DE GRUYTER

Igor M. Rouzine

MATHEMATICAL MODELING OF EVOLUTION

VOLUME 1: ONE-LOCUS AND MULTI-LOCUS THEORY AND RECOMBINATION

SERIES IN MATHEMATICS AND LIFE SCIENCES 8/1

Igor M. Rouzine

Mathematical Modeling of Evolution

De Gruyter Series in Mathematics and Life Sciences

Edited by
Anna Marciniak-Czochra, Heidelberg University, Germany
Benoît Perthame, Sorbonne-Université, France
Jean-Philippe Vert, Mines ParisTech, France

Volume 8/1

Igor M. Rouzine

Mathematical Modeling of Evolution

Volume 1: One-Locus and Multi-Locus Theory and Recombination

DE GRUYTER

Mathematics Subject Classification 2010
Primary: 55XX; Secondary: 60XX

Author
Dr. Igor M. Rouzine, Laboratory of Computational and Quantitative Biology
Institut de Biologie Paris Seine
Sorbonne Université
15404 Case courrier
75005 Paris
France
igor.rouzine@sorbonne-university.fr

ISBN 978-3-11-060789-5
e-ISBN (PDF) 978-3-11-061545-6
e-ISBN (EPUB) 978-3-11-060819-9
ISSN 2195-5530

Library of Congress Control Number: 2020943697

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Preface

The book will benefit readers with a background in population genetics, physical sciences, and applied mathematics, and the interest in mathematical models of genetic evolution. In the first chapter, we analyze several thought experiments based on a basic model of stochastic evolution of a single genomic site in the presence of factors of random mutation, directional natural selection, and random genetic drift. In the second chapter, we present a more advanced theory for a large number of linked loci. In the third chapter, we include the effect of genetic recombination into account and find out the advantage of sexual reproduction for adaptation.

https://doi.org/10.1515/9783110615456-202

Contents

Chapter 1
Basic theory of one-locus evolution

1.1 Introduction

Evolution of organisms occurs due to mutation, selection, recombination, and chance. The changes from one species to another are not sudden, but rather lead to a gradual observable change in the genetic composition of a population of organisms. Since Charles Darwin, experimental and theoretical studies researched the factors of the evolution of various organisms and produced an enormous body of the literature.

In early 1990s, biologists became interested in special problems of virus evolution. The reasons for this interest were as follows. First, the researchers wanted to learn how modern viruses emerged from earlier viruses, both recently and during the long-term coevolution with their hosts. Second, evolution of a virus within a single host or at the level of a host population is capable of creating new populations of viruses with changed properties allowing them to deceive the immune response, acquire the resistance to antivirals, or become more or less virulent. Third, because their replication rates are high, population sizes vary in a broad range, and mutation rates are high, a virus makes an excellent experimental model for testing mathematical models of evolution.

Biological factors dominating the evolution of human immunodeficiency virus (HIV) during a persistent human host infection received a lot of attention at that time. HIV, along with hepatitis C virus (Simmonds, 2004), displays huge genetic variation as well as a very high speed of evolution. In the most variable regions, individual genomes isolated from an infected person can vary by as much as by 3–5% (Balfe et al., 1990; Lamers et al., 1993; Wolfs et al., 1990). The rate of substitutions in the envelope gene is approximately 1% per year (Shankarappa et al., 1999). Due to this variation, the virus can change to infect various organs and tissues (Chavda et al., 1994; Groenink et al., 1992; Keys et al., 1993) and to quickly become resistant to antiviral drugs (Cleland et al., 1996; Lopez-Galindez et al., 1991). Evolution of a virus plays a major role in evading the immune system (Burns and Desrosiers, 1994; Nietfield et al., 1995; Rouzine and Rozhnova, 2018; Takahashi et al., 1989; Wolfs et al., 1991). Furthermore, RNA viruses have relatively high mutation rates, $10^{-6} - 10^{-4}$ per site per generation. Average mutation rate of HIV is approximately $3 \cdot 10^{-5}$ per nucleotide site per replication cycle (Mansky and Temin, 1995), which is much higher than mutation rate ~10^{-9} observed for organisms. Population sizes of viruses vary from one infected cell to 10^{13}. HIV population size in an average untreated patient was estimated to be in the range between 10^7 and 10^8 infected (HIV RNA positive) cells (Haase, 1999). Evolutionary estimates confirm this estimate showing the effective population size of $10^5 - 10^6$ infected cells or more (Pennings et al., 2014; Rouzine and Coffin, 1999a; Rouzine et al., 2014). Another

https://doi.org/10.1515/9783110615456-001

important feature of persisting viruses as HIV and HCV is a continuous steady state within a host. It has been shown that the large majority of productively infected cells dies and is reinfected every day (Ho et al., 1995; Wei et al., 1995). In comparison, common cold and influenza viruses are cleared from patients rapidly and persist only at the level of a population, also due to continuous evolution (Bedford et al., 2015; Luksza and Lassig, 2014; Rouzine and Rozhnova, 2018; Smith et al., 2004). These problems demand mathematical modeling of evolution.

A large toolbox of modeling has been applied to understand evolution of viruses and organisms. These methods fall into two classes: population genetics, which studies the fist-principle mechanisms, and the descriptive methods of statistical genetics. The methods of population genetics developed first were based on one of the two different theoretical frameworks, as follows. First approaches were deterministic, made for infinite population size. Quasi-species theory (Eigen and Biebricher, 1988; Holland et al., 1992) assumed that the frequency of a given allele (genetic variant) at any time is predictable given the initial frequency, the mutation rate, and the selection coefficient. Selection coefficient is defined as the relative fitness difference between the different alleles. One might think that such approaches are justified, at least, for viruses, due to the large number of infected cells at each generation (Haase, 1999). Nevertheless, a number of factors, such as initial establishment of alleles in a population subject to random genetic drift (Chapter 1) aggravated by interference between different alleles (Chapter 2), make stochastic effects surprisingly strong even for extremely large populations. Neutral stochastic models that neglect selection proceed from the opposite assumption: that either the population size is very small, or selection is weak, so that the random drift completely dominates over selection.

Indeed, such "neutral" mutations are very important in the evolution of higher organisms where populations are small, genomes have many untranslated regions (introns) (Kimura, 1989). However, their applicability to virus populations – and to many coding regions in organisms – is fairly limited. Many of the assumptions of neutral theory are not appropriate when there is an uneven ratio of synonymous to nonsynonymous changes in a region of the genome (Burns and Desrosiers, 1994; Lech et al., 1996; Lukashov et al., 1995). Such regions clearly argue against the universal application of neutral theory. Twenty years ago, the inclusion of selection effects acting at many linked loci and recombination into the evolutionary analysis presented a serious mathematical challenge. In this book, we relate some of the progress achieved during these two decades.

To illustrate how deterministic and stochastic approaches differ from each other, consider the fate of an allele slightly deleterious to the ability to reproduce. In other words, this mutated allele decreases the progeny number, which is the definition of Darwinian fitness. In a deterministic system, one can demonstrate that the allelic frequency in the population will eventually stabilize at a small level equal to the mutation rate per generation divided by the selection coefficient (Haldane, 1927). The outcome is different in a stochastic system: the population

will be sometimes completely uniform in deleterious allele, and sometimes uniform in wild-type allele (Watterson, 1975), switching occasionally from one to another. This distinction is important practically, because it concerns a mutation that can, for example, make a virus resistant to an antiviral drug even before treatment, and then amplified in number by that drug (Coffin, 1995). It can occur as a single or several mutations (Hermisson and Pennings, 2005). This scenario is equally relevant for the dynamics of trait diversity in animal populations.

To solve such a problem and many other problems, a general theory was needed that would take into account both selection and random genetic drift, as well as interference between many evolving sites. The aim was to develop, from the first principles, a more general theory that would include all these effects. Below, we describe a model that is appropriate to haploid virus populations and to diploid animal populations for the case when evolution is not neutral, no allelic dominance, and a directed selection takes place. We focus on the struggle between deterministic and stochastic behavior as it occurs in various thought experiments.

In this chapter, based on review (Rouzine et al., 2001), we start from the simplest possible model of that kind: a single genomic site (locus) with only two possible alleles, which evolves under the influence of constant selective pressure in a well-stirred population. In Chapter 2, we show that the simultaneous evolution of multiple loci in asexual populations leads to additional complications due to interference between loci. We do not consider recombination explicitly until Chapter 3, where we analyze how it offsets the effect of interlocus interference. The impact of epistasis (the biological interaction between loci) will be analyzed in the next volume of this book. Although our focus is on haploid populations, the results are directly relevant for diploid populations of animals in the cases where allelic dominance and epistasis are negligible.

Although recombination is not considered explicitly, very strong recombination is implied in this model, for the approximation of an isolated locus to be correct. Also, evolving loci must be spaced sufficiently far apart in the genome, depending on the recombination rate. However, even in the absence of recombination, the one-locus approximation is a useful start for understanding the interaction between selection and stochastic factors qualitatively. Below we introduce a model that is applicable, in principle, at any population size, mutation rate, and selection strength. Despite its simplicity, the model makes a variety of useful predictions. In the extreme limits of small and large population size, predictions cross over to the standard results of deterministic or neutral theory. We demonstrate the existence of a very broad parameter region where evolution exhibits mixed behavior: under certain conditions, stochastic factors win, while in other cases, dynamics is nearly deterministic.

1.2 One-locus model and the Fokker–Planck equation

Before starting our mathematical adventure, a model based on biological knowledge must be developed. Below we describe a model and obtain an equation for the evolutionary dynamics including selection and stochastic factors. We explain the main factors of evolution, as well as the mathematical representation of the model in the form of a diffusion equation. Then, we will state the boundary conditions for this equation, which represent populations that are nearly uniform genetically.

1.2.1 Population model

Let us consider the evolution of a single nucleotide position (site, locus) assuming that each nucleotide has a choice between two variants termed in genetics "alleles." Such a model can be used for a real genome with multiple loci only if the evolving loci are sufficiently distant in DNA, and recombination with other genomes due to sexual reproduction (yes, some viruses have it too) is sufficiently frequent to make their evolution independent. In the absence of efficient recombination, dynamics of close loci interfere with each other, as described in Chapter 2. Following a convention, we will call the more fit allele "wild type" and the less fit allele "mutant." The locus has two constant, small parameters (both much less than 1): the mutation rate per site per generation, μ, and selection coefficient s equal to the relative difference in offspring number (fitness) between the two alleles. The site and mutations at other genomic sites have multiplicative contribution to fitness. That is to say, we neglect epistasis with other sites in genome which exists due to biological interaction between nucleotides of RNA or DNA. In this chapter, we also neglect linkage disequilibrium between loci and assume that different nucleotides evolve independently, which imply sufficiently strong recombination with other genomes. Linkage effects will be considered in Chapter 2, and epistasis in Volume 2. The mutation rate is assumed to be the symmetric in both directions. We will assume it to be the same for all substitutions: A, C, T, G to A, C, T, G. We do not include insertions and deletions (which can be described as asymmetric mutations). The selection coefficient, in real organisms, can vary over a wide margin across sites and depends on the external conditions, but here, for the sake of simplicity, it is assumed to be constant.

The model (Figure 1.1) includes the dynamics of a cell population comprised of two alleles: a fraction, f, of individuals carrying a mutant allele, and the remaining individuals, $1-f$, carrying the wild type allele (Figure 1.1A). The number of each type of individuals (for a virus, infected cells) changes with time, that is, with each new generation. The total number of individuals can vary in real life, but we will assume it to be constant. In each generation, a fixed (large) number of offspring is produced by each individual. Then, every old generation dies out and is replaced by a new generation. The number of offspring capable of establishing a new generation

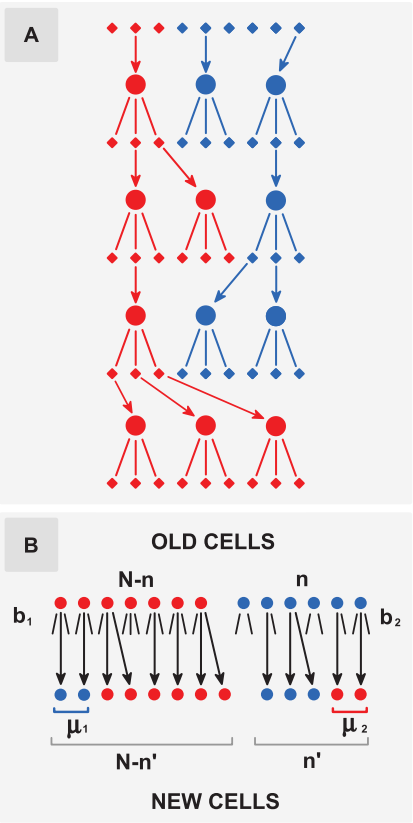

Figure 1.1: Factors of evolution. (a) Genetic drift due to random progeny sampling. Circles and small diamonds: individuals and their progeny. Red and blue denote two alleles. (b) A population model including random drift, selection, and mutation. Two generations are shown. Arrows: progeny which creates a new generation. Mutants (blue) yield fewer progeny per individual than wild-type individuals (red). A small fraction of progeny, μ_1 and μ_2, mutate to the other allele (based on Rouzine et al. (2001)).

differs by a factor of $1-s$ between the two alleles, creating selection for the allele with higher fertility. Because the total population number is fixed, and the offspring number per individual is assumed large, only a small fraction of offspring can establish the next generation. When producing an offspring, it can mutate into the opposite allele with a small probability μ.

Fine model details, such as a fixed number of offspring per individual of each type and the point of the reproduction cycle at which mutation occurs, are not important on the long timescales. We also neglected the time overlap between generations, but it may cause only a change in the rate of random drift factor of 2, which can be absorbed by rescaling the population size (Moran, 1958). By contrast, such assumptions as two alleles per site, the lack of interaction with the other sites in genome and constant directional selection are essential for the results.

The model includes three essential factors of evolution: natural selection, stochastic drift, which exists due to random sampling of progeny, and random mutation. We

now describe briefly the effect of each of these factors and how they affect the composition of the population, as it changes in time:

(i) *Random genetic drift*. The offspring making the new generation is chosen randomly from mutants and wild type. Because of this random sampling of progeny, the allelic frequency exhibits random diffusion in time (Fisher, 1922; Wright, 1931), termed "random drift" (Figure 1.1A). If mutation and selection are absent, an initially diverse population made of an allelic mixture eventually becomes either uniformly wild type or uniformly mutant (see further).

(ii) *Natural selection*. The difference in the number of progeny from individuals with different alleles creates natural selection. As we demonstrate further, selection alone would cause extinction of the less fit allele and drive the system into a uniformly better-fit population.

(iii) *Random mutation*, in contrast to drift and natural selection, diversifies the population. In the absence of two other factors, mutation would bring the system into an equilibrium composition in which forward and reverse mutations balance each other. We assume here that the forward and reverse mutation rates are equal, so that this equilibrium is reached when each allele makes a half population.

In the presence of all three factors, the population will arrive at a steady state where mutation compensates the homogenization effect of selection and random drift. The statistical properties of the population no longer vary in time in the steady state. Although the allelic frequency will fluctuate in time, the momenta of state variables, including the average and the standard deviation, will remain constant.

The full model including the three factors (Figure 1.1B) considers an asexual haploid population of N individual genomes (or diploid population without allelic dominance of $N/2$ individuals) comprising two alleles: n individuals are mutant (less fit), and $N-n$ individuals are wild type (better fit). The total population size N is assumed constant, but mutant number $n(t)$ changes with time t. The frequency of a mutant allele is defined as $f = n/N$. In each generation, a mutant individual produces b_1 mutant offspring, and each wild-type individual produces b_2 wild-type offspring (Figure 1.1B). After reproduction, the parents die of the old age. In the case of virus population, infected cells can die from viral effects or the immune response. The average number of offspring per parent, b_1 and b_2, is assumed to be large, $b_1 \gg 1$, $b_2 \gg 1$ and differs a bit between the two alleles, $b_1 = b_2(1-s)$, where selection coefficient s, such that $s \ll 1$, reflects the small difference in replication ability between wild type and mutant. We also assume that population is well stirred and that the total population N stays constant. Each offspring can mutate into the opposite allele with a small probability μ, $\mu \ll 1$. The population model is a particular case of the Wright–Fisher process. As we noted previously, this model is equally suitable for haploid populations and diploid populations containing two copies of each gene, as long as allelic dominance is weak.

1.2.2 Stochastic evolution

The word "evolution" can be assigned different meaning. For an evolutionary biologist and anthropologist, evolution is about the origin of species and development of organs. Our focus here will be on much shorter timescales and dynamics of the genetic composition of a population. In deterministic dynamics, which applies in very large populations of infected cells, if one knows the initial mutant frequency and has the appropriate equations, one can, in principle, predict the mutant frequency at later times with arbitrary precision. In practice, these equations are never known exactly, since there are too many factors to include, but this is a separate issue [see an example of model selection in Rouzine and Coffin (1999b)]. In contrast, in a limited population, due to the presence of drift and random mutation, one cannot predict the time dependence of the mutant fraction forward except for a very short time. Even if the precise initial value is known, the error of that prediction increases in time. If random factors are strong enough, and they often are the error in the allelic frequency and its mean predicted value become eventually of the same order of magnitude. Thus, the evolution of the genetic composition, although directed by natural selection, is a stochastic process.

Randomness of evolutionary process does not mean, however, that it is completely arbitrary. Very useful predictions are possible about statistical momenta, even if a specific trajectory for the frequency of mutants f cannot be predicted. Instead of the trajectory $f(t)$, one can trace the probability density, $\rho(f,t)$. By the definition, $\rho(f,t)df$ is probability that a population has a mutant frequency within the interval at time t. This quantity that can actually be measured using DNA sequences from parallel evolving populations.

The probability density can be approximated by a histogram made of bins corresponding to the number of times the mutant frequency is observed to fall within a certain range of values. In the limit where both the number of data points and the number of histogram bins are large, the histogram approaches a smooth function, which is the probability density $\rho(f,t)$ up to a constant prefactor. The normalization integral $\int df\rho(f,t)$ is equal to 1. The probability density function informs statistical parameters, such as the mean value and standard deviation, which can be experimentally tested (Section 1.3). For example, the characteristic half-width of the probability density indicates the error of the prediction of mutant frequency.

If we know its form at the present moment, the evolution equation (Figure 1.2A) determines how the probability density changes in time. Knowing the initial probability density $\rho(f,0)$, this equation can be used predict its form at any time in the future, just us the mutant frequency itself would be predicted in a deterministic process. The essential difference is that the state variable is no longer a scalar but a function of f. In the next subsection, we will obtain the master equation from the population model using Markov chain formalism. The rest of the chapter is dedicated to solving this equation for different various conditions and thought experiments.

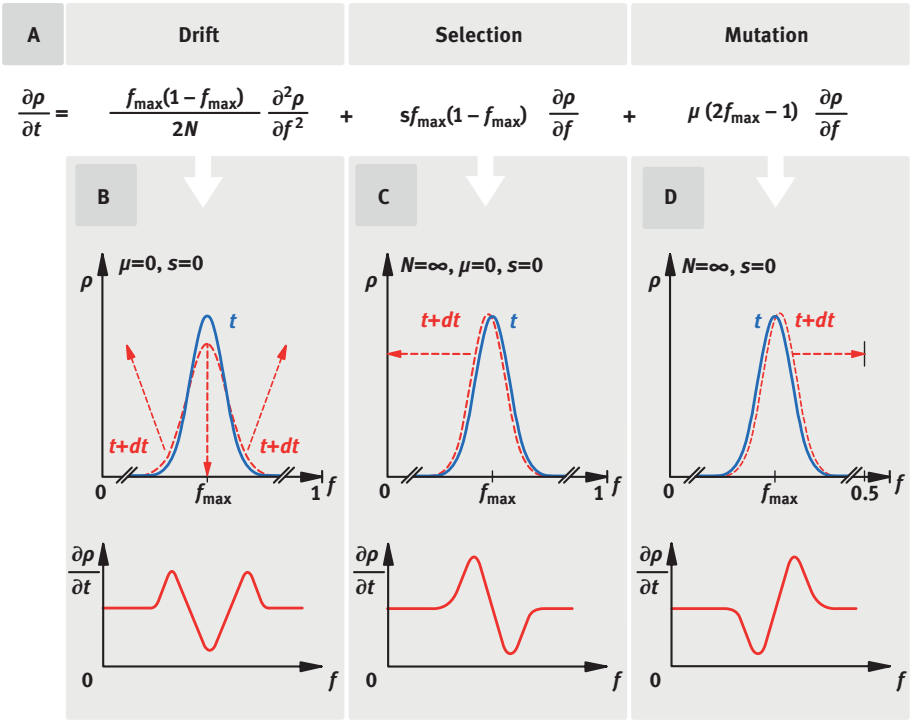

Figure 1.2: Illustration of the stochastic evolution equations (1.1) and (1.2). (A) The equation in the particular case, where the probability density of allelic frequency, $\rho(f,t)$, is a narrow peak. (B to D) Local changes in $\rho(f,t)$ in short time dt corresponding to each part of the right-hand side of the equation in (A). The top row shows the resulting effects: spread (B) and shift (C and D) of the peak. Bottom row: the time derivatives. Blue solid and red dashed lines: two adjacent moments in time (based on Rouzine et al. (2001)).

Figure 1.2A presents the general form of master equation. We will comment on its qualitative meaning. The right hand-side of the master equation is a sum of three terms, which together determine how $\rho(f,t)$, varies in a short time interval, dt (Figure 1.2A). The first term accounts for the effect random drift, the second term includes natural selection, and the third corresponds to random mutation. To explain their respective roles, we consider each term separately, by neglecting the other two terms (Figure 1.2B to D). For an example, we look at a narrow peak of $\rho(f,t)$, around some value f_{max}. Here the second, selection term forces an increase in the probability density to the left from the peak and a decrease to the right . The combination of these changes shifts the probability density to lower mutant frequencies (Figure 1.2C). Thus, the mutant is being selected against, as it should. Term 3 in the equation, corresponding to mutation, also causes a shift, but this time toward 50% composition. This is what mutation with symmetric mutation rates is supposed

to do (Figure 1.2D). Finally, the first term in the equation (drift, diffusion) does not cause a shift. It has a different effect. Due to its presence, the probability density decreases near the maximum and increases in the tails (Figure 1.2B), causing the probability density spread outward. This widening implies that the accuracy within which one can predict the value of mutant frequency decreases. A more general form of the stochastic equation when the probability density, $\rho(f)$, is not necessarily localized in a narrow interval of f, is given in eqs. (1.1) and (1.2).

In the equation (Figure 1.2A), a physicist will recognize Fokker–Planck equation and a mathematician will recognize the forward Kolmogorov equation (Kolmogorov, 1931). This formalism was pioneered in the field of population genetics by Wright (1945) and then used by Kimura (1954, 1955a,b) and Kimura (1994) to study evolution in different situations. As follows from these studies, this diffusion equation is much more general than the model we will use for its derivation in Section 1.2.3. In fact, it is quite versatile and can describe various populations without allelic dominance (Kimura, 1964). Originally, the Fokker–Planck equation was employed in evolution theory based on the analogy with gas kinetics, which we will make use below (Fisher, 1922). Later, its broad applicability was confirmed for different population types (Maynard Smith, 1971; Watterson, 1975). As we already mentioned, the equation is extremely simple and does not include many acting factors of evolution. Depending on the biological scenario, the following factors may or may not be important: epistasis, interference between linked loci, variable-in-time selection coefficient, time-dependent population size, and allelic dominance (Kimura, 1955b). We will address the effects of multi-site linkage in Chapter 2. Epistasis, allelic dominance, and time-dependent selection (e.g., due to the immune response) will be considered in the next volume of this book.

As we mentioned, a good *mathematical analogy* for the evolution equation is a one-dimensional gas of variable density. Assume that the gas is mixed with air and restricted between two walls (Figure 1.3A). Then, the mutant frequency is similar to the coordinate between the walls, and the probability density is similar to the local gas density. Then, the term with the second derivative of the master equation (Figure 1.2A) accounts for the diffusion of the gas particles in the air, and the other two terms combined introduce a force acting on the gas particles in the presence of friction (for example, electric field). Importantly, the coefficient of diffusion depends on the coordinate as $f(1-f)$, as if the air has variable density. Another useful analogy is gel electrophoresis of proteins in a gel matrix. The electrostatic force on the charged polymer molecules and the force of friction from the gel matrix make them move and segregate into bands whose location depends on the length of a molecule. Molecular diffusion leads to finite bandwidths of molecules of the same lengths increasing in time. Although the electrophoresis or the gas system are in no way related to the biology of reproduction or to their evolution, as we shall see further, this formal analogy between the two systems turns out to be very useful for understanding mathematical results.

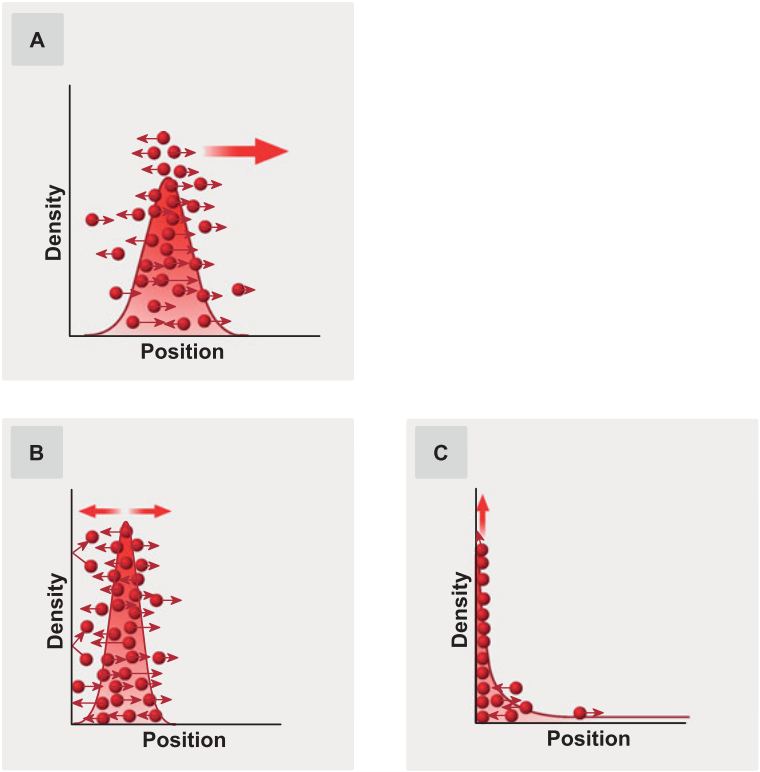

Figure 1.3: Illustration of stochastic evolution equation using the formal analogy between the probability density and gas with a variable local density. *X*-axis: Frequency of deleterious alleles analogous to a position between walls. The walls at $f = 0,1$ correspond to the two uniform population, wild type (better fit) and mutant (less fit). (A) Gas particles are subject to diffusion and a directed force when they are far from the walls. (B, C) Boundary conditions. (B) At very large population sizes, the flux at a wall vanishes, eq. (1.2). (C) At small population sizes, gas particles can condense on a wall, and the total flux at a wall does not need to be 0, eqs. (1.5) and (1.6) (based on Rouzine et al. (2001)).

In Section 1.2.3, we present the master equation for evolution and complement it with boundary conditions. In Section 1.2.4, we consider a Markovian process for the virus population model; in Section 1.2.5, we simplify it using some realistic approximations to obtain a continuous diffusion equation; and in Section 1.2.6, we determine the boundary conditions. In Sections 1.3 to 1.7, we will analyze equations and the boundary conditions in various parameter regions under different initial conditions. Notation used in Chapter 1 is shown in Table 1.1.

Table 1.1: Mathematical notation in Chapter 1.

Symbol	Definition
A, B, C, F	Undetermined constants of functions
D	Relative interpop. distance per site
$\delta(x)$	Dirac delta function of x
δ_{ij}	Kronecker symbol
f	Mutant frequency
G	Gene fixation probability
g	Continuous part of probability density
K	Time correlation function of f
μ	Mutation probability per site per gen.
M	Mean change in f per generation
N	Population size
n	Number of mutant individuals
P_n	Probability of having n mutants
p_0	Probability of a pure wild-type state
p_1	Probability of a pure mutant state
q	Probability density flux
ρ	Probability density of f
s	Selection coefficient
T	Intrapopulation distance per site
t	Time (generation number)
V_x	Variance of x
x	Any parameter
$\bar{x}$	Expectation value (mean) of x
x_{ss}	Value of x in steady state

1.2.3 Evolution equation

The evolution of the virus population is described by a set of differential equations and boundary conditions, which come in two versions, as follows. The choice of the version depends on the population size interval. The first case shows very large and

diverse populations, when many copies of both wild type and mutant are present and $\mu N \gg 1/\log(N)$. The master equation and the boundary conditions have the following form (Kimura, 1955a):

$$\partial\rho/\partial t = -\partial q/\partial f \tag{1.1}$$

$$q(f,t) = -\frac{1}{2N}\frac{\partial}{\partial f}[f(1-f)\rho] - sf(1-f)\rho - \mu(2f-1)\rho \tag{1.2}$$

$$q(f,t)_{f\to 0} = q(f,t)_{f\to 1} = 0 \tag{1.3}$$

Equations (1.1) to (1.3) are valid under the conditions of weak selection, $s \ll 1$, rare mutation, $\mu \ll 1$, and large population size, $\mu N \gg 1/\log N$. In this case, genetic composition changes very slowly, and time and mutant frequency can be approximated with continuous variables. The effect of each term in the right-hand side of eq. (1.2) is illustrated in Figure 1.2.

To continue our gas analogy (Figure 1.3), eq. (1.1) represents the equation of detailed balance. It states that the number of gas molecules is conserved i.e., molecules cannot be born or vanish but can only travel from one place to another. The value $q(f,t)$ defined in eq. (1.2) has the meaning of the probability flux, which is analogous to the flux of gas molecules across a unit area. Thus, eq. (1.1) states that the probability density of allelic frequency, like mass density, is conserved locally. The boundary conditions in eq. (1.3) show that the flow cannot cross the boundaries of the interval in f, just like gas particles cannot cross crossing the confining walls (Figure 1.3B).

As we demonstrate in Section 1.2.6, in small populations, $\mu N \ll 1/\log(1/\mu)$, the population has a significant chance to be genetically uniform, when $f = 0$ or 1. Such a state is analogous to the gas condensate at a cold wall. In this regime, one has to isolate two Dirac delta-function terms from the continuous probability density:

$$\rho(f,t) = p_0(t)\delta(f) + p_1(t)\delta(1-f) + g(f,t) \tag{1.4}$$

where p_0 and p_1 are the probabilities of a population being uniformly wild-type and mutant states, respectively, and $g(f,t)$, such that $f(1-f) \gg 1/N$, is the polymorphic part.

The boundary conditions for these equations have a form

$$\frac{dp_0}{dt} = -q(0,t), \quad \frac{dp_1}{dt} = q(1,t), \quad N\mu\log N \ll 1 \tag{1.5}$$

$$2N\mu p_0 = [fg(f)]_{f\to 0}, \quad 2N\mu p_1 = [(1-f)g(f)]_{f\to 1} \tag{1.6}$$

Boundary conditions (eq. (1.5) describe the dynamics of monomorphic state probability linked to probability flux (analogous to gas condensation or evaporation from the wall, Figure 1.3C). They represent a simple conservation law. The second set of eq. (1.6) shows how mutation can transition between a monomorphic state, $f = 0$ or 1, and a single-copy diverse state, $f = 1/N$ or $(N-1)/N$. Equation (1.6) will be derived from the population model in Section 1.2.6.

The continuous part of the probability density, $g(f,t)$, for small population sizes follows the equation

$$\partial g/\partial t = -\partial q/\partial f$$

$$q(f,t) = -\frac{1}{2N}\frac{\partial}{\partial f}[f(1-f)g] - sf(1-f)g, \quad N\mu \log N \ll 1 \tag{1.7}$$

The reader will notice that, unlike in the expressions at large N, eqs. (1.1) and (1.2), the mutation term is absent. The mutation rate is present only in the boundary conditions, eq. (1.6).

Equations (1.5) to (1.7) are valid in populations of moderate size. We can easily estimate the upper bound on N from the boundary conditions of eq. (1.6). The total probability of having a diverse state, by the definition, is $\int_0^1 g(f)df$. As it follows from eq. (1.6), $g(f)$ diverges near the boundaries, where it is given by $g(f) \approx 2\mu N p_0/f$ and $g(f) \approx 2\mu N p_1/(1-f)$, respectively. The main contribution to the integral of $g(f)$ comes from a small area near borders $f \approx 1$ or 0 and has truncated at $f(1-f) \sim 1/N$, which corresponds to a single copy of either mutant or wild-type allele. Hence, the probabilities of monomorphic states become small, at $\mu N \log N \gg 1$.

As we already mentioned, the validity of these equations does not depend on fine details of reproduction. Many other one-locus, two-allele populations are amenable to eqs. (1.1) and (1.2), as long as they are controlled by the same dominant processes: mutation, directed selection, and random sampling of progeny. If additional factors come on stage, for example, allelic dominance or time fluctuations of selection coefficient, this method can be generalized (Kimura, 1955b). For example, the approach can be generalized for multiple loci using a system of equations with the number equal to the number of haplotype frequencies minus one (Kimura, 1994). In reality, such a generalization method becomes very impractical starting with three loci. We address multi-locus evolution in Chapter 2. Now we will derive the cited equations from the virus replication model described in Section 1.2.

1.2.4 Derivation from a Markovian process

We denote the probability of having n mutant cells at time t, where t is the number of a generation, and n can change from 0 through N, as $p(n,t)$. This system is formalized by a Markovian process:

$$p(n,t+1) = \sum_{n'=0}^{N} P(n|n')p(n',t) \tag{1.8}$$

where $P(n|n')$ is the probability of finding n mutants, given that their number at the previous step is n'. Below, we derive the form $P(n, n')$ for the population model introduced in Section 1.2.1.

First, we obtain the conditional probability $P(n|n')$ in eq. (1.8), in the absence of mutation and denote it $P_0(n|n')$. If the number of mutants in a generation is n', according to the model, the total offspring numbers produced by mutant and wild-type individuals are

$$B_1 = b_1 n', \quad B_2 = b_2(N - n') \tag{1.9}$$

respectively. The offspring numbers per individual, b_1, b_2, are related as

$$b_1 = b_2(1 - s) \tag{1.10}$$

where s is the small selection coefficient. We will consider animals or viruses with a large progeny number, $b_1, b_2 \gg 1$. If n is the number of new mutants, then the numbers of progeny that create a new generation must be n and $N - n$, respectively. The rest of progeny dies before maturity or does not perpetuate. The probability of n new mutants, $P_0(n|n')$, is proportional to the number of possible ways: one can choose n successful mutants from B_1 possible and $N - n$ wild-type individuals from B_2 possible

$$P_0(n|n') = A \frac{(n')^n (N - n')^{N-n} (1 - s)^n}{n!(N - n)!} \tag{1.11}$$

where we used eqs. (1.9) and (1.10). Factor A is determined by the normalization condition, $\sum_n P_0(n|n') = 1$.

Now we include mutation between the two alleles. Suppose that m_1 deleterious and m_2 beneficial mutations occur in n mutants and $N - n$ wild-type individuals, respectively (Figure 1.1B). The resulting number of mutant-infected cells, n'', will be $n'' = n + m_1 - m_2$. The probability of having m_2 beneficial mutations among n successful mutants is given by the Poisson formula with average μn

$$\pi(m_2|n) = \frac{(\mu n)^{m_2}}{m_2!} e^{-\mu n}, \quad m_2 = 0, 1, \ldots \tag{1.12}$$

(This formula is true if $n \gg 1$. Otherwise, eq. (1.12) still can be used for $m_2 = 0$ and 1, which are the only relevant values in this case, because mutation rate is very small $\mu \ll 1$.) Likewise, the probability of having m_1 forward mutations is $\pi(m_1|N - n)$. Hence, for the conditional probability $P(n''|n')$, we get

$$P(n''|n') = \sum_{n=0}^{N} \sum_{m_1=0}^{N-n} \sum_{m_2=0}^{n} \delta_{n'', n+m_1-m_2} \pi(m_1|N - n) \pi(m_2|n) P_0(n|n') \tag{1.13}$$

Here, kernel in the absence of mutation $P_0(n|n')$ is given by eq. (1.11), and Kronecker–Ricci tensor is defined as $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

1.2.5 Diffusion limit

The discrete evolution equation given by eqs. (1.8), (1.11), and (1.13) is convenient for numeric simulation, but quite difficult for analytic treatment. Also, it has imbedded model-dependent details, which do not show in large populations, $N \gg 1$, on long timescales $t \gg 1$ and should be discarded. Focusing on the case when both mutant and wild type have many copies ($n \gg 1$ and $N - n \gg 1$), eq. (1.8) can be transformed to a more convenient and less model-dependent differential form. Since s and μ are small, the conditional probability, $P(n|n')$, changes slowly in n, n', as given by

$$\left|P(n+1|n') - P(n|n')\right| \ll P(n|n'), \quad \left|P(n|n'+1) - P(n|n')\right| \ll P(n|n')$$

Then, matrix $P(n|n')$ can be approximated by a gradual function of its arguments n, n'. Substituting the Stirling formula $n! \approx (2\pi n)^{1/2}(n/e)^n, n \gg 1$, into eq. (1.12), we find that $P_0(n|n')$ has a maximum at $n' \approx n$, and that the maximum is narrow. Next, rewriting $P_0(n|n') = \exp\left[\log\left(P_0(n|n')\right)\right]$ and approximating the argument of the exponential with the second-order Taylor expansion in $n' - n$, we get

$$P_0(n|n') = A \exp\left\{-\frac{\left(n - n' + sn'\left(1 - n'/N\right)\right)^2}{2n'\left(1 - (n'/N)\right)}\right\} \tag{1.14}$$

The characteristic half-width of this function in $|n - n'|$ is much more narrow than the n' but larger than unit, as given by $1 \ll |n - n'| \ll \min(n', N - n')$, which confirms our above assumption that $P_0(n|n')$ can be considered a smooth function.

The last paragraph was based on a derivation with a large parameter. The derivation depends on the assumption valid in many real life cases that the values of $n, N, 1/s, 1/\mu$ are all much larger than 1. In this approach, we made a hypothesis that the function $P_0(n|n')$ is slow in both arguments and will confirm its validity, as well as determine the borders of this approximation, a posteriori, after the result will have been obtained. Another, more formal method would be to calculate probability $P_0(n|n')$ in the limit $N \to \infty, s \to 0, \mu \to 0$, while keeping products μN and sN constant. In our experience, the large-parameter method almost always leads to a correct result and is easier to use, especially when there exists an independent verification of the result (for example, by computer simulation).

