\right] = \mathrm{E}\left[\sum_{k=2}^{n} k\Gamma(k)\right]. \tag{4.28}$$

A sample of size $n = 2$

Consider a sample of size $n = 2$. In this case, the calculations of the expectation and the variance of Γ are quite simple. Since $\mathrm{E}\left[\Gamma\right] = \mathrm{E}\left[\sum_{k=2}^{n} k\Gamma(k)\right] = 2\,\mathrm{E}\left[\Gamma(2)\right]$ and $\Gamma(2)$ is exponentially distributed with parameter $1+2\Lambda_r I_{22}(M)$, we obtain

$$\mathrm{E}\left[\Gamma\right] = \frac{2}{1 + 2\Lambda_r I_{22}(M)}, \tag{4.29}$$

$$\mathrm{Var}\left[\Gamma\right] = \left(\frac{2}{1 + 2\Lambda_r I_{22}(M)}\right)^2, \tag{4.30}$$

i.e., Γ is exponentially distributed with parameter $\frac{1+2\Lambda_r I_{22}(M)}{2}$.

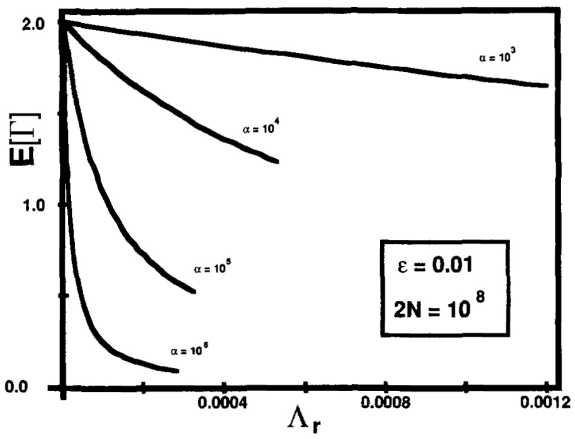

Figure 4.3: The expected length of the ancestral tree $E[\Gamma]$ in the case of recurring selected substitutions as a function of Λ_r.

Note that if $2\Lambda_r I_{22}(M) \ll 1$, then $E[\Gamma] \approx 2$ as in the coalescent of an isolated neutral locus (cf. 2.7a). On the other hand, if $2\Lambda_r I_{22}(M) \gg 1$, then $E[\Gamma] \approx \frac{2}{\Lambda_r I_{22}(M)} < 2$ which is less than in the isolated neutral case. Thus, the effect of hitchhiking on the coalescent, considering recurring selected substitutions in the ancestral history of a sample, depends on the magnitude of $\Lambda_r I_{22}(M)$, i.e., on the number of selected substitutions, the recombination rate and the probability of no coalescence event during the selective phase as a function of the recombinational distance. The larger $\Lambda_r I_{22}(M)$ is, the larger is the influence of hitchhiking on the expected total length of the ancestral tree.

For different values of α, Figure 4.3 shows the expectation of Γ for a sample of size $n = 2$ as a function of Λ_r, the expected number of selected substitutions per crossover. According to Kaplan *et al.* [17], the expectation is insensitive to the population size and the same is true for $I_{22}(M)$ for fixed α. We observe that for large values of α the hitchhiking effect reduces $E[\Gamma]$ substantially even for small values of Λ_r. Since the properties of the number of segregating sites S follow directly from the distribution of Γ, the hitchhiking effect can substantially reduce the number of polymorphisms in the neutral region.

If the expected number of crossovers c decreases, Λ_r increases and $E[\Gamma]$ decreases as shown in Figure 4.3. Thus, in regions where crossovers do not occur frequently the model predicts a lower level of heterozygosity than suggested by the neutral theory.

4.3 Conclusion

In their analysis of the hitchhiking effect Kaplan *et al.* [17] focused on the genealogical history of a sample and the effect on segregating sites. In contrast to Maynard Smith and Haigh ([29], Chapter 3), they also considered the stochastic effects on the frequency of the favourable mutant near the boundaries 0 and 1. Their results support the conclusion of Maynard Smith and Haigh [29] that selective sweeps reduce genetic variation in neutral regions.

The coherence of these two models can be observed considering the case of a single selected substitution. Assume s to be sufficiently small, such that the haploid and the diploid case are effectively equivalent. Then, the heterozygosity obtained in the deterministic model $H_\infty = H_0\Omega(2-\Omega)$ according to (3.37) is approximately equal to the heterozygosity in the stochastic model $H_\infty = H_0 P_{22}$ according to (4.23) if $P_{22} \approx \Omega(2-\Omega)$. Numerical simulations ([17], p. 898) showed that P_{22} is less than $\Omega(2-\Omega)$, i.e., the reduction of heterozygosity in the stochastic model is greater than in the deterministic model. But there is only a quantitative difference, the results are qualitatively the same. In fact, the difference between P_{22} and $\Omega(2-\Omega)$ is not really surprising, since Maynard Smith and Haigh [29] already pointed out that the deterministic approach underestimates the effect on the neutral locus (3.1.3).

In the case of recurring selected substitutions, the results of the stochastic model contrast the neutral theory as well by showing that selection effects neighbouring loci. In the neutral model, the expected length of the ancestral tree T is about two (cf. (2.7a)). According to (2.15a), the expected number of segregating sites follows directly from $\mathrm{E}[T]$ and is given by $\mathrm{E}[S] = \mu\,\mathrm{E}[T]$. Thus, in the neutral theory, $\mathrm{E}[S]$ is directly proportional to $2N$.

In the model of Kaplan *et al.* [17], the expected length of the ancestral tree, denoted by $\mathrm{E}[\Gamma]$, is a decreasing function of α and Λ_r (see Figure 4.3). Therefore, the expected number of segregating sites is no longer directly proportional to $2N$. For example, if we assume Λ_r to be constant, then increasing $2N$ from 10^8 to 10^9 increases $\mathrm{E}[S]$ only by a factor five. Thus, variation on the neutral locus is substantially reduced due to genetic hitchhiking. This result even holds, if the fraction of selected substitutions f_s is negligible as assumed in the neutral theory. For example, if we choose $c = 10^{-8}$, $\mu = 10^{-9}$, $2N = 10^9$, $s = 10^{-3}$ and $\Lambda_r = 2.5 \cdot 10^4$, then we obtain $\mathrm{E}[\Gamma] = 0.096$, although

$$f_s = \frac{\lambda}{\lambda+\mu} \approx \frac{C\Lambda_r}{2N\mu} = 2.5 \cdot 10^{-3}. \tag{4.31}$$

This means that only $\frac{1}{400}$ of selective substitutions is fixed, but $\mathrm{E}[\Gamma]$, and con-

sequently $\mathrm{E}[S]$, is reduced by a factor $\frac{1}{20}$ ([17], p. 898). The only restrictions of the model are the assumptions that α is large and that Λ_r is sufficiently small, such that we do not have to consider simultaneous selective sweeps.