We can obtain a "smooth" expression for the full probability $P(n|n')$ in eq. (1.13) noticing that mutations are very rare ($\mu \ll 1$), therefore, the likely values of m_1 and m_2 are much smaller than those of $N - n$ and n. In the right-hand side of eq. (1.13), we substitute $n'' - m_1 + m_2$ for n in the arguments of both π functions and the argument of P_0 and expand them all in $m_1 - m_2$ up to linear terms. Then, the sums in m_1 and m_2 can be calculated using eq. (1.12), which yields

$$P(n|n') = (1+2\mu)P_0(n|n') + \mu(2n-N)\frac{\partial P(n|n')}{\partial n} \approx P_0(n+\mu(2n'-N)|n') \tag{1.15}$$

Since the probability $p(n,t)$ in eq. (1.8) (with one exception discussed further), is treated as a function of n, it will is more convenient, from now on, to consider the probability density $\rho(f,t) = Np(Nf,t)$ of the mutant frequency $f \equiv n/N$, normalized by the condition $\int_0^1 df\,\rho(f,t) = 1$. In the new notation, the evolution equation, given by eqs. (1.8), (1.14), and (1.15), can be rewritten as

$$\rho(f, t+\epsilon) = \int df'\ \Pi_\epsilon(f|f')\rho(f', \mathrm{t}) \tag{1.16}$$

$$\Pi_\epsilon(f|f') = \left[2\pi\epsilon V(f')\right]^{-\frac{1}{2}} \exp\left\{-\frac{\left[f-f'-\epsilon M(f')\right]^2}{2\epsilon 2V(f')}\right\} \tag{1.17}$$

$$M(f) \equiv -sf(1-f) - \mu(2f-1) \tag{1.18}$$

$$V(f) \equiv -\frac{1}{N}f(1-f) \tag{1.19}$$

where $\epsilon = 1$ is the generation time interval. Due to the continuous-in-t approximation, ϵ in the above expressions can be substituted by any small time interval. In other words, ϵ is the time differential. Notations $M(f)$ and $V(f)$, in eqs. (1.18) and (1.19), have the respective meaning of the average value and of the variance of the change in f per generation. In a more general form

$$M(f') = \epsilon^{-1}\int df(f-f')\Pi_\epsilon(f,f') \tag{1.20}$$

$$V(f') = \epsilon^{-1}\int df\left[f-f'-\epsilon M(f')\right]^2\ \Pi_\epsilon(f,f') \tag{1.21}$$

These formulae can be confirmed by directly inserting eq. (1.17) into eq. (1.20) and (1.21).

As we are about to show now, the integral equations (1.16) and (1.17) can be transformed to the differential form known as forward Kolmogorov in mathematics of Fokker–Planck equation in statistical physics:

$$\frac{\partial\rho}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial f^2}(V\rho) - \frac{\partial}{\partial f}(M\rho) \tag{1.22}$$

which, together with eqs. (1.18) and (1.19), yields the promised master equations (1.1) and (1.2).

We will derive eq. (1.22) from eqs. (1.16) and (1.17) in a most general form, without defining functions $M(f)$ and $V(f)$. The key assumption is that the variance of the change of allelic frequency per generation is small, $V(f) \ll 1$, and that the averages

of $(f-f')^3$, $(f-f')^4$, ... of the conditional $\Pi_\epsilon(f|f')$ in eq. (1.17) are powers of ϵ higher than 1. As one can check using eqs. (1.17) to (1.19), these assumptions are valid if selection is weak, $s \ll 1$ (Section 1.2.1).

Consider any observable quantity, $A(f)$, localized far the end of the interval $0 < f < 1$, so that $A(f)$ and its first derivative can be neglected at the ends of the interval, $f = 0, 1$. Then, its population average is given by

$$\bar{A}(t) = \int df\ A(f)\rho(f,t) \tag{1.23}$$

Multiplying right-hand side and left-hand side of eq. (1.16) by the factor of $A(f)$ and integrating in f, we get

$$\bar{A}(t+\epsilon) = \int df' \rho(f',t) \int df\ A(f)\rho(f,t) \tag{1.24}$$

Since the characteristic width of $\Pi_\epsilon(f|f')$ in terms of $f-f'$ is small, we are allowed to approximate $A(f)$ in the integrand in eq. (1.24) by a linear expansion in $f-f'$. Evaluating the resulting integral in f and discarding terms of higher than the first order in ϵ, we get

$$\bar{A}(t+\epsilon) = \bar{A}(t) + \epsilon \int df' \rho(f',t) \left[\frac{V(f')}{2} \frac{\partial^2 A(f')}{\partial f'^2} - M(f') \frac{\partial A(f')}{\partial f'} \right] \tag{1.25}$$

where we used eqs. (1.20) and (1.21). Next, we integrate the integral in f' in eq. (1.25) by parts and replace the difference $\bar{A}(t+\epsilon) - \bar{A}(t)$ with time derivative

$$\frac{d\bar{A}}{dt} = \int df' A(f') \left[\frac{1}{2} \frac{\partial^2}{\partial f'^2} \left(V(f')\rho(f',t) \right) - \frac{\partial}{\partial f'} \left(M(f')\rho(f',t) \right) \right] \tag{1.26}$$

Finally, we arrive at the promised evolution equation (1.22), by choosing $A(f') = \delta(f-f')$ and using eqs. (1.23) and (1.26). The width of the "delta –function" is supposed to be much larger than $\sqrt{V(f)} \ll 1$, but smaller than the scale of f for any significant change in $\rho(f,t)$. We assume that $\rho(f,t)$ is sufficiently smooth and changes little when f changes by amount $V(f)$. In the examples studied in the following sections, this condition is usually satisfied.

1.2.6 Derivation of the boundary conditions

The value of f cannot take values less than 0 or greater than 1. Thus, eqs. (1.1) and (1.2) are incomplete without describing the system behavior near ends of the interval $f = 0$ and 1. Figure 1.2A shows schematically the case where of a large number of both allele copies (f is not near 0 or 1). In this case, the mutant frequency f can be considered a continuous variable. There are important cases, however, when the dynamics

of a few copies of the minority variant has to be considered. For this end, we need to derive the boundary conditions for f near 0 and 1 from the virus population model described in Sec 1.2.1. We derive it below and demonstrate that the boundary conditions depend on a parameter region.

An important parameter featuring in the boundary conditions is the probability flux q, which is similar to the gas flow (Figure 1.3). When a population is sufficiently large, (Figure 1.3b), the correct boundary conditions must stipulate that the probability flux vanishes in the two uniform states, that is, completely mutant or completely wild type, eq. (1.3). If a population is small (Figure 1.3C), the border flux is not zero, eqs. (1.5) and (1.6), because the population can be found in a completely uniform state with a significant probability, and that probability can either increase or decrease in time, creating flux from or to the border. This is similar to the gas which evaporates or condenses to liquid form on a wall.

We can interpret two different types of boundary conditions biologically. In a large population that can be viewed as almost deterministic, a genetically uniform state is unlikely, because it is quickly destroyed by mutations. In a small population, mutations are rare, so that a genetically uniform state can take place with a finite ($\gg 1/N$) probability. This also demonstrates that the effect of mutation on evolution depends on the population number. In a large population, mutations may be important even in a very polymorphic state (for example, if selection is weak). In small populations, the role of mutations is only to create a copy or a few copies of a new allele in a uniform population; once the copy is there, mutation events can be forgotten until the population becomes uniform again due to random drift or natural selection. As discussed in Section 1.4, random drift makes a new allele extinct soon after its emergence, but new mutations restore genetic diversity.

In a population larger than the inverse mutation rate, $N \gg 1/\mu$, the boundary conditions, eq. (1.3), state that the flux of the probability density, $q(f,t)$, must be zero at boundaries $f = 0, 1$. The reason is the continuity condition and the obvious fact that a uniform state, $f = 0$ or 1, is quite unlikely when the population is much larger than the inverse mutation rate ($N\mu \gg 1$) due to many mutation events occurring each generation. As we show now, the flux does not vanish at $f = 0$, $f = 1$ in smaller systems; hence, boundary conditions (1.3) should be modified accordingly.

We use a proof *ad absurdum*. Suppose that boundary conditions (1.3) are, in fact, valid at small population sizes. Then, as it follows from eqs. (1.2) and (1.3), probability density $\rho(f,t)$ diverges at the boundaries, provided condition $N\mu < 1/2$ is met. Solving the equation $q(f,t) = 0$ near $f = 0$ and near $f = 1$ separately, one gets

$$\rho(f,t) = \begin{cases} C_0 f^{2\mu N - 1} & f \ll 1 \\ C_1 (1-f)^{2\mu N - 1} & 1 - f \ll 1 \end{cases} \tag{1.27}$$

where C_0 and C_1 are any constants. Integrating eq. (1.27), the first from $f = 0$ to $f = 1/2$ and the second from $f = 1/2$ to $f = 1$, one finds that the regions that contributes most

to the two integrals are the narrow regions near the interval borders such that $\log(1/f) \sim 1/\mu N$, $\log[1/(1-f)] \sim 1/\mu N$. If the population is larger than $1/[\mu \log(1/\mu)]$, these ranges of f correspond to many copies of an allele: $f \gg 1/N$ and $1-f \gg 1/N$. Therefore, the probability of monomorphic states is small, and eq. (1.3) gives the boundary conditions.

If, however, the population is smaller than $1/[\mu \log(1/\mu)]$, the values of f most contributing to normalization are $f \ll 1/N$ and $1-f \ll 1/N$, that is, much less than a single copy per population. We conclude that a population of the statistical ensemble a significant probability to be in a uniform state, $f = 0$ or 1. To include these states into consideration, we have to isolate the corresponding corresponding terms in $\rho(f)$, as given by eq. (1.4), and derive new boundary conditions which reflect this change. Because two more time-dependent variables, p_0 and p_1 exist than in a large population, four (rather than two) conditions at the boundaries are required. The first pair of equations in eq. (1.5) describes the continuity condition that the flux of probability to a uniform state is equal to its rate of change in time in a uniform state. We now obtain the second pair.

We need to return to the discrete model, with $p(n)$ and eq. (1.8). We can consider only one boundary regions in n, for example, $n \ll N$; the conditions for the opposing region, $N-n \ll N$, are similar. Probability $p(n)$ has two distinct components: probability of uniform wild type, $p(0,t) \equiv p_0(t)$ and a small component $p(n,t), n \neq 0$, which varies slowly with n at $n \gg 1$. [Strictly speaking, $p(n,t)$ is diverging as n^{-1} at $n \to 0$, eq. (1.27), but divergence of the integral $\int p(n,t)dn$ is logarithmic, which is slow enough for our aim.]

Equations (1.11) and (1.13) can be simplified using the condition $n \ll N$. Equation (1.11) becomes

$$P_0(n|n') = \frac{(n')^n (1-s)^n}{n!} e^{-n'(1-s)} \tag{1.28}$$

Inequality $n \ll N$ allows one also to neglect deleterious mutations in eq. (1.13) and keep only terms with $m_2 = 0$. Next, the condition $N\mu \log N \ll 1$ implies that even a single mutation is rare in a population. All terms with $m_1 \geq 1$ in eq. (1.13) can be neglected except for the term with $m_1 = 1$ for specific values $n'' = 1$, $n' = 0$. The transition between these two states can occur only by mutation. As a result, eq. (1.13) simplifies to

$$P(n,n') = (1-\mu N)P_0(n|n') + \mu N \delta_{n,1}\delta_{n',0} \tag{1.29}$$

For solving the discrete-n equation for $p(n,t)$, eq. (1.8), it is convenient to use the characteristic polynomial of the probability function $\varphi(x,t)$:

$$\varphi(x,t) = \sum_{n=0}^{N} p(n,t)(1-x)^n \equiv p_0(t) + \phi(x,t) \tag{1.30}$$

where $0 < x < 1$, and $\phi(x,t)$ is a sum over the polymorphic part of $p(n,t)$, $n \neq 0$. The evolution equation for $\varphi(x,t)$ follows from eqs. (1.8) and (1.28) to (1.30):

$$\varphi(x,t+1) = (1-\mu N)\varphi\left(1-e^{-x(1-s)},t\right) + \mu N(1-x)p_0(t) \tag{1.31}$$

Since integral of $p(n,t)$ diverges logarithmically, as we just noted, the characteristic number of alleles n for $n \neq 0$ is large, $n \gg 1$. Therefore, the reciprocal scale of x for function $\phi(x,t)$ is small ($x \ll 1$). Using this fact, we expand the right-hand side of eq. (1.31) to the quadratic terms in x and obtain

$$\frac{dp_0}{dt} + \frac{\partial\phi}{\partial t} = -\left(sx + \frac{x^2}{2}\right)\frac{\partial\phi}{\partial x} - x\mu N p_0 \tag{1.32}$$

Here we substituted $\varphi = p_0 + \phi$ and employed the strong inequalities $s \ll 1$, $\mu N \ll 1$, and $\phi \ll p_0$. We note that at $x \ll 1$, the function $\phi(x,t)$ can be replaced with integral in n and thus represents the Laplace transform of $p(n,t)$ at $n > 0$:

$$\phi(x,t) = \int_{0+}^{\infty} dn\ e^{-xn}p(n,t) \equiv \mathscr{L}_x\{p(n,t)\} \tag{1.33}$$

Using Laplace transform, one can rewrite eq. (1.32) in the form

$$\mathscr{L}_x\left\{\frac{\partial p(n,t)}{\partial t} + \frac{\partial q(n,t)}{\partial n}\right\} = \left[-q(n,t)_{n\to 0} - \frac{dp_0}{dt}\right] + \frac{x}{2}\{\lim_{n\to 0}([n\ p(n,t)]) - 2\mu N p_0\} \tag{1.34}$$

where $q(n,t)$ is the probability flux

$$q(n,t) = -\frac{1}{2}\frac{\partial(np)}{\partial n} - snp \tag{1.35}$$

which coincides with its definition in eq. (1.7) in the limit $f = n/N \ll 1$.

Note that neither the probability function, $p(n,t)$, $n \neq 0$, nor its derivatives contain a delta function or its derivatives. Delta function has already been separated in the uniform-state term p_0. As it is well known, a Laplace transform of an analytic function can neither be constant nor increase in the limit of large x. Therefore, both bracketed terms in eq. (1.34) must be zero, and we arrive at the desired boundary condition at $f \to 0$ given by eqs. (1.5) and (1.6). Since the left-hand side of eq. (1.34) is identically zero, the braced term in eq. (1.34) must be zero as well. As a result, we obtain the promised differential evolution equation (1.7) at $f \ll 1$. The boundary conditions at another boundary, $f \to 1$, are obtained in a similar manner. Thus, Laplace transformation allows us to obtain both the evolution equation and its boundary conditions in one step.

1.3 Thought experiments and observable parameters

Now that we have both the evolution equation and its boundary conditions, we need to know the state of the population at the initial moment. The initial state, obviously, depends on a specific problem or an experiment. Now we introduce several "thought experiments," which are important for a broad range of populations and practical applications. We will also define quantitative parameters suitable for experimental comparison. The following experiments are relevant in various situations:

(i) *Accumulation of deleterious alleles*. The initial state is a purely wild-type population, $f = 0$. The evolutionary process consists of accumulation of deleterious alleles due to random mutation. Their level, as we will demonstrate, is eventually limited by negative selection.

(ii) *Adaptation*. If a population has experienced a strong change in environmental conditions or migrated into a new environment, it will initially be poorly adapted, $f = 1$. Over time, accumulation of beneficial mutations will occur, until a new mutation–selection balance is established. The process of accumulating beneficial mutations is called "adaptation," and is the speed of adaptation is the rate of fitness increase.

(iii) *Growth competition*. Consider a population comprising equally –mixed wild type and mutant, $f = 0.5$ or another strongly diverse population). Competition between two alleles ensues and leads to the decrease of the less-fit allele until a new mutation–selection balance is established.

(iv) *Gene fixation*. This important problem inspired considerable work in classical population genetics (Fisher, 1930; Haldane, 1927; Kimura, 1962; Kimura and Ohta, 1969; Wright, 1931) and is still very useful for understanding other stochastic experiments (Chapters 2 and 3). Suppose, a single advantageous allele is added to a uniform population, $f = 1/N$. The allele can have one of two fates: either it will survive random drift and expand to the entire population (be fixed), or it will go extinct (Figure 1.1a). We need to estimate (i) the fixation probability, (ii) if the average time to reach the level where allele survival is ensured, and (iii) the chance that the new lineage will reach a given size before it goes extinct.

(v) *Steady state*. After a sufficient time, the system passes to a steady state, in which any statistics are constant.

(vi) *Genetic divergence*. Let us split a steady-state population into two isolated parts. Initially, both populations have identical genetic composition, and then they evolve independently and diverge genetically. As time goes on, their respective genetic compositions correlate less and less. At which characteristic time does the loss of correlation occur?

(vii) *Time correlation of f*. This experiment studies stochastic fluctuations of f in time in the steady state. The parameter of interest is the average timescale associated with these fluctuations.

1.3.1 Observable parameters

The probability density $\rho(f)$ of the mutant frequency f predicted by the stochastic equation is an observable parameter. However, to measure it directly, one would need to make a histogram of mutant frequencies from a large ensemble of populations. More convenient for experimental testing is the average (mean, expectation) value, which requires a smaller number of populations to measure. Let parameter $A(f)$ be any deterministic function of mutant frequency f. Its mean value $\bar{A}$ and the variance V_A are defined by

$$\bar{A}(t) = \int_0^1 df\ A(f)\rho(f,t) \tag{1.36}$$

$$V_A(t) = \overline{\left(A-\bar{A}\right)^2} = \overline{A^2 - \bar{A}^2} \tag{1.37}$$

Useful observable parameters whose statistical properties can be measured in real experiments and compared with the theoretical predictions will be introduced later.

The first observable parameter is the mutant frequency itself, f. It can be compared with the experimental value if one knows which allele is better fit.

The second observable is the pairwise genetic distance within population, T, defined as the probability that a randomly sampled sequence pair differs at the locus. Although there are other metrics of intrapopulation variability (e.g., entropy), we will use this simple metrics (Nei, 1972) usually called Hamming distance. This is coincident with the standard definition of the average number of nucleotide differences for two randomly sampled genomes, except applied to a single base. Genetic distance T can be expressed as

$$T = 2f(1-f) \tag{1.38}$$

which varies between 0 ($f = 0$ or 1) and 0.5 ($f = 0.5$). Unlike the mutant frequency f, the genetic distance estimate does not depend on wild-type knowledge.

The mean values of parameters f and T are given by eqs. (1.36) and (1.4) to (1.7), with $A(f) = f$ and $A(f) = 2f(1-f)$, respectively. The variance V_f and the two averages $\bar{f}, \bar{T}$ are related as

$$V_f = \bar{f}\left(1-\bar{f}\right) - \frac{\bar{T}}{2} \tag{1.39}$$

In addition to intrapopulation genetic distance, which characterizes population diversity, we also introduce the distance T_{12} between two populations defined in the same way as T, except that the two compared sequences are taken randomly from two separate populations

$$T_{12} = f_1(1-f_2) + f_2(1-f_1) \tag{1.40}$$

where f_1, f_2 are the respective mutant allelic frequencies. The interpopulation distance may change from 0, when the two populations consist completely from same allele, to 1, when the two populations are uniform in opposite alleles.

The interpopulation distance must be equal or larger than the mean of the corresponding intrapopulation distances. Therefore, it is handy to introduce also the relative distance, D, equal to

$$D = T_{12} - (T_1 + T_2)/2 = (f_1 - f_2)^2 \tag{1.41}$$

The value of D varies between 0 (two populations have the same genetic composition f) and 1 when one is purely mutant and another is completely wild type). If the two populations have diverged for a long time and are almost statistically independent, one can easily check that $\bar{D} = (\overline{f_1} - \overline{f_2})^2 + V_{f_1} + V_{f_2}$. Alternative definitions of the genetic distance could be used (Nei, 1972). We prefer our nucleotide difference definition (Hamming distance), because its statistical momenta are easy to evaluate.

We turn now to the genetic divergence experiment. Suppose that a parental population has been split into two populations at $t = 0$. The daughter populations are assumed to grow rapidly to the original size, while their initial composition is inherited from the parental population ($f = f_0$). The value f_0 is of course random and obeys distribution $\rho(f_0)$. To monitor how the two population diverge, we calculate how distance $\bar{D}$ increases after the separation. The average value, $\bar{D}$, is given by

$$\bar{D} = \int_0^1 df_1 \int_0^1 df_2 \int_0^1 df_0 (f_1 - f_2)^2 \rho(f_1, t|f_0)\rho(f_2, t|f_0)\rho(f_0) \tag{1.42}$$

where $\rho(f_0)$ denotes the probability density of the initial allele frequency, eqs. (1.4) to (1.7), and $\rho(f, t|f_0)$ obeys our evolution equation and meets the initial condition

$$\rho(f, 0|f_0) = \delta(f - f_0) \tag{1.43}$$

By evaluating integrals in eq. (1.42), $\bar{D}(t)$ can also be expressed in terms of the evolving variance of f, $V_f(t|f_0)$

$$\bar{D} = 2\int_0^1 df_0 V_f(t|f_0)\rho(f_0) \tag{1.44}$$

The variance $V_f(t|f_0)$ increases from 0 at $t = 0$ to its equilibrium value at infinite time. Thus, eq. (1.44) expresses the relative distance between two diverging populations in terms of the variance in a single population.

All the above observables can be measured at a single time point, including dynamic experiments (the first three) and experiments in the steady state.

Our next parameter compares the state of population at two moments in time; we will define it for a steady-state population. The time correlator $K(t)$ determines

how fast the population erases memory from a preceding random fluctuation of the mutant frequency f

$$K(t) = \frac{1}{V_f^{\rm ss}}\left[\overline{f(0)f(t)} - \overline{f_{\rm ss}}^2\right] \tag{1.45}$$

The choice of the initial moment $t = 0$ in eq. (1.45) in the steady state is arbitrary. The function $K(t)$ varies from maximum correlation 1 at $t = 0$ to 0 at $t = \infty$. The same correlator can also be written in terms of the mean frequency $\bar{f}(t|f_0)$:

$$K(t) = \frac{1}{V_f^{\rm ss}}\int_0^1 df_0\, f_0\, \bar{f}(t|f_0)\rho_{\rm ss}(f_0) - \bar{f}_{\rm ss}^{\,2} \tag{1.46}$$

Here conditioned mean $\bar{f}(t|f_0)$ is defined by eq. (1.43) with $A(f) = f$ and under the initial condition given by eq. (1.36). The time correlation function $K(t)$ is maximum and equal to 1 at $t = 0$ and vanishes at $t \to \infty$. The time at which $K(t)$ decays by 50% represents the characteristic half-period of random oscillations of f.

Now we are all set. We have master equations to solve, boundary conditions to use, experiments to investigate, and observables to predict. In the next Sections 1.4–1.7, we will derive all observables for the listed thought experiments in various parameter regions. We will verify and illustrate our analytic results with stochastic Monte-Carlo simulation.

1.4 Steady state

After a sufficiently long journey, any population under constant conditions arrives at a steady state, where its statistics no longer change, although state variables keep fluctuating around their plateaux. We will discuss further the statistical properties of steady state in various intervals of the population number.

1.4.1 General case

We start from eq. (1.1) to (1.3), which apply when populations are large, $N\mu \gg 1/\log N$. From the steady-state condition, $\partial\rho/\partial t \equiv 0$, one gets

$$q(f) \equiv 0 \tag{1.47}$$

where $q(f)$ is given by eq. (1.2). We can separate variables f and p in the resulting differential equation. By integrating it, we obtain (Wright, 1931)

$$\rho_{\rm ss}(f) = C[f(1-f)]^{-1+2\mu N} e^{-2Nsf}, \quad N\mu \log N \gg 1 \tag{1.48}$$

where C is a normalization constant to ensure that the full integral of $\rho_{\rm ss}(f)$ is 1.

As discussed in Section 1.2, in small populations such that $N\mu \log N \ll 1$, the probability density has delta-function components, eq. (1.4), and obeys eqs. (1.5)–(1.7). Now the steady-state conditions have the form

$$\partial g/\partial t = 0, \quad dp_0/dt = dp_1/dt = 0$$

From these conditions and with $q(f)$ given by eq. (1.2) with $\mu = 0$, since mutations enter only boundary condition, we get

$$\rho_{\rm ss}(f) = g_{\rm ss}(f) + \frac{1-p_{\rm pol}}{1+e^{-2Ns}}\left[\delta(f) + e^{-2Ns}\delta(1-f)\right] \tag{1.49}$$

$$g_{\rm ss}(f) = \frac{2\mu N}{1+e^{-2Ns}}\frac{e^{-2Nsf}}{f(1-f)}, \quad f(1-f) \gg 1/N, \quad \mu N \log N \ll 1 \tag{1.50}$$

$$p_{\rm pol} \approx 2\mu N \log\left[\min\left(N, \frac{1}{s}\right)\right] \tag{1.51}$$

where $p_{\rm pol} \approx 1$ is the total probability of having a polymorphic population.

At small population sizes, both forms of probability density, eqs. (1.48) and (1.49), are singular at $f = 0$ and $f = 1$, though in a different way. The two versions can be shown, for certain purposes, to be interchangeable for small population sizes. Specifically, eqs. (1.48) and (1.49) can be shown to predict, with error $O(\mu N)$, the same lower momenta of f, that is, the expectation value and variance.

The version in eq. (1.49), although longer, is generally more practical in this regime. The form of eq. (1.48), on the other hand, applies at large populations as well and is more suitable for the studies of the transition to the deterministic limit.

1.4.2 Steady state in the selectively neutral case: $s \ll \mu$

If the selection coefficient is much smaller than the mutation rate, selection is negligible. (This is not the only scenario when selection can be neglected, see Section 1.6.) In this limit, $s \ll \mu$, the transition between stochastic and deterministic behavior occurs due to competition between mutation, which acts as a deterministic factor in large populations, and genetic drift. For this reason, we have to recount the basics of the selectively neutral theory. The fundamental prediction of any stochastic model is that the fluctuations of mutant frequency decrease with the population size. In other words, the probability density is distributed broadly in small populations and makes narrow maxima in large populations. The transition between the two limits is controlled by product μN, which we have already seen in our boundary conditions. That product represents the mutation rate per population of genomes. For example, for most RNA viruses, $\mu N = 1$ when the infected cell number, N, is in the range $10^4 - 10^6$. For DNA organisms, μ is much smaller due to the DNA editing enzymes, and the population has to count in billions to be in that range.

With increasing mutation rate per population μN, the probability density becomes more narrow, as illustrated in Figure 1.4 (Wright, 1931). The change in width is a result of the competition between mutations, which diversify the system, and random drift, which forces the system toward a uniform state. When μN is much smaller than 1 (the case termed "drift regime," Table 1.2), random drift is much stronger, and a typical population is either uniform or weakly diverse. Therefore, the probability density $\rho(f)$ is U-shaped, with a minimum at the half-and-half composition. At the smallest values of μN [the condition is given in eq. (1.5)], the system is most likely to be either uniformly mutant or uniformly wild type, without a single opposite allele present. The net probability of any polymorphic state will be on the order of μN, which is much smaller much smaller than 1. That estimate can be interpreted as the frequency of diverse (in genetics, "segregating") sites in a population.

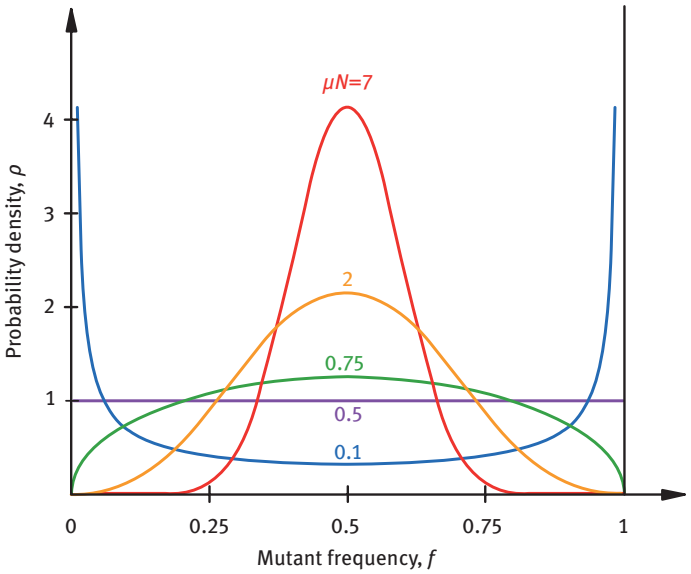

Figure 1.4: Probability density in the steady state in the selectively neutral regime, $s \ll \mu$. Curves show $\rho_{ss}(f)$ at different values of μN shown at curves (based on Rouzine et al. (2001)).

Now we turn to larger populations. With increasing μN, the U-shape of probability density gradually flattens out (Figure 1.4). The valley at $f = 50\%$ turns into a maximum, when μN crosses the value of $1/2$. The peak becomes more and more narrow as μN becomes much larger than 1. The last fact implies that the mutant frequency approaches its deterministic limit of $1/2$, due to the balance between reverse and forward mutations. We will call this limit of large population sizes "mutation regime" (Table 1.1).

Table 1.2: Classification of the regimes of genetic evolution in the one-locus model.

Regime	**Neutral limit** ($s \ll \mu$)				**In the presence of selection** ($s \gg \mu$)			
	Population interval	**Behavior**	**Factors in steady state**	**Factors in diverse state**	**Population interval**	**Behavior**	**Factors in steady state**	**Factors in diverse state**
Drift	$N \ll 1/\mu$	Stochastic	Drift, mutation	Drift	$N \ll 1/s$	Stochastic	Drift, mutation	Drift
Selection drift					$1/s \ll N \ll 1/\mu$	Stochastic	Drift, mutation, selection	Selection
Selection					$N \gg 1/\mu$	Determin.	Mutation, selection	Selection
Mutation	$N \gg 1/\mu$	Determin.	Mutation	Mutation				

To describe the selectively neutral case formally for both small and large populations, we will use the form of eq. (1.48) for the distribution density. Putting $s = 0$ in eq. (1.48) and normalizing the resulting expression to 1, we get (Wright, 1945)

$$\rho_{ss} = \frac{\Gamma(4\mu N)}{\Gamma^2(2\mu N)}[f(1-f)]^{-1+2\mu N}, \quad s \ll \mu \tag{1.52}$$

where $\Gamma(x)$ is the Euler gamma function, $\Gamma(x) = \int_0^\infty dy\, y^{x-1}e^{-y}$, and we plugged in the identity for beta function (Abramowitz and Stegun, 1964)

$$\int_0^1 df\, f^{x-1}(1-f)^{y-1} = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{1.53}$$

Equation (1.52) is plotted at different values of μN in Figure 1.4.

The mean values and the variance of f and intrapatient distance T can be obtained from eq. (1.52) and eqs. (1.36) to (1.38):

$$\bar{f} = \frac{1}{2}, \quad V_f = \frac{1}{4(1+4\mu N)}, \quad \bar{T} = \frac{2\mu N}{1+4\mu N}$$

$$V_T = \frac{2\mu N}{4(1+4\mu N)^2(3+4\mu N)}, \quad s \ll \mu \tag{1.54}$$

Here, to evaluate the integrals over f in eqs. (1.36) and (1.37), we used eq. (1.53) and $\Gamma(x+1) = x\Gamma(x)$ (Abramowitz and Stegun, 1964). For small populations ($\mu N \ll 1$), eq. (1.54) yield well-known results of the selectively neutral theory:

$$\bar{f} = 1/2, \quad V_f = 1/4, \quad \mu N \ll 1, \quad sN \ll 1 \tag{1.55}$$

$$\bar{T} = 2\mu N, \; V_T = \frac{2\mu N}{3}, \quad \mu N \ll 1, \quad sN \ll 1 \tag{1.56}$$

In intuitive agreement with Figure 1.4, the scaled standard deviation, $V^{1/2}/\bar{f}$, is on the order of 1 at $\mu N \leq 1$ and much smaller at large μN, eq. (1.54) (Figure 1.6b).

1.4.3 Steady state with selection: $\mu \ll s \ll 1$

If selection coefficient is superior to mutation rate (but still much less than 1), Darwinian selection comes into play. Then, selection factor can be neglected only in a very small population $Ns \ll 1$, which scenario has the same behavior as the drift regime we just discussed. At larger population sizes, selection is critically important, because it causes the probability density, eqs. (1.48) or (1.49) to (1.51), to be asymmetric in favor of a predominantly wild-type population (Figure 1.5).

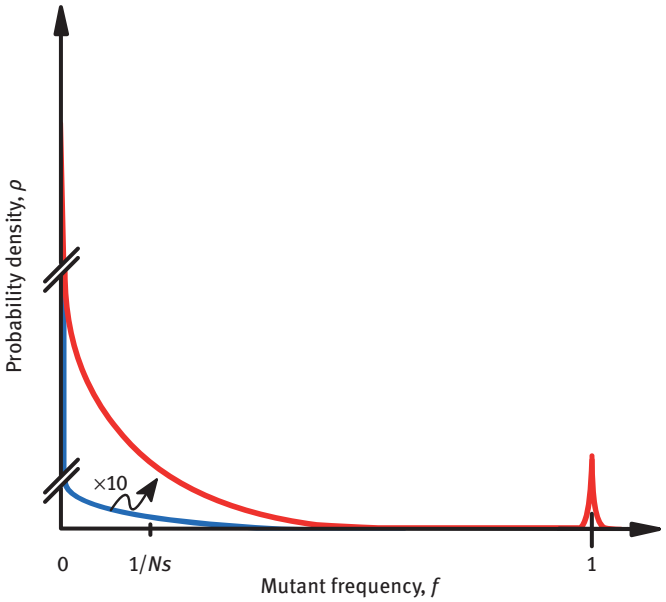

Figure 1.5: Schematic plot of the steady-state probability density $\rho_{ss}(f)$ in the selection-drift regime, $1/s \ll N \ll 1/\mu$. Blue curve is the full distribution and the red curve is magnification of its tail. Note delta functions at $\rho = 0$ and 1, together with the exponential tail extending from pure wild type, $f = 0$ (based on Rouzine et al. (2001)).

As in the neutral case, $\rho(f)$ contracts as N increases, but selection due to the factor $\exp(-2Nsf)$ in eqs. (1.48) and (1.50) causes asymmetry of $\rho(f)$. The existence of the exponential tail indicates that natural selection and drift are the main competing factors, and that both have the same order of magnitude at the characteristic frequency

$$f_{\text{stoch}} \sim \frac{1}{Ns}$$

This famous equation, which we do not even number because it is so short, is called "stochastic threshold." It is important not only for the steady state, but also in many other situations, including gene fixation and adaptation discussed in this and the following sections. Another important difference from neutral evolution is the existence of an additional broad interval in N. In the limit of very large populations, when μN is much larger than 1 ("selection regime" in Table 1.2), the probability density is a narrow peak localized near its deterministic value. This value is given by the mutation-selection balance ratio, μ/s.

Between these two limits, there exist a wide interval in population size between the inverse mutation rate and the selection coefficient, termed "selection drift" in Table 1.2, in which all three evolutionary forces are important. Specifically, mutations

create diversity, selection keeps mutants at a low level, and drift causes fluctuations of f. The function $\rho(f)$ comprises three components, as follows (Figure 1.5). (i) A large delta function located at $f = 0$ implies that a population is, most likely, uniformly wild type. (ii) An exponential in the interval of f [0 1] with a small magnitude and scale in $f \sim 1/(Ns) \ll 1$ (Wright, 1931) shows that the chance of a population being genetically diverse is low, and that if a population happens to be diverse, the proportion of mutants is small. The chance of a diverse population given by the area under the curve is small in parameter μN, eq. (1.51). (iii) A tiny peak at $f = 1$ is significant only when population size is near the lower boundary, $N \sim 1/s$.

The selection-drift regime has mixed properties which combine stochasticity and determinism. On the one hand, the form of the probability density suggests a very stochastic behavior. On the other hand, the average mutant frequency and the average genetic distance are given, over most of the regime, by their deterministic values calculated for much larger populations.

The mean values and scaled standard deviations for f and T are shown in Figure 1.6 as a function of the population size. As it is expected from results in Figure 1.5, in the selection-drift regime, the relative standard deviations are much larger than 1 indicating that their fluctuations are stronger than their mean values (Figure 1.6B). At the same time, remarkably, the mean values stay the same as in the selection regime $N \gg 1/\mu$, where fluctuations are much smaller (Figure 1.6A). We need to emphasize that the magnitude of fluctuations strongly exceeds the Poisson statistics prediction. In Sections 1.6 and 1.7, we will illustrate a typical steady-state process by stochastic simulation. Examples of such simulations, for each interval of N, are plotted in Figure 1.6C (Rouzine and Coffin, 1999a).

To obtain analytic expressions for very small populations, $N \ll 1/s$ (drift regime), we omit s in eq. (1.48) and arrive at the results obtained in Section 1.4.2 for the drift regime. In the opposite limit of large populations $N \gg 1/\mu$ ("selection regime," Table 1.2), $\rho_{ss}(f)$ in eq. (1.48) has a sharp maximum due to mutation-selection balance, $f = \mu/s$. Expanding $\log \rho_{ss}$ in f near the maximum, we obtain a Gaussian form (Karlin and McGregor, 1964):

$$\rho_{ss}(f) \approx Ce^{-\frac{Ns^2}{\mu}\left(f-\frac{\mu}{s}\right)^2}, \quad \mu N \gg 1, \quad s \gg \mu \tag{1.57}$$

Here the maximum position, $f = \mu/s$, is the mean steady-state value in the deterministic limit. For the average values and variances of f and T, eqs. (1.36) to (1.38), we have

$$\bar{f} = \frac{\mu}{s}, \quad V_f = \frac{\mu}{2Ns^2}, \quad \bar{T} = \frac{2\mu}{s}, \quad V_T = \frac{2\mu}{Ns^2}, \quad \mu N \gg 1, \quad s \gg \mu \tag{1.58}$$

We observe that in the intermediate interval, $1/s \ll N \ll 1/\mu$, eq. (1.58) yields $V_f \gg \bar{f}^2$. Therefore, fluctuations are strong in this case, but we cannot neglect selection as in the drift regime. Both drift and selection are important in this case. We will analyze $\rho_{ss}(f)$

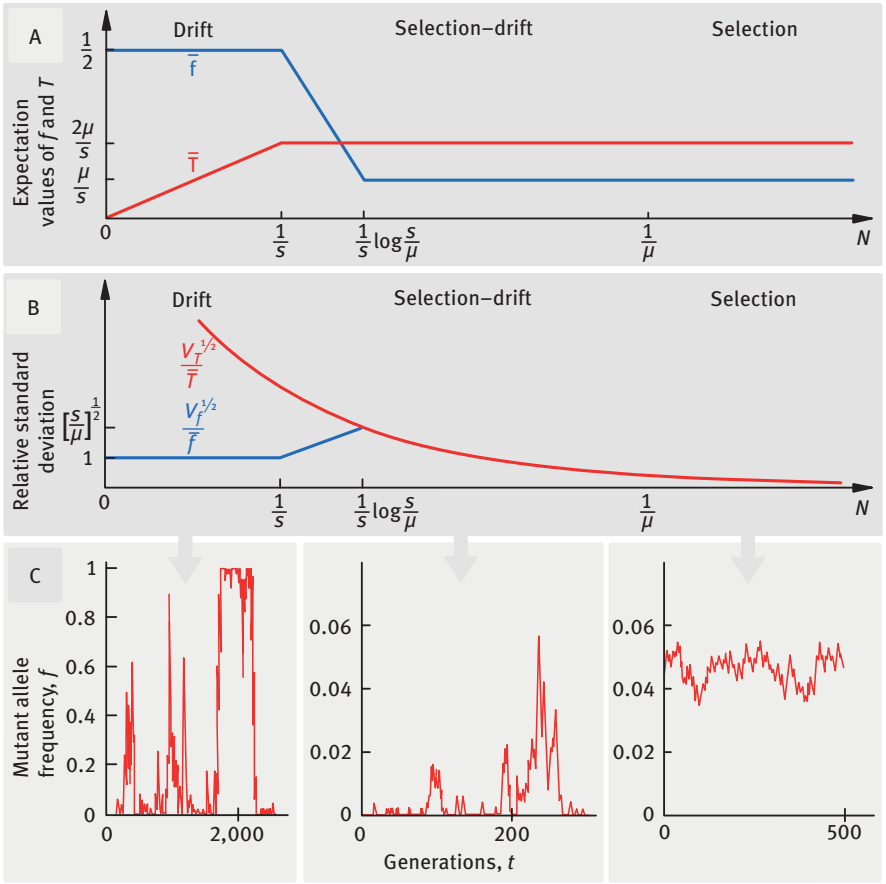

Figure 1.6: Dependence of the observable parameters in the steady state on the population size in the three main intervals of population size, *N*. (A) Average mutation frequency $\bar{f}$ and genetic distance, $\overline{T}$. (B) Relative standard deviations of the two parameters. (C) Representative Monte-Carlo simulation runs in the respective intervals of *N* (details in the legends to Figures 1.10 to 1.12) (based on Rouzine et al. (2001)).

and its lower momenta using the form given by eqs. (1.49) and (1.50). As we already mentioned, function $\rho_{ss}(f)$ has three components (Figure 1.5): two peaks at $f = 0$ and 1 that correspond to the uniform states, and an exponential with scale $f \sim 1/Ns$, which describes the distribution of diverse states under selection and drift. The probability that a population is diverse, eq. (1.51), is small, $p_{pol} \approx 2\mu N \log(1/s)$. To obtain the first two momenta for mutant frequency and the distance, eq. (1.49) is substituted into eqs. (1.36) and (1.37). Using the fact that values $f \sim 1/Ns$ mostly determine the integrals of $g_{ss}(f)$ we obtain

$$\bar{f} = \frac{\mu}{s} + e^{-2Ns}, \quad V_f = \frac{\mu}{2Ns^2} + e^{-2Ns}$$
$$\bar{T} = \frac{2\mu}{s}, \quad V_T = \frac{2\mu}{Ns^2}, \quad \frac{1}{s} \ll N \ll \frac{1}{\mu} \tag{1.59}$$

At the transition point in population size, $N \sim 1/s$, these four values match, by an order of magnitude, their neutral–limit values in eqs. (1.55) and (1.56). At a higher population size, $N \sim (1/s)\log(s/\mu)$, they match the quasi–deterministic results in eq. (1.58) derived for $N \gg 1/\mu$. Curiously, in most of the regime, $(1/s)\log(s/\mu) \ll N \ll 1/\mu$, all the averages and variances happen to coincide with their respective values in the deterministic limit, even though the relative standard deviations, $V_f/\bar{f}^2$ and $V_T/\bar{T}^2$, are much larger than 1, indicating that fluctuations of mutant frequency and genetic distance exceed their respective averages (Figure 1.6B).

1.5 Boundaries of deterministic approximation

As we have just shown, the deterministic behavior of equilibrium state is reached when μN is much larger than 1. In this section, we will study the transition between stochastic behavior and deterministic approximation in the more general case, where the statistical properties of a population are time–dependent.

1.5.1 Deterministic limit

As we already mentioned, stochastic and deterministic models work with different types of state variables. The deterministic models consider a scalar, the time-dependent frequency of mutants, and the second use a whole function, a probability density that changes in time. We need to check that these two different approaches match in very large populations. In this limit, both must predict deterministic evolution. Here we will solve Kolmogorov equation (1.1), in $N \to \infty$ limit. We will demonstrate that the resulting probability density is, indeed, a very narrow peak around a time-dependent mutant frequency (Figure 1.7b), which satisfies a deterministic evolution equation we derive later.

A deterministic equation can make a prognosis for mutant frequency as a function of time if the initial value is known. Examples of plots for three types of initial conditions, corresponding to adaptation, accumulation of deleterious mutants, and competition of two variants (Section 1.3) are given in Figure 1.8. In each case, the population approaches the same steady state with $f = \mu/s$, after a time interval proportional to the inverse selection coefficient $1/s$ (Section 1.4). The time of adaptation is somewhat longer compared to the two other experiments. The reason for the delay is that, before crossing the entire interval $f = [0\ 1]$, beneficial alleles have to be

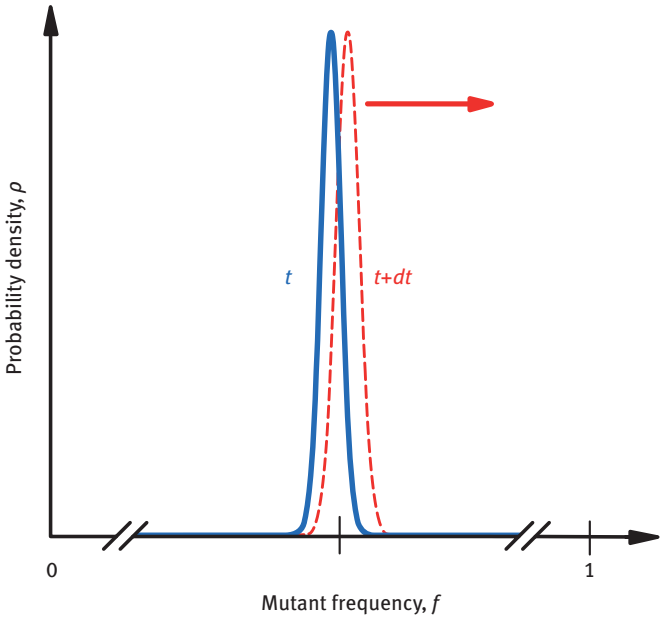

Figure 1.7: In the deterministic limit, the probability density $\rho(f,t)$ of the mutant frequency f is a moving delta function. Blue solid and red dashed curves show two consecutive moments of time.

generated for selection to operate on. For this reason, the initial slope of the time course of the mutant frequency is small in accumulation and adaptation experiments (Figure 1.8). Selection starts to dominate over mutation causing the time dependence $f(t)$ to curve only after a growing lineage frequency exceeds μ/s.

1.5.2 Deterministic equations

In this section, the equation of evolution in the deterministic limit is derived by two methods, first, from deterministic first principles and, second, as a limiting case of the stochastic equation at $N \to \infty$. Then we will verify that both methods give the same result, for arbitrary initial conditions. Then the boundaries of the parameter region of deterministic approach are found.

1.5.2.1 Main results

In the limit of large populations $N \gg 1/\mu$, the time-dependent probability density has the form

$$\rho(f,t) = \delta(f - f_d(t)) \tag{1.60}$$

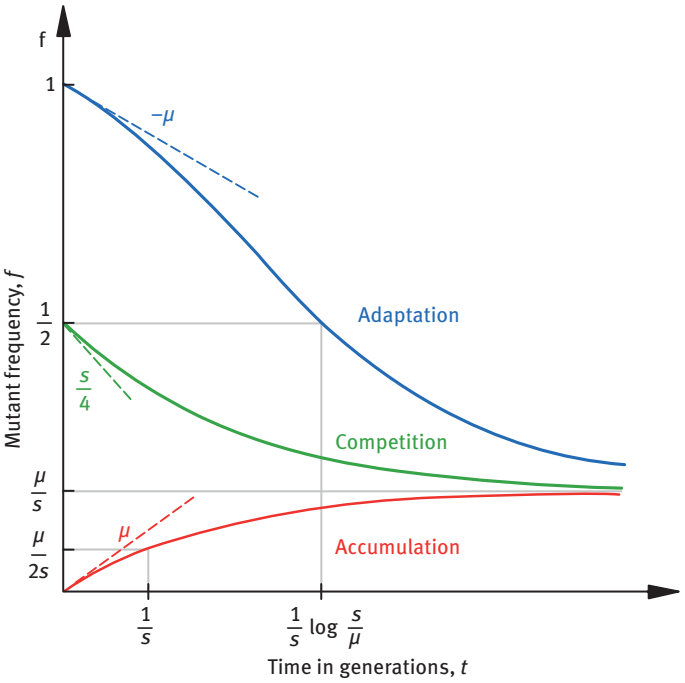

Figure 1.8: Schematic dependence of the mutant frequency, f, on time t in the deterministic limit. The three curves correspond to three experiments with different initial values of $f(0)$. Red: accumulation of deleterious alleles, $f(0) = 0$; green: growth competition, $f(0) = 1/2$; blue: adaptation process, $f(0) = 1$. The value of the ratio μ/s used in the figure may be unrealistically high and is used for clarity of plot only. Dashed lines show initial slopes (based on Rouzine et al. (2001)).

$$\frac{df_d}{dt} = M(f_d) = -sf_d(1-f_d) - \mu(2f_d - 1) \tag{1.61}$$

Equation (1.61) represents the deterministic evolution equation for the moving peak in f (Figure 1.7). These expressions agree with the meaning of $M(f)$, defined by eq. (1.18), as the mean change in f in each generation, eq. (1.20). The actual change in f coincides with the average change, due to the absence of random drift and random factor in mutation in this limit. The first term in the right-hand side of eq. (1.61) describes natural selection against the existing mutant allele and hence vanishes in a uniform population, where selection stops working since there is nothing to select from $f_d = 0$ or 1. The mutation term in eq. (1.61) is not zero at $f_d = 0$ or 1, since mutation, unlike selection or drift, takes place even in a completely uniform population. The term vanishes at $f_d = 1/2$, where reverse and forward mutations cancel each other. (Here we assume that direct and reverse mutations have the same mutation

rate. It is a simplification, because, in real biological populations, mutation is rarely symmetric. Forward and back rates can easily differ by an order of magnitude. In this case, the mutation-selection equilibrium will not be reached at a half but at 10% or 90%. This detail can easily be taken into account while preserving the spirit of the results.)

Equation (1.61) can easily be solved in the general cases by separating variables. In two limits, the case with selection and the neutral case, the solution has an especially simple form:

$$f(t) = \begin{cases} \frac{1}{2} + \left(f_0 - \frac{1}{2}\right)e^{-2\mu t}, & \mu \gg s \\ f_{ss} + \frac{(f_0 - f_{ss})e^{-st}}{1 + f_{ss} - f_0 + (f_0 - f_{ss})e^{-st}}, & \mu \ll s \end{cases} \tag{1.62}$$

where $f_{ss} \equiv \mu/s$ is the famous mutation-selection steady-state value (Haldane, 1924, 1927). The second part of eq. (1.62) represents a sigmoidal dependence (Figure 1.8).

First, we derive eq. (1.61) from first-principles neglecting random factors. The starting equations suitable for the model have the form

$$\frac{dn_1}{dt} = (1-\mu)(1-s)\kappa n_1 + \mu\kappa n_2 - \omega n_1 \tag{1.63}$$

$$\frac{dn_2}{dt} = \mu(1-s)\kappa n_1 + (1-\mu)\kappa n_2 - \omega n_2 \tag{1.64}$$

where numbers n_1 and n_2 correspond to mutant and wild-type infected cells, respectively, κ is the reproduction coefficient for the wild type, and $1/\omega$ is the average life span of an individual (infected cell). To match our population model with discrete generations in time intervals of $\Delta t = 1$, we choose $\omega = 1$. Using our notation $f = n_1/(n_1 + n_2)$, calculating derivative df/dt, and using eqs. (1.63) and (1.64), we arrive at a single equation (1.61). The equation is nonlinear in f, because f depends in nonlinear way on n_1, n_2. Note that it applies regardless on whether the population size is constant, as we assumed in the rest of this chapter. It might be as well expanding or compressing.

In large but finite populations random drift creates a finite width of probability density has, $w(t)$ (Section 1.5.3). As long as the relative width $w(t)/\{f_d(t)[1-f_d(t)]\}$ remains much less than 1, deterministic approximation is a good starting point. In the neutral limit, $\mu \gg s$, we show below that the boundary in N of the selection regime is the same as in the steady state: $N \sim 1/\mu$. In the presence of selection, $\mu \ll s$, that boundary depends on the initial conditions set in the experiment. Results for

the first three thought experiments (Section 1.3) (Figure 1.8) are listed below, assuming that the initial value f_0 is known:

$$w = f(1-f) \times \begin{cases} \dfrac{1}{\sqrt{N\mu}} & \text{acc. del. } f_0 = 0, \quad f \ll f_{ss} \\ \dfrac{1}{\sqrt{N\mu}} & \text{adapt. } f_0 = 1, \quad f_{ss} \ll f \\ \dfrac{1}{\sqrt{Ns}} \sqrt{\dfrac{1-2f}{f(1-f)} + 2\log\dfrac{1-f}{f}} & \text{gr.comp. } f_0 = \frac{1}{2}, \quad f_{ss} \ll f < \frac{1}{2} \end{cases} \tag{1.65}$$

where $f \equiv f_d(t)$. Note that, if we start from a uniform initial population, the criterion of determinism is the same as we obtained in the steady state in Section 1.4, $N \gg 1/\mu$. For the growth competition experiment which starts from a diverse population, it is much softer, $N \gg 1/s$, provided the mutant frequency is high above the steady-state value, f_{ss}.

1.5.3 Derivation from the stochastic equation

At large population numbers, $N \to \infty$, the diffusion term with the second derivatives in the right-hand side of eq. (1.2) is negligibly small. Consequently, the probability density $\rho(f,t)$ must be narrow in f, and we will use of this fact below. Let us can present the master equations, eqs. (1.1) and (1.2), in an equivalent form

$$\frac{\partial \rho}{\partial t} + M(f)\frac{\partial \rho}{\partial f} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}[f(1-f)\rho] - \frac{dM}{df}\rho \tag{1.66}$$

where $M(f)$ is given by eq. (1.18). Each term on the right-hand side of eq. (1.66) is much smaller than any term on the left-hand side and can be considered small perturbation. The first term is small in $1/N$, where N is very large, and $(dM/df)\rho$ is much smaller than $M(f)(\partial\rho/\partial f)$, because $\rho(f)$ is very narrow and hence changes faster in f than $M(f)$. Hence, in the lowest approximation in $1/N$, we can set the left-hand side of eq. (1.66) to 0. A partial solution is a delta–function with a center moving in time, eqs. (1.60) and (1.61) (Figure 1.7). One can check this fact by a simple substitution.

The general solution for $\rho(f,t)$ is a linear combination of solutions of the form of eq. (1.60), each solution with its own initial condition $\rho(f,0) = \rho_0(f)$. If the initial value of f is known with a high accuracy, $\rho(f,t)$ is a single delta–function. Then, we solve eq. (1.66) in the next approximation in $1/N$ to find the finite width of $\rho(f,t)$.

Deterministic dynamics $f(t)$ can be obtained in the general form from eq. (1.61). Rewrite eq. (1.61) as

$$\frac{1}{s}\frac{df}{dt} = (f - f_{ss})(f - f_*) \tag{1.67}$$

$$f_{\mathrm{ss},*} = \frac{1}{2} + \frac{\mu}{s} \mp \sqrt{\frac{1}{4} + \left(\frac{\mu}{s}\right)^2} \tag{1.68}$$

Here the plus and minus signs corresponds to f_* and f_{ss}, respectively. Below we drop the subscript "d" in f_d. The values of parameters f_{ss} and f_* are in the intervals $0 < f_{\mathrm{ss}} < 1/2$ and $f_* > 1$, respectively. The value of f_{ss} is the mutant frequency in mutation-selection balance. The value of f_* is just a formal value. The value of f_{ss} approaches the values of $1/2$ and μ/s in the limits $\mu \gg s$ and $\mu \ll s$, respectively (Section 1.4). Equation (1.67) is an ordinary differential equation with separating variables and can be integrated to obtain dynamics of mutant explicitly:

$$f(t) = f_{\mathrm{ss}} + \frac{(f_0 - f_{\mathrm{ss}})(f_* - f_{\mathrm{ss}})e^{-(f_* - f_{\mathrm{ss}})st}}{f_* - f_0 + (f_0 - f_{\mathrm{ss}})e^{-(f_* - f_{\mathrm{ss}})st}} \tag{1.69}$$

where $f_0 = f(0)$ is the initial condition. From eq. (1.69), $f(\infty) = f_{\mathrm{ss}}$. Note that the function $f(t)$ either only increases or only decreases, depending on the initial values, and never crosses its asymptotic value f_{ss}. Asymptotics of eq. (1.69) with and without selection are given in eq. (1.62). The characteristic time required to come halfway to the steady state is given by $1/s$ and $1/\mu$ in the two limits, respectively. Schematic plots of $f(t)$ for $\mu \ll s$ of initial conditions, $f_0 = 1,\ 1/2,$ and 0 (adaptation, growth competition, deleterious accumulation) are shown in Figure 1.8. In particular, the formula for $f(t)$ in the accumulation experiment simplifies to

$$f(t) = \frac{\mu}{s}\left(1 - e^{-st}\right), \quad f_0 = 0, \qquad \frac{\mu}{s} \ll 1 \tag{1.70}$$

1.5.4 Boundaries of the deterministic approximation

The frequency of mutants fluctuates around its deterministic value due to random genetic drift, which is present even in very large populations. At sufficiently small populations, fluctuation magnitude becomes comparable to the average frequency of the minority allele (either mutant or wild type), and the deterministic description becomes a poor approximation. The corresponding boundary in N depends on the initial conditions. As we show further, when the initial state is composed entirely of the better-fit or less-fit variant, the deterministic criterion is met when $\mu N \gg 1$. A much smaller population (Section 1.3) can follow the deterministic path as long as it is initially diverse. The criterion on diversity is that mutant frequency f must be larger than "stochastic threshold" $1/Ns$, which determines the scale of the tail of $\rho(f)$ at the steady state (Figure 1.5). This is because a weakly diverse population is sensitive to random mutation events, while natural selection controls a strongly diverse population, mutation being a small correction. Therefore, as long as selection is stronger than the drift, determinism wins.

To obtain the validity range of deterministic description, we need to consider finite N and estimate the width of the probability density. For this end, we employ the method of perturbation in $1/N$ to solve eq. (1.66) in the next approximation. We use an ansatz of the automodel form

$$\rho(f,t) = \frac{1}{w(t)} F\left(\frac{f - f_d(t)}{w(t)}\right) \tag{1.71}$$

where $F(u)$ is a normalized function, $\int du\, F(u) = 1$. The width of the probability density $w(t)$ is assumed to be much less than f_d, which is the validity condition for deterministic approximation. We will find the interval of N where this assumption actually holds further. Since $\rho(f,t)$ changes in f much faster than $M(f)$, we can expand $M(f)$ in the left-hand side of eq. (1.66) linearly in $f - f_d$. In the right-hand side, $O(1/N)$, we retain only the largest terms in $1/N$ by approximating $f(1-f) \approx f_d(1-f_d)$ and $M(f) \approx M(f_d)$. Substituting eqs. (1.71) into (1.66), we obtain

$$\begin{aligned} -f_d'(t) + M[f_d(t)] &= \left\{ \frac{s f_d(t)[1 - f_d(t)](1 - f_d)}{2w(t)} \right\} \frac{F''(u)}{F'(u)} \\ &+ \frac{1}{\sqrt{Ns}} \left\{ w'(t) - M'[f_d(t)] w(t) \right\} \frac{F(u) + uF'(u)}{F'(u)} \end{aligned} \tag{1.72}$$

where primes denote the first derivatives of the corresponding functions. To solve eq. (1.72), we make an observation that the braced terms on the right-hand side and the left-hand side depend only on time, while the factors multiplying the braced terms are functions only of u. Since u and t are two independent variables, it stands to reason that eq. (1.72) can be correct only if the left-hand side is always zero, and the ratio in braces is a constant. We denote it λ. Therefore, eq. (1.72) splits into three separate differential equations: eq. (1.61) for $f_d(t)$ and the equations for $F(u)$ and $w(t)$:

$$\lambda F'' + uF' + F = 0 \tag{1.73}$$

$$\frac{dw}{dt} - M'(f_d) w - \frac{\sqrt{Ns^3}}{2\lambda} \frac{f_d(1 - f_d)}{w} = 0 \tag{1.74}$$

Without the loss of generality, constant λ in the above equations can be simply set to 1 by rescaling $u \to \lambda^{1/2} u$, $w \to \lambda^{-1/2} w$, which leaves the probability density, eq. (1.71), unchanged. The solution of eq. (1.73) is a Gaussian

$$F(u) = \frac{1}{\sqrt{\pi}} \exp\left[-\frac{(u - C)^2}{2} \right] \tag{1.75}$$

where the normalization prefactor ensures that $\int du\, F(u) = 1$. We can set $C = 0$, since any other choice is equivalent to a shift in the definition of f_d in eq. (1.71), which keeps the probability density invariant.

To solve eq. (1.74) for width $w(t)$ this equation can be reduced to two simpler equations by substituting $w(t) = y(t)\phi(t)$ and demanding that $y(t)$ satisfies equation

$$\frac{dy}{dt} - M'(f_d)y = 0 \tag{1.76}$$

After solving the resulting equation for $\phi(t)$ and solving eq. (1.76), we obtain the general solution of the form

$$w(t) = \exp\left[\int dt\, M'(f)\right]\sqrt{C' + \frac{1}{N}\int dt\, f(1-f)\exp\left[-2\int dt\, M'(f)\right]} \tag{1.77}$$

where $f \equiv f_d(t)$, we remind, is a time-dependent function. We can change the variable of integration from t to f by using eq. (1.61). With an initial condition $w(0)$, the width w becomes a function only of f

$$w = w(0) + \frac{|M(f)|}{\sqrt{N}}\left|\int_{f_0}^{f} d\phi\, \frac{\phi(1-\phi)}{M^3(\phi)}\right|^{\frac{1}{2}} \tag{1.78}$$

$$(f - f_{ss})(f_0 - f_{ss}) > 0 \tag{1.79}$$

The condition in eq. (1.79) ensures convergence of the integral in f in eq. (1.78). It also follows from the fact that $f(t)$ never crosses its steady-state level (Figure 1.8).

To test the quasi-deterministic criterion, we now calculate the relative fluctuation of f given by ratio $w/[f(1-f)]$ and require it to be much less than 1. We start from the case of the steady state. The problem here is that the integral in eq. (1.78) diverges when the upper limit of the integral f is set to f_{ss}. Hence, we need to calculate it as a limit $f \to f_{ss}$. We consider f to be close to f_{ss}, then expand $M(f) \approx M'(f_{ss})(f - f_{ss})$, and then evaluate the limit $f \to f_{ss}$. Then, two large terms cancel, and we obtain

$$w_{ss} = f_{ss}(1-f_{ss}) \times \begin{cases} \dfrac{1}{\sqrt{2N\mu}}, & \mu \ll s \\ \dfrac{1}{\sqrt{N\mu}}, & \mu \gg s \end{cases} \tag{1.80}$$

Please note that the steady-state allele frequency, f_{ss}, is different in the two limits in eq. (1.80). We find that the deterministic criterion, $w_{ss} \ll f_{ss}(1-f_{ss})$, is satisfied when $N \gg 1/\mu$, which corresponds to selection regime or selection-mutation regime in Table 1.2 In selectively-neutral regime, $\mu \gg s$, as one can demonstrate from eq. (1.78), the deterministic criterion is the same, even when the population is far

from steady state. The condition on N is more complex if we are far from steady state and selection is important, $\mu \ll s$. In this case, evaluation of eq. (1.78) at $f \gg f_{ss}$ produces promised eq. (1.65).

1.6 Stochastic dynamics in the selectively neutral limit

As we found out when considering the steady state, selection can be neglected altogether at the smallest population sizes, $N \ll 1/s$. In this section, we consider the nonequilibrium dynamics in this neutral regime. We will consider all the thought experiments from Section 1.3: growth competition in a diverse population, fixation of a beneficial allele, transition from a uniform population to the steady state, divergence of populations, and time correlations.

1.6.1 Dynamics of diverse populations and gene fixation

In a diverse population, mutations are negligible due to their small rate, less than once a generation: $\mu N \ll \mu/s \ll 1$. Therefore, genetic drift is the only factor causing the change in mutant frequency with time (Figure 1.1A). The mutant frequency follows a diffusion trajectory until the population becomes completely uniform in one or anther allele, which happens with an equal probability (almost equal if you take into account corrections from selection, which we neglect in Sec. 1.6). Computer simulation in Figure 1.9B illustrates a representative random process. The average generation number that takes for a population to lose diversity (for one allele to become fixed) is equal, as we show below, to the population size, N (Kimura, 1955a; Wright, 1931). This fixation time fluctuates between individual population, and the distribution function has a simple exponential form. The random process can be understood also from the probability density evolution. The latter, initially located near $f = 0.5$, gradually expands until it occupies the interval $[0\ 1]$ and then decays, leaving only two peaks at the uniform states, $f = 0$ and 1 (Figure 1.9A).

Thus, in a random time on the order of N, the population arrives at a uniform state. This fact has a serious impact on ancestral relationship of genomes. Let us classify all individuals in a population into two equal classes and paint each class by a distinct color. Next, we divide each class into two equal subclasses and label them by two distinct shades. After that, we split each subclass into two subsubclasses and paint them by two hues, and so on and so forth. If we repeat this for a sufficient number of times ~ $\log N$, all the individuals in the population will eventually have different color tags.

Consider now evolution of each class and all subclasses. According to the above result, in a time not exceeding a few multiples of the class size, one of the two subclasses within each group will vanish. Likewise, the surviving subclass contains two

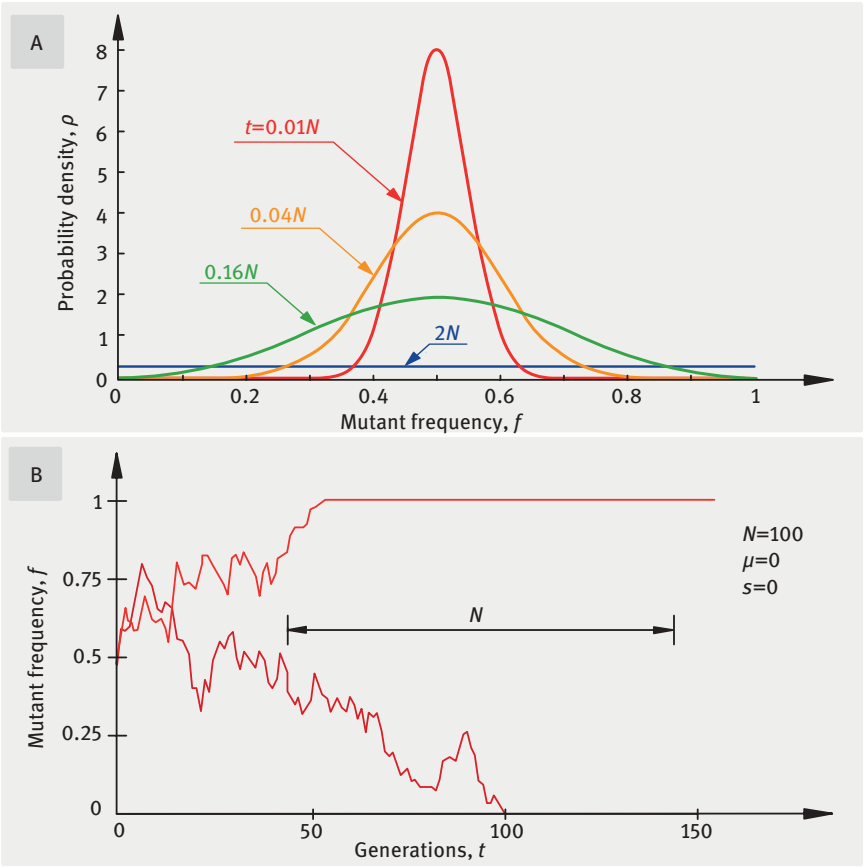

Figure 1.9: Decay of genetic diversity in the drift (selectively neutral) regime. A growth competition experiment for the initial condition $f_0 = 0.5$ and $Ns \ll 1$ is shown. (A) Change in the probability density $\rho(f,t)$ in time according to analytic theory, eqs. (1.81) and (1.82). Moments of time are shown. (B) Two representative dependences $f(t)$ obtained by stochastic simulation of the discrete Wright–Fisher process described in Section 1.2.1 (Rouzine et al., 2001). Parameters are shown.

smaller sub subclasses, one of which also becomes extinct in even shorter time. Hence, in a time interval comparable with population size N, all the individuals will have the same color. In other words, all the individuals will comprise descendants of a single individual. We conclude that any two individuals in a population descend from a single common ancestor, and that this ancestor lived ~N generations ago, times a random factor on the order 1. Rigorous analysis by the technique of the branching process supports this estimate (Kingman, 1982a,b; Rodrigo and Felsenstein, 1999; Rodrigo et al., 1999).

Another very important in practice experiment is fixation of an allele. Suppose a mutation event adds an new allele into an otherwise uniform population at $t = 0$. The allele faces two possible outcomes. After a while, it will either become extinct due to random drift, and this is the most probable outcome, or if it is really lucky, it can spread to the entire population. We need to answer the following questions: (i) What is fixation probability? (ii) Assuming that the allele becomes fixed, what is the average fixation time? As we show later, the fixation probability is on the order of $1/N$ (Kimura, 1962), and the time to fixation is on the order of the population size.

We can also ask more general questions. What is the probability that the progeny (lineage) of an allele will ever exceed a given size of n copies? What is the average growth time to that copy number? As it turns out, the results are analogous to the results for full fixation, except that the subpopulation size n substitutes for the total population size N. These estimates allow us to understand, in a qualitative way, many important results on stochastic dynamics.

1.6.1.1 Main results

In small populations such that

$$N \ll \min\left(\frac{1}{s}, \frac{1}{\mu}\right)$$

natural selection is a small correction and can be neglected (Section 1.4). In most of the interval, mutation enters the equations only through the boundary condition and are negligible in the state, which is already genetically diverse. The listed experiments display two, widely different timescales: a shorter scale associated with random drift, $t \sim N$, and a much longer time, related to mutation rate, $t \sim 1/\mu$.

Let us consider an evolving, very diverse population and focus on the shorter timescale, $t \sim N$. As discussed earlier, $f(t)$ follows diffusion trajectory until it hits the rock bottom, a monomorphic state (Figure 1.9B). We can understand this process in terms of probability dynamics, as follows. The probability density, $g(f, t)$, given by eq. (1.4) spreads out from the point $f = f_0$ until it will have occupied the whole interval of f (Figure 1.9A) (Kimura, 1955a; Wolfs et al., 1990):

$$g(f,t) = \sqrt{\frac{N}{2\pi f_0(1-f_0)t}} \exp\left[-\frac{N(f-f_0)^2}{2f_0(1-f_0)t}\right], \quad 1 \ll t \ll N \tag{1.81}$$

Then it slowly decays in the entire interval, redistributing probability integral towards the two uniform states, $f = 0, 1$:

$$g(f,t) = 6f_0(1-f_0)e^{-\frac{t}{N}}, \quad t \gg N \tag{1.82}$$

The relation between the displacement in f and time t following from eq. (1.81), $f - f_0 \sim \sqrt{tN}$, is similar to that in a standard diffusion process. At $t \gg N$, the net

probability of having a diverse population $\int df\, g(f,t)$ gradually decreases with time in an exponential fashion, eq. (1.82). Using the gas analogy, the probability of polymorphism is being absorbed by the two monomorphic states $f = 0$ and 1, just as gas can form liquid on cold walls.

If initial population is weakly diverse, $f_0 \ll 1$, dynamics of the spread of density $g(f)$ deviates from classical diffusion due to the boundary proximity effect:

$$g(f,t) = Af_0 \frac{2N}{t^2} e^{-\frac{2Nf}{t}}, \quad \sqrt{Nf_0} \ll t \ll N \tag{1.83}$$

where $A \sim 1$ is a constant.

For the allele fixation problem, the probability $G(f)$ that the lineage of a single new allele will reach level f and the average growth time $t_G(f)$ are given by

$$G(f) \sim \frac{1}{Nf}, \quad t_G(f) \sim Nf \tag{1.84}$$

respectively. In particular, we get $G(1/N) \sim 1$, since it is given that a single allele was present in the beginning. The gene fixation probability from eq. (1.84) is $G(1) \sim 1/N$, with the corresponding time $t_G(f) \sim N$ (Kimura, 1962).

1.6.1.2 Derivation

At small population sizes $N \ll 1/s$ considered here, we have use the master equation in the form of eqs. (1.4)–(1.7). Without selection term, eq. (1.7) becomes

$$\frac{\partial g}{\partial t} = \frac{1}{2N} \frac{\partial^2}{\partial f^2} [f(1-f)g] \tag{1.85}$$

Equation (1.85) should be solved with boundary conditions from eqs. (1.5) and (1.6) and specific initial values for $p_0(0),\ p_1(0),$ and $g(f,0)$.

We start from a polymorphic population with a known mutant frequency $f_0 \neq 0$ or 1, which implies

$$p_0(0) = p_1(0) = 0, \quad g(f,0) = \delta(f - f_0) \tag{1.86}$$

Mutations that enter the problem via boundary conditions, eq. (1.6), become important at much longer timescales than those involved in random drift considered here. Hence, we can set $\mu = 0$ in eq. (1.6), which means that $g(f,t)$ either does not diverge at the boundaries of f at all or diverges more slowly than $1/f$ and $1/(1-f)$. Equation (1.85) can be solved for $g(f,t)$ in the general form of a sum over eigenfunctions $h_i(f)$ (Kimura, 1955a):

$$g(f,t) = \sum_{i=0}^{\infty} a_i h_i(f) e^{-\frac{\lambda_i t}{N}} \tag{1.87}$$

$$-2\lambda_i h_i(f) = \frac{\partial^2}{\partial f^2}[f(1-f)h_i(f)] \tag{1.88}$$

$$a_i = \int_0^1 df\, f(1-f)h_i(f)g(f,0) \tag{1.89}$$

The eigenvalues λ_i corresponding to nondiverging solutions of eq. (1.88) and the eigenfunctions $h_i(f)$ are given by

$$\lambda_i = 1 + \frac{i(i+3)}{2}, \quad i = 0,\ 1,\ 2,\ \ldots$$

$$h_i(f) = 2\sqrt{\frac{2i+3}{(i+1)(i+2)}}\, C_i^{(\frac{3}{2})}(1-2f) \tag{1.90}$$

where $C_i^{(3/2)}(x)$ are Gegenbauer polynomials (Abramowitz and Stegun, 1964). The set of functions $[h_i]$ is orthonormal, as given by $\int_0^1 df\, f(1-f)h_i(f)h_j(f) = \delta_{ij}$. Function $g(f,0)$ will be derived below in asymptotic limits in time, for two cases: strong and weak initial polymorphism.

1.6.1.3 Decay of strong polymorphism

Consider a strongly diverse population, that is, $f_0 \sim 1 - f_0 \sim 1$ in eq. (1.86). In the beginning of the evolution process, $g(f,t)$ has a narrow peak at $f = f_0$, so that the factor $f(1-f)$ in eq. (1.85) can be then approximated by constant $f_0(1-f_0)$. Instead of using the general solution in eq. (1.87) which looks very formidable, we treat the simplified equation. Since the initial density distribution, $g(f,t)$, is assumed far from either boundary and very narrow initially, there no characteristic scale in f to compare with. Therefore, it is expected to assume an auto-model form as it evolves in time. Substituting ansatz $g(f,t) = B(t)^{-1}F[B(t)(f-f_0)]$ into eq. (1.85), we solve it for $F(u)$ and $B(t)$ by the same method that we employed for an automodel solution in Section 1.5.4 and arrive at eq. (1.81).

This automodel solution works only in a limited time interval, $t \ll N$, while the probability density peak still remains narrow, as given by $B(t) \gg 1$. In the opposite limit, $t \gg N$, we make use of the eigenvector series in eq. (1.87), where all terms but the first, $i = 0$, can be dropped. Finding λ_0 and $h_0(f)$ from eq. (1.90) and a_0 from eqs. (1.86) and (1.89), we arrive at eq. (1.82).

1.6.1.4 Gene fixation and weak polymorphism

Consider now a slightly diverse population, $f(0) = f_0 \ll 1$. The value of $f_0 = 1/N$ corresponds to a single individual genome in a uniform population. We aim to evaluate the small probability, $G(f)$, of having the lineage of this individual to reach the number of Nf copies before it becomes extinct, and, should this rare event occur, the average time

of growth, $t_G(f)$. A lot of effort has been dedicated to this important problem in various models and scenarios (Barton and Rouhani, 1991; Fisher, 1930; Good et al., 2012; Haldane, 1924; Kimura, 1962; Kimura and Ohta, 1969; Neher et al., 2010). The use of the backward-in-time Kolmogorov equation solves this problem by considering the target frequency, f, fixed and the initial frequency, $f(0)$, as a state variable dependent on time in the past (Kimura, 1962). To be consistent with the rest of this chapter, we will use a semiquantitative derivation based on the forward Kolmogorov equation, which yields the same result within a numerical factor on the order one. Even in the "backward" method, since we deal with range $f \sim 1/N$ corresponding to a few allele copies in a population, the numerical factor in G depends on finer details of a population model, which vary between organisms.

The decay of weak polymorphism can easily be obtained from eq. (1.85). At t such that $1 \ll t \ll N$, the density $g(f,t)$ is not small at $f \ll 1$. Unlike for strongly diverse population, only $1-f$ in eq. (1.85) can be approximated by a constant, 1, while the factor f has to be kept as is. As a result, seeking solution of eq. (1.85) in an automodel form, we arrive at $g(f,t)$ in eq. (1.83).

The total probability of polymorphism in generation t can be obtained right away from eq. (1.83)

$$p_{\text{pol}}(t) = \int_0^\infty df\, g(f,t) = \frac{A}{t}, \quad t > 1 \tag{1.91}$$

As eq. (1.91) confirms, the new lineage will likely disappear after a few generations due to random genetic drift. The factor A is obtained from the equality $p_{\text{pol}}(1) \sim 1$, which yields $A \sim 1$. This value is an estimate within a numeric factor ~1, since the continuous-in-f approach we use breaks down at $t \sim 1$ and $f \sim 1/N$. In the same way, probability $G(f,t)$ that the new lineage will exceed frequency f at time t is given by

$$G(f,t) = \int_f^\infty df' g(f',t) \sim \frac{1}{t} e^{-\frac{2Nf}{t}} \tag{1.92}$$

The probability $G(f,t)$ reaches maximum at $t = 2Nf$ generations. The height and position of this maximum represent the promised estimates for the probability of having the new lineage grow to frequency f, $G(f)$, and for the average time of that growth, $t_G(f)$, respectively, eq. (1.84).