The stochastic model provided by Kaplan *et al.* [17] confirms that selected substitutions reduce the number of polymorphic sites in contradiction to the neutral theory and, therefore, support the contentions of Maynared Smith and Haigh [29].

Chapter 5

A diffusion approximation

The last model that will be presented here considers both the effect of a single selected substitution (cf. Chapter 3) and the effect of recurring selected substitutions (cf. Chapter 4) by applying the framework of diffusion theory. It has been presented in 1992 by Stephan *et al.* [41] in order to obtain a model that can be generalized to describe polymorphisms on non-neutral loci. This approach has been motivated by contemporaneous studies that suggested that a significant fraction of mutations at the molecular level is not selectively neutral ([41], p. 239).

Stephan *et al.* [41] based their model on the moment analysis of Otha and Kimura [36] presented in 1975. However, the original model considers a different biological situation; indeed, Otha and Kimura described the effect on a newly arising neutral mutant while a favourable allele on the selected locus is on its way to fixation. They derived a set of ordinary differential equations to describe the change in expected heterozygosity on the neutral locus. Stephan *et al.* adapted the model to the biological situation of our study that considers the effect of a newly arising selectively favourable mutant on a preexisting neutral polymorphic locus. The latter biological situation seems to be more relevant since strongly selected mutations fix rapidly.

5.1 The model

Consider two linked loci. The first locus is denoted by A and we assume that a selectively favourable mutant A_1 arises at time $\tau = 0$. The wild type is designated by A_2. The other locus, referred to as the B locus, is selectively neutral and its two different allelic types are denoted by B_1 and B_2, respectively. The selective advantage of A_1 is given by $s > 0$ and we

genotypes	A_1A_1	A_1A_2	A_2A_2
fitnesses	$1+2s$	$1+s$	1

Table 5.1: The fitnesses of the three possible genotypes.

assume additive fitness. The fitnesses of the three different genotypes are shown in Table 5.1 (cf. Table 4.1 and Table 3.2). Assume a randomly mating diploid population of size N and define $\alpha = 2Ns$. We require that the dynamics on the selected locus are independent of the dynamics on the neutral locus. As in 4.1.1, the frequency of A_1 is given by the stochastic process $X(\tau)$, $\tau > 0$, and we assume α to be large, such that the frequency process can be modelled deterministically as long as it stays away from the boundaries 0 and 1 ([25]). Near the boundaries, the frequency of A_1 needs to be treated stochastically.

Consider $\varepsilon > 0$ small. Then, the representation of the X process (cf. (4.1)-(4.3)) is given by

$$X(\tau) = \begin{cases} X(\tau) & \text{if } 0 \leq \tau \leq \tau_\varepsilon \\ x(\tau) & \text{if } \tau_\varepsilon \leq \tau \leq \tau_{1-\varepsilon} \\ X(\tau) & \text{if } \tau > \tau_{1-\varepsilon}, \end{cases} \tag{5.1}$$

where

$$\tau_\varepsilon = \inf\{\tau : X(\tau) = \varepsilon\}, \qquad \tau_{1-\varepsilon} = x^{-1}(1-\varepsilon), \tag{5.2}$$

and $x(\tau)$ is determined by the differential equation

$$\frac{\mathrm{d}x(\tau)}{\mathrm{d}\tau} = sx(\tau)\left(1 - x(\tau)\right), \tag{5.3}$$

$$x\left(\tau_\varepsilon\right) = \varepsilon, \qquad \tau \geq \tau_\varepsilon,$$

which is a classical continuous time selection model, that describes the change in frequency of an allele on a selected diallelic locus in the case of no dominance forward in time.

As in 3.1, the fraction of B_1 alleles among genes carrying A_1 is denoted by Q_{A_1}, and the fraction of B_1 alleles among genes carrying A_2 is denoted by Q_{A_2}. These fractions are time-dependent frequencies, but in order to reach a higher legibility we will drop the superscripts (τ) if not necessary and write

Q_{A_1} and Q_{A_2} instead of $Q_{A_1}^{(\tau)}$ and $Q_{A_2}^{(\tau)}$. The total frequency of B_1 in the whole population at time $\tau > 0$ is given by

$$Q = Q_{A_1} x(\tau) + Q_{A_2} \left(1 - x(\tau)\right). \tag{5.4}$$

The solution of (5.3) is given by

$$x(\tau) = \frac{\varepsilon}{\varepsilon + (1-\varepsilon) e^{-s(\tau - \tau_\varepsilon)}}. \tag{5.5}$$

For convenience, we introduce a new variable $\bar{\tau} = \tau - \tau_\varepsilon$, which measures the time since the frequency of A_1 has reached ε, i.e., at $\bar{\tau} = 0$ it is clear that the mutant A_1 is going to fixation. This implies that all formulas based on this new timescale $\bar{\tau}$ only consider the cases where A_1 gets fixed. The time that the X process needs to increase from ε to $1 - \varepsilon$ is

$$\hat{\tau} = \tau_{1-\varepsilon} - \tau_\varepsilon = -2 \frac{\ln \varepsilon}{s}. \tag{5.6}$$

Let $\phi\left(Q_{A_1}, Q_{A_2}, \bar{\tau}\right)$ denote the joint probability density function of Q_{A_1} and Q_{A_2} at time $\bar{\tau} > 0$, and let R denote the recombination rate between the two loci. Then, we define the expectation of any polynomial f of Q_{A_1} and Q_{A_2} at time $\bar{\tau}$ by

$$\mathrm{E}[f, \bar{\tau}] = \int_0^1 \int_0^1 f\left(Q_{A_1}, Q_{A_2}\right) \phi\left(Q_{A_1}, Q_{A_2}, \bar{\tau}\right) \, \mathrm{d}Q_{A_1} \mathrm{d}Q_{A_2}, \tag{5.7}$$

which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}} \mathrm{E}[f, \bar{\tau}] = \mathrm{E}\left[L(f), \bar{\tau}\right], \tag{5.8}$$

where L denotes the differential operator

$$\begin{aligned} L = & \frac{Q_{A_1}\left(1 - Q_{A_1}\right)}{4N x(\bar{\tau})} \frac{\partial^2}{\partial Q_{A_1}^2} + R(1 - x(\bar{\tau}))\left(Q_{A_2} - Q_{A_1}\right) \frac{\partial}{\partial Q_{A_1}} \\ & + \frac{Q_{A_2}\left(1 - Q_{A_2}\right)}{4N(1 - x(\bar{\tau}))} \frac{\partial^2}{\partial Q_{A_2}^2} + R x(\bar{\tau})\left(Q_{A_1} - Q_{A_2}\right) \frac{\partial}{\partial Q_{A_2}}. \end{aligned} \tag{5.9}$$

The differential operator L formally equals the differential operator of the Kolmogorov backward equation in multiple variables (see, for example, [9], Chapter 4.8). For a derivation of (5.8) and (5.9) see [41], Appendix 1. Originally, it has been introduced into the field of population genetics by Otha and Kimura in 1969 ([35], for details see also [21], pp. 183-190), who pointed out that it is especially useful for a moment analysis of the frequency distribution in the case of two linked loci.

Choosing f in (5.8) successively as Q_{A_1}, Q_{A_2}, $Q_{A_1}^2$, $Q_{A_1}Q_{A_2}$ and $Q_{A_2}^2$ leads to a set of ordinary differential equations for the first- and second-order moments of the frequencies of B_1 on its backgrounds A_1 and A_2 at time $\bar{\tau}$ with respect to ϕ given by

$$\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\,\mathrm{E}\left[Q_{A_1}\right] = R(1-x(\bar{\tau}))\,\mathrm{E}\left[Q_{A_2}-Q_{A_1}\right], \tag{5.10a}$$

$$\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\,\mathrm{E}\left[Q_{A_2}\right] = Rx(\bar{\tau})\,\mathrm{E}\left[Q_{A_1}-Q_{A_2}\right], \tag{5.10b}$$

$$\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\,\mathrm{E}\left[Q_{A_1}^2\right] = \mathrm{E}\left[\frac{Q_{A_1}\left(1-Q_{A_1}\right)}{2Nx(\bar{\tau})} + 2R(1-x(\bar{\tau}))Q_{A_1}\left(Q_{A_2}-Q_{A_1}\right)\right], \tag{5.11a}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\,\mathrm{E}\left[Q_{A_1}Q_{A_2}\right] = \mathrm{E}\big[&R(1-x(\bar{\tau}))Q_{A_2}\left(Q_{A_2}-Q_{A_1}\right)\\ &+Rx(\bar{\tau})Q_{A_1}\left(Q_{A_1}-Q_{A_2}\right)\big],\end{aligned} \tag{5.11b}$$

$$\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\,\mathrm{E}\left[Q_{A_2}^2\right] = \mathrm{E}\left[\frac{Q_{A_2}\left(1-Q_{A_2}\right)}{2N(1-x(\bar{\tau}))} + 2Rx(\bar{\tau})Q_{A_2}\left(Q_{A_1}-Q_{A_2}\right)\right]. \tag{5.11c}$$

Note that since these equations are formulated on the timescale $\bar{\tau}$, we do not consider the cases where A_1 becomes extinct. It is remarkable that equations (5.10) for the first-order moments do not contain second-order terms. Furthermore, no equation depends on higher-order moments. The equations in (5.10) and (5.11) are identical to those derived by Otha and Kimura [36]. Their solutions are conditioned expectations depending on the initial conditions Q_{A_1} and Q_{A_2} at time $\bar{\tau}=0$ which equals $\tau=\tau_\varepsilon$. The choice of these initial conditions is determined the biological situation.

Our goal is to obtain an unconditioned expectation of an arbitrary function f of Q_{A_1} and Q_{A_2}. Therefore, we distinguish between the two cases where the selectively favourable mutant A_1 arises either on background B_1, i.e., $Q_{A_1}^{(0)}=1$, or B_2, i.e., $Q_{A_1}^{(0)}=0$ (cf. (3.37)). Then, the unconditioned expectation $\bar{\mathrm{E}}$ of f can be calculated by

$$\bar{\mathrm{E}}[f] = Q_{A_2}^{(0)}\,\mathrm{E}\left[f|Q_{A_1}^{(0)}=1\right] + \left(1-Q_{A_2}^{(0)}\right)\mathrm{E}\left[f|Q_{A_1}^{(0)}=0\right], \tag{5.12}$$

where $Q_{A_2}^{(0)}$ and $1-Q_{A_2}^{(0)}$ are the probabilities that A_1 arises on background B_1 or B_2, respectively. The conditioned expectations $\mathrm{E}\left[f|Q_{A_1}^{(0)}=1\right]$ and $\mathrm{E}\left[f|Q_{A_1}^{(0)}=0\right]$ in (5.12) are formulated on the timescale τ. Since the set of

ordinary differential equations of (5.11), which we will use to calculate these expectations, is formulated on the timescale $\bar{\tau}$, we need to approximate the initial frequencies $Q_{A_1}^{(0)}$ and $Q_{A_2}^{(0)}$ to obtain corresponding initial conditions $Q_{A_1}^{(\tau_\varepsilon)}$ and $Q_{A_2}^{(\tau_\varepsilon)}$. Therefore, we assume (cf. 4.1.5)

$$Q_{A_1}^{(\tau_\varepsilon)} = Q_{A_1}^{(0)} \qquad \text{and} \qquad Q_{A_2}^{(\tau_\varepsilon)} = Q_{A_2}^{(0)}. \tag{5.13}$$

The first assumption requires that no crossing over involving A alleles takes place during the initial phase from 0 to τ_ε. This is reasonable if the recombination rate is very low. The second assumption is valid for large populations, since there it is unlikely that random genetic drift changes the frequency of A_2B_1 substantially during the short initial phase ([7], p. 383). Furthermore, simulations showed that the reduction of heterozygosity does not depend on the initial frequency $Q_{A_2}^{(0)}$ of B_1 ([41], p. 246).

5.2 The effect of a single substitution on heterozygosity

Let $H_{1-\varepsilon}$ denote the expected heterozygosity on the neutral locus at time $\tau_{1-\varepsilon}$, i.e., at the end of the selective phase. It can be calculated by inserting $f = 2Q(1-Q)$ into (5.12) (cf. 3.2.3), where Q denotes the total frequency of B_1 in the whole population defined in (5.4). Thus,

$$\begin{aligned} H_{1-\varepsilon} =& 2\bar{\mathrm{E}}\left[Q(1-Q)\right] \\ =& 2Q_{A_2}^{(0)}\,\mathrm{E}\left[Q(1-Q)\,|Q_{A_1}^{(0)}=1\right] + 2\left(1-Q_{A_2}^{(0)}\right)\mathrm{E}\left[Q(1-Q)\,|Q_{A_1}^{(0)}=0\right]. \end{aligned} \tag{5.14}$$

For further investigations, we start by considering only the deterministic part of the model, which equals the deterministic model of Chapter 3. Thereafter, we will consider the more general case including the stochastic effects caused by the finite population size.