1.6.2 Transition from a uniform population to the steady state

Our next thought experiment is the accumulation of alleles starting from a genetically uniform state. Which allele we chose does not really matter, since selection is neglected anyway, so we choose the starting population comprised of pure wild type

($f_0 = 0$). As the system evolves, mutants are generated, most of them go extinct, but eventually one of them will become fixed, as described in the previous subsection, and population will lose wild type and become uniformly mutant. Then, new wild-type alleles will be injected again, then go extinct, them again until the population will eventually go back to the purely mutant state. In the long run, the population will be going back and forth between the two uniform states (Figure 1.10B). In the long run, after a few switches, statistical properties will no longer depend on the starting allele, so that the probabilities of the two uniform states will equalize at $1/2$.

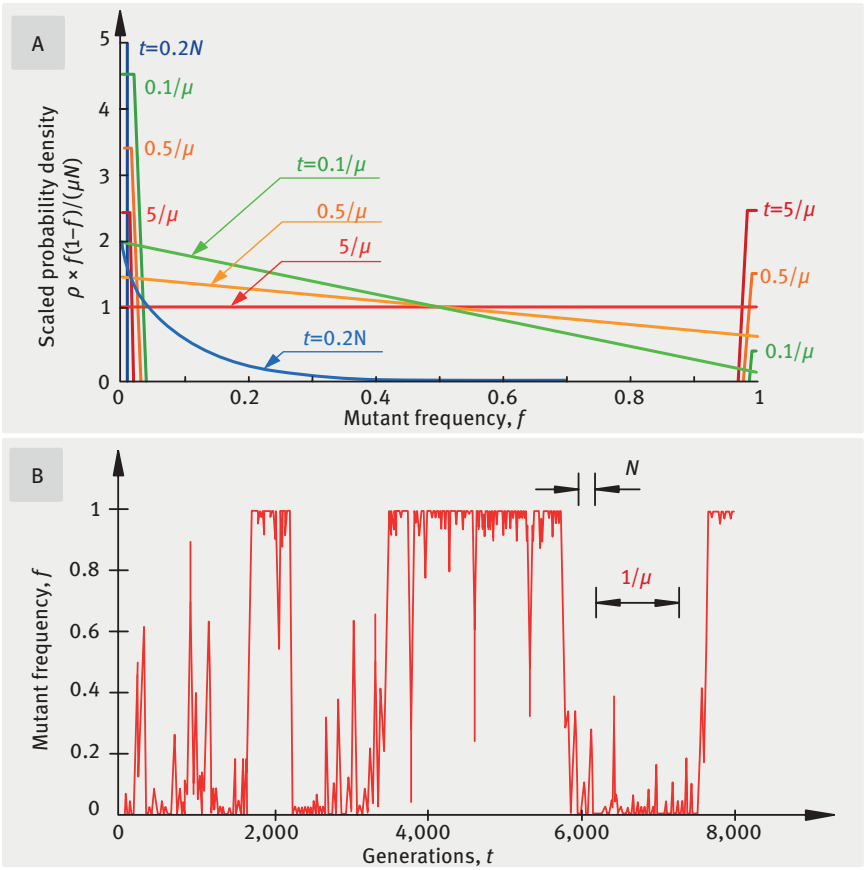

Figure 1.10: Time dependence of the mutant frequency $f(t)$ in the drift regime, $Ns \ll 1$, on a long timescale starting from a uniform state, $f(0) = 0$. (A) Change in the probability density in time, eqs. (1.85) to (1.87). Peaks at $f = 0$ and 1 correspond to the uniform states; their probabilities are shown by the relative peak heights (arbitrary units). Different moments of time are shown by curves of different color. (B) A representative Monte-Carlo process. The double-pointed arrows show the average time between peaks and peak width. Parameters are shown (based on Rouzine et al. (2001)).

The corresponding dynamics of the probability density, $\rho(f,t)$, is depicted schematically in Figure 1.10A. The initial state is a narrow peak at $f=0$. As time goes on, the probability density spreads into the interval between 0 and 100%, and a new small peak at a fully mutant population appears, $f=1$, which represents rare events of an early fixation. Then, the one peak decays and the other grows concurrently until becoming nearly equal; thus the steady state is established asymptotically in the long run (Figure 1.4).

This is, again, similar to a gas system with the liquid condensed on the left wall, which gradually evaporates, diffuses as a gas to the other wall, and condenses again on the other wall. The system is close to equilibrium when both walls have about the same amount of condensate with residual gas remaining between the walls.

In addition to the formal analysis, it is useful for intuition to look at the transition to the steady state from the angle of a typical random process. Such process can be generated by a Monte-Carlo algorithm based on the discrete-time, discrete-n model in Section 1.2. If $\rho(f,t)$ is compared to the local gas density, a dependence of the mutant frequency on time can be compared to a gas particle trajectory. A simulation of such a stochastic process, together with the relevant timescales, is shown in Figure 1.10B (Rouzine and Coffin, 1999a). The process resembles a corrupted telegraph signal flipping back and force between 0 and 1. The numerous random spikes on the bottom and on the top are due to new lineages created by mutations that became extinct.

Monte-Carlo simulation and probability density dynamics demonstrate the existence of two different timescales . The average waiting time for a switch from pure wild type to pure mutant or back is on the order inverse mutation rate $1/\mu$. In agreement with this fact, on this timescale, the probability density equilibrates between mutant and wild type (Figure 1.4A). The time spent on a successful sweep from 0 to 1 is much shorter, on the order of population size N. This corresponds to the time of the formation of the tail of probability density. Analytically, the two timescales can be obtained formally from the evolution equations (1.4) to (1.6), or estimated in the language of allele fixation. The two results are in agreement with each other and are confirmed by simulation (Figure 1.10B). The total probability to find a diverse state is, at any time, small and on the order of μN, as is also the case in the steady state (Section 1.4). Importantly, this probability is established after approximately N generations, which timescale is associated with genetic drift. This happens much faster than the full equilibration, where mutation sets the timescale.

1.6.2.1 Main results

Transition from a genetically uniform state with $f=0$ to a steady state has two phases (Figure 1.10A). In the first, fast phase, which takes place on timescale $t \sim N$, the density $g(f,t)$ grows a thin tail in the interval $0<f<1$. The probability of uniform state, p_0,

remains close to 1 at all times. This means that mutant alleles are found only in rare populations:

$$g(f,t) = \frac{2\mu N}{f} e^{-\frac{2Nf}{t}}, \quad t \ll N \tag{1.93}$$

In the second slow phase, $t \sim 1/\mu$, $p_0(t)$ and $p_1(t)$ decay and increase, respectively, both converging to $1/2$:

$$g(f,t) = \frac{\mu N}{f(1-f)} \left[1 + (1-2f)e^{-2\mu t}\right], \quad t \gg N \tag{1.94}$$

$$p_{0,1} = \frac{1 \pm e^{-2\mu t}}{2} + O(\mu N) \tag{1.95}$$

The expectation value and variance of f change in time as given by

$$\bar{f}(t) \approx p_1 = \frac{1}{2}\left[1 - e^{-2\mu t}\right] \tag{1.96}$$

$$V_f(t) = \frac{1}{4}\left[1 - e^{-4\mu t}\right] \tag{1.97}$$

They both saturate, as expected, at their respective steady-state values we evaluated in Section 1.4, eq. (1.55). Interestingly, the time dependence of $\bar{f}(t)$ in eq. (1.96) is exactly the same as in the deterministic limit under selective neutrality, eq. (1.62), despite of enormous fluctuations between realizations. Interestingly, the mean intrapatient distance with its variance, T, V_T obtained from eq. (1.94), do not depend on time on this timescale and are given by their steady-state values, eq. (1.56). This is because genetic distance is drift–driven and reaches its steady level much earlier than the occurrence of full equilibration at $t \sim N$.

These two timescales can be understood intuitively based on the gene fixation results, eq. (1.84), as follows. Mutant genomes are generated with a small probability, μN, per generation. The lucky ones are fixed with a small probability, $G(1) \sim 1/N$. Therefore, a typical time interval between the switches from full wild type to uniform mutant and back to wild type on the order of $\sim N/(\mu N) = 1/\mu$. The short transition time over which the population sweeps from one end to another can be estimated from the fixation time, $t_G(1) \sim N$, eq. (1.84) (Figure 1.10A).

1.6.2.2 Derivation of the transition from a uniform population to the steady state

Let the population consist initially of one allele only, for example, $f = 0$. We will solve eq. (1.85) with the initial conditions

$$p_0(0) = 1, \quad p_1(0) = 0, \quad g(f,t) \equiv 0 \tag{1.98}$$

In a short time span, $g(f,t)$ remains mostly at $f \ll 1$, and we assume $p_0 \approx 1, p_1 \approx 0$ which will be confirmed below. Hence, diffusion equation (1.85), with boundary conditions (1.6), takes a simplified form

$$\frac{\partial g}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}(fg), \quad [fg]_{f\to 0} = 2\mu N \tag{1.99}$$

At the initial conditions given by eqs. (1.98), eq. (1.99) has the automodel solution in eq. (1.93). One can obtain this solution by using ansatz $g(f,t) = A(f)F[B(t)f]$ and separating it into two equations, as described in Section 1.5.4.

At $t \sim N$, the probability density spreads onto the entire interval of f. To confirm our assumption about a small change in p_0, the decrease in probability p_0 can be estimated by integrating the first part of eq. (1.5) from $t \sim 1$ to $t \sim N$, which yields $1 - p_0 \sim \mu N \log N \ll 1$. This confirms our initial assumption that p_0 remains close to 1 in this interval of time. From eq. (1.93), average frequency f and genetic distance T defined in eqs. (1.36) and (1.38) are

$$\bar{f}(t) \approx \mu t, \quad \overline{T}(t) \approx 2\mu t$$

At much longer times, $t \ll N$, the probabilities p_0 and p_1 are slowly decaying and increasing, respectively, with the speed depending on the small mutation rate. Since the characteristic diffusion time is short, $t \sim N$, the density $g(f,t)$ is in quasi-equilibrium at all times quickly adjusting to the relatively slow variation in $p_0(t)$ and $p_1(t)$. Putting $\partial g/\partial t = 0$ in eq. (1.99), we obtain the general solution

$$f(1-f)g(f,t) = C_1(t) + C_2(t)f$$

Finding functions $C_1(t)$ and $C_2(t)$ from boundary conditions, eqs. (1.5) and (1.6), and using $q(f,t) = C_2(t)$ obtained from eq. (1.2) with $s = \mu = 0$, $\rho = g$, we arrive at the desired solution for $g(f,t)$ and $p_{0,1}(t)$, eqs. (1.94) and (1.95).

1.6.3 Population divergence and the time correlator

We now consider the divergence of two populations separated from a parental population in steady state at time $t = 0$ and the time correlation function of mutant frequency the memory in random fluctuation in the steady state (Section 1.3 above). As it turns out, they both are characterized by the long timescale of neutral dynamics, $1/\mu$, that we have obtained in Section 1.6.2. In the genetic divergence experiment, the genetic distance, $D(t)$, increases in time from 0 to the maximum corresponding to the limit when the two populations become statistically independent.

1.6.3.1 Main results

After separation at $t = 0$ from the same population, the relative genetic distance between populations, eq. (1.44), increases as

$$\bar{D}(t) = \frac{1}{2}\left(1 - e^{-4\mu t}\right) \tag{1.100}$$

The time correlation function for a steady state of a single population, eq. 1.45, decays with time exponentially, as given by

$$K(t) = e^{-2\mu t} \tag{1.101}$$

The three characteristic timescales, the half-time of divergence in $\bar{D}(t)$, the half-time of the correlation function decay in $K(t)$, and the time of equilibration c (previous section) are all on the same order, the inverse mutation rate, $1/\mu$. The reason for this similarity is that all three timescales are equal to the time it takes for an allele to emerge and cross stochastic threshold to escape extinction.

1.6.3.2 Derivation

We start from a stochastic, steady-state population of size N that happens to have, at some moment, which we denote $t = 0$, allele frequency $f = f_0$. Then, the population is split in two populations, which are assumed to grow quickly to the parental size, N. In Section 1.3, we expressed the average relative distance between the two populations, $\bar{D}$, in terms of the conditional variance $V_f(t|f_0)$ for the given initial value of f_0, as given by eq. (1.44). We can simply the procedure of finding $\bar{D}$ using the fact that, in the drift regime, the system is usually genetically uniform, either fully mutant or fully wild type, except for relatively short sweeps between the two sides (Figure 1.10B). Therefore, we can neglect with the polymorphic part of the distribution $g(f)$ and approximate the distribution density $\rho_{ss}(f_0)$ with a sum of two delta–functions

$$\rho_{ss}(f_0) \approx \frac{1}{2}\left[\delta(f_0) + \delta(1 - f_0)\right] \tag{1.102}$$

Hence, we need to know the value of variance, $V_f(t|f_0)$, at only, $f_0 = 0$ and 1. The expression for $V_f(t|0)$ has already been derived in eq. (1.97). From the symmetry between the two alleles in the drift regime, we have $V_f(t|0) = V_f(t|1)$. We arrive at the desired result, eq. (1.100), by combining eqs. (1.44) with (1.102).

Note that this approximation cannot be used to calculate the average polymorphism $\bar{T}$, since it would yield 0. Polymorphism is on the order of μN, eq. (1.56). For calculating $\bar{f}$ and V_f, however, this approximation works.

Function $K(t)$ that quantitates the timescale of fluctuation of f in a single steady-state population can be expressed in terms of the conditional expectation value $\bar{f}(t|f_0)$, eq. (1.46). As in the case of the interpopulation distance, the value of f_0 which mostly defines the integral in eq. (1.46) is $f_0 = 1$. We have $\bar{f}(t|1) = 1 - \bar{f}(t|0)$ from allelic

symmetry in drift regime, and $\bar{f}(t|0)$ was already found in eq. (1.96). Substituting $V_f^{ss} = 1/4$ from eq. (1.55), we obtain the desired equation (1.101).

1.7 Dynamics in the selection-drift regime

In this section, we study nonequilibrium behavior in the most theoretically interesting and biologically important interval of population sizes equally relevant for viruses, bacteria, plants, and animals, $1/s \ll N \ll 1/\mu$ termed "selection-drift regime" in Table 1.2. As we will discover later, the relative roles played by natural selection and stochastic diffusion in the dynamics of population depend strongly on the initial genetic composition of the population, f_0. Specifically, the dynamics of growth competition, $f_0 \approx 0.5$, is almost deterministic, so that this experiment need not be discussed again, as it has already been studied in Section 1.5. In the accumulation experiment, $f_0 = 0$, however, the overall dynamics is very stochastic, except for the average values of the mutant frequency and the intrapopulation distance, which are, remarkably, the same as in the corresponding deterministic conditions. In the adaptation experiment, $f_0 = 1$, the half-time of adaptation is much longer than in the selection regime and strongly fluctuates between realizations. In this last two experiments, stochastic effects are as important as natural selection.

1.7.1 Accumulation of deleterious mutations

Just we discussed for the drift regime in Section 1.6, accumulation of mutants corresponds to the extension of the probability density $\rho(f,t)$ initially localized as a delta-function peak at $f = 0$, which corresponds to a uniformly wild-type population, onto the interval between 0 and 1. In contrast to the drift regime, the end steady–state in selection-drift regime is very asymmetric with respect to $f = 1/2$ (Figure 1.5). Mutation and drift opposed by negative selection form a thin exponential tail in $\rho(f,t)$ at small f indicating that diverse populations are rare and only weakly diverse (Figure 1.5). As a result, the probability of wild-type state p_0 remains close to 1 and the time to the steady state is the selection scale, $1/s$, as in the selection regime (Section 1.5) which is much faster than in the drift regime where it is very long, $1/\mu$.

Figure 1.11 shows a representative Monte-Carlo simulation based on the Wright Fisher model (Section 1.2). A single deleterious allele is generated and starts a new lineage. The growth of lineage initially occurs under the condition that random drift overwhelms selection. The maximum frequency that can be reached by this clone at equilibrium is limited by negative selection and is on the order of $1/(Ns)$, stochastic threshold, which corresponds to the clone size of $1/s$ copies (Figure 1.11 and 1.5). This frequency is much higher than $f = \mu/s$ in deterministic limit (Section 1.4), and is determined by the balance between selection and drift.

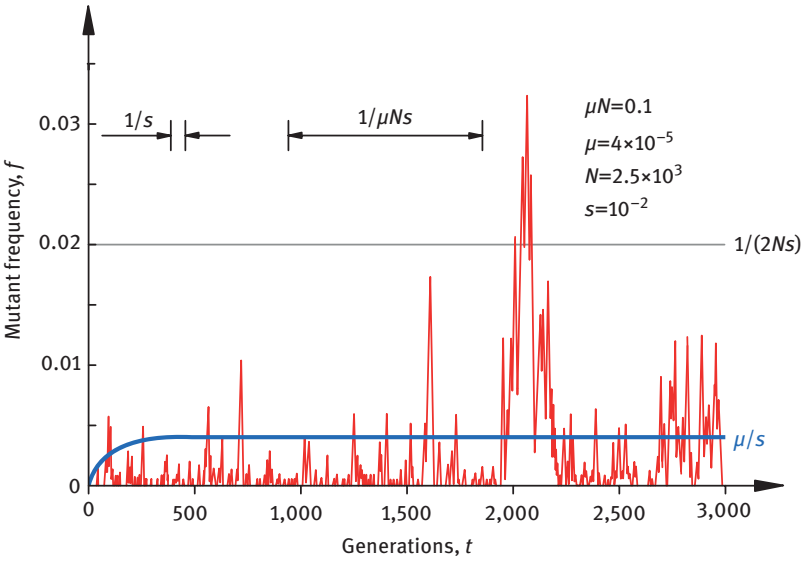

Figure 1.11: Simulated accumulation of deleterious alleles in the selection-drift regime. A random Monte-Carlo run is shown for $1/s \ll N \ll 1/\mu$ and the initial condition: $f_0 = 0$. The double-pointed arrows show the average time between peaks and peak width. Thin horizontal line shows the stochastic threshold. The solid smooth line shows the deterministic dependence for comparison. Parameters are shown (based on Rouzine et al. (2001)).

Above this stochastic threshold, selection is the dominant factor. The further growth of a deleterious allele lineage cannot occur due to negative selection, and it soon starts to shrink and becomes extinct. Soon, a new allele emerges due to mutation and repeats the exercise. As a result, in the long term, we observe an irregular series of random sparse peaks, most are very small, with rare peaks reaching to the tail of the probability density, $1/(Ns)$ (Figure 1.11). The half-life of a mutant subpopulation, which determines the large peak width, is $1/s$ generations as well.

The average time interval between the tall peaks, $\sim 1/(\mu Ns)$, is much longer than their width, $\sim 1/s$, which is why they are so far apart (Figure 1.11). The longer time is the time that it takes for a new allele to emerge and succeed in reaching the stochastic threshold $1/s$. The shorter time is the time of growth and extinction of the lucky clone before that it disappears into oblivion. We can obtain all useful estimates two ways, either from the evolution equation (1.99), or from the more intuitive gene fixation approach, eq. (1.84). The probability of population being diverse can be estimated either as the area under the exponential in Figure 1.5 or as the ratio of the two times, and both methods yield $\sim \mu N$. For comparison, an accumulation experiment in the selection regime ($\mu N = 20$) is simulated in Figure 1.12.

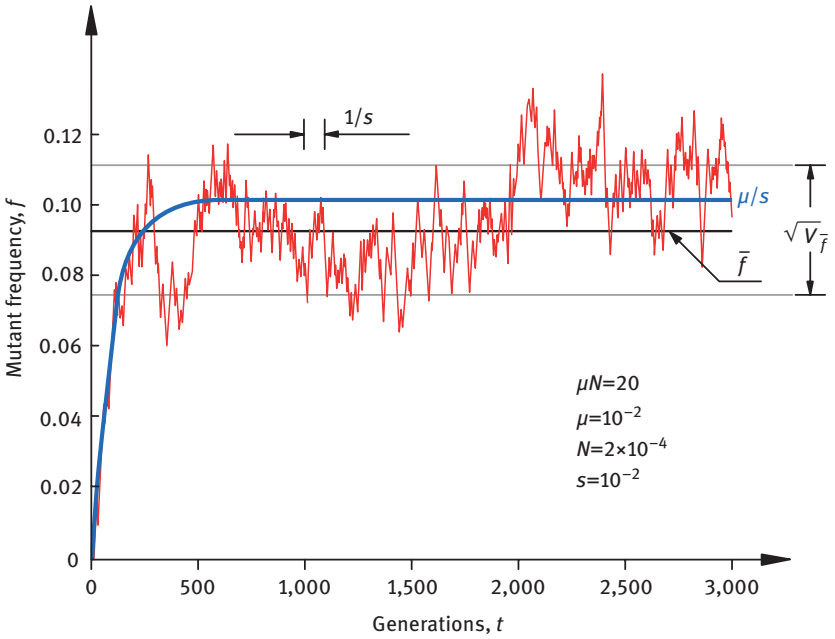

Figure 1.12: Accumulation of deleterious mutants in the selection regime, $\mu N \gg 1$. Wiggly curve is a representative run of Monte-Carlo simulation of the model described in Section 1.2. Horizontal lines show the average mutant frequency $\bar{f}$ and the standard deviation $\sqrt{V_{\bar{f}}}$ in the steady state, eq. (1.58). Parameters are shown (based on Rouzine et al. (2001)).

1.7.1.1 Main results

As mentioned already, in the accumulation experiment probability density $\rho(f)$ sprouts a weak exponential into the interval $0 < f < 1$ (Figure 1.5). The relevant scales can be estimated from the allele fixation argument, eq. (1.84). We consider a stochastic process $f(t)$ which starts from a purely wild-type population and ends up in a steady state (Figure 1.11). A single mutant individual appears and usually gets extinct; sometimes, it grows by random diffusion in f, to a stochastic threshold (see Section 1.4.3):

$$f \sim 1/(Ns)$$

Further growth is efficiently stopped by negative selection. The timescale that it takes to grow to that level can be estimated from eq. (1.84):

$$t_G(1/(Ns)) \sim 1/s$$

Mutation injects new alleles into the population at rate μN per generation. The probability for a new lineage to grow to the stochastic threshold is $G(f) \sim s \ll 1$, eq. (1.84). Hence, the average time interval between such events (high peaks in Figure 1.11) is

$[1/(\mu N)](1/s) = 1/(\mu N s)$. Because $1/s$ is a lifespan of such lineage, A steady-state population can be found in a polymorphic state with probability ~ μN (cf. Section 1.4).

The expressions for the time dependence of mean frequency $\bar{f}$ and variance V_f derived below from the evolution equation have a form

$$\bar{f}(t|0) = \frac{\mu}{s}\left[1 - e^{-st}\right] \tag{1.103}$$

$$V_f(t|0) = \frac{\mu}{2Ns^2}\left[1 - e^{-st}\right]^2 \tag{1.104}$$

In the long term, the two parameters transition to their steady-state values, eq. (1.59). The average intrapatient distance and its variance are given by

$$\bar{T} \approx 2\bar{f} \text{ and } V_T \approx 4V_f$$

Remarkably, the expectation value of the frequency $\bar{f}(t|0)$ in eq. (1.103) exactly coincides with its respective deterministic value, eq. (1.70), although its fluctuations are very large and strongly exceed the average

$$V_f(t|0)/\bar{f}(t|0)^2 = 1/(2N\mu) \gg 1$$

This interesting result is a formal consequence of the linear function $M(f)$ in Fokker–Planck equation (1.22), in the limits $f \ll 1$ or $1 - f \gg 1$. In our population model (Section 1.2), the linearity condition is met asymptotically in a weakly diverse state, including the case of deleterious allele accumulation. In a diploid population with a strong allelic dominance (not considered in our book), mean change per generation $M(f)$ does not need to be linear even at very small f but rather quadratic (Kimura, 1962; Kimura and Ohta, 1969). Then, the average f at small N is not equal to its deterministic value.

1.7.1.2 Derivation

To simplify derivations, we will focus on an interval in N, a bit more narrow than the selection-drift interval (Table 1.2). Specifically, we assume $(1/s)\log(s/\mu) \ll N \ll 1/[\mu \log(1/s)]$. The aim of this additional restriction from below and above is that we can, at the same time, use the more convenient formalism, eqs. (1.5) to (1.7), and also neglect the second probability density peak at $f = 1$ (Figure 1.5). The crossover from the drift regime to the selection-drift regime occurs in the interval, $1/s \ll N \ll (1/s)\ \log\ (s/\mu)$, which is relatively narrow, in the logarithmic sense, and is not considered further (Section 1.4).

As usual, we split the probability density into two parts, corresponding to pure wild-type and diverse populations

$$p_{\text{tot}}(f,t) = p_0(t)\delta(f) + g(f,t) \tag{1.105}$$

where $g(f,0) \equiv 0$ and $p_0(t) = 1$ define the initial conditions. In the long term, density $g(f,t)$ is supposed to cross over with time to the steady-state form in eq. (1.50):

$$g(f,\infty) = (2\mu N/f)\exp(-2Nsf)$$

The dynamic equation and the boundary condition describing this process are

$$\frac{\partial g}{\partial t} = \frac{1}{2N}\frac{\partial^2}{\partial f^2}(fg) + s\frac{\partial}{\partial f}(fg) \tag{1.106}$$

$$(fg)_{f\to 0} = 2\mu N \tag{1.107}$$

These expressions follow from eqs. (1.6) and (1.7) assuming $f \ll 1$ and $p_0 \approx 1$.

In the beginning, $t \ll 1/s$, allelic frequency is far below the stochastic threshold, $f \ll 1/(Ns)$, so that the second (selection) term in the right-hand side of eq. (1.106) is negligible. The dominant terms are mutation introduced via the boundary condition and random genetic drift described by the diffusion term. Hence, in the beginning, $g(f,t)$ for the drift regime applies, found from eq. (1.85). On longer times when stochastic threshold is becoming near, we have to take into account negative selection against the new lineage. In principle, one could solve eq. (1.106) in the entire time interval using an eigenfunction series based on Laguerre polynomials (Abramowitz and Stegun, 1964). Lower momenta of $\rho(f)$, however, can be obtained without a handbook of special functions. After multiplying both sides of eq. (1.106), first by f and second by f^2 and integrating over f, we get a system of two linear ordinary differential equations for $\bar{f}$ and $\overline{f^2}$:

$$\frac{d\bar{f}}{dt} = \mu - s\bar{f} \tag{1.108}$$

$$\frac{d\overline{f^2}}{dt} = \frac{\bar{f}}{N} - s\overline{f^2} \tag{1.109}$$

To obtain eqs. (1.108) and (1.109), we integrated the right-hand side of eq. (1.106) by parts and used the boundary condition in eq. (1.107). Solving first eq. (1.108) and then eq. (1.109) and using the initial conditions $f(0) = V_f(0) = 0$, we obtain the desired expectation value $\bar{f}(t)$ and variance $V_f(0) \approx \overline{f^2}$ given by eqs. (1.103) and (1.104).

1.7.2 Populations divergence and correlations in time

We consider here the divergence of two populations isolated at $t = 0$ and the correlation function of fluctuations of the mutant frequency in time. Both timescales, as we show below, are $1/s$, the inverse selection coefficient. These problems are related, because they both show for how long, on average, the system remembers its previous random fluctuation of genetic composition. The answer obtained below is that this memory is very short; it lasts only an average lifespan of a mutant lineage before it disappears.

1.7.2.1 Main results

If a population is split into two populations at $t = 0$, their interpopulation distance averaged over the initial mutant frequency f_0 is given by

$$\bar{D}(t) \approx V_f(t|0) = \frac{\mu}{Ns^2}\left[1 - e^{-st}\right]^2 \tag{1.110}$$

Correlation function of $f(t)$ in a single steady-state population has a form

$$K(t) = e^{-st} \tag{1.111}$$

Note that the timescale, $1/s$, is much smaller than the value in the drift regime, $1/\mu$ (Section 1.6). The transition between these two values occurs in a relatively narrow (from the logarithmic point of view) interval, $1/s \ll N \ll (1/s)\log(s/\mu)$, which we have already met in Section 1.4.3 when discussing the steady state (Figure 1.6). The rapid crossover is controlled by the dynamics of the probability of a purely mutant population, p_1 (the small peak in Figure 1.5), that rapidly becomes exponentially small when N exceeds $(1/s)\log(s/\mu)$.

1.7.2.2 Derivation

The average relative distance $\bar{D}(t)$ between two split populations increases with divergence time as given by eq. (1.44), in which $V_f(t|0))$ is defined by eqs. (1.36) and (1.37) with $A(f) = f$ and the initial condition $\rho(f, 0) = \delta(f - f_0)$. As follows from the density function in eq. (1.105) and the following equation for g(s),, the uniformly wild-type state, $f_0 = 0$, contributes most to the integral in f_0, eq. (1.44). Using this fact and eq. (1.104), we arrive at eq. (1.110).

In contrast, for time correlation function $K(t)$, as follows from eq. (1.46), only the polymorphic initial states, $f_0 \neq 0$ and $f_0 \neq 1$ are important. Promised eq. (1.111) is obtained by substituting g_{ss} from eqs. (1.50) at $Ns \gg 1$, $\bar{f}(t, f_0) = f_0 \exp(-st)$following from eq. (1.62) at $1 \gg f_0 \gg f_{ss}$, and variance V_f^{ss} from eqs. (1.59) into (1.46).

1.7.3 Adaptation process

Now we consider the most exciting "thought experiment," which created all the diversity of life on the Earth under changing conditions: the adaptation process. We will stick to our basic model with constant conditions and a single locus. Now the initial population is uniformly mutant and acquires beneficial alleles. We will encounter the same timescales and the scale of f as in the case of accumulation of deleterious alleles considered in Section 1.7.2. As in the latter case, natural selection and random drift dominate larger and smaller minority clones, respectively. The crucial difference is that, in this case, natural selection is positive and hence it speeds up rather than slows down the growth of a new lineage.

Again, in order to be fixed, a new allele has to survive initial genetic random drift and reach the frequency of stochastic threshold

$$f \sim 1/Ns$$

at which point natural selection and random drift have the same order of magnitude effect (Sections 1.4 and 1.7.1). The small probability of that event, $G(f)$, is on the order of selection coefficient, s (eq. (1.84)). More precisely, for the virus model with Poisson distribution of progeny number per individual, the probability of fixation is (Haldane, 1927)

$$G = 2s$$

However, if an allele can make it above the threshold, $f > 1/(Ns)$, natural selection will take care of the rest and fix an allele with a probability close to 1, over the deterministic timescales, $t \sim 1/s$ (Section 1.5). Therefore, the weak link in the adaptation process is $f(t)$ growing above the stochastic threshold while drifting randomly. After that, the new lineage will reach the number close to 100% and establish a new steady state, as we have described in Section 1.7.1. Stochastic dynamics below the critical size is the same as for accumulation experiment. Therefore, the average waiting time for adaptation to start is calculated as the product of the fixation probability, ~s, and of the rate at which mutation makes new alleles each generation, μN. The result is the waiting time ~$1/(\mu Ns)$. This result gives the same timescale as a typical time between two high peaks in the mutant accumulation regime (Figure 1.11). Examples of simulated adaptation are shown in Figure 1.13. Figure 1.14 shows dynamics of the probability density, including the diverse populations (Figure 1.14a) and the uniform states (Figure 1.14b).

1.7.3.1 Main results

We start from the uniform mutant state, $f(0) = f_0 = 1$, and monitor how population transitions are close to a pure wild type, $f \sim 1/(Ns) \ll 1$. Probability density $\rho(f, t)$ changes in two phases that occur over two timescales, $t \sim 1/s$ and $t \sim 1/\mu Ns$ (cf. the case of mutant accumulation, Section 1.7.2). During the first phase, $t \sim 1/s$, rare populations become diverse with $Nf \sim 1/s$ copies of minority alleles. Accordingly, the probability density sprouts a thin tail of $\rho(f, t)$ at $f < 1$ (Figure 1.14a). The second, longer phase, $t \sim 1/\mu Ns$, is associated with waiting for a sweep to the wild-type side (Figure 1.13). This timescale manifests in the decay of the probability of the purely mutant state, $p_1(t)$(Figure 1.14b). In the sweep phase, the probability density components change in time

$$p_1(t) = e^{-2\mu Nst}, \quad p_0(t) \approx 1 - p_1(t), \quad t \gg \frac{1}{s} \tag{1.112}$$

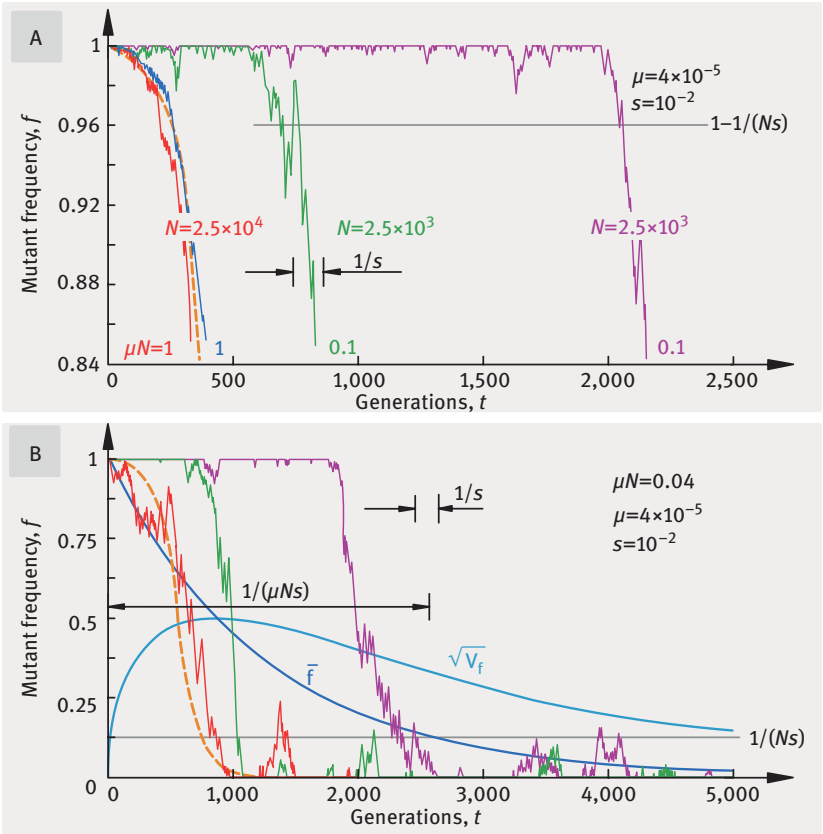

Figure 1.13: Monte-Carlo simulation of adaptation process (fixation of the better-fit allele) in the selection-drift regime, $1/s \ll N \ll 1/\mu$. Adaptation in the deterministic limit, $N = \infty$, is shown by the dashed orange curves for comparison. Parameters are shown. (A) Beginning of adaptation. Two random Monte-Carlo runs are shown for each of two population sizes (shown). (B) Full adaptation at a smaller population size, $N = 1000$. Three random runs are shown. Solid lines show the analytic results for the average and the standard deviation of the mutant frequency, eq. (1.114). Double arrows show two timescales: average waiting time for a switch and its length (based on Rouzine et al. (2001)).

$$g(f,t) = \frac{2\mu N}{f(1-f)}\left\{p_1(t) + [1-2p_1(t)]e^{-2Nsf}\right\} \tag{1.113}$$

where we made a small relative error, $O(\mu N)$. These results are plotted in Figure 1.14. The expectation value and variance of parameters f, T are

$$\begin{aligned} \bar{f}(t) &= p_1(t) \\ V_f(t) &= p_1(t)[1-p_1(t)] \\ \bar{T}(t) &= 4\mu N p_1(t), \quad t \gg \tfrac{1}{s} \end{aligned} \tag{1.114}$$

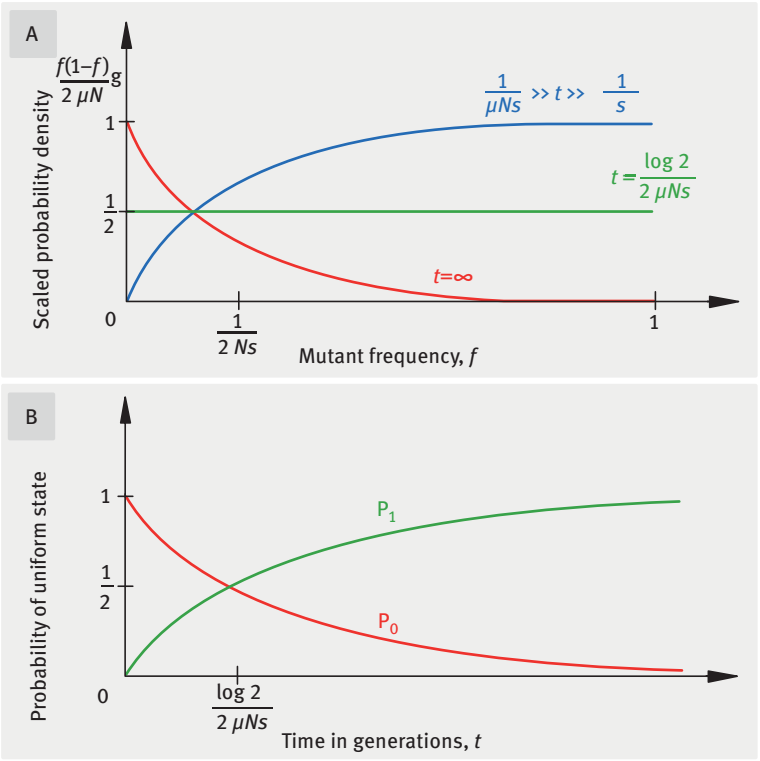

Figure 1.14: Adaptation (fixation of an advantageous variant) in the selection-drift regime, $1/s \ll N \ll 1/\mu$. (A) Dynamics of the scaled probability density of mutant frequency $g(f,t)$ for diverse populations, $0 < f < 1$, at three times t(shown). (B) Probabilities of uniform states, wild-type p_0 and mutant p_1, as a function of time, eqs. (1.112) and (1.113) (based on Rouzine et al. (2001)).

with $p_1(t)$ taken from eq. (1.112).

Equation (1.114) has a finite accuracy as well. Strictly speaking, the three observables do not become zero in the long term, as it follows from eq. (1.114), but cross over to small steady-state values, eq. (1.59) (Figure 1.5). Also, eq. (1.114) has an apparent discrepancy: although the initial population is uniformly mutant with $\bar{T}(0) = 0$, eq. (1.114) predicts a finite distance at small times $t \sim 1/s$. The reason for the difference is that the average distance $\bar{T}(t)$ increases from zero and reaches a plateau relatively early at $t \sim 1/s$. Equation (1.104) is derived for longer timescales, as shown.

We emphasize that the average waiting-for-adaptation time, $1/\mu Ns$, is much longer than in the deterministic regime, $(1/s)\log(s/\mu)$ (Figure 1.13). Therefore, stochastically small populations have slower adaptation.