5.2.1 The deterministic case

The deterministic behaviour of the model is determined by (5.10) ignoring the higher-order moments of (5.11). If we assume $s \ll 1$, these equations correspond to those of Maynard Smith and Haigh (cf. (3.27) and (3.28)). Their solutions are given by

$$\mathrm{E}\left[Q_{A_1}\right] = Q_{A_1}^{(\tau_\varepsilon)} + R\left(Q_{A_2}^{(\tau_\varepsilon)} - Q_{A_1}^{(\tau_\varepsilon)}\right)\int_0^{\bar{\tau}} \frac{(1-\varepsilon)e^{-(s+R)y}}{\varepsilon + (1-\varepsilon)e^{-sy}}\,\mathrm{d}y, \tag{5.15a}$$

$$\mathrm{E}\left[Q_{A_2}\right] = Q_{A_2}^{(\tau_\varepsilon)} + R\left(Q_{A_1}^{(\tau_\varepsilon)} - Q_{A_2}^{(\tau_\varepsilon)}\right)\int_0^{\bar{\tau}} \frac{\varepsilon e^{-Ry}}{\varepsilon + (1-\varepsilon)e^{-sy}}\,\mathrm{d}y. \tag{5.15b}$$

The frequency of A_1 at time $\tau_{1-\varepsilon}$ is given by $x(\tau_{1-\varepsilon}) = 1-\varepsilon$ (see (5.2)). Therefore, the total frequency of B_1 in the whole population at time $\tau_{1-\varepsilon}$ (see (5.4)) is given by

$$Q^{(\tau_{1-\varepsilon})} = Q_{A_1}^{(\tau_{1-\varepsilon})} + \varepsilon\left(Q_{A_2}^{(\tau_{1-\varepsilon})} - Q_{A_1}^{(\tau_{1-\varepsilon})}\right). \tag{5.16}$$

The time $\tau = \tau_{1-\varepsilon}$ corresponds to the time $\bar{\tau} = \hat{\tau} = -2\frac{\ln\varepsilon}{s}$. The values of $\mathrm{E}\left[Q_{A_2}^{(\hat{\tau})}\right]$ and $\mathrm{E}\left[Q_{A_1}^{(\hat{\tau})}\right]$ can be calculated by using (5.15) with $\hat{\tau} = -2\frac{\ln\varepsilon}{s}$. Therefore, the expectation of Q at time $\hat{\tau}$ is given by

$$\begin{aligned}\mathrm{E}\left[Q^{(\hat{\tau})}\right] &= \mathrm{E}\left[Q_{A_1}^{(\hat{\tau})}\right] + \varepsilon\left(\mathrm{E}\left[Q_{A_2}^{(\hat{\tau})}\right] - \mathrm{E}\left[Q_{A_1}^{(\hat{\tau})}\right]\right)\\ &= \mathrm{E}\left[Q_{A_1}^{(\hat{\tau})}\right] + \left(Q_{A_2}^{(\tau_\varepsilon)} - Q_{A_1}^{(\tau_\varepsilon)}\right)\varepsilon^{1+\frac{2R}{s}}\end{aligned} \tag{5.17}$$

In order to calculate the conditioned expectation $\mathrm{E}\left[Q^{(\hat{\tau})}|Q_{A_1}^{(0)}\right]$ we insert the solution of $\mathrm{E}\left[Q_{A_1}^{(\hat{\tau})}\right]$ given in (5.15a) into (5.17) and choose the initial condition as $Q_{A_1}^{(\tau_\varepsilon)} = Q_{A_1}^{(0)}$ according to (5.13).

The expected heterozygosity $H_{1-\varepsilon}$ at the end of the selective phase at time $\hat{\tau}$ is given by

$$\begin{aligned}H_{1-\varepsilon} =& 2Q_{A_2}^{(\tau_\varepsilon)}\,\mathrm{E}\left[Q^{(\hat{\tau})}|Q_{A_1}^{(0)}=1\right]\left(1-\mathrm{E}\left[Q^{(\hat{\tau})}|Q_{A_1}^{(0)}=1\right]\right)\\ &+ 2\left(1-Q_{A_2}^{(\tau_\varepsilon)}\right)\mathrm{E}\left[Q^{(\hat{\tau})}|Q_{A_1}^{(0)}=0\right]\left(1-\mathrm{E}\left[Q^{(\hat{\tau})}|Q_{A_1}^{(0)}=0\right]\right),\end{aligned} \tag{5.18}$$

and, using the formulas for the conditioned expectation $\mathrm{E}\left[Q^{(\hat{\tau})}|Q_{A_1}^{(0)}\right]$ we derived above, we find

$$\frac{H_{1-\varepsilon}}{2Q_{A_2}^{(\tau_\varepsilon)}\left(1-Q_{A_2}^{(\tau_\varepsilon)}\right)} = RI(\hat{\tau})\left(2-RI(\hat{\tau})\right) + O\left(\varepsilon^{(1+\frac{3R}{s})}\right), \tag{5.19}$$

where

$$I(\bar{\tau}) = \int_0^{\bar{\tau}} \frac{e^{-(R+s)y}}{\varepsilon + e^{-sy}}\,\mathrm{d}y. \tag{5.20}$$

Since the favourable allele A_1 arises in a single copy at time $\tau = 0$, the initial frequency of B_1 in the whole population is determined by the initial fraction $Q_{A_2}^{(0)}$ of B_1 on background A_2. According to (5.13), the initial fractions $Q_{A_1}^{(0)}$ and $Q_{A_2}^{(0)}$ of B_1 on its respective backgrounds A_1 and A_2 do not change before τ_ε. Therefore, the weighted expectation $\bar{\mathrm{E}}[Q^{(\tau_\varepsilon)}]$ of the total frequency of B_1 at time τ_ε, i.e., at the begin of the selective phase, is $\bar{\mathrm{E}}[Q^{(\tau_\varepsilon)}] = Q_{A_2}^{(\tau_\varepsilon)}$, and thus, the initial heterozygosity H_ε on the neutral locus at time τ_ε is given by

$$H_\varepsilon = 2Q_{A_2}^{(\tau_\varepsilon)} \left(1 - Q_{A_2}^{(\tau_\varepsilon)}\right). \tag{5.21}$$

Hence, (5.19) describes the reduction rate of heterozygosity on the neutral locus due to a single substitution.

The integral defined in (5.20) determines the time-dependent change of $\mathrm{E}\left[Q_{A_1}^{(\bar{\tau})}\right]$. It can be approximated ([41], pp. 243-244) by

$$RI(\hat{\tau}) \approx 1 - \varepsilon^{\frac{R}{s}}. \tag{5.22}$$

For $\varepsilon = \frac{1}{2N}$, the value of (5.22) corresponds to the approximation of the reduction of the frequency of B_1 in the deterministic model of Maynard Smith and Haigh (see (3.31)).