1.7.3.2 Derivation

We will start with master equations (1.4)–(1.7) and the initial conditions that describe this case:

$$g(f,0) \equiv 0, \quad p_0(0) = 0, \quad (0) = 1$$

We will focus on the second longer phase of evolution with the timescale, ~$1/(\mu Ns)$, where the probability of purely mutant state $p_1(t)$ decreases from 1 to nearly 0 and, respectively, the probability of purely wild type $p_0(t)$ increases from 0 to almost 1. We will use the fact that the equilibration of a polymorphic state, $f \neq 0$, $f \neq 1$, does not depend on the slow mutation process and take into account natural selection and genetic drift only. Therefore, probability density of a polymorphic state, $g(f,t)$, rapidly adjusts to the slow changes in $p_0(t)$, $p_1(t)$. Setting the condition of quasi-equilibrium $\partial g/\partial t \approx 0$ in eq. (1.7) and solving the resulting equation, we get the general solution in the form

$$q(f,t) \equiv q(t)$$

$$g(f,t) = \frac{1}{f(1-f)}\left[-\frac{q(t)}{s} + A(t)e^{-2Nsf}\right] \tag{1.115}$$

where coefficients $q(t)$ and $A(t)$ change in time very slowly compared to selection timescale $1/s$. Finally, substituting eq. (1.115) into the boundary conditions in eqs. (1.5) and (1.6) and finding from resulting equations functions $q(t)$, $A(t)$, $p_0(t)$, and $p_0(t)$, we obtain desired equations (1.112) and (1.113).

Chapter 2
Multi-locus theory of asexual populations

2.1 Clonal interference and genetic background effects strongly modify evolutionary dynamics

In Chapter 1, we have analyzed a simplest model of evolution assuming an isolated genomic site. Real-life populations have a large number of genetically diverse sites inherited together. Even RNA viruses with their short genomes, such as HIV and influenza, have hundreds of variable loci. The genetic difference between two randomly sampled humans is $\sim 0.1\%$, which corresponds to several millions of variable sites. The multi-site context requires another, more general theory. Earlier work in population genetics offered overwhelming theoretical arguments (Felsenstein, 1974; Fisher, 1930; Hill and Robertson, 1966; Maynard Smith, 1971; Muller, 1932) and ample experimental evidence (Rice, 2002) that, in the absence of recombination intrinsic for sexual reproduction, genetic compositions at different loci interfere with each other strongly, so that the one-locus model discussed in Chapter 1 is not directly applicable. Quantitative results obtained in the one-site framework, although useful in some special cases, cannot be directly applied to multi-site evolution due to various interference effects. These effects result from the pervading factor of co-inheritance (linkage), the fact that genetic information is passed from a parent to the offspring altogether, rather than spreading it across the population. (Such a spread is possible, to an extent, in bacteria.)

"Clonal interference," also known as Fisher–Muller effect (Fisher, 1930; Muller, 1932), is one of the most important consequences of linkage. Clonal interference (CI) requires three evolutionary factors to be present: (i) natural selection, (ii) the absence of frequent recombination, and (iii) limited population size. Essentially, it can be viewed as competitive exclusion. Individuals with different beneficial alleles grow in number and compete with each other for space in a population (Figure 2.1). The winner is the lineage with the largest fitness gain added over beneficial alleles offset by the presence of deleterious alleles. The winning lineage eliminates the other lineages from the population.

However, natural selection is not enough for clonal interference to take place. If populations are extremely large, exponentially large in the number of loci, it disappears (Rouzine et al., 2003). In this limit, selection exists without clonal interference: the frequent formation of countless nested clones with new mutations within already expanding clones effectively unlinks the site. Indeed, in infinite population, mutation at multiple sites can rapidly generate any possible sequence. Thus, instead of competing alleles, they are produced almost instantaneously at all sites as a system of deeply nested clones, and each clone is large.

https://doi.org/10.1515/9783110615456-002

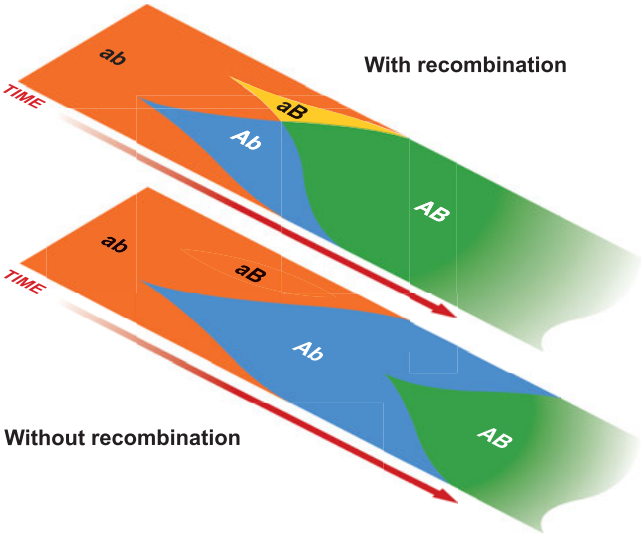

Figure 2.1: Clonal interference is opposed by recombination. Bottom: Beneficial mutations at two different genomic sites in the original variant ab generate better-fit haplotypes, Ab and aB. These new lineages interfere with each other's growth. The fitter clone is taking over the population. Eventually, a second mutation generates fittest haplotype AB. Top: In the presence of frequent recombination of genomes (sexual reproduction), the two clones are quickly combined to form the fittest haplotype AB (based on Imhof and Schlotterer (2001)).

Therefore, the interference is a cocktail of three essential ingredients: natural selection, weak or no recombination, and a limited population size.

A task of great practical importance is predicting the average rate of evolution of a population as a function of system parameters. This can be measured by the average substitution rate and the adaptation rate, which both can be either positive or negative. Accumulation of beneficial alleles leads to the increase in fitness, that is, adaptation. The adaptation rate determines the speed of average fitness increase. Due to the effect of CI, the adaptation rate in a multi-locus system is slowed down compared to a system of independent loci described by a one-site model in Chapter 1 (Maynard Smith, 1971; Rouzine et al., 2003). In contrast to the independent-site case, the adaptation rate does not increase proportionally with the number of loci, L, population size, N, or the mutation rate per site, μ. In the extreme scenario without nested clones considered in Section 2.2, beneficial alleles at different sites have to spread through the population one by one. The speed of progression of the "traffic jam" does not depend on the queue length. Nested clone formation partly offsets the adverse effect of linkage and slightly accelerates the traffic jam, similar to cars that could sometimes drive over each other.

The technical difficulty with incorporating nested clones into the model is that the number of possible sequences increases exponentially with the locus number,

2^L for binary sequences and 4^L if all bases are possible at each locus. Historically, three approximations described in Sections 2.2, 2.3–2.7, and 2.9, respectively, approached this difficulty in a different way incorporating an increasing degree of biological realism.

2.2 Two-clone approximation of clonal interference

Gerrish and Lenski (1998), who proposed the term "clonal interference," considered a model with a positive selection coefficient, which varies among sites according to an exponential distribution. The model considered two competing clones created by beneficial mutation from the same original strain at two randomly chosen sites and neglected mutation at other sites. The mutation with the larger s spreads to the population, and the one with the smaller s becomes extinct. This pioneering approach successfully explained the observed dependence of the adaptation rate of bacteria *E. coli* on the population size, N, and the mutation rate per genome, U_b (Arjan et al., 1999; Gerrish and Lenski, 1998). However, because the two-clone approximation neglects multiple competing clones and nested mutations, it does not apply at very large N. Adding a third site somewhat extends the interval of the applicability of this approach (Schiffels et al., 2011). In Sections 2.3–2.7, we will consider interference events at a large number of sites with a fixed value of s. A unifying approach with variable s and multiple loci will be described in Section 2.8.

The argument of Gerrish and Lenski (1998) unravels as follows. A population is assumed to be genetically uniform until a beneficial mutation appears at time $t = 0$, $f(0) = 1 - 1/N$. The beneficial mutant, being better fit, slowly displaces the ancestral variant until taking over the entire population, assuming the absence of any interfering mutation. The mean number of interfering alleles is the expected number of new alleles emerging during this process that can cross stochastic threshold (Chapter 1) and has a larger fitness than the first beneficial allele. The total number of new alleles is

$$NU_b \int_0^\infty f(t)dt = \frac{U_b}{s} N \log N \tag{2.1}$$

where

$$f(t) = \frac{f(0)}{f(0) + [1 - f(0)]e^{st}}$$

is the frequency of the original variant. The last equation represents a standard one-locus result, eq. (1.62), with $f_{ss} = 0$.

We assume that the effects of beneficial mutations at different sites are sampled for an exponential distribution, as is frequently observed in experiment (Acevedo

et al., 2014; Imhof and Schlotterer, 2001; Kassen and Bataillon, 2006; Stern et al., 2014). The probability density for s is assumed to be $(1/s_0)e^{-s/s_0}$, where s_0 is the average mutational effects that may be determined from the empirical data .

The probability that a beneficial allele has a selection coefficient, s', larger than s and, at the same time, can cross stochastic threshold (Section 1.7.3):

$$\int_s^\infty (2s')(1/s_0)e^{-s'/s_0}ds' = 2(s+s_0)e^{-s/s_0}$$

where $2s$ is the probability of an allele to survive random genetic drift for the virus model (Haldane, 1927) (Chapter I).

Because the loss of the emerging lineage during genetic drift occurs during the first generations, while its loss due to the selective competition is more likely to occur later, when the new clone occupies already a large part of population (Section 1.7), we can make the assumption that these two processes are independent. Therefore, in eq. (2.1), the expected number of mutations that are better than a beneficial allele with selective coefficient s and survive genetic drift is

$$\lambda(s) = 2NU_b \log N \frac{s+s_0}{s} e^{-\frac{s}{s_0}} \approx (2NU_b \log N)e^{-\frac{s}{s_0}}, \quad s \gg s_0 \tag{2.2}$$

This is the mean number of mutations interfering with the new emerging clone. Later, we consider the case of large $s \ll s_0$, whose interval, as we show below, is relevant for large population sizes.

A beneficial allele will be fixed if it survives genetic drift and is not prevented to grow by a better allele in the time interval required for fixation of the first allele. The probability that a beneficial allele with fitness s will be established is the product

$$P_{\text{fix}}(s) = 2se^{-\lambda(s)} \tag{2.3}$$

where the exponential implies independent random events that obey Poisson's statistics. At $NU_b \gg 1$, eq. (2.3) experiences a sharp double-exponential decrease at small s (see eq. (2.2). The last inequality sets the low boundary of the regime of clonal interference. Taking into account the probability density for selective advantage s is $(1/s_0)e^{-s/s_0}$, the probability that a random beneficial allele will become fixed is

$$\langle P_{\text{fix}} \rangle = \left(\frac{1}{s_0}\right) \int_0^\infty 2s\ e^{-\lambda(s)-\frac{s}{s_0}} ds \tag{2.4}$$

With the fixation probability given by eq. (2.4), the average substitution rate of beneficial alleles is

$$V = NU_b \langle P_{\text{fix}} \rangle \tag{2.5}$$

We note that, in the regime of clonal interference, $NU_b \gg 1$, the integrand in eq. (2.4) has a sharp maximum at s equal to

$$s_{\max} = s_0 \log(2NU_b \log N) \tag{2.6}$$

Therefore, the integral in eq. (2.4) can be calculated analytically by expanding the log of the integrand near the maximum in the Taylor series and making it into a Gaussian. For the average accumulation rate of beneficial mutations V, we obtain

$$V = s_0 \frac{\sqrt{2\pi}}{e} \frac{\log(2NU_b \log N)}{\log N} \tag{2.7}$$

This answer has to be compared with the result of the independent-site model, $V_{1\text{site}} = 2s_0 NU_b$, where $2s$ is the probability of fixation (Chapter 1). In contrast, the speed given by eq. (2.7) is not increasing linearly either with population size or mutation rate. In fact, the evolution rate increases very slowly (logarithmically) with N and U_b. At very large population sizes, evolution rate saturates at the maximum value $s_0 \frac{\sqrt{2\pi}}{e}$, far below the independent-locus value, $V_{\{1\text{site}\}}$. The reason for this behavior is that there are more and more interfering mutations in a larger population, so that $\langle P_{\text{fix}} \rangle$ is decreasing as $V/(NU_b)$, nearly compensating the increase in the number of mutational events. In the next sections, Sections 2.3–2.8, we will show that the existence of a system of nested clones with multiple mutations relaxes the saturation of the substitution rate and allows it to keep climbing until finally reaching the one-site limit.

2.3 Traveling-wave method for multiple loci and clones

An early precursor of this approach classified all genomes according to their fitness (Kessler et al., 1997; Tsimring et al., 1996) and applied deterministic dynamics to the fitness classes. They showed the existence of a distribution traveling in an imaginary genetic space, with speed determined by $1/N$ cutoff of the distribution.

A more realistic version of this approach is described in the following sections. It allows to take into account fitness effects of mutations, to make a correct description of stochastic effects, and to accurately derive the speed of the evolution (Rouzine et al., 2008; Rouzine et al., 2003). This method is able to consider an arbitrary number of clones at any level of nesting, as well as account for the effect of genetic background and the other linkage effects.

The number of sequence variants in a population is exponentially large if the number of sites L is large. To account for the immense number of possible sequences, we need to classify them, first, into large groups with the same fitness. We put binary sequences into discrete bins, each bin with a fixed number of less-fit alleles, denoted k, somehow distributed among L sites (Figure 2.2a). The value of k determines

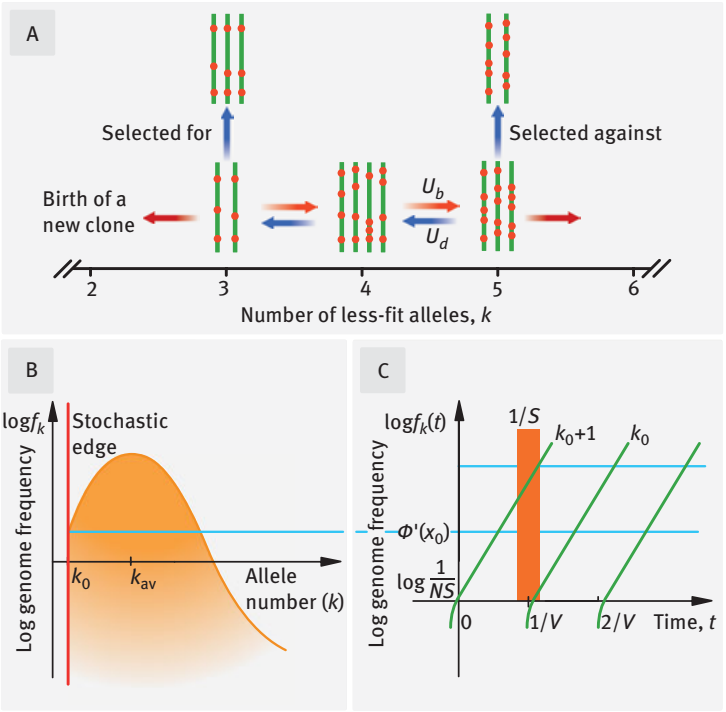

Figure 2.2: Asexual adaptation of multiple linked loci represents a traveling wave in fitness whose speed is determined by stochastic dynamics of the high-fitness tip. (A) Sequences (green lines) are classified into discrete fitness classes according to number of deleterious alleles k (red circles). The evolutionary factors acting on each class size are beneficial and deleterious mutation with respective rates U_b and U_d per genome, random genetic drift and natural selection. (B) The semi-deterministic approach (Rouzine et al., 2008; Rouzine et al., 2003). The dependence of the logarithm of genome frequency f_k on k is determined from a deterministic equation (2.8). The solution is a solitary wave with a cutoff at low mutation numbers, $k = k_0$, and a moving maximum, $k_{av}(t)$. (C) To determine the speed $V = -dk_{av}/dt$, stochastic dynamics of the fittest class $f_{k0}(t)$ is considered (here $V > 0$). A new class is likely to become extinct when its frequency is below stochastic threshold $\sim 1/NS$ (based on Rouzine and Weinberger (2013).

sequence fitness, $-sk$, with respect to the fittest sequence with $k = 0$. Each mutation event at a site is assumed to have a fixed fitness effect s or $-s$, depending on the allele existing at the site: a better-fit allele can have a deleterious mutation to the less-fit allele, fitness effect $-s$, with probability $\mu \ll 1$ per generation, and a less-fit allele can be converted to a better-fit allele, with same probability.

The main approximation confirmed later on, is that, at sufficiently large population sizes (how large, we will find out in the end), almost any nonempty fitness class is large enough in size to be treated deterministically, that is, without taking into account random genetic drift and replacing mutation with its average cumulative effect (Figure 2.2b). The only exception to this approximation will be the fittest

class with the smallest $k = k_0$, which is small and must be treated stochastically, as described in the next sections. The next-to-fittest class, $k = k_0 + 1$, is assumed to be sufficiently big to be approximately deterministic (Rouzine et al., 2008; Rouzine et al., 2003). This approach breaks down the very difficult problem into several relatively simple steps, which we will follow through in Sections 2.3.1 to 2.3.5:

(i) Write a deterministic balance equation that controls dynamics of all fitness classes with the exception of the best-fit class.
(ii) Solve the equation to obtain a traveling-wave solution with an undetermined speed, which has the meaning of the mean substitution rate.
(iii) Demonstrate that the high-fitness end of the distribution ends abruptly at a location depending on the wave speed.
(iv) Express the edge-to-center difference in fitness density in terms of the wave speed.
(v) Find the center value from the normalization condition.
(vi) Identify the deterministic cutoff point of the wave at the high fitness end with the stochastic edge.
(vii) Estimate the average frequency of the edge class from a one-locus-style stochastic consideration.
(viii) Match this result to the deterministic result and thus find the wave speed.

2.3.1 Deterministic equation for fitness classes

We start from the equation of deterministic dynamics (1.61), which was derived in Chapter 1 for two alleles. Now, instead of two alleles, we have many fitness classes comprised of diverse sequences, and we are going to apply this equation to each fitness class., The frequency of class with k less-fit alleles in a population, $f_k(t)$, is described by a finite-difference equation of a form

$$f_k(t+1) - f_k(t) = U_d\, f_{k-1}(t) + U_b\, f_{k+1}(t) - \{U_d + U_b + s(k - k_{\text{av}})\} f_k(t) \qquad (2.8)$$

where k changes from 0 to L and $f_{-1}(t) \equiv f_{L+1}(t) \equiv 0$. By definition, $\sum_k f_k(t) = 1$ any time t. Here

$$U_d = \mu(L - k_{\text{av}}), \quad U_b = \mu k_{\text{av}}$$

are the deleterious and beneficial mutation rates per genome per generation, respectively, and

$$k_{\text{av}}(t) = \sum_{k=0}^{L} k f_k(t)$$

is the average allele number, a term which makes eq. (2.8) nonlinear. Here we neglect multiple mutations per genome and assume $s \ll 1$. Taking into account multiple mutations leads to the same equation (Rouzine et al., 2008; Rouzine et al., 2003).

Since treating a discrete equation is difficult, we will approximate eq. (2.8) with a differential equation in partial derivatives. Because the mutation rates and selection coefficient are all small, by analogy with the one-locus model in Chapter 1, we assume that $f_k(t)$ evolves slowly, so that we can approximate $f_k(t+1) - f_k(t) \approx \frac{\partial f_k}{\partial t}$. A continuous approximation for $f_k(t)$ as a function of k is less trivial. In fact, the dependence of $f_k(t)$ on k can be quite sharp far from the center of the distribution. However, its logarithm changes slowly with k, as given by

$$f_{k+1}/f_k \approx \exp[\partial \log f(k,t)/\partial k]$$

After introducing rescaled time $d\tau = (U_b + U_d)dt$ and rescaled selection coefficient

$$\sigma = s/(U_b + U_d)$$

Equation (2.8) can be reduced to a form continuous in both k and τ:

$$\frac{\partial \log f_k(t)}{\partial \tau} = (1-\alpha)e^{-\frac{\partial \log f_k(t)}{\partial k}} + \alpha e^{\frac{\partial \log f_k(t)}{\partial k}} - \sigma(k - k_{\text{av}}) - 1 \tag{2.9}$$

where

$$\alpha \equiv U_b/(U_b + U_d) = k/L$$

is the fraction of deleterious alleles in a genome. The nonlinear differential equation in partial derivatives, eq. (2.9), is the deterministic master equation for the fitness distribution dynamics.

Note that, strictly speaking, both mutation rates, U_d and U_b (and hence, α and σ) depend on k. The dependence is, however, relatively weak in the limit of a large genome with many loci, $L \gg 1$. We consider the regime when the traveling wave is far away from the best-fit sequence, $k = 0$, $|k - k_{\text{av}}| \ll k_{\text{av}}$. Therefore, both mutation rates change little within the wave interval of k, and we are allowed to replace them with their values at average $k = k_{\text{av}}$.

Equation (2.9) has traveling wave solutions of the form

$$\log f(k,\ t) \equiv\ \phi(x), \quad x = k - k_{\text{av}}(\tau) \tag{2.10}$$

After substituting eq. (2.10) into eq. (2.9), one obtains (Rouzine et al., 2003)

$$\sigma x = (1-\alpha)e^{-\phi'(x)} + \alpha e^{\phi'(x)} + v\phi'(x) - 1 \tag{2.11}$$

Here $\phi'(x) \equiv d\phi(x)/dx$, and new notation v is the scaled speed of the wave movement

$$v = (U_b + U_d)^{-1} dk_{\text{av}}(t)/dt$$

This value can be either negative (adaptation) or positive (accumulation of deleterious mutations) and is measured with respect to the genomic mutation rate.

For the given value of v, eq. (2.11) and boundary condition $\phi(0)$ fully define the form of the fitness distribution, $\exp[\phi(x)]$. Although it is not possible to solve eq. (2.11) analytically, we can use it to find the wave velocity, v, in the general form. We will show that the form of the wave is approximately Gaussian near the center, but deviates from that dependence far from the center, and has a cutoff at high fitness ($x_0 < 0$). The high-fitness tail length ("lead") will be related, in the general form, to the evolution rate, v, and the model parameters.

2.3.2 Width and speed of the traveling wave

Our next step is to derive the maximum value, $\phi(0)$. The main approximation is that the logarithm of $\phi(x)$ is a slow function of x. This is true if the characteristic width of $\phi(x)$ in x is much larger than unit. The condition, as we shall see, is always met when population has many evolving sites, which is the case of present analysis. Because the fitness distribution itself, $\exp[\phi(x),]$, is not approximated with a slow function of x, it does not need to be broad. In the process of adaptation, the wave can be either broad or narrow, depending on a population size. However, the width affects normalization properties.

For a narrow wave, $\mathrm{Var}[k] = \mathrm{Var}[x] \ll 1$, the distribution is mostly localized near its center, $k \approx k_{\mathrm{av}}$. From the normalization condition $\sum_{k=0}^{L} f_k(t) = 1$, we get

$$\phi(0) \approx 0 \tag{2.12}$$

The only exception is the rare case where k_{av} is almost half-integer, so that the neighbor classes, $k_{\mathrm{av}} \pm 1/2$ have similar sizes.

When the wave is broad, $\mathrm{Var}[k] \gg 1$, we can approximate the normalization sum, $\sum_{k=0}^{L} f_k(t)$, with an integral, $\int e^{\phi(x)} dx$. The range of $|x|$ that contributes most to the integral is close to the maximum of distribution, where the derivative of $\phi(x)$ is still small, $|\phi'(x)| \ll 1$. Therefore, we can expand the exponentials in the right-hand side of eq. (2.11) linearly in $\phi'(x)$. After integration in x, we find that

$$\phi'(x) = -\frac{\sigma x}{1 - 2\alpha - v}, \quad |x| \ll (1 - 2\alpha - v)/\sigma \tag{2.13}$$

Integrating this expression in x and taking into account normalization condition $\int dx \exp[\phi(x)] = 1$, for $\phi(x)$ near its maximum at $x = 0$ we obtain

$$\phi(x) = \log \sqrt{\frac{\sigma}{2\pi(1 - 2\alpha - v)}} - \frac{\sigma x^2}{2(1 - 2\alpha - v)} \tag{2.14}$$

In particular, the expression of interest for $\phi(0)$ is

$$\phi(0) = \begin{cases} \log\sqrt{\dfrac{\sigma}{2\pi(1-2\alpha-v)}} & \text{if } \phi(0) < 0 \\ 0 \end{cases} \tag{2.15}$$

Equation (2.14) implies that genome distribution near $k = k_{\text{av}}$ can be approximated with a Gaussian, with the variance

$$\text{Var}[k] = \frac{1-2\alpha-v}{\sigma} \tag{2.16}$$

Due to the condition that the variance in eq. (2.16) is positive, the scaled wave speed, v, has to be less than $1-2\alpha$. In other words, if we are in a regime where deleterious mutations accumulate, which corresponds to $v > 0$, the rate of their accumulation cannot exceed this value. In addition, the Gaussian form is valid in the wave center only if its width is large, $\text{Var}[k] \gg 1$. The actual width of the wave, given the population size and other parameters, will be obtained in the following sections.

Equation (2.16), known as the Fisher fundamental theorem (FFT), connects the width of the genome distribution in fitness to the substitution rate. We note that this FFT is quite general and can be obtained directly from discrete equation (2.8) even if $\text{Var}[k] \ll 1$ (Rouzine et al., 2008). Qualitatively, the theorem states that a broader wave has larger fitness differences and hence, a stronger effect of positive selection on the substitution rate.

2.3.3 High-fitness edge

In Sections 2.3.1 and 2.3.2, we have derived a deterministic wave equation that describes how the fitness distribution moves in fitness space. However, the set of solution is continuous, with a continuous interval of the wave speed, and the wave speed remains undefined. As it turns out, the evolution speed of the entire population is determined by a very small class of best-fit individuals. Behavior of these individuals, due to their small number is stochastic and subject to random mutation and random drift.

Importantly, the deterministic wave profile ends at a point in the high-fitness tail, as shown in Figure 2.3. The biological justification for the existence of the cutoff is the stochastic factor, which is strong on the wave edge. Highly fit genomes at the edge are either gradually gained in the regime of adaptation or gradually lost in the Muller's ratchet regime when deleterious mutations build up. Thus, the cutoff must coincide with the stochastic edge of the wave, when it gradually retreats or advances (Figure 2.3). The rate of the gain or loss of the fittest class, which depends on the mutation rate and the population size, is the limiting factor of the progress

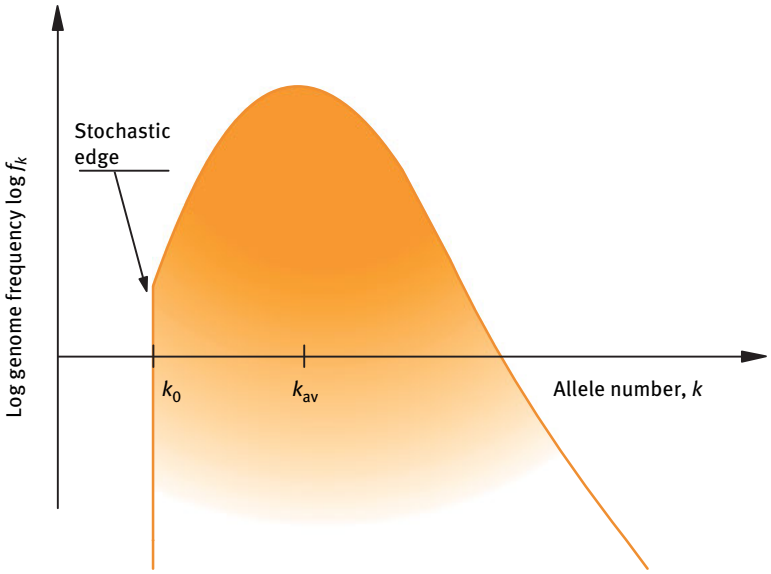

Figure 2.3: Schematic illustration of the solitary wave profile in the log scale, $\log f_k(t)$, Figure 2.2B. There are no genomes with fewer than k_0 deleterious alleles present in the population at this moment. The speed of the wave $V = -dk_{av}/dt$ is determined by how fast alleles are gained or lost at the stochastic edge, $k = k_0$ (based on Rouzine et al. (2008)).

(or regress) of the wave. This logic is extremely different from the logic of independent site models, often used in population genetics. In the next section, the deterministic description of the edge will be matched to stochastic dynamics of the fittest.

We now determine the edge location with respect to the wave center by considering the deterministic bulk of the population. Although we cannot solve eq. (2.11) for $\phi(x)$, explicitly we can extract the important information about the location of the stochastic edge, as follows. We will re-interpret eq. (2.11), instead of equation for $\phi'(x)$, as an equation for function $x(\phi')$. If function $x(\phi')$ has an absolute minimum, it implies that $\phi(x)$ must have an end point at the stochastic edge (Figure 2.4).

Note that the value of function $x(\phi')$, eq. (2.11), is positive infinite for both negative and positive infinite ϕ'. To locate minima, we calculate the derivative and obtain

$$0 = \sigma \frac{dx(\phi')}{d\phi'} = -(1-\alpha)e^{-\phi'(x)} + \alpha e^{\phi'(x)} + v \tag{2.17}$$

Solving this equation for $e^{\phi'(x)}$, we obtain the only positive solution

$$e^{\phi'(x_0)} \equiv u = \frac{1}{2\alpha}\left[-v + \sqrt{v^2 + 4\alpha(1-\alpha)}\right] \tag{2.18}$$

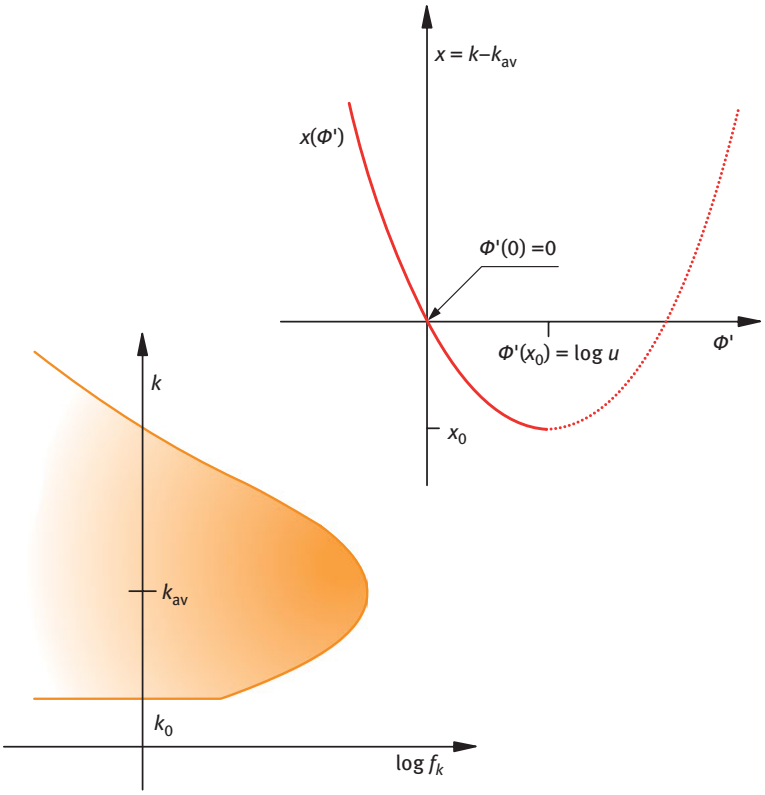

Figure 2.4: The minimum of the function $x(\phi')$determines the location of the stochastic edge, k_0 (based on Rouzine et al. (2008)).

The corresponding value of $x(\phi') \equiv x_0$ follows from eq. (2.11)

$$x_0 = -\frac{1}{\sigma}(1 - 2\alpha u - v\log u - v) \tag{2.19}$$

where new notation u is defined in eq. (2.18), and we have made use of the identity $(1-\alpha)/u = \alpha u + v$ that follows from eq. (2.17). Any mutation class corresponding to $x < x_0$ is empty; there are no genomes beyond the edge.

Now we can verify the validity of our main approximation that the lead |x_0| is long and hence the log probability density of fitness is smooth in x. At a modest speed, $|v| \sim 1$, the condition $|x_0| \gg 1$, implies that $\sigma \ll 1$, or $s \ll U_b + U_d$. For most organisms, genomic mutation rate $U_b + U_d$ is in the range $0.01 - 0.1$. Thus, we consider small selection strength, as everywhere in the book.

2.3.4 Difference between the wave edge and its center

Our aim is to find a general formula that links the wave speed, v, to total mutation rate $U_b + U_d$, population size N, selective coefficient s, and fraction of less-fit sites α. To achieve this goal, we will consider the difference $\phi(0) - \phi(x_0)$. Specifically, we will evaluate the difference in two alternative ways, from deterministic bulk and stochastic edge, and from equating them, obtain the speed.

We can trivially write

$$\phi(0) - \phi(x_0) = \int_{x_0}^{0} \phi'(x)dx \tag{2.20}$$

Then, we calculate the integral by making substitution $x = x(\phi')$, integrating by parts, and using the fact that $\phi'(x_0) = \log u$ and $\phi'(0) = 0$. The former condition is the definition of u, while the second follows from eq. (2.11) (Figure 2.4). Equation (2.20) then takes the form

$$\phi(0) - \phi(x_0) = -x_0 \log u - \int_{\log u}^{0} x(\phi')d\phi' \tag{2.21}$$

The integral in eq. (2.21) can be calculated by integrating both sides of eq. (2.11) in $d\phi'$. After substituting x_0 from eq. (2.19), we get

$$\phi(0) - \phi(x_0) = \frac{1}{\sigma}\left\{1 - 2\alpha - \frac{v}{2}\left[\log^2(eu) + 1\right] - 2\alpha u \log u\right\} \tag{2.22}$$

where e is Euler's constant, $\log e = 1$.

We have now calculated difference $\phi(0) - \phi(x_0)$ from the deterministic equation. In addition, we have found $\phi(x)$ directly at 0, as given by eq. (2.15). The quantity $\phi(x_0)$ represents the mean logarithm of best-fit class frequency at the stochastic edge. Since its dynamics is dominated by random mutation and stochastic drift, we cannot evaluate $\phi(x_0)$ from the deterministic consideration alone. Below we use the one-locus stochastic theory from Chapter 1 to estimate the expected log frequency of the edge class.

2.3.5 Stochastic treatment of the fittest class

In Section 2.3.3, we introduced a deterministic equation, which has a continuous set of solutions in the form of solitary waves with various speeds, that is, average substitution rates. Now, to choose the correct solution, we have to analyze stochastic dynamics of the fittest class, $k = k_0$. The idea is to treat it within the framework of the one-locus two-variant model analyzed in Chapter 1 as a beneficial minority allele

evolving in the presence of natural selection, asymmetric mutation and random genetic drift, while the rest of population is considered a less-fit majority allele with the fitness equal to the average population fitness (Rouzine et al., 2003). In Figure 2.5, we compare dynamics of fitness classes predicted by such semi-deterministic approach realized numerically to the full stochastic simulation. In agreement with our analytic prediction, either type of simulation produces a traveling wave which moves to higher or to lower fitness (lower or high deleterious allele number k) depending on

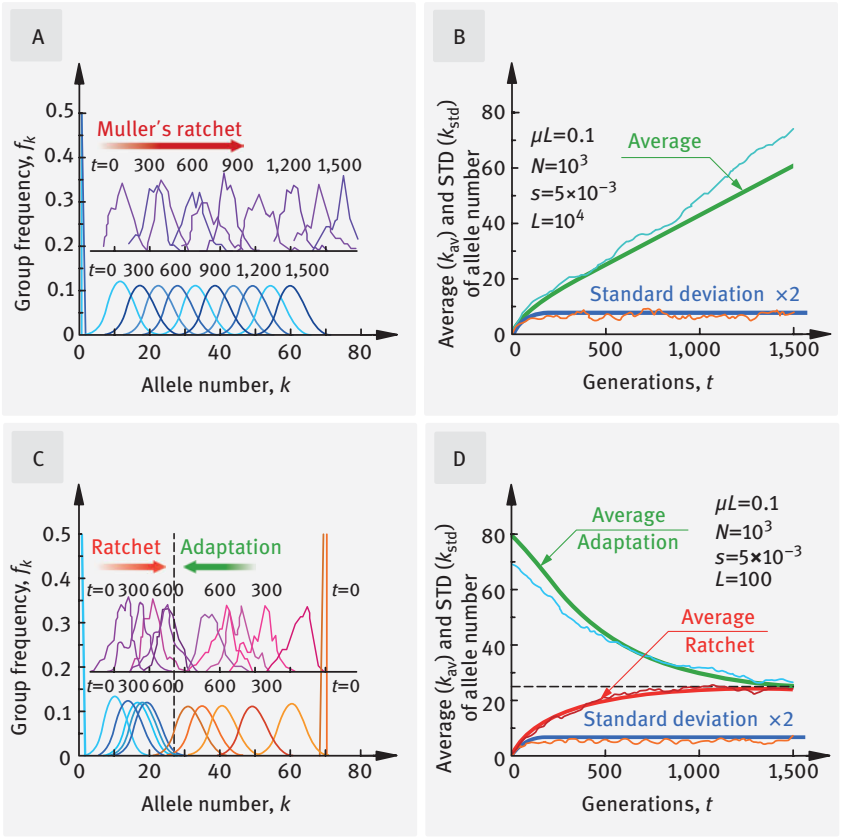

Figure 2.5: Two numerically obtained examples of the evolution of a population in a long (A and B) and a relatively short (C and D) genome. Parameter values are shown in (B) and (D), respectively. (A and C) The frequency of sequences with k mutant loci at different times (shown on the curves). Fat arrows show the direction of evolution. Ragged curves obtained by pseudorandom simulation correspond to either Muller's ratchet (red arrow), initial value $k(0) = 1$ or to adaptation (green arrow), initial $k(0) = 70$. The smooth curves below (blue for ratchet, brown for adaptation) are obtained by another type of simulation using a deterministic approximation for the bulk and stochastic treatment for the fittest class. (B and D) The corresponding time dependence for the average and the standard deviation of k (wave width) for the two methods of simulation. Dashed lines show the steady-state value of k_{av} (based on Rouzine et al. (2003)).

model parameters: population size N, selection coefficient s, total locus number L, and genomic mutation rate $\mu L = U_d + U_b$. The difference between two simulation methods is that, in the full simulation, the wave is ragged, but the speed of the wave is quite similar. Our task is to calculate the evolution rate analytically.

Analysis of the fittest class, $f_{k_0}(t)$, is based on the simple generalization of diffusion equation (1.1), accounting for asymmetry of mutation rates. A simplified treatment ensure good accuracy, as follows.

We remind the reader the important property of the Fokker–Planck equation we obtained in Chapter 1: selection dominates over random drift when the minority allele count, in our case $Nf_{k_0}(t)$, is higher than the stochastic threshold, $1/|S|$; in this case, beneficial allele is established in the population. Below that threshold, random drift rules, and the minority will probably be lost. In our case, S is the effective selection coefficient of the fittest class in this case. If $f_{k_0}(t)$ is much larger than $1/(N|S|)$, but much smaller than 1, we can use deterministic equation (1.61), which, in our notation, has the form

$$\frac{\partial f_{k_0}}{\partial t} = M(t) + Sf_{k_0}(t),$$

$$M(t) = U\alpha f_{k_0+1}(t)$$

$$S = s(k_{\text{av}} - k_0) - (U_b + U_d) \tag{2.23}$$

In the selection coefficient S of the best-fit class, the first positive term is due to a better-than-average fitness, and the second negative term is due to mutations causing decay of the class. $M(t)$ has the meaning of the effective mutation rate in the two-allele model.

As we will find out later, the sign of S is determined by the sign of the speed v, which is positive in the adaptation case, $v < 0$, and negative in the case of accumulation of deleterious alleles, $v > 0$ (Muller ratchet). Indeed, it stands to reason that, in the case of adaptation, the fittest class k_0 is born from a beneficial mutation with average rate M. Then this new class is subject to diffusion and, if it is lucky, becomes established, that is, survives random drift and grows further with probability on the order of 1. Once again, for that to happen, its frequency in a population must reach the stochastic threshold, $1/N|S|$. In contrast, in the ratchet regime, the best-fit class decreases exponentially in time until it passes below that threshold and is almost certainly lost. Due to this treatment of a characteristic frequency as a sharp threshold, we will acquire the error of a numeric prefactor ~1 at population size, N. The error is acceptable, because the speed of evolution, as we shall see, depends on N only logarithmically, and N is assumed everywhere here large.

To match the deterministic bulk to the edge class, we start by noting that the fittest class size depends on time in a saw-like fashion, because the mutational load k_0 changes abruptly in time by unit when a new class is established/lost. In the Muller's ratchet regime, the class decays until lost, and the next class becomes the fittest class,

and in the regime of adaptation, the fittest class expands until a beneficial mutation within it give birth to a new fittest class (Figure 2.2c). To match the stochastic edge to the bulk, we require that $\log f_{k_0}(t)$ averaged over one period of the saw is equal to that the logarithm of the deterministic solution at the edge, $\phi(x_0)$ (Rouzine et al., 2008)

$$\phi(x_0) = \frac{1}{2}\left[\log\frac{1}{N|S|} + \log f_{\max}\right] \tag{2.24}$$

where $f_{\max}$ is the maximum frequency of best-fit class (Figure 2.2c), when the class is about to switch to the next one. In Sections 2.4 and 2.5, we obtain the expression for $f_{\max}$ and determine $\phi(x_0)$ from eq. (2.24) for adaptation, Muller ratchet, and steady state.

2.4 Adaptation due to accumulation of beneficial mutations

Adaptation starts when there is a sudden change in environmental conditions or a population is introduced into a new environment due to migration. Initially, the population will be adapted in a less-than-optimal way. Gradually, beneficial mutations will accumulate until the mutation–selection balance is reached (or conditions change again). In our notation, adaptation corresponds to negative wave speeds, $v < 0$, because it moves towards smaller numbers of deleterious alleles, k. Furthermore, we will consider here the case far from the mutation–selection balance, such that $|v| \gg 1$. In this regime, deleterious mutation is a negligible correction to selection coefficient S. Only beneficial mutation matters by creating new alleles. In this particular case, it is more convenient to return, from the scaled velocity, v, back to the accumulation rate of beneficial alleles $V > 0$:

$$V = -(U_b + U_d)v \gg U_b + U_d \tag{2.25}$$

and to beneficial mutation rate $U_b = \alpha(U_b + U_d)$ instead of α. We will assume also that s is larger than U_b, which is typically quite small, $10^{-5} - 10^{-1}$, and that $V \gg s/\log(s/U_b)$, which is met for sufficiently large populations. Under this condition, as we will show, the lead is long and the entire multiple clone regime applies. In this regime, we will consider two subintervals of population size, which correspond to (i) moderate adaptation rates, when $s/\log(s/U_b) \ll V \ll s$ and the fitness distribution is narrow, and (ii) large adaptation rates, $V \gg s$, where it is broad.

In the case of a large negative v, the expressions for the log derivative of probability density at the edge, the lead, and the drop in log distribution, eqs. (2.18), (2.19), (2.22), simplify

$$\phi'(x_0) = \log u = \log\frac{V}{U_b} \tag{2.26}$$

$$x_0 = -\frac{V}{s}\left[\log\frac{V}{U_b} - 1\right] \tag{2.27}$$

$$\phi(0) - \phi(x_0) = \frac{V}{2s}\left[\log^2\frac{V}{eU_b} + 1\right] \tag{2.28}$$

Again, we remind that the approach assumes a continuous log distribution, $\phi'(x_0)/\phi''(x_0) \gg 1$, which is equivalent to a long lead, $|x_0| \gg 1$. Even though the tail is long, the half-width of the wave may be smaller or larger than 1. For large and intermediate population sizes, from eq. (2.15), we obtain

$$\phi(0) = \begin{cases} -\frac{1}{2}\log\frac{2\pi V}{s}, & V \gg s \\ 0, & s/\log\left(\frac{V}{U_b}\right) \ll V \ll s \end{cases} \tag{2.29}$$

When populations are so small that V is smaller than $s/\log(V/U_b)$, the lead becomes short, $|x_0| \ll 1$, the continuous-in-k approximation breaks down, and we pass from the multiple-clone regime to the regime of pairwise clonal interference (Section 2.2). In that case, the constant s approximation also will not work, and we need to introduce the distributed values of s.

According to our method, we match the bulk cutoff to the average log of the fittest class. Above the stochastic threshold, the best-fit class frequency has dynamics given in eq. (2.23) with

$$S = V\,\log\frac{V}{U_b}, \quad M = U_b\, f_{k_0+1}(t) \tag{2.30}$$

Because $V > U_b$, selection coefficient S is positive, which implies the exponential expansion of the fittest class. The role of beneficial mutation is to create this class, and then add more and more clones to the class. Additional clones result in a time-dependent pre-factor. However, the growth is mostly exponential, and the log derivative in time is mostly S. To check this approximation, from eqs. (2.26) and (2.23), we find that the edge value $f_{k_0}(t) \gg M$ if $V \gg U_b$. Therefore, the mutation term M in the log derivative of $f_{k_0}(t)$ in eq. (2.23) can be neglected, and we approximate it with S.

Above the stochastic threshold $1/NS$ where eq. (2.23) applies, the dynamics of the current fittest class $f_{k_0}(t)$ represents a nearly periodic saw-shaped dependence in log scale (Figure 2.2C). A new fittest class emerges due to a mutation event in one of genomes of the current fittest class. If it reaches a size on the order of the stochastic threshold, which happens rarely, the new class will be established. Thus, the wave moves one notch in k. By the definition of adaptation rate V, two consecutive classes emerge in time interval $1/V$.

Let us estimate the maximal frequency $f_{\max} = f_{k_0}(t = 1/V)$ at the moment of time when a new established allele appears. The total number of mutational opportunities between clicks (the number of genomes that may potentially generate a beneficial mutation) is N multiplied by the time integral of $f_{k_0}(t) \propto \exp(St)$ over one saw period, $0 < t < 1/V$, which can be estimated as $Nf_{\max}/S$. Indeed, because of the exponential increase of $f_{k_0}(t)$ in time, the integral

$$\int_0^{1/V} f_{k_0}(t)dt$$

is mostly contributed by times within interval $1/S$ when $f_{k_0}(t)$ is near the maximal value $f_{\max}$ (Figure 2.2C, orange strip). This time interval, $1/S = (1/V)/\log(V/U_b)$, is much shorter than the saw period $1/V$ due to assumption $V \gg U_b$.

Next, the mean number of fittest genomes generated during a saw period is the product of the beneficial mutation rate U_b and of the above number of mutational opportunities. A lineage can survive genetic drift with a small probability $2S$ (Haldane, 1927; Kimura, 1962) (Chapter 1), so that the mean number of alleles established during one period is

$$2SU_bNf_{\max}/S = 2U_bNf_{\max}$$

The desired value of $f_{\max}$ is found from the condition that this number is equal to unit. Indeed, by the definition, exactly one new fitness class is established per period, so that

$$f_{\max} \approx \frac{1}{2U_bN} \tag{2.31}$$

According to eq. (2.24), we match $\phi(x_0)$ to the average between the minimum and the maximum value $f_{\max}$ under deterministic growth (Figure 2.2C). With eq. (2.31), this yields

$$\phi(x_0) = -\log\left[N\sqrt{VU_b \log\left(\frac{V}{U_b}\right)}\right] \tag{2.32}$$

This result, as we mentioned, is based on approximating the log of the fitness distribution with a continuous function in k. The next correction to this approximation due to the discreteness of fitness classes was obtained by Rouzine et al. (2008). They showed that term $\log|x_0|$ has to be added to the difference $\phi(0) - \phi(x_0)$ to account for the corrections to $\phi(x)$ near the edge caused by discreteness of k. Adding the correction term $\log|x_0|$, substituting eqs. (2.32) and eq. (2.15) into eq. (2.28), and neglecting numerical constants multiplying N inside a large logarithm, we obtain the desired relation between the evolution rate and system parameters (Rouzine et al., 2008):

$$\log N \approx \frac{V}{2s}\left(\log^2\frac{V}{eU_b}+1\right)-\log\sqrt{\frac{s^3U_b}{V^2\log\left(\frac{V}{U_b}\right)}}\,,\quad V\gg s \tag{2.33}$$

Because $\log N \gg 1$, the second term is relatively small. The expression in interval $V \ll s$ is quite similar

$$\log N \approx \frac{V}{2s}\left(\log^2\frac{V}{eU_b}+1\right)-\log\sqrt{\frac{s^3U_b}{V^2\log\left(\frac{V}{U_b}\right)}}\,,\quad \frac{s}{\log\left(\frac{V}{U_b}\right)}\ll V\ll s \tag{2.34}$$

The two results differ by a factor of $\sqrt{V/s}$ in the logarithm. At large population size N, the numeric effect of this factor is small. The correction for discreteness of k accounts for a 10–15% correction in V.

To make use of eq. (2.33) or (2.34) one can calculate V iteratively at every N and the iterations converge fast due to the logarithmic dependence on V in the right-hand side. (Alternatively, N can be plotted as a function of V.) At very large population sizes, iterating eq. (2.33) we obtain

$$V \approx \frac{2s\,\log\left(N\sqrt{sU_b}\right)}{\log^2\left[(s/U_b)\log\left(N\sqrt{sU_b}\right)\right]} \tag{2.35}$$

This finding confirms the prediction of previous approximations (Gerrish and Lenski, 1998; Kessler et al., 1997; Tsimring et al., 1996) that adaptation rate is greatly suppressed by linkage of many sites. As in the clonal interference regime (Section 2.2), the adaptation rate, V, is not linearly proportional to the genomic mutation rate or the population size, as it would be for independent sites. Instead, it increases with all these parameters slowly, logarithmically (Figure 2.6). Unlike in the pairwise clonal interference model (Section 2.2), the adaptation rate does not saturate at large N, but keeps increasing logarithmically. This is the consequences of including nested clones at multiple sites, which partly offset interference.

The first prediction of the evolution rate in this model was proposed by (Rouzine et al., 2003). Although it was asymptotically accurate at large N, it had a modest error in the prefactor at N inside a large logarithm in eq. (2.35). Subsequent studies confirmed (and improved upon) the accuracy of that approach, including an upgrade in (Rouzine et al., 2008) which we discussed here, as well implementation of a branching process (Brunet et al., 2008; Desai and Fisher, 2007). The results of various approaches (Brunet et al., 2008; Desai and Fisher, 2007; Rouzine et al., 2008; Rouzine et al., 2003) are compared with each other and with computer simulation in Figure 2.6. Thus, the accuracy of the initial findings on the speed of asexual adaptation (Rouzine et al., 2003) has been confirmed and improved upon in subsequent work.

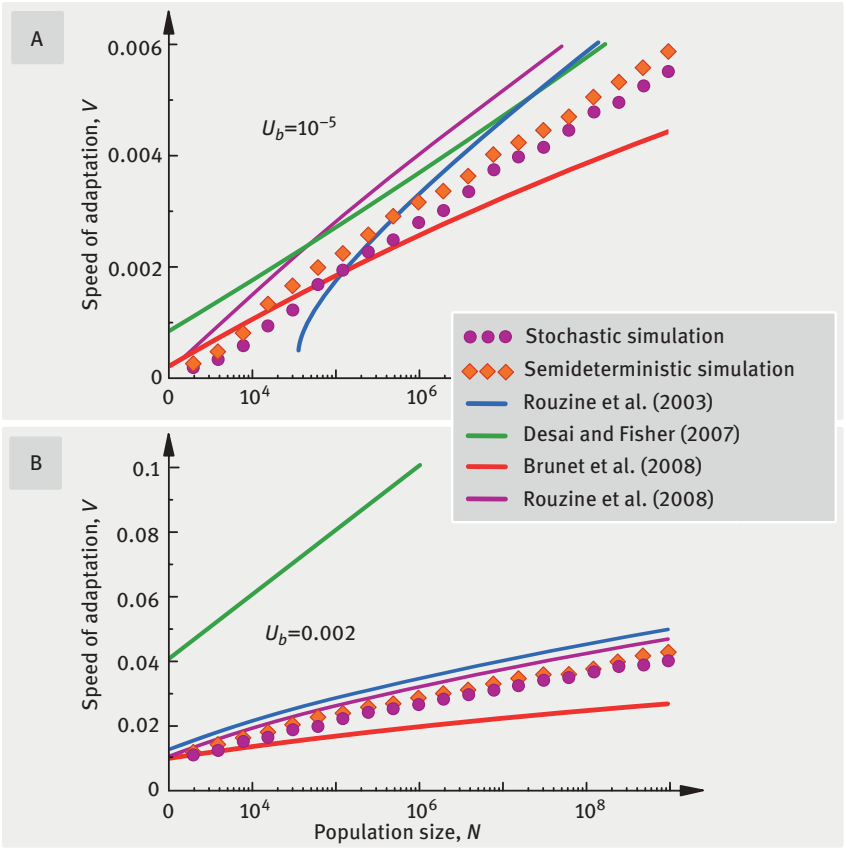

Figure 2.6: Four analytic approximations to calculate the speed of asexual adaptation are compared to two types of stochastic simulation. The asexual adaptation rate increases logarithmically with the population size N. (A, B) Curves: analytic adaptation rate predicted by two different versions of traveling wave theory and their upgrades: (Rouzine et al., 2003) (blue) and upgrade (Rouzine et al., 2008) (purple); (Desai and Fisher, 2007) (green) and upgrade (Brunet et al., 2008) (red). Symbols: full stochastic simulation (purple circles), and semideterministic simulation in which only the fittest class is treated stochastically (orange diamonds). Parameters: selection coefficient $s = 0.01$, beneficial mutation rate per genome U_b is (A) 10^{-5} and (B) 2×10^{-3}, deleterious mutation is absent (based on Brunet et al. (2008)).

2.5 Accumulation of deleterious mutations (Muller's ratchet)

In Section 2.5, we have considered a case without deleterious mutations, because we assumed that population is large, the average fitness is small, and we are far from a steady state. In real genomes, most potential mutations are deleterious or neutral (the latter are not considered here). Most of these deleterious mutations kill the fitness of the strain, hence, they are quickly lost from the population and do not

need not be considered. However, a large portion of deleterious mutations have a sufficiently small effect $|s| \ll U_d$ to accumulate in the genome to high levels offsetting the adaptation process.

In the extreme case when a population starts from the maximal fitness, $k = 0$, any mutation can only decrease fitness, so that average fitness will decrease until selection and beneficial mutation will stop this process. Below we consider such a scenario, where all mutations have a small negative fitness effect, $s < 0$, $|s| \ll U_d$. We will show that the linkage of deleterious mutations to each other results in the accumulation of deleterious mutations at higher rates than it would be for independent sites. This effect was first predicted by Muller to explain the advantage of sexual reproduction, and later termed "Muller's ratchet" (Felsenstein, 1974; Muller, 1932).

The mechanics of the ratchet is, as follows. Again, we assume that beneficial mutation is absent. In this case, the system described by eq. (2.8) is initially in equilibrium termed "mutation-selection balance" given by

$$U_d f_{k-1} = \{U_d + s(k - k_{\text{av}})\} f_k$$

This expression describes a curve with the maximum at $k = k_{\text{av}}$. Essentially, mutation pushes the system towards large k and selection works in the opposite direction. Now, we turn on the random drift due to finite population size. Then, the best-fit class with smallest k will, sooner or later, be lost due to random genetic drift, and this loss is irreversible (hence, the term "ratchet"). After that, the entire fitness distribution will shift towards larger k by unit (a ratchet click). When the population is very large, $\log N > U_d/s$, clicks occur very rarely. Between clicks, the fitness distribution assumes the transient mutation-selection equilibrium. Therefore, the average ratchet rate is exponentially small (Gordo and Charlesworth, 2000; Haigh, 1978; Stephan et al., 1993). In contrast, for moderate population sizes, $\log N < U_d/s$, ratchet clicks are so frequent that the distribution does not have enough time to reach equilibrium and shifts in a quasi-continuous way, except for the leading edge. At the stochastic edge, the time dependence is saw-like, similar to the adaptation case in Figure 2.2C but running in the opposite direction: decay of the fittest class, its loss, switch to the next, and so on and so forth. Below in Section 2.6 we will show that even a small rate of beneficial mutations can stop Muller's ratchet (Goyal et al., 2012; Rouzine et al., 2003). However, when the population is far from equilibrium and very highly fit, close to the best-fit sequence, beneficial mutations are simply too few to be important. Below, we calculate the speed of the ratchet neglecting beneficial mutations.

In our notation, the ratchet regime corresponds to the limit $\alpha \to 0$. In this case, $U_b = 0$ and the scaled speed v of the wave is positive, that is, the wave moves toward more deleterious alleles, $0 < v < 1$. The following derivation requires that $s \ll U_d$, which inequality ensures that the lead is long, that is, corresponds to multiple alleles. As we will see, the same condition implies a broad wave, $\text{Var}[k] \gg 1$ and hence the smooth behavior in k.

Let us simplify the deterministic part for this case. Setting $\alpha = 0$ in the expression for u, eq. (2.18), we find

$$u = 1/v \tag{2.36}$$

Using this expression, for the lead, eq. (2.19), we obtain

$$x_0 = -\frac{1}{\sigma}[1 - v\,\log(e/v)] \tag{2.37}$$

The difference in log genome frequency, eq. (2.22), now simplifies

$$\phi(0) - \phi(x_0) = \frac{1}{\sigma}\left\{1 - \frac{v}{2}\,\left[\log^2\left(\frac{e}{v}\right) + 1\right]\right\} \tag{2.38}$$

We remind that the continuous-in-k approach is based on the assumption that the lead is long, $|x_0| \gg 1$, which reduces to the conditions that $\sigma \ll 1$ and v is not too close to unit (large population sizes). Thus, for the rapid Muller's ratchet to exist, the selection coefficient has to be much smaller than the total mutation rate.

In this case, the wave is also broad, and hence $\phi(0)$ is small, as given by

$$\phi(0) = \log\sqrt{\frac{\sigma}{2\pi(1-v)}} \tag{2.39}$$

which is eq. (2.15) at $\alpha = 0$.

Following the drill, we now obtain $\phi(x_0)$ based on the stochastic consideration of the edge. Because we neglect beneficial mutation in this limit, we set $M(t) \equiv 0$ in the dynamic equation for the edge class, eq. (2.23). For the effective selection coefficient of the edge class, $S = -U_d(1 + \sigma x_0),$ we obtain

$$S = -U_d v \log\left(\frac{e}{v}\right) \tag{2.40}$$

where we have made use of eq. (2.37). We have $S < 0$ at $v < 1$, which implies that the best-fit class is selected against, in this regime. After integrating eq. (2.23), the fittest class decays in time as

$$f_{k_0}(t) = f_{k_0}(0)e^{-|S|t} \tag{2.41}$$

where $t = 0$ denotes the time of the loss of the previously best-fit class with $k_0 - 1$ alleles. Equation (2.41) applies until the copy number $Nf_{k_0}(t)$ remains higher than the stochastic threshold $1/|S|$ (Chapter 1). Below the threshold, drift becomes dominant, and the k_0 class is as good as lost. (We remind that the stochastic "threshold" is smeared out, which fact creates an error on the order of 1 multiplying population size N. The error is acceptable, because N is assumed large and it enters in the argument of a logarithm, see Sec 2.4 on the adaptation regime.)

To couple the fittest class to the deterministic bulk, per Section 2.3.5, we need to equate the bulk value $\phi(x_0)$ to $\log f_{k_0}(t)$ averaged over the period when it is above the threshold. Using eq. (2.24) and the equality $f_{\max} = f_{k_0}(0)$ based on eq. (2.41), we obtain

$$\phi(x_0) = \frac{1}{2}\left[\log\frac{1}{N|S|} + \log f_{k_0}(0)\right] \tag{2.42}$$

By the definition of velocity v, time period $t_{\text{click}} = 1/(U_d v)$ corresponds to one click in k. The time when the fittest class passes the threshold, t_{loss}, can be obtained from eq. (2.41):

$$\frac{1}{|S|N} = f_{k_0}(0)e^{-|S|t_{\text{loss}}} \tag{2.43}$$

After equating the time to the loss of class and the ratchet click, $t_{\text{click}} = t_{\text{loss}}$, we find:

$$\log f_{k_0}(0) = \log\frac{1}{N|S|} + \frac{|S|}{vU_d} \tag{2.44}$$

Using eqs. (2.42) and (2.40), we obtain

$$\phi(x_0) = -\log\left[Nv^{\frac{3}{2}}U_d \log\left(\frac{e}{v}\right)\right] \tag{2.45}$$

Because the accuracy of stochastic threshold $1/(N|S|)$ is limited by a numerical coefficient, we omit the numerical coefficient in the argument of the logarithm.

We have neglected stochasticity of the next-fit class, $k = k_0 - 1$, because it is always bigger than best-fit class. The average size ratio can be estimated, as given by $f_{k_0-1}/f_{k_0} = \exp(\phi'(x_0)) = u = 1/v$, eq. (2.36). Therefore, in the general case when v is not too close to 1, the next-fit class is several-fold larger than the best-fit class. The error of this approximation is, again, a number at N, which we omit anyway in eq. (2.45).

We remind the reader that, to obtain $\phi(0) - \phi(x_0)$ in eq. (2.38), we have used the continuous-in-k approximation based on the assumption that the lead $|x_0|$ is long, which is the case in large populations. There exist, however, a correction due to discreteness, $\log|x_0|$, which has to be added to $\phi(0) - \phi(x_0)$ (Rouzine et al., 2008). Combining eqs. (2.38), (2.39), and (2.45) and adding the correction term, we arrive at our final result for the ratchet rate

$$\sigma\log\left(NU_d\sigma^{\frac{3}{2}}\right) \approx \left[1 - \frac{v}{2}\left(\log^2\frac{e}{v} + 1\right)\right] - \sigma\log\left[\sqrt{\frac{v^3}{1-v}}\,\frac{\log(e/v)}{1 - v\log(e/v) + 5\sigma/6}\right] \tag{2.46}$$

The second term in eq. (2.46) is supposed to be a small (but not negligible in practice) correction to the first term in the limit $\sigma \to 0$. This transcendental equation for the normalized ratchet rate, $v = (1/U_d)dk_{\text{av}}/dt$, relates it to the selection strength

$\sigma = s/U_d$, population size N, and mutation rate U_d. We can evaluate eq. (2.46) in two ways. We can either plot N as a function of v and the other model parameters (solid lines in Figure 2.7) or we can solve it iteratively for v at each given value of N.

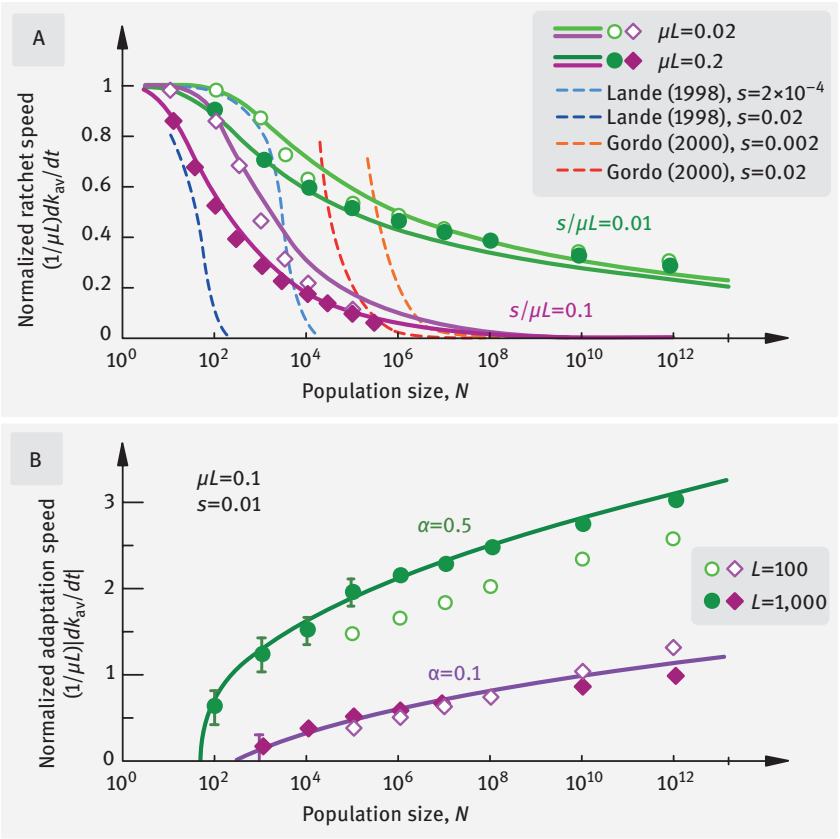

Figure 2.7: Examples of the analytically predicted rescaled substitution rate at specific parameter values compared with simulation results. (A) Normalized ratchet speed as a function of population size N: analytic results (solid line) versus simulation (symbols) (Rouzine et al., 2003). Beneficial mutations are absent, $\alpha = 0$. Purple: Results for $\sigma = s/U_d = 0.1$, eq. (2.46). The dashed lines are the large-N asymptotics [(Gordo and Charlesworth, 2000), eq. (3a) and (3b)] and small-N asymptotics [(Lande, 1998), eq. (2c), multiplied by NU_d]. Green: Results for $\sigma = s/U_d = 0.01$. Parameters are shown. (B) Normalized adaptation speed as a function of N in the presence of both analytic results (solid line) versus simulation (symbols). Green and purple curves correspond to two different values of the less-fit allele fraction α (shown). Parameters including full mutation rate μL, selection coefficient s, and total locus number L are shown (based on Rouzine et al. (2003)).

2.6 General case and mutation-selection equilibrium

In Sections 2.4 and 2.5, we assumed that a population is either high below or high above the steady state in terms of average fitness. When a population is not very far from steady state, both deleterious and beneficial mutations are important for the evolutionary dynamics. In the general case, the accumulation rate v can be expressed in terms of two composite parameters, the fraction of beneficial (compensatory) genomic mutations, α, and the normalized log population size [(Rouzine et al., 2003), Appendix, eqs. (19)–(21)]. The method is the same as described in the previous sections. We have

$$\sigma \log\left(\frac{N}{N^*}\right) \approx 1 - 2\alpha - v - \frac{v}{2}\log^2 u - v\log u - 2\alpha u \log u, \;\; v < 1 - 2\alpha \tag{2.47}$$

$$N^* = \frac{\sqrt{2\pi \mathrm{Var}[k]}}{\mu L \xi(\alpha, v)} \tag{2.48}$$

$$\xi(\alpha, v) \sim \begin{cases} 1 & \alpha \sim |v| \sim 1 \\ v\log\left(\frac{e}{v}\right) & \alpha, \;\; 0 < v < 1 \\ \alpha \log^2\left(\frac{|v|}{\alpha}\right), & v < 0, \;\; |v| \gg \sqrt{\alpha} \end{cases} \tag{2.49}$$

Equation (2.48) an approximate estimate for prefactor N^*; for more accurate expressions in the case of adaptation or ratchet, see the last terms in eqs. (2.33) and (2.46), respectively. The lead of fitness wave x_0 is given by

$$x_0 = -(1/\sigma)(-2\alpha u - v\log u + 1 - v) \tag{2.50}$$

Analytic predictions of evolution speed at different values of less-fit allele fraction α and population size N are compared with results of stochastic simulation in Figure 2.7B.

Steady state. If the population size is not too large, $\sigma \log(N/N^*) < 1$, the population eventually arrives a steady state at the value of α, where the processes of adaptation (Section 2.4) and ratchet (Section 2.5) exactly balance each other (Figure 2.5C and D). (A steady state exists also at larger populations, but due to beneficial mutation, it is very low in k and is close to the one-site model equilibrium discussed in Section 1.4.3.) Setting $v = 0$ in eqs. (2.47) to (2.50), we obtain the equilibrium position of the distribution center, $k_{\mathrm{av}} = L\alpha$ (Rouzine et al., 2003)

$$\sigma \log\left(\frac{N}{N^*}\right) = 1 - 2\alpha - \sqrt{\alpha(1-\alpha)}\,\log\frac{1-\alpha}{\alpha}, \;\; v = 0 \tag{2.51}$$

where $N^* = \alpha\sqrt{sU_d}$ from eq. (2.49), and the lead is

$$x_0 = -\left(\frac{1}{\sigma}\right)\left[1 - 2\sqrt{\alpha(1-\alpha)}\right] \tag{2.52}$$

Numerically, except at very low population sizes, these values of α predicted by eq. 2.51 are quite small, $\alpha \ll 1$. In other words, very few beneficial mutations are sufficient to stop Muller ratchet [see (Rouzine et al., 2003), Figure 2B].

This analytic result has been confirmed in (Goyal et al., 2012) who used a similar method, with the only difference in the parameter $N^* \sim s$ inside of the large logarithm [see their Appendix, eq. (13) and Figures. 4B and S1]. In an experimental and modeling work on poliovirus evolution, Xiao et al. (2017) tested both estimates of N^* and found that the difference is numerically minor in a broad parameter range. These last authors also confirmed the validity of the analytic result (2.51) with the help of Monte-Carlo simulations and showed that the two analytic results converge in the limit of $(s/\mu L)^2 \ll \alpha$.

2.7 Transition to the one-locus model at large *N*

Figure 2.8 presents schematically the phase diagram in N and α, which shows the dominant evolutionary forces and the evolution direction based on results obtained in Sections 2.4 to 2.6. Random genetic drift, natural selection, and linkage are all important in the regions of the delayed adaptation and Muller's ratchet. In a broad interval of population sizes N, $\log(1/s) < \log N \ll k\log(\sigma L)$, the adaptation rate is small and depends on the population size very slowly as compared to the one-locus dynamics (Chapter 1), In the third area, which stretches at large N and small α (Figure 2.8), $\log N > U_d/s$, linkage of loci is negligible, and the accumulation dynamics of deleterious mutations crosses over to the one-locus prediction (Chapter 1). The transition takes place when the high-fitness edge of the wave reaches to the fittest possible sequence, $k = 0$, that is, at $x_c = -k_{\mathrm{av}}$.

We can conclude that the simple model from Chapter 1 can be used either in the limit of recombination with very frequent crossovers, or when only one locus at a time is strongly diverse. Only in the limit of extremely large N, the adaptation rate V in eq. (2.25) has a transition to the one-locus deterministic result

$$V = s\bar{k}$$

which is eq. (1.61) with $k/L \ll 1$, $\mu \ll 1$. This transition is caused by the fact that, at infinite N, every genetic variant can be found in a population due to frequent mutation events which break down linkage disequilibrium and make loci independent. In agreement with this fact, models with $N = \infty$ and no epistasis do not generally find any advantage for recombination (Felsenstein, 1974; Kondrashov, 1993). These

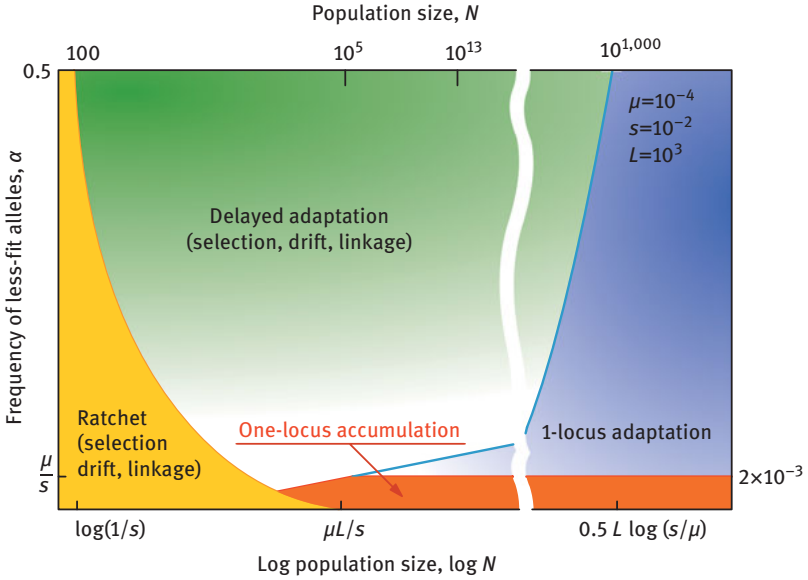

Figure 2.8: Schematic phase diagram of the overall direction and dominant factors of evolution. The upper axis and the parameter values (shown) are representative for RNA viruses. For organisms, the mutation rate per site is smaller, $\mu \sim 10^{-9}$, and the number of loci is larger, $L > 10^6$–10^7 (based on Rouzine et al. (2003)).

models, however, have little realism. The transition to independent-locus regime happens at asexual population sizes that are unrealistically large (Rouzine et al 2003)

$$N_{\text{one-locus}} \sim (\sigma L)^{Ck}, \quad C \sim 1 \tag{2.53}$$

We remind that the number of segregating (diverse) loci in most populations counts in millions and even in many RNA viruses in hundreds. Therefore, the estimate in equation (2.53) may easily exceed the number of protons in visible universe. Recombination, as we shall demonstrate in Chapter 3, is hugely advantageous to adaptation, precisely because of interference effects and a limited population size.

2.8 Mutation with a variable effect on fitness

In the previous sections we assumed that all mutations have a fitness effect, either positive or negative, but of the same magnitude, s. In real organisms, a broad variation in s among loci is observed. A priori it is not obvious whether many sites simply average out to an effective value of s (Hegreness et al., 2006). Early Monte-Carlo studies (Fogle et al., 2008) showed that the approximation of an average s

may be valid when the distribution of s decays faster than an exponential at large s. Simulation showed that if the tail of the distribution decays more slowly, the approximation of effective s fails and a broad range of svalues must be analyzed explicitly.

Significant progress in understanding these results was reached in two analytic works (Good et al., 2012; Schiffels et al., 2011). Schiffels et al. (2011) generalized the two-clone interference approach (Gerrish and Lenski, 1998) (Section 2.2) to include a third, nested clone, which extend its validity to larger populations. The second approach is more aligned with our aims and is described in detail in this section (Good et al., 2012). The model considers a large number of loci and applies a version of the traveling wave method described in in the previous Sections. This approach employs a method adopted from (Neher et al., 2010) and is confirmed by the tunable approach (Hallatschek, 2010).

The main conclusion shows that the adaptation mechanism and the dependence of adaptation rate on model parameters, indeed, essentially depends on the form of the distribution tail of s at large s, as well as on a population size range (Figure 2.9). If decay of the tail is slower than exponential, the important mutations are the few mutations that have the largest fitness effect, similar to the clone interference models, Section 2.2 (Gerrish and Lenski, 1998; Schiffels et al., 2011). For faster-than-exponential decay of the distribution, multiple sites are important, and the result is reduced to that for adaptation at fixed s, Section 2.4, with some effective s. The

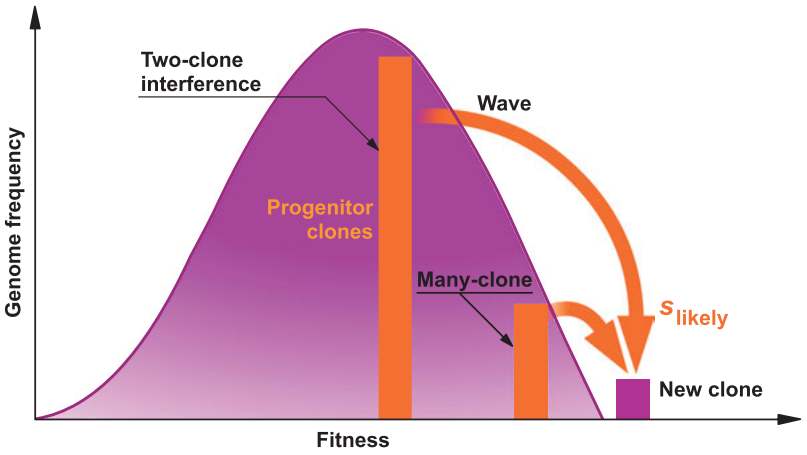

Figure 2.9: Mechanism of asexual adaptation when selection coefficient s varies among sites resembles either the two-clone interference model (Section 2.2) or the traveling wave model with fixed s (Sections 2.3 to 2.6) depending on how steep is the tail of the distribution and how large is the population size (Good et al., 2012). A distribution of the selection coefficient of the form $\exp[-s/\sigma]$ is considered in the text, where σ is a constant parameter. Purple curve: the traveling wave of the frequency of fitness classes. Orange bars: the most likely progenitor class. Purple bar: a new fittest clone (based on Rouzine and Weinberger (2013)).

most interesting case is the exponential decay of the distribution of s. In this regime, both regimes take place, depending on the population size. Below we consider this interesting case in detail.

2.8.1 Approach

As the models in the previous sections, the model in (Good et al., 2020) considers a population of N individual genomes with beneficial mutation rate U_b per genome. These mutations are assumed to take place in a large number of loci, each with its own beneficial effect s which contributes linearly to the log fitness of a genome. We will approximate the fitness landscape by a continuous distribution of selection coefficients among sites $\rho(s)$. Epistasis is neglected. We neglect deleterious mutation with a small effect, which is correct when population size is sufficiently large and the system is far from the steady state (Section 2.4). Deleterious mutations with very strong effect do not contribute to evolution, for the obvious reason.

The method described below applies to a wide range of the forms of $\rho(s)$ (Good et al., 2012). Nevertheless, for the sake of simplicity and practicality, we will focus on the case of the exponential distribution often observed in experiment (Acevedo et al., 2014; Imhof and Schlotterer, 2001; Kassen and Bataillon, 2006; Stern et al., 2014; Wrenbeck et al., 2017):

$$\rho(s) = \frac{1}{\sigma} e^{-\frac{s}{\sigma}} \tag{2.54}$$

where average selection coefficient σ is assumed to be much larger than the mutation rate U_b.