Inserting (5.22) into (5.19) leads to an approximation of the reduction rate of heterozygosity on the neutral locus given by

$$\frac{H_{1-\varepsilon}}{H_\varepsilon} \approx 1 - \varepsilon^{\frac{R}{s}}. \tag{5.23}$$

5.2.2 The more general case

According to (5.16), the total frequency of B_1 at the end of the selective phase is approximately equal to the frequency of B_1 alleles among genes carrying A_1, i.e., $Q^{(\tau_{1-\varepsilon})} \approx Q_{A_1}^{(\tau_{1-\varepsilon})}$. Thus, in order to calculate the expected heterozygosity $H_{1-\varepsilon}$ at the end of the selective phase, we derive a differential equation for $\mathrm{E}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right]$ instead of $\mathrm{E}\left[Q\left(1 - Q\right)\right]$. Therefore, we subtract (5.11a) from (5.10a) and obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}} \mathrm{E}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right] + \mathrm{E}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right] \left[\frac{1}{2Nx(\bar{\tau})} + 2R(1 - x(\bar{\tau}))\right] \\ = R(1 - x(\bar{\tau}))\left[2\,\mathrm{E}\left[Q_{A_1}\left(1 - Q_{A_2}\right)\right] - \mathrm{E}\left[Q_{A_1} - Q_{A_2}\right]\right].\end{aligned} \tag{5.24}$$

Using (5.15), the expectation of $Q_{A_1} - Q_{A_2}$ can be calculated by

$$\mathrm{E}\left[Q_{A_1} - Q_{A_2}\right] = \mathrm{E}\left[Q_{A_1}\right] - \mathrm{E}\left[Q_{A_2}\right] = \left(Q_{A_1}^{(\tau_\varepsilon)} - Q_{A_2}^{(\tau_\varepsilon)}\right) e^{-R\bar{\tau}}. \tag{5.25}$$

Hence, according to (5.12), the weighted expectation of $Q_{A_1} - Q_{A_2}$ is given by

$$\bar{\mathrm{E}}\left[Q_{A_1} - Q_{A_2}\right] = Q_{A_2}^{(\tau_\varepsilon)}\left(1 - Q_{A_2}^{(\tau_\varepsilon)}\right) e^{-R\bar{\tau}} - \left(1 - Q_{A_2}^{(\tau_\varepsilon)}\right) Q_{A_2}^{(\tau_\varepsilon)} e^{-R\bar{\tau}} = 0. \tag{5.26}$$

Therefore, taking the weighted expectation of (5.24) leads to

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\bar{\mathrm{E}}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right] + \bar{\mathrm{E}}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right]\left[\frac{1}{2Nx(\bar{\tau})} + 2R(1 - x(\bar{\tau}))\right] \\ &\quad = 2R(1 - x(\bar{\tau}))\bar{\mathrm{E}}\left[Q_{A_1}\left(1 - Q_{A_2}\right)\right]. \end{aligned} \tag{5.27}$$

Equations (5.10a) and (5.11a) indicate that the expected heterozygosity only changes substantially if $\bar{\tau} < \frac{1}{2}\hat{\tau}$, i.e., as long as the frequency of the favourable mutant A_1 is low. Furthermore, (5.10b) suggests that during this time period the frequency of B_1 on background A_2 remains almost constant at its initial value $Q_{A_2}^{(\tau_\varepsilon)}$. Therefore, the expectation of $Q_{A_1}Q_{A_2}$ can be approximated by $\mathrm{E}\left[Q_{A_1}Q_{A_2}\right] \approx Q_{A_2}^{(\tau_\varepsilon)}\,\mathrm{E}\left[Q_{A_1}\right]$ for $\bar{\tau} < \frac{1}{2}\hat{\tau}$. The accuracy of this approximation has been justified by numerical integrations of (5.10) and (5.11) ([41], p. 244). We use this approximation and (5.15a) to calculate the weighted expectation of $Q_{A_1}\left(1 - Q_{A_2}\right)$ according to (5.12), and derive

$$\bar{\mathrm{E}}\left[Q_{A_1}\left(1 - Q_{A_2}\right)\right] = \left(1 - Q_{A_2}^{(\tau_\varepsilon)}\right)\bar{\mathrm{E}}\left[Q_{A_1}\right] = Q_{A_2}^{(\tau_\varepsilon)}\left(1 - Q_{A_2}^{(\tau_\varepsilon)}\right). \tag{5.28}$$

Thus, for $\bar{\tau} < \frac{1}{2}\hat{\tau}$ the differential equation (5.27) can be approximated by

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\bar{\tau}}\bar{\mathrm{E}}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right] + \bar{\mathrm{E}}\left[Q_{A_1}\left(1 - Q_{A_1}\right)\right]\left[\frac{1}{2Nx(\bar{\tau})} + 2R(1 - x(\bar{\tau}))\right] \\ &\quad = 2R(1 - x(\bar{\tau}))Q_{A_2}^{(\tau_\varepsilon)}\left(1 - Q_{A_2}^{(\tau_\varepsilon)}\right). \end{aligned} \tag{5.29}$$

Stephan *et al.* ([41], Appendix 2) solved (5.29) approximately, and calculated the reduction rate of heterozygosity on the neutral locus due to a single hitchhiking event according to (5.14),

$$\frac{H_{1-\varepsilon}}{H_\varepsilon} = \frac{2R}{s}\alpha^{-\frac{2R}{s}}\Gamma\left(-\frac{2R}{s}, \frac{1}{\alpha}, \frac{1}{\alpha\varepsilon}\right), \tag{5.30}$$

where $\Gamma\left(x, y, z\right)$ denotes the generalized incomplete gamma function defined by $\Gamma(x, y) - \Gamma(x, z)$ ([11], Equation (8.350.2)).

$-\log\left(\frac{R}{s}\right)$	(5.19)	(5.23)	(5.30)	Kaplan *et al.*	(5.10), (5.11)
3.0	0.027250	0.027253	0.021631	0.021637	0.021626
2.8	0.042839	0.042847	0.034062	0.034079	0.034050
2.6	0.067033	0.067052	0.053437	0.053477	0.053409
2.4	0.104120	0.104167	0.083341	0.083376	0.083279
2.2	0.159879	0.159989	0.128786	0.128580	0.128650
2.0	0.241173	0.241422	0.196182	0.195660	0.195883
1.8	0.354090	0.354624	0.292346	0.291868	0.291707
1.6	0.499418	0.500456	0.421466	0.419731	0.420186
1.4	0.665391	0.667132	0.579071	0.576768	0.576761
1.2	0.822766	0.825075	0.744880	0.740796	0.741370
1.0	0.934787	0.936904	0.883579	0.878784	0.879500
0.8	0.986376	0.987465	0.965524	0.961843	0.962341
0.6	0.998804	0.999032	0.994531	0.992985	0.993085