As in Sections 2.3–2.6, a population develops into a solitary wave moved in fitness coordinate with a constant average rate $v = d\bar{X}(t)/dt$ and shape $f(x)$, where $x = X - \bar{X}(t)$ is the relative fitness of an individual X with respect to the average log fitness $\bar{X}(t)$ (Figure 2.10). In the previous section with constant s, we had $X = -sk$, where k is the integer number of deleterious alleles existing against the background of best-fit possible genome under given conditions. Here we use instead fitness notation X and consider it any real number.

The shape of the traveling wave is rigorously determined, for fixed s, by the method of stochastic threshold used in the previous sections (Rouzine et al., 2008; Rouzine and Coffin, 2005, 2007, 2010; Rouzine et al., 2003). Alternatively it can be obtained by using other methods of traveling-wave theory, such as tunable constraint models (Hallatschek, 2010) or stochastic calculations of the best-fit class (Brunet et al., 2008; Desai and Fisher, 2007). However, all these approaches are less convenient to apply for the case of distributed s. For our purposes, it will be sufficient to employ an approximation to the true shape of the fitness profile. As we

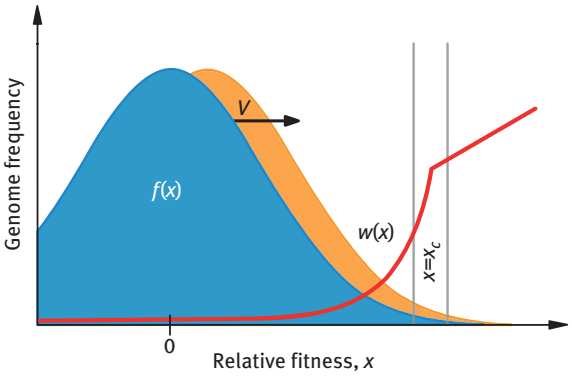

Figure 2.10: The process of adaptation (schematics). Genome frequency profile with given fitness, $f(x)$, moves at a constant rate, v. The fixation probability, $w(x)$, increases sharply with x until it reaches a thin border layer $x = x_c$ after which it transitions to the result of one-locus theory (Chapter 1), $w(x) = x$ (based on Good et al. (2012)).

showed in Section 2.3.2, the wave becomes close to a Gaussian with the variance given by the speed of adaptation v, eqs. (2.14), (2.10):

$$f(x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}} \tag{2.55}$$

where we rescaled notation $-\sigma v \to v$, $\sigma x \to x$ and assumed small mutation rate, $v \gg U$. The meaning of new notation, adaptation v, is the speed of the change in fitness, as opposed to the substitution rate of less-fit alleles we used in previous sections.

2.8.2 Probability of lineage establishment

Suppose we have a beneficial allele occurring in a genome and want to trace its fate. It may either become extinct or survive. Competition of the allele in genome with fitness X against the remaining population depends on time dependence of the average fitness $\bar{X}(t)$, which increases at a constant speed, v, in the stationary process. As long as the allele lineage remains smaller than the population size, we can analyze its behavior in terms of a branching process with birth rate $B(X,t) = 1 + X - \bar{X}(t)$ and death rate 1, by the choice of time unit. We also include mutation from X to fitness $X + s$ in interval $[s,\ s + ds]$ with rate $U_b \rho(s) ds$ per genome per generation.

The rate of evolution is determined by the fixation (nonextinction) probability $w(X,t)$ of a lineage of a genome with fitness X at time t. In the simplest case, when the population is small, and sweeps at different loci are rare and hence independent, it is given by $w = x$ (Haldane, 1927), see Chapter 1. When multiple fixation

events occur at the same time, they interfere with each other, as discussed in Section 2.2. When a lineage that does not become extinct early eventually takes a large portion of the population, the branching-process formalism starts to break down because of inter-lineage interaction effects. We assume that all lineages that rise to such a size are already guaranteed to fix, so that we can equate the probability of the nonextinction at a low level, $w(X,t)$, with the probability that a lineage is fixed.

To continue uninterrupted, a lineage has to undergo multiple mutations to higher fitness values. This is necessary to escape the constant increase in the average fitness as the population. Such a process has been described in the case of recombination (Neher et al., 2010) discussed in Section 3.4 and to the mutational surfing of genes in populations expanding geographically (Excoffier and Ray, 2008; Hallatschek and Nelson, 2008).

In next sections, we will demonstrate that fixation probability $w(x)$ satisfies the equation

$$v\ \frac{dw}{dx} = xw(x) - w(x)^2 + U_b \int_0^\infty ds\ \rho(s)[w(x+s) - w(x)] \tag{2.56}$$

2.8.3 Self-consistency condition for the evolution rate

The survival of allelic lineages is linked to traveling wave by the condition that the population adapts by generating new mutations that manage to get fixed. The probability of fixation of a single mutation of effect s, denoted $\pi(s)$, can be obtained from fixation probability of a lineage of individual x, $w(x)$, by averaging over the distribution of fitness backgrounds, $f(x)$, in which it could have occurred

$$\pi(s) = \int_{-\infty}^{\infty} dx\ w(x) f(x-s) \tag{2.57}$$

Consistency requires that adaptation rate is given by the average fixation rate of new mutations fix weighted by their fitness effect, as given by

$$v = NU_b \int_{-\infty}^{\infty} ds\ s\ \pi(s)\rho(s) \tag{2.58}$$

The distribution of the fixed mutations in their fitness effect is

$$\rho_f(s) \propto \pi(s)\rho(s) \tag{2.59}$$

When taken together, eqs. (2.54)–(2.58) determine the distribution of fixed mutations in s and the adaptation dynamics. We write down and analyze their solution in the following section.

2.8.4 Fixation probability and adaptation rate

As shown in Section 2.8.5, when population sizes are large and mutation infrequent, $w(x)$ found from eq. (2.56) experiences a sharp change at a threshold fitness, $x = x_c$, above which it approximates the one-locus result, $w = x$ (Figure 2.10) and below which it decays exponentially. One can approximate the fixation probability piecewise as

$$w(x) \approx \begin{cases} 0 & \text{if } x < 0 \\ x_c e^{\frac{x^2 - x_c^2}{2v}} & \text{if } 0 < x < x_c \\ x & \text{if } x > x_c \end{cases} \tag{2.60}$$

where x_c is determined by the condition

$$2 = U_b \int_0^\infty ds\ \rho(s)\ \frac{e^{\frac{sx_c}{v}} - 1}{s} e^{-\frac{s^2}{2v}} + \frac{U_b}{v x_c} \int_{x_c}^\infty dx\ x e^{\frac{x_c^2 - x^2}{2v}} \int_0^\infty ds\ \rho(s) e^{-\frac{s^2}{2v} + \frac{xs}{v}} \tag{2.61}$$

Intuitively, below the transition point, $x < x_c$, allele fixation probability $w(x)$ is proportional to the integral of the lineage size in time, which is the total number of mutational opportunities for the lineage. Fixation, in this case, is limited by the rare event that the lineage mutates again. In contrast, fixation at $x > x_c$ is determined by the probability that the lineage survives random genetic drift. The point x_c, which is obtained from eq. (2.61), has the intuitive interpretation as the boundary in fitness above which clonal interference does not decrease the fixation probability significantly.

To calculate the average fixation probability $\pi(s)$, we substitute $w(x)$ from eq. (2.60) and $f(x)$ from eq. (2.55) into eq. (2.57) and then integrating yields

$$\pi(s) \propto \frac{e^{\frac{sx_c}{v}} - 1}{s} e^{-\frac{s^2}{2v}} + \frac{e^{\frac{x_c^2}{2v}}}{v x_c} \int_{x_c}^\infty dx\ x\ e^{-\frac{(x-s)^2}{2v}} \tag{2.62}$$

This function, $\pi(s)$, determines the shift in s of those mutations that fix, as compared to their raw distribution, $\rho(s)$. The distribution of fixed mutations is product $\rho_f(s) \propto \pi(s)\rho(s)$. Interestingly, $\pi(s)$ has an interval of effective neutrality for $s < v/x_c$, where alleles fix with a probability approximately equal to $\pi(s) = 1/N$ (Section 1.6.1) (Schiffels et al., 2011). Above this characteristic value, the probability of fixation exponentially increases in s before reaching the one-locus limit at $s > x_c$.

The approximate expressions for $w(x)$ and $\pi(s)$ contain the rate of adaptation v, which is self-consistently obtained by substituting eq. (2.62) into condition (2.58). This substitution produces a second relation between x and v. Explicit calculations are carried out in Section 2.8.5, and we summarize the main results below. In Figures 2.11–2.13, we compare these analytical predictions to Wright–Fisher simulations.

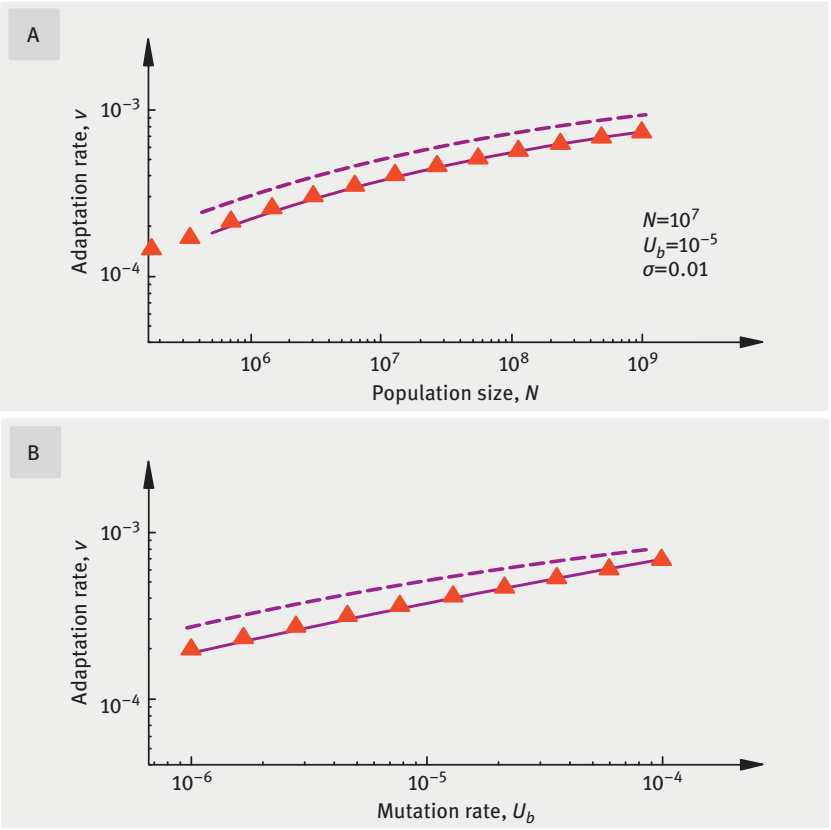

Figure 2.11: Adaptation rate, v, as a function of the population size, N (A), and the beneficial mutation rate, U_b (B), in the case of relatively small NU_b. Parameters are shown. $\sigma = 0.01$ is the average selection coefficient, eq. (2.54). Symbols and solid line denote simulation and analytic results, respectively (see text). Dashed line: two-clone interference theory prediction (Section 2.2) (based on Good et al. (2012)).

Assuming the exponential distribution of fitness effects, eq. (2.54), the integrals over the selection coefficient s in eqs. (2.61) and (2.58) are sharply peaked at value $s = s^*$

$$s^* = x_c - \frac{v}{\sigma} \tag{2.63}$$

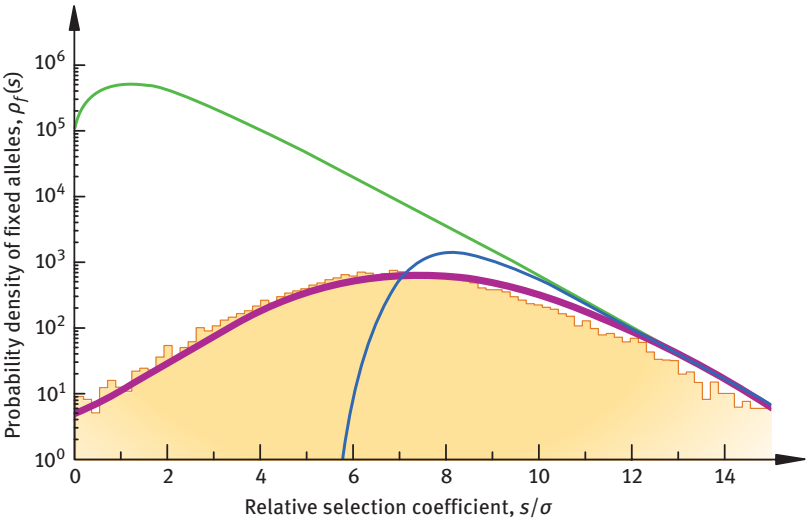

Figure 2.12: Distribution of fitness effects of fixed mutations, $\rho_f(s)$. Solid purple line: the analytic prediction. Blue dashed line: the prediction of the clonal interference theory (Section 2.2). Green line: prediction of the one-locus model (Chapter 1). Parameters: $N = 10^7$, $U_b = 10^{-5}$, $\sigma = 0.01$ (based on Good et al. (2012)).

which represents the dominant value of the selection coefficient. This fact results to two coupled equations for v and x_c (Good et al., 2012)

$$2 = \frac{U_b}{\sigma}\sqrt{\frac{2\pi\sigma^2}{v}}\left[1 + \frac{\sigma}{x_c} + \frac{v}{\sigma x_c - v}\right] e^{\frac{\left(x_c - \frac{v}{\sigma}\right)^2}{2v}} \tag{2.64}$$

$$1 = NU_b\left[\frac{x_c{}^2}{v} - 1 + \frac{2x_c\sigma}{v} + \frac{2\sigma^2}{v}\right] e^{-\frac{x_c - \frac{v}{2\sigma}}{\sigma}} \tag{2.65}$$

In the general case, numerical solution is required for this system of equations, but asymptotic analytical expressions for these quantities are feasible in two important limits, as follows.

If the dominant fitness effect size is relatively large and comparable to the lead of distribution, $s^* \sim x_c$, which takes place at intermediate population sizes (see further), then the dominant mutations represent large jumps in fitness compared to the lead of distribution, $|x - x_c| \sim x_c$ on the order of magnitude (Figure 2.9, left orange bar). In this case, one can solve eqs. (2.64) and (2.65) by iterations and obtain

$$v \approx \frac{\sigma^2 \log^2(NU_b)}{2\log\left(\frac{\sigma}{U_b}\right)}, \quad \log(NU_b) \ll 2\log\left(\frac{\sigma}{U_b}\right) \tag{2.66}$$

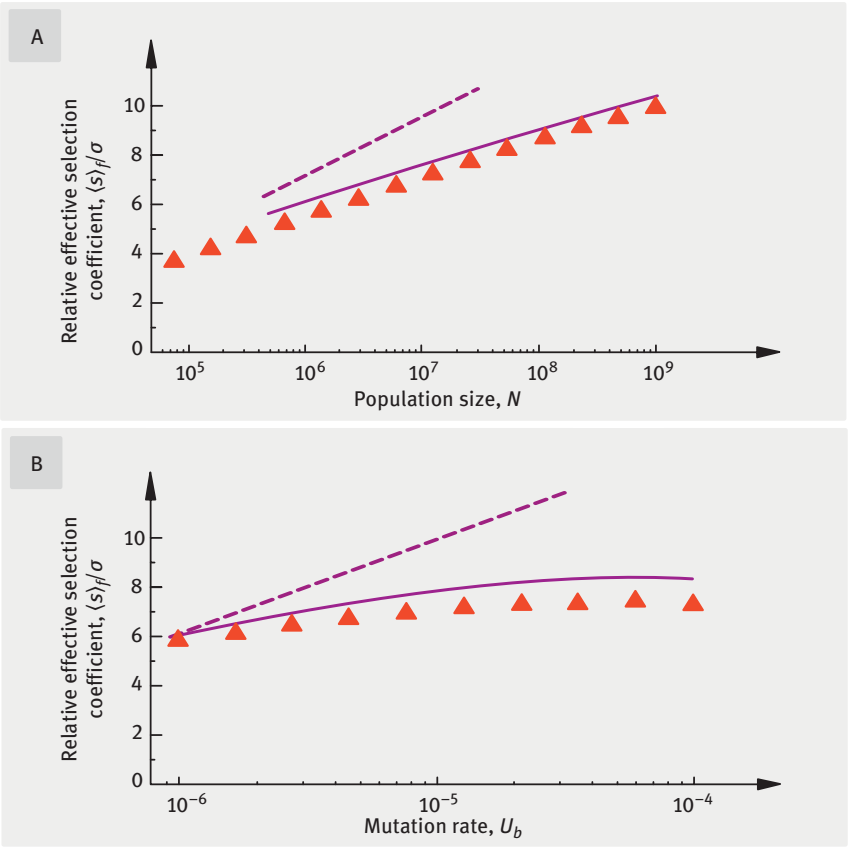

Figure 2.13: The average fitness effect of a fixed mutation as a function of the population size, N (A), and beneficial mutation rate, U_b (B). Orange triangles and the purple solid curve show simulation and analytic results, respectively. Dashed purple line: two-clone interference model results (Section 2.2.). Parameters are as in Figure 2.12 (based on Good et al. (2012)).

which holds at intermediate NU_b. Here we neglected logarithmic factors inside of large logs. For sufficiently large NU_b, in contrast, most successful mutations will have a relatively small effect compared to the lead and occur in genomes close to the high-fitness transition point ($s^* \ll x_c$) (Figure 2.9, right orange bar). In this case, the approximate expression for the adaptation rate from eqs. (2.64) to (2.65) has a form

$$v = 2\sigma^2 \log[NU_b \log(NU_b)], \quad \log(NU_b) \gg 2 \log\left(\frac{\sigma}{U_b}\right) \tag{2.67}$$

Therefore, there are two distinct regimes of adaptation in the two intervals of population size. One type of adaptation proceeds by large jumps in fitness comparable to the lead of distribution, and it maps to two-clone interference model discussed in Section 2.2

[which also assumed, we remind, an exponential mutation spectrum, eq. (2.54)]. Another type of adaptation occurs by relatively small jumps, and maps to the results of the multi-locus model with a constant fitness effect of mutation, as long as selection coefficient s and beneficial mutation rate U_b are replaced with the effective values

$$s_{\text{eff}} = s^*, \quad U_{\text{eff}} \sim U_b \sqrt{2\pi v}\, \rho(s^*) \tag{2.68}$$

The second type of mapping can be shown to exist in the more general case than the exponential mutation spectrum assumed here (Good et al., 2012). In the next section, we derive these results formally.

2.8.5 Derivation of fixation probability

Our derivation is based on a mean-field approximation that individual lineages become extinct or fix independently on each other, with fitness distribution of genomes given by its mean. As we have seen previously for other models, the details of reproduction model have no impact on statistical quantities such as v and $\rho_f(s)$ (Chapter 1 and previous sections of this chapter). For dynamics of individual lineages, we can use a branching process technique in continuous time. The process has birth rate $B(X,t) = 1 + X - X(t)$, death rate $D = 1$, and mutations from fitness X to fitness $X + [s,\ s + ds]$ occurring at rate $U_b \rho(s) ds$.

We will use $p(n, X, t)$ to denote the probability of extinction of a lineage that starts from n individuals with fitness X at time t. The backward master equation including birth, death, and mutation events reads

$$\begin{aligned} p(n,X,t-dt) = {} & \left[1 - n\ dt\left(2 + X - \bar{X}(t) + U_b\right)\right]\ p(n,X,t) + n\ dt\left[1 + X - \bar{X}(t)\right]\ p(n+1,X,t) \\ & + n\ dt\ p(n-1,X,t) + \ n\ dt\ U_b \int_0^\infty ds\ \rho(s)\ p(1,X+s,t)\ p(n-1,X,t) \end{aligned} \tag{2.69}$$

Here the four terms represent the probabilities of nothing happening, birth, death, and mutation, respectively. Passing to the continuous time limit, we obtain integro-differential equation

$$\begin{aligned} -\frac{1}{n}\frac{\partial}{\partial t} p(n,X,t) = {} & -\left(2 + X - \bar{X}(t) + U_b\right) p(n,X,t) + \left[1 + X - \bar{X}(t)\right] p(n+1,X,t) \\ & + \ p(n-1,X,t) + \ U_b \int_0^\infty ds\ \rho(s) p(1,X+s,t) p(n-1,X,t) \end{aligned} \tag{2.70}$$

By assumption, each lineage becomes extinct independently. Therefore, we seek a solution in the form $(n,X,t) = p(1,X,t)^n$. We easily verify Equation (2.70) that has a solution in this form, and $p(1,X,t)$ satisfies equation

$$-\frac{\partial}{\partial t}p(1,X,t) = -\left[2+X-\bar{X}(t)+U_b\right]p(1,X,t)+\left[1+X-\bar{X}(t)\right]p(1,X,t)^2 + 1+U_b\int_0^\infty ds\ \rho(s)p(1,X+s,t) \tag{2.71}$$

The fixation probability of one genome, $w(X,\ t),$ is related to the extinction probability $p(1,X,t)$ in the obvious way

$$w(X,t) = 1-p(1,X,t) \tag{2.72}$$

which can be substituted into eq. (2.71) to produce

$$-\frac{\partial w(X,t)}{\partial t} = \left[X-\bar{X}(t)\right]w(X,t)-\left[1+X-\bar{X}(t)\right]w(X,t)^2 + U_b\int_0^\infty ds\ \rho(s)\left[w(X+s,t)-w(X,t)\right] \tag{2.73}$$

Due to the translational symmetry of the right-hand side, the fixation probability of lineage depends on absolute fitness X and time t as a function of the relative fitness, $x = X-\bar{X}(t)$. Hence, we can replace the time derivatives with the derivatives in x and arrive at the desired ordinary differential equation for $w(x)$ given above

$$v\ \frac{dw}{dx} = xw(x)-w(x)^2+U_b\int_0^\infty ds\ \rho(s)\left[w(x+s)-w(x)\right] \tag{2.56}$$

Here, we assumed $x \ll 1$. In other words, all relevant selection pressures are small, and evolution is gradual.

Now, we need to solve eq. (2.56) for $w(x)$ to obtain eqs. (2.60) and (2.61). The left-hand side of eq. (2.56) represents the wave movement, and the three terms in its right-hand side are due to selection, a nonlinear effect, and mutation. Although we cannot obtain a general analytic solution, we can approximately solve it in different intervals of x. For large positive x, a typical lineage expands relatively fast. Therefore, its fate will be sealed (die or be established) long before additional mutations or the advance of the average fitness will be able to significantly impact its growth. Thus, for large x, the first and second terms in the right-hand side of eq. (2.56) dominate, which yields the standard one-locus result with x instead of s (Chapter 1):

$$w(x) \approx x \tag{2.74}$$

This approximation remains valid if the other terms remain relatively small, which occurs when $x^2 \gg v$, $x^2 \gg U_b\bar{s}$.

With x decreasing, eventually, the advance of average fitness $X(t)$ will start having a significant adverse effect on lineage dynamics on w. In this regime of interference,

the fixation probability is greatly reduced, and the quadratic term w^2 becomes small. If we neglect also the mutation term, because mutation is rare, we obtain

$$v\,\frac{dw}{dx} \approx x\,w(x)$$

which yields

$$w(x) \approx Ae^{x^2/(2v)} \tag{2.75}$$

where A is a constant of integration. The transition point between the interference regime and the drift regime, $x = x_c$, can be calculated by matching eqs. (2.74) to (2.75) at x_c, which yields

$$A = x_c e^{-x_c{}^2/(2v)} \tag{2.76}$$

After substituting this solution, eq. (2.75), into eq. (2.56), the condition that the quadratic term, w^2, is negligible:

$$\frac{x_c}{x} e^{\frac{x^2 - x_c^2}{2v}} \ll 1 \tag{2.77}$$

This inequality holds, as long as $x_c \gg \sqrt{v}$ and $(x_c - x)/x_c \gg v/x_c^2$. The exponential slope in x in the left-hand side of eq. (2.75) defines the characteristic width of the boundary layer between the two regimes, $\delta \sim v/x_c$, which is very narrow relatively to the lead for $x_c \gg \sqrt{v}$. Also, our assumption that the mutation term in eq. (2.56) is negligible at $x = x_c$ is valid as long as $x_c \gg \sqrt{U_b \bar{s}}$.

At even smaller fitness values, eq. (2.75) cease to apply, because it makes the biologically meaningless prediction that fixation probability $w(x)$ increases at fitness values below the average. Fortunately, as long as inequality $x_c \gg \sqrt{v}$ holds, fixation probability $w(x)$ for these fitness values is so low that we can effectively approximate it by zero. The dominant contribution from $w(x)$ comes from highly-fit genomes, therefore, does not depend on the form of $w(x)$ in this low-fitness region. Combining all three segments of x, we obtain the desired formula for the probability of fixation, $w(x)$, eq. (2.60).

While we have shown the existence of a sharp transition at $x = x_c$, we have not yet determined its location. Although the transition point location is determined by eq. (2.56), it cannot be captured by the above approximate analysis. In order to extract x_c from eq. (2.56), we use a method, as follows. We multiply its both sides by the Gaussian factor $\exp\left(-x^2/2v\right)$ and integrate over x. Then the main terms related to the wave speed and natural selection cancel and we get an equality

$$\int_{-\infty}^{\infty} dx\; w(x)^2 e^{-\frac{x^2}{2v}} = U_b \int_{-\infty}^{\infty} dx\; w(x) \int_0^{\infty} ds\; \rho(s) \left[e^{-\frac{(x-s)^2}{2v}} - e^{-\frac{x^2}{2v}} \right] \tag{2.78}$$

From the form of $w(x)$ in eq. (2.60), the left-hand side of eq. (2.78) can be evaluated by dividing the integration interval into two, one to the left from x_c, and another to the right. Because Gaussian functions change very rapidly near x_c due to our assumption $x_c \gg \sqrt{v}$, their arguments can be approximated with their linear expansion in $x - x_c$. As a result of integration in Eq. (2.78), we arrive at the condition for x_c given by eq. (2.61).

Now we need to calculate integrals in s in eq. (2.61) using eq. (2.54). We notice that the dependence of both integrands on s is mostly determined by an exponential with an argument

$$g(s) = -\frac{s^2}{2v} + \frac{x}{v}s - \frac{s}{\sigma} \tag{2.79}$$

Note that $g(s)$ reaches a narrow maximum at

$$s = s^* \equiv x_c - \frac{v}{\sigma}$$

as given by eq. (2.63). At this maximum point

$$g(s^*) = -\frac{(x - v/\sigma)^2}{2v}, \quad g''(s^*) = -\frac{1}{v} \tag{2.81}$$

Hence, we can approximate these integrals with the integral of a Gaussian. If we also assume $s^* \gg \sqrt{v}$, we readily arrive at eq. (2.64), the first of coupled equations.

Finally, in order to calculate wave velocity v, we substitute eq. (2.60) for the fixation probability, and eq. (2.55) for the fitness profile, into the velocity consistency condition, eq. (2.58), and evaluate the integrals using the same method. We arrive at the second coupled equation, eq. (2.65). Together, eqs. (2.64) and (2.65) allow us to solve for v and x_c as explained above. This derivation was repeated for a more general form of $\rho(s)$ (Good et al., 2012).

Chapter 3
Multi-site evolution with recombination

3.1 Two roles of recombination in adaptation

The main difference between asexual and sexual organisms is the presence of recombination. During mating, genetic information is combined from two parental DNA, a half from each, and is passed to the progeny. The evolutionary function of recombination is to create better-fit genomes. Natural selection amplifies them further. As we demonstrate in this chapter, recombination can effectively diminish clonal interference and expedite adaptation by orders of magnitude, even if the recombination rate is relatively small. For example, in an average HIV-positive individual off therapy, only ~1% of infected cells are infected with two different viruses and hence can undergo recombination (Batorsky et al., 2011; Neher and Leitner, 2010). There are 3 – 10 recombination crossovers per genome (Levy et al., 2004). Yet, even such a minor rate of recombination is enough to speed up the rate of HIV adaptation several–fold (Batorsky et al., 2011).

A new key parameter is the recombination rate per genome, r, defined as the probability per generation that an individual has recombination with another randomly chosen genome. It is also called the "outcrossing rate." Another new parameter is the average number of crossovers per genome, M. For example, in the human genome, which has 100% sexual reproduction, we have $r = 1$, $M = 2.5$ crossovers per chromosome, and 23 pairs of chromosomes.

Two limiting scenarios of adaptation exist, based on two main roles of recombination, as follows. A role of recombination is to bring together beneficial alleles to the same fitness background (and, conversely, excise deleterious alleles). As illustrated in Figure 2, this operation counteracts the effect of clonal interference. Another role of recombination is to facilitate the establishment of new beneficial alleles into a population. As shown in Chapter 2, in asexual populations, new beneficial mutations can survive in future populations and contribute to adaptation, only if they occur in the fittest individuals. In sexual populations, recombination can transfer alleles from modest genetic backgrounds to better-fit genomes and, in this way, alleviate the negative factor of "background selection" (Rice, 2002). These two roles of recombination are best studied in limiting scenarios of adaptation:

Role 1 is to generate new, better-fit sequences. This scenario corresponds to the evolution on moderate timescales, driven by recombination and selection alone. In Section 3.2, we develop a model, which assumes that small frequencies of beneficial alleles exist at all sites in the very beginning (e.g., due to previous mutation or migration events), but on different individual sequences (Rouzine and Coffin, 2005, 2007, 2010). Further mutation is assumed to be negligible. These individual alleles are joined

https://doi.org/10.1515/9783110615456-003

by recombination into pairs and triplets, and so on, which eventually results in a traveling wave in fitness coordinate. Due to selection, the wave propagates toward higher fitness values. The wave speed, as usual, is limited by the extension of the stochastic edge where establishment of new rare recombinants occurs (Figure 3.1). Then, phylogenetic relation of sequences kicks in decreasing the efficacy of recombination. The frequencies of beneficial alleles at certain sites are gradually increased, at other sites, beneficial alleles are eventually lost due to random drift and clonal interference. Eventually, the wave stops due to inbreeding, when all genomes become the same. We will consider this scenario in Sections 3.2 and 3.4.

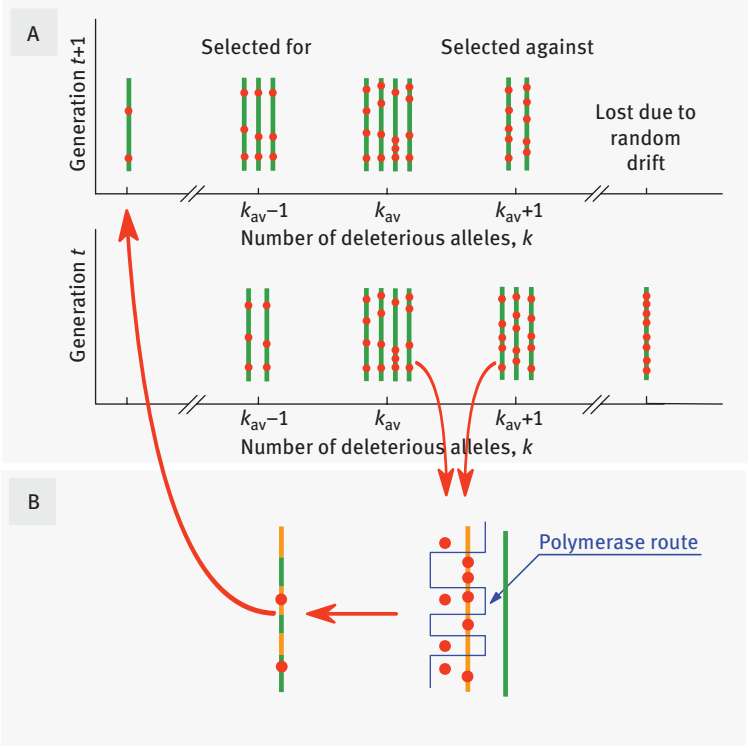

Figure 3.1: A model of evolution in the presence of selection, recombination, and random drift. (A) Haploid population in two consecutive generations. Green lines: genomes. Red circles: less-fit alleles. (B) The recombination mechanism. Blue broken line: the route of RNA/DNA polymerase between the two parental genomic templates (based on Rouzine and Coffin (2005)).

Role 2: After the wave in Scenario 1 runs out of diverse sites, because all alleles at each site are either lost or fixed, a new regime ensues where mutation is necessary to infuse new alleles. Alternatively, the wave may not be formed to begin with, because allelic copies per site are too few, or recombination is too rare to combine them

quickly. In either scenario, adaptation requires continuous production of new beneficial alleles. Adaptation in this case is very slow (Neher et al., 2010), because only several sites have enough diversity to contribute to the recombination effect. In that scenario, which we consider in Section 3.3, mutation and the fixation probability of a beneficial allele are two factors limiting infusion of new alleles into a population. Recombination assists with fixation of new alleles. In purely asexual populations, only mutations occurring in the most-fit genomes survive and do not succumb to extinction clonal interference. Mutated alleles in less-fit genomes do not establish lineages, a consequence of clonal interference known as the "background selection" effect (Rice, 2002). However, even very rare recombination events expand the interval of probable fixation from the high-fitness edge toward the bulk of distribution. This expansion occurs due to repeated leaps of good alleles from one genome to another that is better-fit, a process termed "gene surfing" (Neher et al., 2010; Rice, 2002).

The biological roles of recombination are not restricted to the two listed roles. The other roles include stopping Muller's ratchet (Muller, 1932; Muller, 1964) and compensating gamete defects. However, here we will consider only the limit of adaptation, will neglect deleterious mutation, and focus on these two roles related to beneficial mutations.

The value of outcrossing rate, r, and the length of evolution decide whether Scenario 1 or Scenario 2 is the better description of adaptation process. Scenario 2 (allele-fixation) applies either when recombination rate r is small, near the crossover region to the purely asexual regime, or for any r in the long-term stationary process. Scenario 1 applies in a broad range of recombination rates r but in transient (although, possibly, quite long) process. (In principle, one can imagine a situation where both mechanisms of recombinations work in synergy.)

For example, peoples who speak Indo-European languages and have diversity $T \sim 0.1\%$, did not need new mutations to detectably diverge from the common roots over the last 6,000 years, because the combination of migration, founder effects, selection, and sexual reproduction is enough for evolution on that timescale, ~ 300 generations. However, on much longer timescales, such as tens of thousands of years, we observe a substantial change of phenotype due to new adaptive mutations. We consider both scenarios in Sections 3.2, 3.3, and 3.4.

3.2 Recombination and natural selection (no mutation)

3.2.1 Approximation of uncorrelated genomes

As we mentioned, the short-range evolution of a diverse population is driven by natural selection and recombination alone. Our basic model (Figure 3.1) (Rouzine and Coffin, 2005) is very similar to the asexual model in Section 2.2.

It considers a population of N haploid genomes that have a large number of loci L (or a diploid population of $N/2$ individuals without allelic dominance). All loci have two alleles, as in our previous models, with a small fitness difference $s \ll 1$. Generations do not overlap: all the genomes die and are replaced with their offspring. The genome fitness with respect to the best possible fitness (progeny number) is given by $\exp(-sk)$ where k is the number of less-fit alleles. By the definition, the best possible genome has $k = 0$. The last expression assumes that epistasis (interaction between loci) is absent, and that all alleles have identical fitness effect. The effects of epistasis will be considered in Volume II of this book.

3.2.1.1 Model of recombination

The mechanism of recombination depends on an organism. For the purpose of this work, we will focus on assumptions and parameters relevant to a broad range of viruses. With some reservation, our results will also be relevant for short segments of animal genomes, as long as the probability of recombination, r, is properly rescaled and allelic dominance is absent; but we will talk, for a moment, specifically about viruses. The effective size of virus population is given by the total number of virus genomes inside of cells, N, which produce new infectious virus particles able to reach new cells. An infected cell generates virus particles with RNA (or DNA) copies of the viral genome, and these particles find new cells to infect. A cell coinfected with two particles can produce some particles that contain recombined pairs of their genomes (Figure 3.1, bottom). Hence, fraction r of all genomes will undergo recombination with another virus genome, while fraction $1-r$ represents a copy of a single parental genome. Parameter r is often called "the outcrossing rate." Recombination between the two genomes occurs due to random crossovers of the polymerase protein between the two RNA templates (Levy et al. 2004).

We assume further that the number of crossovers per genome M is large and fluctuates according to Poisson distribution, and that the recombinant is composed of and equal mixture of each parental genome. Sometimes, the recombination rate is defined not per genome but between two genome sites, r_2. It is related to the genomic recombination rate as $r_2 = rM\Delta L/L$, where ΔL is the number of base pairs between the two sites.

The critical assumption of this section, which will be lifted in Section 3.4, is that, given the total number k of less-fit loci, they are distributed uniformly and randomly among L available sites, and that their locations in different genomes do not correlate. Note that we do not assume full statistical independence of genomes, because variance of k between genomes, $\mathrm{Var}[k]$, will be shown to be smaller than the Poisson value, $\sqrt{\bar{k}}$, where $\bar{k}(t)$ is the population average of k. However, genomes will correlate with each other only in the value of k. In Section 3.4, this approximation will be lifted and site-by-site correlations will be taken into account. We also assume random mating: any pair of genomes has an equal probability to recombine.

This basic model does not include mutation events, because we assume that all necessary one-site alleles already exist in the beginning. We are interested in the case when adaptation is recombination-driven and is much faster than asexual adaptation due to mutation.

3.2.1.2 Validity range

The following derivation of the results in Section 3.2.2 is asymptotically exact in a broad range of parameters given by strong inequalities, as follows:

$$1 \ll \overline{k} \ll \frac{1}{s^2}, \quad s \ll r \ll s\sqrt{\overline{k}}$$

$$1 \ll \ln(Nr) \ll \min[\overline{k}, 1/(s^2\overline{k}), \quad \overline{k}(s/r)^2 \log\left(s\sqrt{\overline{k}}/r\right) \tag{3.1}$$

As it easy to observe, for $\overline{k} = 100 - 1,000$ and $s = 0.01 - 0.05$, the range of r, N is quite broad. In what follows, we will neglect small terms with the use of eq. (3.1).

The most important, as we show later, is double inequality $s \ll r \ll s\sqrt{\overline{k}}$. At very small recombination rates, $r \ll s$, recombination is not important. In the limit $r \gg s\sqrt{\overline{k}}$, clonal interference effects are fully destroyed by recombination, and the adaptation rate is equal to the deterministic independent-locus result, $V = s\overline{k}(1 - \overline{k}/L)$ (Chapter 1). This interval exists and is broad, since real populations are usually far from the best-fit sequence, so that the number of alleles, $\overline{k}$, is usually much larger than 1.