Table 5.2: The table compares the reduction rates of heterozygosity due to a single hitchhiking event for $2N = 10^8$, $\alpha = 10^5$, $\varepsilon = 10^{-6}$ and different values of R. The first two columns represent the deterministic case showing the real value of (5.19) and the approximation of (5.23). The next column represents the reduction rate in the stochastic case determined by (5.30). We compare these values to the reduction rate of the coalescent approach provided by Kaplan *et al.* [17], which are obtained by a linear interpolation of P_{22}-data (cf. (4.23)). The last column contains the reduction rate calculated from the numerical solutions of the set of ordinary differential equations given in (5.10) and (5.11) by applying the Runge-Kutta method. ([41], Table 1)

According to Stephan *et al.* ([41], p. 245), in the case where $\varepsilon \leq \frac{1}{\alpha}$, the reduction of heterozygosity depends only weakly on ε and the reduction rate can be approximated by

$$\frac{H_{1-\varepsilon}}{H_\varepsilon} \approx \frac{2R}{s}\alpha^{-\frac{2R}{s}}\Gamma\left(-\frac{2R}{s}, \frac{1}{\alpha}\right). \tag{5.31}$$

Table 5.2 compares the reduction rates of heterozygosity due to a single hitchhiking event in the deterministic and the stochastic case to the results of the coalescent approach of Kaplan *et al.* ([17], Chapter 4) and the exact values obtained by solving the system of ordinary differential equations (5.10)-(5.11) applying the Runge-Kutta method. The values of the stochastic case in the diffusion approximation agree well with those of Kaplan *et al.*. For ε small, the deterministic model overestimates the true value of the reduction rate, i.e., the reduction of heterozygosity in the deterministic case is less than in the stochastic models.

5.3 The effect of recurring substitutions on heterozygosity

In the case of recurring selected substitutions, the reduction rate of average heterozygosity on the neutral locus is given by the probability that the neutral locus escapes the hitchhiking effect, i.e., the probability that the neutral locus will not reach homozygosity during a selective sweep.

As in 4.2, let Λ denote the expected number of selected substitutions per nucleotide site per chromosome per generation (which equals the expected number of selected substitutions per nucleotide site per unit of time). Assume that selected mutations arise according to a time-homogeneous Poisson process with mean Λ and are located randomly on the genome. Let m denote the physical distance of a selected substitution to the neutral locus, measured in nucleotide sites, and let ρ denote the expected number of crossovers (cf. $\rho = 2NC$ in 4.2). Then, the recombinational distance of a selected substitution to the neutral locus is given by ρm. We define a maximum value, $\bar{M}$, of the recombinational distance, i.e., $\rho m \leq \bar{M}$, within which selected substitutions effect the neutral locus (cf. $2NM = \bar{M}$ in 4.2). In order to apply the theory of a single selected substitution to the case of recurring substitutions, we assume $\Lambda_r = \frac{2N\Lambda}{\rho}$ to be sufficiently small, such that at any time at most one selected substitution that affects the considered neutral locus is on its way to fixation (for a specification of $\bar{M}$ and Λ_r we refer to [17], p. 894).

Let k_h denote the rate of selected substitutions that lead to fixation on the neutral locus, and let $h(R)$ denote the reduction rate of heterozygosity in the case of a single substitution given by the right-hand side of (5.19), (5.23), (5.30) or (5.31) as a function of the recombination rate R. Since $h(R)$ can be interpreted as the probability of escaping the hitchhiking effect of a single substitution (cf. 4.1.5), $1-h(R)$ will be understood as the probability that the neutral locus reaches homozygosity during a single selective sweep.

Consider a chromosomal region within a physical distance from m to $m+\Delta$ nucleotide sites to the neutral locus. The expected number of selected substitutions per generation that occur in this region of Δ nucleotide sites is given by $2N\Lambda\Delta$. Therefore, the rate of selected substitutions occurring within a recombinational distance of ρm in a region of Δ nucleotide sites that drag the neutral locus to fixation is given by

$$2N\Lambda(1-h(\rho m))\Delta. \tag{5.32}$$

Averaging (5.32) over all possible values of m and rescaling from ρm back to

R leads to

$$k_h = 2\Lambda_r \int_0^{\bar{M}} (1 - h(R))\, \mathrm{d}R. \tag{5.33}$$

Consider a sequence of substitutions occurring within a physical distance $\frac{M}{\rho}$. Since k_h is the number of selected substitutions that lead to fixation per occurring selected substitution, the probability that one of the selected substitutions takes along the neutral locus is

$$\frac{k_h}{1 + k_h}, \tag{5.34}$$

and the probability that none of these substitutions effect the neutral locus, i.e., that it escapes the hitchhiking effect, is

$$\frac{1}{1 + k_h}, \tag{5.35}$$

which is the reduction rate of average heterozygosity in the case of recurring substitutions.

Let $H_{neutral}$ denote the expected heterozygosity of the neutral locus according to the neutral theory. Then, the expected heterozygosity H in the case of recurrent selected substitutions is given by

$$H = \frac{H_{neutral}}{1 + k_h}. \tag{5.36}$$

This corresponds to Equation (4.29) in the stochastic model of Kaplan *et al.* [17].

To evaluate the integral in (5.33), we use the series representation of the incomplete gamma function ([11], Equation (8.354.2)) and neglect the higher-order terms. Therefore, we obtain for the right-hand side of (5.31)

$$h(R) \approx 1 - \alpha^{-\frac{2R}{s}} \Gamma\left(1 - \frac{2R}{s}\right), \tag{5.37}$$

and, using the special function $\lambda(.,.)$ ([11], Equation (9.640.5)) to express the integral over $\alpha^{-\frac{2R}{s}} \Gamma\left(1 - \frac{2R}{s}\right)$, we derive

$$k_h \approx -\alpha \frac{\Lambda}{\rho} \lambda\left(\frac{1}{\alpha}, -\frac{2\bar{M}}{s}\right). \tag{5.38}$$

Thus, the reduction rate of expected heterozygosity on the neutral locus compared to the neutral value in the case of recurring selected substitutions

is given by

$$\frac{H}{H_{neutral}} \approx \frac{1}{1 - -\alpha\frac{\Lambda}{\rho}\lambda\left(\frac{1}{\alpha}, -\frac{2\bar{M}}{s}\right)}. \tag{5.39}$$

The results of the diffusion approximation agree well with those of Kaplan *et al.* shown in Figure 4.3.

5.4 Conclusion

Stephan *et al.* [41] chose the diffusion approach in order to obtain a general model that could readily be adapted to study the effect of strongly selected mutations on weakly selected loci or on slightly deleterious alleles. The presented diffusion approximation is based on the moment analysis of Otha and Kimura [36].

The original model considered the effect of a single selected substitution on its way to fixation on a newly arising neutral mutant. Therefore, Otha and Kimura distinguished between positive hitchhiking, in the case where the neutral mutant B_1 arises on the favourable allele A_1, and negative hitchhiking, in the case where B_1 arises on the deleterious allele A_2. Then, the total average heterozygosity has been calculated as the weighted mean of those two effects, where $\frac{1}{2N}$ and $1-\frac{1}{2N}$ are the probabilities that B_1 arises on background A_1 or A_2. In this way, they showed that the hitchhiking effect is in general unimportant, since the positive and the negative hitchhiking effect tend to cancel each other out. Furthermore, they pointed out that the reduction rate of heterozygosity is a rapidly decreasing function of NR. Therefore, the hitchhiking effect would at most be important if the reduction rate is much less than the selective advantage of A_1, which seems to be an biological unimportant situation. Their conclusions turned out to be wrong, but they provided a useful set of ordinary differential equations that describe the dynamics of the fist- and second-order moments of the fractions of B_1 on its respective backgrounds A_1 and A_2.

Stephan *et al.* [41] remodeled the provided system by changing its initial conditions and thereby the biological situation. Their analytical results have been approved by numerical simulations carried out by J. Gillespie. The simulations were based on a Wright-Fisher model and consider the initial phase ($\tau < \tau_\varepsilon$) and the phase prior to fixation ($\tau > \tau_{1-\varepsilon}$) as well. The initial frequency of the favourable allele A_1 has been assumed as $\frac{1}{2N}$, i.e., the mutant arises in a single copy. The simulations results for a single substitution are shown in Table 5.1, and compared to the theoretical values of (5.31). If α is sufficiently large, the expected and simulated values agree well. For smaller

	Simulation				
$-\log_{10}(c/s)$	$Q^{(0)}=0.5$		$Q^{(0)}=0.1$		Expected
			$\alpha = 2\times 10^3$		
3.0	0.0124	(0.0034)	0.0128	(0.0016)	0.0139
2.0	0.1208	(0.0123)	0.1302	(0.0075)	0.1308
1.0	0.7315	(0.0363)	0.7378	(0.0078)	0.7456
0.0	0.9919	(0.0130)	0.9910	(0.0006)	0.9990
			$\alpha = 2\times 10^2$		
3.0	0.0051	(0.0021)	0.0050	(0.0025)	0.0094
2.0	0.0928	(0.0123)	0.0944	(0.0192)	0.0899
1.0	0.5342	(0.0139)	0.5240	(0.0367)	0.5977
0.0	0.9320	(0.0440)	0.9115	(0.0306)	0.9902

Figure 5.1: The table compares the simulation results of the reduction rates of heterozygosity due to a single hitchhiking event to the theoretical results of (5.31) for $N = 10^4$, different values of α and different values of the initial frequency $Q^{(0)}$ of the neutral allele B_1. The results of (5.31) are shown in the column "Expected". The simulation values are based on 400 substitution events. The standard error is shown in parentheses.

values of α, the simulations indicate that the diffusion approximation holds only asymptotically.

The results of the diffusion approximation have been compared to prior models, more precisely to the deterministic model of Maynard Smith and Haigh (Chapter 3) and the stochastic model of Kaplan *et al.* (Chapter 4). Their results agree well with those of the coalescent approach of Kaplan *et al.* in the case of a single substitution as well as in the case of recurring substitutions. The deterministic model overestimates the reduction rate of heterozygosity as shown in Table 5.2.

Chapter 6

Conclusion

In 1974, Maynard Smith and Haigh [29] presented a pioneering work about genetic hitchhiking. It has been stimulated by experimental data (for example [6, 27]) that evidenced a surprisingly low level of heterozygosity in abundant species, contrasting the neutral theory [20]. Furthermore, these data showed a constant level of heterozygosity between different species. These facts indicated that there might be another variation-regulating process for neutral loci besides random genetic drift. To prove their contention, Maynard Smith and Haigh [29] developed a deterministic model that describes the loss of heterozygosity at a neutral locus due to a single selected substitution on a linked locus. At least for populations of size equal or greater than 10^6, the model successively confirmed that hitchhiking is more important for the regulation of heterozygosity than random genetic drift. The reduction rate is determined by the ratio $\frac{r}{s}$ and reaches its lowest values if the selective advantage, s, of the favourable mutant is large. In the special case of polymorphisms maintained by weak selection, Maynard Smith and Haigh showed that for $r \leq \frac{1}{N}$ there is a reasonably high chance of a total loss of heterozygosity in a small region around the selected mutation. Since the deterministic model omits stochastic effects that are caused by the finite population size and occur when the selectively favourable mutant is either rare or close to fixation, the deterministic approach leads to an underestimate of the hitchhiking effect.

Kaplan *et al.* [17] remedied the weaknesses of the deterministic model by presenting a stochastic model that considers not only the effects of a single strongly selected substitution but also the effect of recurring strongly selected substitutions. They applied the framework of coalescent theory and argued that the number of segregating sites in a sample of genes is determined by the sum of all selective substitutions that occurred since the MRCA. Therefore,

in their analysis of the hitchhiking effect, Kaplan *et al.* [17] focused on the genealogical history of a sample and the effects on segregating sites.

In the case of a single substitution, the results of the stochastic model support the conclusion of Maynard Smith and Haigh [29] that selective sweeps reduce genetic variation in neutral regions. Their results are quantitatively different, but qualitatively similar. In the case of recurring substitutions, Kaplan *et al.* showed that, even if the fraction of selected substitutions is negligible, hitchhiking reduces heterozygosity. In contrast to the neutral theory, the expected length of the ancestral tree and consequently the number of segregating sites is no longer directly proportional to the population size (see Figure 4.3). The maximal effect will be achieved in regions of restricted recombination.

The diffusion approximation of Stephan *et al.* [41] provides an alternative approach to the previous two cases of single and recurring substitutions. This was motivated by the desire to obtain a more general model about genetic hitchhiking that can directly be extended to the case of non-neutral loci. The model is based on a set of ordinary differential equations that describe the dynamics of the first- and second-order moments of the fractions of the deleterious allele B_1 on its respective backgrounds A_1 and A_2. These equations have been used to derive the reduction rate of heterozygosity. Their results agree well with those of Kaplan *et al.* [17]. A comparison of the three models is shown in Table 5.2.

The results of Kaplan *et al.* [17] and Stephan *et al.* [41] are concordant with a wide range of experimental data that prove a reduced nucleotide diversity in regions of restricted recombination. In 1989, Aguadè *et al.* [1], for example, observed a region in *Drosophila melanogaster*, where the 20-fold lower recombination rate per physical length compared to euchromatin corresponds to a tenfold reduction of average heterozygosity. Stephan and Langley [39] obtained similar results for the base of the X chromosome in *Drosophila ananassae*, where recombination is heavily restricted. Evidence of the importance of hitchhiking was not only found in *Drosophila* (see also [2, 4]) but in plants [40], mice [30], selfing plants [28] and humans [31].

Bibliography

[1] Aguadé, M. and Miyashita, N. and Langley, C.H. Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. *Genetics*, 122:607–615, 1989.

[2] Aquadro, C.F. and Begun, D.J. and Kindahl, E.C. Selection, recombination, and DNA polymorphism in Drosophila. *Non-neutral evolution: theories and molecular data*, pages 46–56, 1994.

[3] Barton, N.H. Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355:1553–1562, 2000.

[4] Begun, D.J., and Aquadro, C.F. Natural selection and recombination in *Drosophila*: Genetic hitchhiking in the distal portion of the X chromosome. *Genetics*, 129:1147–1158, 1991.

[5] Bürger, R. *Einführung in die mathematische Populationsgenetik*. Lecture notes, 2008.

[6] Clegg, M.T. and Kidwell, J.F. and Kidwell, M.G. and Daniel, N.J. Dynamics of correlated genetic systems. I. Selection in the region of the glued locus of *Drosophila melanogaster*. *Genetics*, 83(4):793–810, 1976.

[7] Crow, J.F. and Kimura, M. and others. *An Introduction to Population Genetics Theory*. Harper & Row, New York, 1970.

[8] Donnelly, P. *Interpreting genetic variability: the effects of shared evolutionary history*. 1996.

[9] Ewens, W.J. *Mathematical population genetics*, volume 1. Springer Verlag, 1979.

[10] Gillespie, J.H. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics*, 155(2):909–919, 2000.

[11] Gradshteyn, I.S. and Ryzhik, I.M. *Table of Integrals, Series, and Products*. Academic Press, New York, 1980.

[12] Griffiths, R.C. Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology*, 17:37–50, 1980.

[13] Griffiths, R.C. and Marjoram, P. An ancestral recombination graph. *Progress in Population Genetics and Human Evolution*, Donnelly, P. and Tavaré, S., eds. Springer Verlag, New York:257–270, 1997.

[14] Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201, 1983.

[15] Hudson, R.R. and Kaplan, N.L. The coalescent process in models with selection and recombination. *Genetics*, 120:831–840, 1988.

[16] Kaplan, N.L. and Darden, T. and Hudson, R.R. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.

[17] Kaplan, N.L. and Hudson, R.R. and Langley, C.H. The "hitchhiking effect" revisited. *Genetics*, 123:887–899, 1989.

[18] Kimura, M. Diffusion models in population genetics. *Journal of Applied Probability*, (2):177–232, 1964.

[19] Kimura, M. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[20] Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903, 1969.

[21] Kimura, M. and Ohta, T. *Theoretical Aspects of Population Genetics*. Princeton Universtiy Press, 1971.

[22] Kingman, J.F.C. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[23] Kojima, K.I. and Schaffer, H.E. Survival process of linked mutant genes. *Evolution*, 21:518–531, 1967.

[24] Krone, S.M. and Neuhauser, C. Ancestral processes with selection. *Theoretical Population Biology*, 51:210–237, 1997.

[25] Kurtz, T.G. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7:49–58, 1970.

[26] Kurtz, T.G. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8:344–356, 1971.

[27] Lewontin, R.C. *The Genetic Basis of Evolutionary Change*. Columbia University Press New York, 1974.

[28] Liu, F. and Zhang, L. and Charlesworth, D. Genetic diversity in *Leavenworthia* populations with different inbreeding levels. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1393):293–301, 1998.

[29] Maynard Smith, J. and Haigh, J. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23:23–35, 1974.

[30] Nachman, M.W. Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics*, 147(3):1303–1316, 1997.

[31] Nachman, M.W. and Bauer, V.L. and Crowell, S.L. and Aquadro, C.F. DNA variability and recombination rates at X-linked loci in humans. *Genetics*, 150(3):1133–1141, 1998.

[32] Neuhauser, C. and Krone, S.M. The genealogy of samples in models with selection. *Genetics*, 145:519–534, 1997.

[33] Nordborg, M. Coalescent theory. *Handbook of Statistical Genetics*, pages 843–877, 2007.

[34] Norman, M.F. A central limit theorem for Markov processes that move by small steps. *The Annals of Probability*, 2:1065–1074, 1974.

[35] Ohta, T. and Kimura, M. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, 63:229–238, 1969.

[36] Ohta, T. and Kimura, M. The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). *Genetical Research*, 25:313–325, 1975.

[37] Schneider, K. *Mathematical population genetics*. Lecture notes, 2011.

[38] Simonsen, K.L. and Churchill, G.A. A Markov chain model of coalescence with recombination. *Theoretical Population Biology*, 52:43–59, 1997.

[39] Stephan, W. and Langley, C.H. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. *Genetics*, 121:89–99, 1989.

[40] Stephan, W. and Langley, C.H. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. *Genetics*, 150(4):1585–1593, 1998.

[41] Stephan, W. and Wiehe, T.H.E. and Lenz, M.W. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology*, 41:237–254, 1992.

[42] Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

[43] Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7:256–276, 1975.

[44] Wiuf, C. and Hein, J. Recombination as a point process along sequences. *Theoretical Population Biology*, 55:248–259, 1999.

List of Figures

Curriculum Vitae

Personal Data

Name:	Theresa Dvorak BSc
Date of Birth:	14^{th} of April, 1986
Place of Birth:	Vienna, Austria
Citizenship:	Austria

Education

since 10/2011	**Technical University of Vienna**, A-1040 Vienna Master Studies in Architecture
10/2004-08/2012	**University of Vienna**, A-1090 Vienna Diploma Studies in Mathematics
10/2006-07/2011	**Technical University of Vienna**, A-1040 Vienna Bachelor Studies in Architecture, passed with distinction (BSc)
09/2008-07/2009	**Università di Pisa**, I-56126 Pisa Diploma Studies in Mathematics (Erasmus)
10/1996-06/2004	**BG Bernoullistraße 3**, 1220-Vienna

Work Experience

since 09/2011	**Hafen Wien**, A-1020 Vienna Project developer, Architect, Graphic Designer
02/2011-08/2011	**ÖISS**, Österreichisches Institut für Schul- und Sportstätten, A-1040 Vienna Author, "School Architecture in Austria from 1995-2010" on behalf of the BMUKK
04/2011-09/2011	**BFI**, Berufsförderungsinstitut Wien, A-1020 Vienna Tutor in Mathematics
12/2009-05/2011	**IFS**, Institut für Studentenkurse, A-1070 Vienna Tutor in Mathematics