3.2.1.3 Dynamic equations

We denote with $f(k,t)$ the average frequency of genomes with k deleterious alleles with respect to the best-fit possible sequence. For our model (Figure 3.1), the dynamic equation for $f(k,t)$ has a form

$$f(k,t+1) - f(k,t) = \left\{e^{-s\left[k-\overline{k}(t)\right]} - 1\right\} f(k,t) + r[R(k,t) - f(k,t)] \tag{3.2}$$

where t is time in generations, $e^{-s\overline{k}(t)} \equiv \int dk e^{-sk} f(k,t)$. Here notation $rR(k,t)$ has the meaning of the increase of the class with k alleles due to recombination, specific form of $R(k,t)$ is defined later, and $-rf(k,t)$ is the loss of sequences from class k due to recombination. Normalization condition has a form

$$\int R(k,t)dk = \int f(k,t)dk = 1.$$

Here we study the evolution of genetic variation that already exists in the beginning ("standing variation"). Hence, eq. (3.2) neglects new mutation events, whose role will be considered in Section 3.3. In the above parameter range, eq. (3.1), we have strong inequality $s\left|k - \overline{k}\right| \ll 1$ for all relevant k; therefore, the exponential in eq. (3.2)

can be approximated with the linear expansion in $k-\bar{k}$. In addition, $f(k,t)$ can be approximated with a function continuous in t (Section 3.2.5, Notes 1 and 2). As a result, we have

$$\frac{\partial f}{\partial t} = -s\left[k-\bar{k}(t)\right]f(k,t) + r[R(k,t)-f(k,t)]$$

$$\bar{k}(t) \equiv \int f(k,t)k dk \tag{3.3}$$

The form of the recombination gain function $R(k,t)$ in eq. (3.3) varies between organisms. For example, bacteria exchange genomic segments, while viruses and eukaryotes have crossover recombination with 50% of each parental genome passed to the progeny. We will focus on the latter, although the technique can be generalized for bacteria as well. We will stick to our main assumption that alleles given k are scattered randomly within a genome. We assume that the frequency of less-fit alleles per locus is small, $\bar{k}(t) \ll L$. When two genomes with k_1 and k_2 alleles recombine, they make an offspring genome with $k = (k_1 + k_2)/2 + \varepsilon_1 + \varepsilon_2$ mutations, where $\varepsilon_{1(2)}$ is the deviation of the random allele number in the copied half of a parental genome from the average, restricted by the condition that the total allele number in the genome is fixed and equal to $k_{1(2)}$. Because all k are large, the probability distribution of k has a Gaussian form centered at $k = (k_1 + k_2)/2$. The variance $<\varepsilon_{1(2)}{}^2>$, is calculated from the average number of alleles in a half of a genome, $k_{1(2)}/2$, and the additional factor $1/2$ is caused by the above restriction. Because fluctuations of alleles are independent in the two parents (the central approximation), and, in the stated parameter range, the wave is narrow, $|k_1 - k_2| \ll \bar{k}$ (Section 3.2.5, Note 3), we have $(\varepsilon_1 + \varepsilon_2)^2 = (k_1 + k_2)/4 \approx \bar{k}/2$. The resulting expression for $R(k,t)$ has a form

$$R(k,t) = \frac{1}{\sqrt{\pi \bar{k}}} \int dk_1 \int dk_2 f(k_1,t) f(k_2,t) e^{-[k-(k_1+k_2)/2]^2/\bar{k}} \tag{3.4}$$

Both fitness distribution $f(k,t)$ and recombination function $R(k,t)$ are illustrated in Figure 3.2.

3.2.2 Main results

We start by describing our main results graphically. Their derivation will be given in Sec 3.2.3. Just in asexual populations (Chapter 2), most fitness classes k, except for genomes with smallest k and largest fitness, averaged over realizations, can be treated in a deterministic way. We confirm this approximation further by Monte-Carlo simulation down to very small population sizes, $N \sim 10^2$. Just as in asexual case, the deterministic equation predicts (Section 3.2.3) a moving solitary wave in

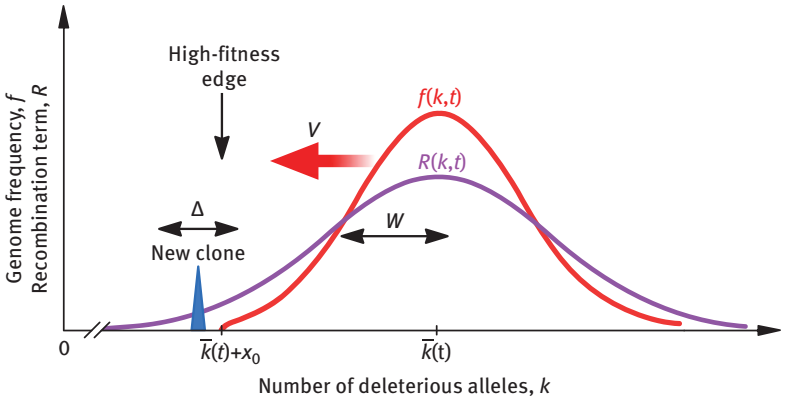

Figure 3.2: Schematic of the moving solitary wave. Red and purple curves: and thin lines, fitness class frequency, $f(k,t)$, and the recombination generator function $R(k,t)$, respectively. Spike: a new recombinant clone generated beyond the wave edge; Δ, interval where most such clones are generated; w and V, the width and the speed (evolution rate) of the wave, respectively (based on Rouzine and Coffin (2005)).

the fitness coordinate with an almost constant profile (Figure 3.2). The mean substitution rate of beneficial alleles is the wave speed, $V = -d\bar{k}/dt$.

We start with the deterministic limit of infinite population, $N = \infty$. In this limit, the wave profile is Gaussian. The variance of k is given by the Poisson value $\bar{k}$ which shows that loci evolve independently of each other at infinite population size. However, if population size is finite, Gaussian approximation works until a point, our familiar high fitness edge, where the wave ends at a finite value of k (Figure 3.2).

As in the asexual case, the stochastic edge at small k deserves a special treatment. Genomes beyond the leading edge (small k) are absent, simply because they did not exist in the beginning. They are acquired gradually during the process of evolution. (The genomes beyond the trailing edge are absent too, because they are already extinct, but that edge is not interesting.) Rare recombinants born just outside the leading edge that escape extinction limit the wave speed (Figure 3.2). To estimate fitness and the average time to generation of the edge class, we again apply the two-allele one-locus model. Again, the emerging recombinant lineage is a minority variant, and the bulk of the population is the majority variant. Further, we will obtain the expressions for the wave width $w$and the wave speed V by a self-consistency condition, matching the time to a new successful recombinant to the shift time of the wave. The result has a form

$$w^2 = p\bar{k}, \quad V = ps\bar{k}$$

$$p = \frac{\ln(Nr)}{\ln\left(Ns^2\bar{k}/r\right)} < 1, \quad 1/N \ll r \ll s\sqrt{\bar{k}} \tag{3.5}$$

Note that the width and speed of the wave are related by Fisher's theorem $V = sw^2$. Formula (3.5) neglects logarithms in the arguments of the two large logarithms, which create a minor error in p.

In a finite population, the variance w^2 in eq. (3.5) is smaller than the Poisson value, $\bar{k}$, by a factor of $p < 1$, because linked loci do not evolve independently. The width is related to the wave speed, as given by $V = sw^2$. In agreement with Fisher Theorem, the wave speed is smaller than its deterministic limit, $V = s\bar{k}$, corresponding to the one-locus result. This is a manifestation of clonal interference partly compensated by recombination (Chapter 2).

Equation (3.5) implies a critical value of population size, $N \sim 1/r$, below which evolution by the described mechanism is not possible. The fair accuracy of eq. (3.5) is confirmed by Monte-Carlo simulation at realistic parameter values (Figure 3.3). In the range of very strong recombination, $(r/s)^2 \gg \bar{k}$, the transition in r from $V = 0$ to maximal speed $V = s$ becomes very sharp and cannot be described by this method.

Equation (3.5) is valid when less-fit loci are rare, $\bar{k} \ll L$. In the case when the external condition sharply changes, the population can be less fit at most of L loci, except for a minority of better-fit alleles. The fraction of alleles in genome, $\bar{k}/L$, decreases gradually during adaptation almost from almost 1 to 0 and, in the middle of the process, is not small. In this case, we can easily generalize the expression for V in eq. (3.5) by replacing $\bar{k}$ in the variance of parental inheritance with $\bar{k}(1 - \bar{k}/L)$. This gives, again, the independent-locus result from Chapter 1 reduced by the factor of p.

The analytic result is compared to Monte-Carlo simulation for representative parameter values in Fig 3.4 (Rouzine and Coffin, 2005). The details of this simulation carried out in the same approximation of uncorrelated genomes are described in Section 3.2.4. The fitness classes with k alleles are shown at different times (Figure 3.4A and D). The average allele number $\bar{k}$ decreases in time (Figure 3.4B, E). The scaled slope $V/(s\bar{k})$ as well as the scaled variance $w^2/\bar{k}$ is compared in Figure 3.3 with the analytic result for p, eq. (3.5) . We observe that the analytic result somewhat underestimates the accumulation rate.

If we consider a random realization of a wave (a simulation run), we observe that it breaks down into separate peaks that become more dense as N increases (Figure 3.4A and D). Nevertheless, the averaged and centered shape of the wave agrees very well with the analytic result (Figure 3.4C and F). In particular, the profile is slightly skewed, with a steep high-fitness cutoff, as expected from the theory.

If population becomes smaller than a critical size, $N < 1/r$, the wave does not get far, because it collapses into a single clone of identical sequences. Then the wave stops, because recombination cannot generate new sequences (Figure 3.4G). (As we will show in Section 3.4, the same stop eventually happens, in the long term, at all parameter values, and even earlier than predicted in Figure 3.4G, due to increasing correlations between genomes caused by inbreeding neglected here.) If

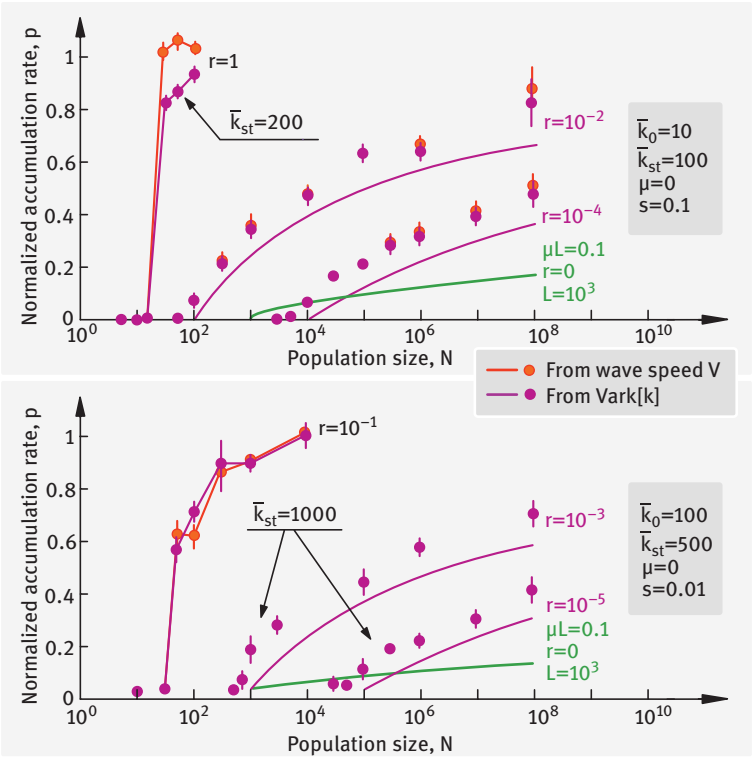

Figure 3.3: Analytic predictions for the evolution speed compared with the Monte Carlo simulation results, both obtained in in the approximation of uncorrelated loci. The average speed, V and squared width, w^2, of a solitary wave without mutation ($\mu = 0$) as a function of the population size, N, for $s = 0.1$ (top) and $s = 0.01$ (bottom). Both quantities are scaled by their respective values in the limit of infinite size. Orange circles: rescaled wave speed $(d\bar{k}/dt)/(s\bar{k})$; purple circles: width squared, $w^2/\bar{k}$. Vertical bars: 67% statistical interval for the estimate of the average; purple lines, analytic results (eq. (3.5). The average allele numbers at the start, k_{st}, and at the sampling time, k_0, are shown. The values of r are on the curves. Simulation results are averaged over 40 random runs (top) and 10 runs (bottom). Green lines: results for an asexual population for $\mu = 10^{-4}$ and the total locus number $L = 10^3$ (based on Rouzine and Coffin (2005)).

the factor of mutation is included, the evolution speed is finite even below the critical population size (green curve in Figure 3.3).

We conclude that even relatively infrequent recombination is quite effective in driving evolution compared to a purely asexual regime. The asexual evolution speed is given by eq. (2.35), which is smaller than the speed driven by recombination by the large factor of $\bar{k}$. The result for the asexual speed transitions to the independent-site result $V_{1-\text{locus}} = s\bar{k}$ only at populations that are exponentially large in parameter $\bar{k}$, as given by eq. (2.53). In contrast, evolution with recombination given by eq. (3.5) reaches a half of the independent-loci rate already at a moderate

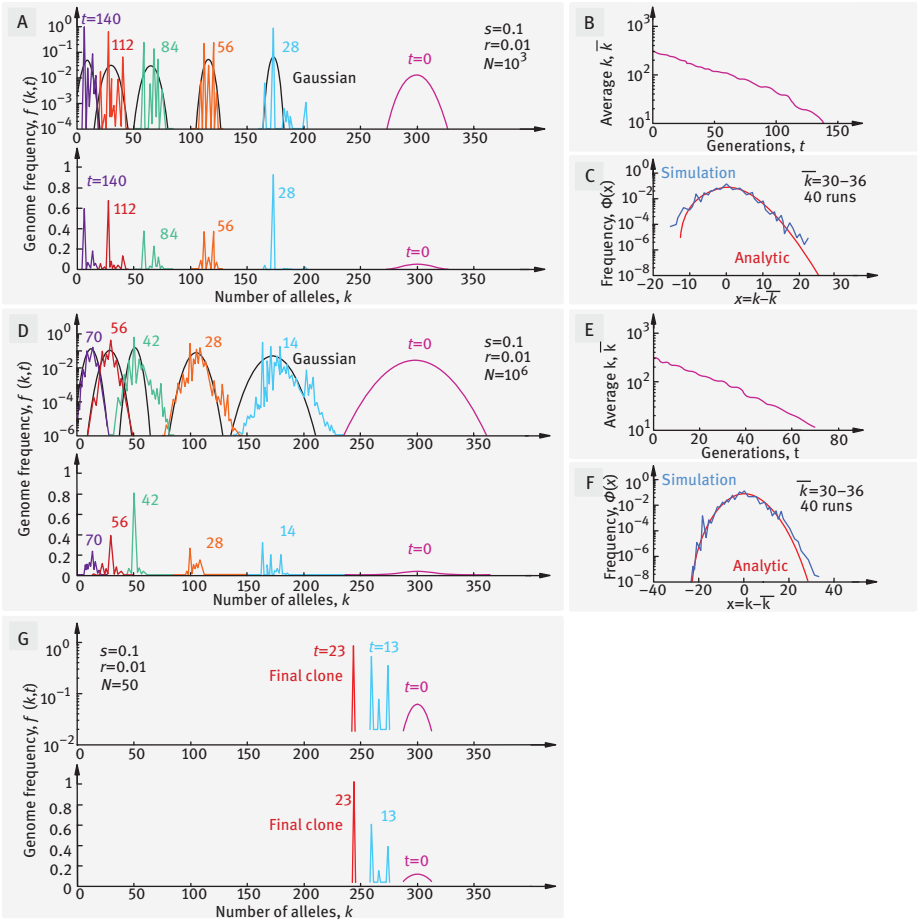

Figure 3.4: Examples of Monte-Carlo simulation of the traveling wave for the case with recombination and no mutation and in the absence of inter-sequence correlations, $f(k,t)$. (A) Top and bottom: logarithmic and linear scales. Curves in alternating colors: fitness class frequency $f(k,t)$ at different times (shown). Black lines: a fit with a Gaussian function. Model parameters are shown. (B) Population-averaged less-fit number $\bar{k}$ as a function of time. (C) Averaged centered fitness class frequency $\phi(x)$. Blue curve: the simulation result averaged over an interval of k (shown) and over 40 random runs. Red line: analytic result (eq. (3.2.19)). (D–F) Analogous results for a larger population size, $N = 10^6$. (G) Simulation below the critical population size, $Nr < 1$ (based on Rouzine and Coffin (2005).

population size, such that $N \sim (1/r)\left(s\sqrt{\bar{k}}\,/r\right)^2$. Therefore, the short-term adaptation under (even rare) recombination and natural selection is much more rapid than the evolution due to new mutations and selection in the absence of recombination provided the necessary better-fit alleles exist in the beginning. This result illustrates the fundamental evolutionary advantage of sexual reproduction.

3.2.3 Derivation

3.2.3.1 Solitary wave solution

Now we present derivation of the analytic results. We start with considering the limit of infinite population size, $N = \infty$, and then move on to finite population size. A partial solution of eqs. (3.3) and (3.4) has the form of a traveling wave

$$f(k,t) = \phi\left[k - \bar{k}(t)\right], \quad R(k,t) = \rho\left[k - \bar{k}(t)\right] \tag{3.6}$$

The traveling wave solution, eq. (3.6), describes the gradual decrease in the average number of deleterious alleles per genome $\bar{k}(t)$ due to recombination and selection, that is, the increase in the number of beneficial alleles. Substituting eq. (3.6), into (3.3) and (3.4), we get

$$V\frac{\partial\phi}{\partial x} = -sx\phi(x) + r[\rho(x) - \phi(x)] \tag{3.7}$$

$$\rho(x) = \frac{1}{\sqrt{\pi\bar{k}}}\int dx_1 \int dx_2 \; \phi(x_1)\phi(x_2)e^{-\frac{\left[x - \frac{x_1 + x_2}{2}\right]^2}{\bar{k}}} \tag{3.8}$$

where we introduced notation $x = k - \bar{k}(t)$, and

$$V = -d\bar{k}/dt$$

is the average substitution rate. In eq. (3.7) we neglected the fact that the wave profile slowly varies in time, which approximation is justified, if the wave is far from the origin, $k = 0$ (Section 3.2.5, Note 4).

The general solution of eq. (3.7) has a form

$$\phi(x) = \frac{b}{w^2}e^{-(x+b)^2/(2w^2)}\int_{x_0}^{x} dx'\rho(x')e^{(x'+b)^2/(2w^2)} \tag{3.9}$$

where x_0 is a constant that we determine further, and new notation b and w^2 is introduced

$$b \equiv \frac{r}{s} \tag{3.10}$$

$$w^2 \equiv V/s \tag{3.11}$$

At infinite population size, eq. (3.9) must be valid at any value of x, even very far from the wave center. Hence, we have $x_0 = -\infty$, otherwise, the function $\phi(x)$ at $x < x_0$ would be negative. The only solution of eqs. (3.8) and (3.9), for which the integral in eq. (3.9) does not diverge at $x' = -\infty$, has the Gaussian form

$$\phi(x) = \rho(x) = \frac{1}{\sqrt{\pi \bar{k}}} e^{-\frac{x^2}{2\bar{k}}} \tag{3.12}$$

$$w^2 = \bar{k} \tag{3.13}$$

For the wave speed, V, we have

$$V \equiv -\frac{d\bar{k}}{dt} = s\bar{k}, \quad N = \infty \tag{3.14}$$

Equation (3.12), which can be verified by direct substitution into eqs. (3.7) and (3.8), shows that the Var[k] is equal to the Poisson value $\bar{k}$, that is, that different alleles are distributed independently between genomes, and different sites are independent on each other. As it turns out, if the outcrossing rate r is sufficiently large (but still may be smaller than 1), loci can become effectively independent at finite N as well. Equation (3.14) is the deterministic result for the one-locus model, eq. (1.61) with $f \ll 1$ in Chapter 1.

3.2.3.2 Finite populations: stochastic edge

In the beginning, highly fit sequences do not exist yet; they are added gradually at the leading edge, which is located at a negative value, $x = x_0$. The value $-sx_0$ is the average relative fitness of the best-fit sequence present in the population (Figure 3.2).

As in the asexual model, at sufficiently large N, stochastic effects from random mutation and genetic drift are negligible for fitness classes k located far from the wave tips. Therefore, eq. (3.9) holds for fitness density $\langle \delta f(k,t) \rangle$ averaged over the statistical ensemble. In other words, in the right-hand side of eq. (3.3), we neglect correlations $\langle \delta f(k,t) \delta \bar{k} \rangle$, where $\delta f(k,t)$ and $\delta \bar{k}$ are fluctuations of the respective values between realizations. Monte-Carlo simulation demonstrates that this approach predicts the average values of $\phi(x)$, V, and w^2 fairly accurately at $N = 1,000$ and larger (see Figures 3.3 and 3.4)), despite of strong fluctuations.

At finite x_0, the integral in eq. (3.9) does not need to converge at $x = -\infty$, and values of w^2 less than $\bar{k}$ are possible. In Section 3.2.5 (Note 5), we show that (i) the lead of distribution $\phi(x)$ is much longer than the wave width, $|x_0| \gg w$, and (ii) in the interval $x_0 < x < |x_0|$, a small edge region, $x - x_0 \sim \delta x \ll |x_0|$, determines the integral in x' in eq. (3.9). Therefore, in this interval of x, eq. (3.9) takes a Gaussian form

$$\phi(x) = \frac{1}{w\sqrt{2\pi}} e^{-\frac{(x+b)^2}{2w^2}}, \quad w^2 < \bar{k}, \; x_0 < x < |x_0| \tag{3.15}$$

$$\frac{b\sqrt{2\pi}}{w}\int_{x_0}^{x} dx'\rho(x')e^{\frac{(x'+b)^2}{2w^2}} = 1 \tag{3.16}$$

where the second equation follows from the normalization condition $\int dx\, \phi(x) = 1$ and eq. (3.9).

Therefore, w^2 is the population variance of k, which can be lower than the Poisson value $\bar{k}$. In other words, the distribution of k across genomes is narrower than the Poisson distribution, since genomes are not fully independent and compete for fitness. We have already seen this effect in Chapter 2 for asexual populations.

Substituting eqs. (3.15) into eq. (3.8) and integrating over x_1 and x_2, for the recombination gain function we obtain

$$\rho(x) = \frac{1}{\sqrt{\pi(\bar{k}+w^2)}}e^{-\frac{(x+b)^2}{\bar{k}+w^2}}, \quad \bar{k} > w^2 \tag{3.17}$$

which is valid at any x. We can use asymptotics (3.15) for $\phi(x)$, because the integrals in x_1 and x_2 in eq. (3.8) converge at $|x_{1(2)}| \sim w \ll |x_0|$ (Section 3.2.5, Note 5).

The edge location x_0 can be linked to the value of w from the normalization condition for fitness density, eq. (3.16). Substituting eq. (3.17) into eq. (3.16), expanding the logarithm of integrand in eq. (3.16) linearly in $x' - x_0$ (Section 3.2.5, Note 5), and integrating in x', we obtain

$$x_0^2 \approx \bar{k}\frac{2p(1+p)}{1-p}\log\left[\frac{s\sqrt{\bar{k}(1-p)}}{r}\right], \quad r \ll s\sqrt{\bar{k}(1-p)} \tag{3.18}$$

where we have neglected logarithmic factors in the argument of the large logarithm.

Substituting eq. (3.17) into eq. (3.9) yields

$$\phi(x) = \frac{b}{\bar{k}^2 p\sqrt{\pi(1+p)}}e^{-\frac{(x+b)^2}{2kp}}\int_{x_0}^{x} dx' e^{\frac{(1-p)(x'+b)^2}{2p(1+p)k}} \tag{3.19}$$

$$p \equiv \frac{w^2}{\bar{k}}, \quad 0 < p < 1 \tag{3.20}$$

We note eq. (3.19) represents a generalization of the Gaussian in eq. (3.15) and applies at any $x > x_0$. In the four intervals of x, function $\phi(x)$ has the form

(i) $x < x_0$, $\phi(x) = 0$;

(ii) $0 < x - x_0 \ll \delta x$ (Section 3.2.5, Note 5), $\phi(x) \propto x - x_0$ from eq. (3.19)

(iii) $x_0 < x < |x_0|$ and $|x_0| - |x| \gg \delta x$, $\phi(x)$ is given by the Gaussian in eq. (3.15);

(iv) $x > |x_0|$, $\phi(x) \propto \rho(x)$ in eq. (3.17).

Deterministic equation (3.18) relates the lead $|x_0|$, to the standard deviation of allele number, w, and hence to the evolution speed, eq. (3.11). In order to obtain a second equation for $|x_0|$ and w, the stochastic dynamics at the edge has to be considered, as follows.

3.2.3.3 Stochastic high-fitness edge

Figure 3.2 illustrates how the extension of the high-fitness edge with time occurs. As in Chapter 2, we consider two genetic variants. The minority variant is a new clone forming near the wave edge, which has an effective selection coefficient, $S = s|x_0|$. The rest of population is the majority variant. A new genome is created beyond the fitness edge by recombination, at $x < x_0$. The rate per generation is, by the definition, $rN\rho(x)$, where $\rho(x)$is given by eq. (3.17). As already mentioned, in large populations, the lead is relatively long, $|x_0| \gg w$. Most beyond-edge genomes are born in a narrow interval of x with a width Δ given by

$$\Delta \sim \left|\frac{d\log\rho}{dx}\right|^{-1}_{x=x_0} \sim \bar{k}/|x_0| \ll |x_0| \tag{3.21}$$

The value

$$G \sim rN\rho(x_0)\Delta$$

is the total generation rate of these recombinants. After a recombinant is born, it will probably become extinct in a few generations due to random drift. However, if it is lucky to grow into a lineage above stochastic threshold $fN \sim 1/S$, which takes place with a small probability $\sim S$, the lineage will be established and extend the wave forward. The mean time to a successful clone is

$$t_{\text{seed}} \sim 1/(Gs) \sim \frac{1}{Nsr\sqrt{\bar{k}}}\, e^{x_0^2/\left(w^2+\bar{k}\right)} \tag{3.22}$$

where we have substituted eq. (3.21) for Δ and eq. (3.17) for $\rho(x_0)$ into the expression for G. From the consistency condition for the evolution speed, the time to such a successful recombinant is the same as the time in which the traveling wave moves by interval Δ, as given by

$$t_{\text{seed}} \sim \frac{\Delta}{V} \sim \frac{1}{|x_0|sp} \tag{3.23}$$

where we used eqs. (3.11) and (3.20) for V and eq. (3.21) for Δ. From eqs. (3.22) and (3.23), we obtain the desired second equation for x_0^2

$$x_0^2 \approx \bar{k}(1+p)\log\frac{Nr}{p} \tag{3.24}$$

A logarithmic factor in the argument of the large logarithm was neglected. Solving eqs. (3.18) and (3.24) together for x_0^2 and p, we arrive at the desired equation (3.5).

The above derivation applies only if the total number of new recombinants per generation is large, $Nr \gg 1$. Indeed, at $Nr \sim 1$, from eqs. (3.24) and (3.21), we obtain $x_0^2 \approx \bar{k}$ and $\Delta \sim |x_0|$, so that our assumptions that the lead is long and that new clones are generated in the vicinity of the high-fitness edge are no longer valid. In this regime, the wave has to stop. Indeed, if a new clone is generated far ahead of the wave, the wave becomes extinct due to fitness difference. The entire population becomes now one clone of identical sequences. Because recombination cannot make any new genomes, evolution comes to an end, until new mutations are produced. We conclude that, in the absence of mutation, there exists a critical point in product Nr below which evolution is not possible. In agreement with this, eq. (3.5) extrapolates to $V = 0$ at this point.

3.2.4 Monte-Carlo simulation

To test and illustrate stochastic mechanics, we discuss the computer simulation used in Figures 3.3 and 3.4. The algorithm developed in Rouzine and Coffin (2005) considered the same model of population and monitored dynamics of discrete fitness classes k and keeps the frequency of deleterious alleles k/L to be small. Effectively, this corresponds to a simulation where genomes in each class are randomized each generation to kill inter-genome correlations and make distribution of alleles random in a genome with given k. This matches the model of recombination to the analytic model, with one correction. To exclude self-recombination of a identical sequences, self-recombination of each fitness class is prohibited. This introduces an error, because a fitness class can have many different sequences that could recombine and produce high-fit progeny. However, because the width of the wave w is large above the critical point in Nr, eq. (3.1), the error is modest.

The simulation shows that the results obtained for the wave should be interpreted in the statistical sense. Unless Nr is very large, the wave in each realization consists of rare fitness classes with sparsely situated values of k within an ensemble-averaged envelope (Figure 3.4). This is because most fitness classes represent separate lineages born from infrequent recombinants.

The probability of having two clones within a class is $1/\Delta k$, where $\Delta k \gg 1$ is the average distance between adjacent classes. Therefore, the exclusion of the self-recombination of a class is equivalent to the exclusion of clonal self-recombination. In contrast, at very large N, the fitness classes are densely packed at adjacent integer values of k. The self-recombination correction, in this case, is inaccurate, because it throws away productive recombination between many clones within a class; however, this is not very important, because the correction is small anyway: the

probability of intraclass recombination is ~$1/w \ll 1$. Therefore, this approximation always works, at any population size.

The simulation stores the integer sequence number $n(k,t)$ for each class k at generation t, where $n(k,t) = Nf(k,t)$. When a generation changes, the mean value $\langle n(k,t+1)\rangle$ is calculated for all $k = 1, \ldots, L$, from the deterministic equation (3.3). The recombinant generator function $R(k,t)$, given by the discrete version of eq. (3.4), is corrected for the absence of recombination of each class with itself and re-normalized back to 1. All classes with mean size $\langle n(k,t+1)\rangle$ smaller than a set low threshold, $n_{\text{emp}} \ll 1$, are emptied in the next generation. To make stochastic simulation faster, new sizes of nonempty classes, $n(k,t+1)$, are generated by one of the two methods:

Method 1. If the mean size of a class, $\langle n(k,t+1)\rangle$, is smaller than a set high threshold $n_{\text{stoch}} \gg 1$, and the total frequency of such groups in a population is less than a set value $f_{\text{tot}} < 1$, these classes are treated as stochastic. Their numbers $n(k,t+1)$ are generated with the use of random generator to obey Poisson distribution with the calculated averages $\langle n(k,t+1)\rangle$. The remaining large groups are treated as deterministic, and their size is calculated as $n(k,t+1) = \langle n(k,t+1)\rangle$.

Method 2. If the total frequency of stochastic classes exceeds a set threshold f_{tot}, all nonempty classes are calculated stochastically from multi-nominal random distribution, as follows. N random points are generated within the interval $[0,1]$, which is then split into subintervals corresponding to classes k with widths proportional to $\langle n(k,t+1)\rangle$. The new numbers $n(k,t+1)$ are set to be the numbers of these random points within interval k.

Method 1 greatly enhanced the speed of the algorithm, without a significant loss in accuracy, at $\bar{k} < 500$ and arbitrarily large N. Both methods produced very similar results when thresholds in Method 1 varied within intervals: n_{emp} within $10^{-4} - 10^{-5}$, n_{stoch} within $500 - 1,000$, and f_{tot} below 0.2 (Rouzine and Coffin, 2005).

For Monte-Carlo run, the time dependence of wave center $\bar{k}(t)$, the logarithm of evolution speed $\log[V(t)] = \log\left[\bar{k}(t) - \bar{k}(t+1)\right]$, the normalized average variance $w^2(t)/\bar{k}(t)$, and the centered wave profile $\phi(x) = n(k,t)/N$, where $x = k - \text{round}\left[\bar{k}(t)\right]$ were calculated. The last three values were averaged over the time interval where $\bar{k}_0 < \bar{k}(t) < 1.2\,\bar{k}_0$, where $\bar{k}_0$ denotes the allele number of sampling, and then over $10 - 40$ random computer runs with a different initial seed of random generator (Figure 3.3). The combined averaging over time and realizations ensured a small statistical error for the estimate of $\langle p\rangle$ (see vertical bars in Figure 3.3). The analytic shape of the wave, eq. (3.15), was used as the initial condition to minimize transitional time to the quasi-stationary simulated wave. Choosing the initial wave center anywhere within $(5 - 10)\,k_0$, one suppresses the residual effect of the choice of the initial condition below the statistical error (Rouzine and Coffin, 2005).

Figure 3.4 shows some examples of simulated waves $f(k, t)$. The log-averaged adaptation rate $V_{\text{av}} = e^{\langle \log V\rangle}$ and the wave width square $w_{\text{av}}^2 = w^2/\bar{k}_0$, scaled to their respective independent-loci values, $s\bar{k}_0$ and $\bar{k}_0$, are shown in Figure 3.3 as a function

of N and r. In accordance with the fundamental Fisher Theorem, eq. (3.11), the normalized values of $w_{\text{av}}^2/\bar{k}_0$ and $V_{\text{av}}/(s\bar{k}_0)$ are very similar; they are also fairly close to the analytic prediction, including The predicted critical point in Nr, where the evolution stops. The analytic dependence V versus N, eq. (3.5), is reproduced with a sufficient accuracy to be practically useful. The analytic theory somewhat underestimates the simulation result due, probably, to the continuous-in-k approximation used in analytic formula (similar error was observed in the asexual model in Chapter 2). At very large recombination parameters, $r \sim s\sqrt{k_0}$ and larger, the above analytic theory does not apply. Instead, we observe a steep rise in V from 0 to 1 when population size crosses point $N \approx 30/r$ (Figure 3.3).

To conclude, we have derived an expression for the substitution rate of advantageous mutations in the case of moderate-term evolution, when the initial population has standing variation at multiple loci, and mutation can be neglected. Based on our findings, we predict that even very small recombination alone is more efficient for adaptation than mutation alone.

An important limitation of the model we considered is that it neglects with correlation between genomes due to common ancestry of some loci. An appropriate technique taking this factor into account will be presented in Section 3.4. Also, when beneficial alleles do not pre-exist in a population, or in the long term evolution, both mutation and recombination are essential, and another approach has to be used. Such an approach will be described in Section 3.3.

3.2.5 Approximations used

Here we will explain some approximations made in Section 3.2.2.

Note 1

In eq. (3.3), we assumed that, for all relevant k, $s|k-\bar{k}| \ll 1$. Because the low-fitness tail of distribution at large k is not important, we need to check this condition only at the high-fit end, $k-\bar{k} = x_0 < 0$. Using eq. (3.24) for x_0, we obtain the validity condition,

$$\log(Nr) \ll 1/(s^2\bar{k}) \tag{3.25}$$

Note 2

We assumed in eqs. (3.2) and (3.3) that $f(k,t)$ can be replaced with a function continuous in t, which implies $V|d\log\phi/dx| \ll 1$. At negative x, $|d\log\phi/dx|$ reaches its maximum at $x = x_0$, eq. (3.15), where we have

$$V\left|\frac{d\log\phi}{dx}\right|_{x_0} \approx ps\bar{k}\frac{|x_0|}{w^2} \sim \sqrt{\bar{k}\log(Nr)} \tag{3.26}$$

where we used eq. (3.5) for V and w and eq. (3.24) for x_0. We arrive at the validity condition in eq. (3.25).

Note 3

When deriving eq. (3.5), the traveling wave was assumed to be narrow compared to its distance from the point $k = 0$, as given by $|x_0| \ll \bar{k}$. Using eq. (3.24) for $|x_0|$, the validity condition becomes

$$\log(Nr) \ll \bar{k} \tag{3.27}$$

Note 4

When calculating $\partial f/dt$ in eq. (3.3) to obtain Eq. (3.7), we neglected the implicit dependence of ϕ on t. This action is justified, if

$$\left|\frac{dw}{dt}\frac{\partial\phi}{\partial w}\right| \ll \left|V\frac{\partial\phi}{\partial x}\right| \tag{3.28}$$

From eqs. (3.20) for w^2 and (3.5) for p, we get

$$\frac{d(w^2)}{dt} = pV + \bar{k}\frac{dp}{dt} \approx pV \tag{3.29}$$

From eq. (3.15), we have

$$\frac{\partial\phi}{\partial x} = -\frac{x}{w^2}\phi, \quad \frac{\partial\phi}{\partial w} = -\frac{2x^2}{w^3}\phi \tag{3.30}$$

Substituting eqs. (3.29) and (3.30) into eq. (3.28) and using $p = w^2/\bar{k}$, eq. (3.28) a form $|x| \ll \bar{k}$. It is sufficient that this condition is valid at $x = x_0$, which is equivalent to the condition that the wave is narrow, eq. (3.27).

Note 5

When deriving eq. (3.15), we assumed that $|x_0| \gg w$, that is, the lead is longer than the width of distribution. Using eqs. (3.18) and (3.5), we have

$$|x_0|/w \sim \log^{1/2}\left[(s/r)\sqrt{\bar{k}(1-p)}\right]$$

which is much larger than 1, if

$$1 - p \gg (r/s)^2/\bar{k}.$$

Using eq. (3.5) for p, the validity condition takes a form

$$r \ll s\sqrt{\bar{k}}, \quad \log(Nr) \ll \left(\frac{s\sqrt{\bar{k}}}{r}\right)^2 \log\frac{s\sqrt{\bar{k}}}{r} \tag{3.31}$$

Over most of the interval $|x| < |x_0|$, the integral in x' in eqs. (3.9), (3.16), and (3.19) was assumed to be contributed from a narrow interval, $x' \approx x_0$. To conform it, at $|x_0| - |x| \gg \delta x$, where

$$\delta x \sim p\bar{k}/[(1-p)|x_0|]$$

the integral in eq. (3.19) is mostly contributed from a region $x' - x_0 \sim \delta x$. Using eq. (3.18), we have $\delta x/|x_0| \sim (w/x_0)^2 \ll 1$, which, again, yields the conditions in eq. (3.31).

3.3 Stationary evolution with recombination and mutation

In the previous section, we considered the case of evolution with pre-existing genetic variation and no mutation. Here, we will consider another scenario (Neher et al., 2010), where beneficial alleles are introduced all the time by new mutation events and form a stationary traveling wave, acting in accord with recombination. The stationary process is established after a sufficiently long time when initial diversity and recombination are no longer sufficient to maintain steady evolutionary process and stationary cooperation between new mutation and recombination is required.

When deleterious mutations can be neglected, which is the case in a sufficiently large population and not too close to fitness maximum (Chapter 2), the rate of adaptation is the product of the total rate of beneficial mutations NU_b, the magnitude of their beneficial effect in fitness, s, and their fixation probability. As in Chapter 2, we assume that the fixation probability is the probability that the allele becomes established, that is, that its lineage grows to sufficiently high levels in a population and will not become extinct due to stochastic fluctuations. In a population where all individuals have exactly the same fitness, a beneficial mutation with selective advantage s has the probability of establishment, $P_e \approx s$ and $2s$ in the continuous and discrete generation models, respectively (Moran, 1958). In a population with fitness variance, however, a new beneficial mutation can occur on different genetic backgrounds, so that its establishment probability will be greater if the genetic background is better-fit. But even highly-fit genotypes are soon outcompeted by the other beneficial mutation clones (clonal interference, Chapter 2). To avoid the fate of extinction, the descendants of the allele have to jump to higher backgrounds by the means of recombination (Rice, 2002) (Figure 3.5). As a result, the probability of survival for a new lineage decreases as the average adaptation rate increases, because it is harder for the moving allele to keep up with evolving population. At the same time, the speed of adaptation

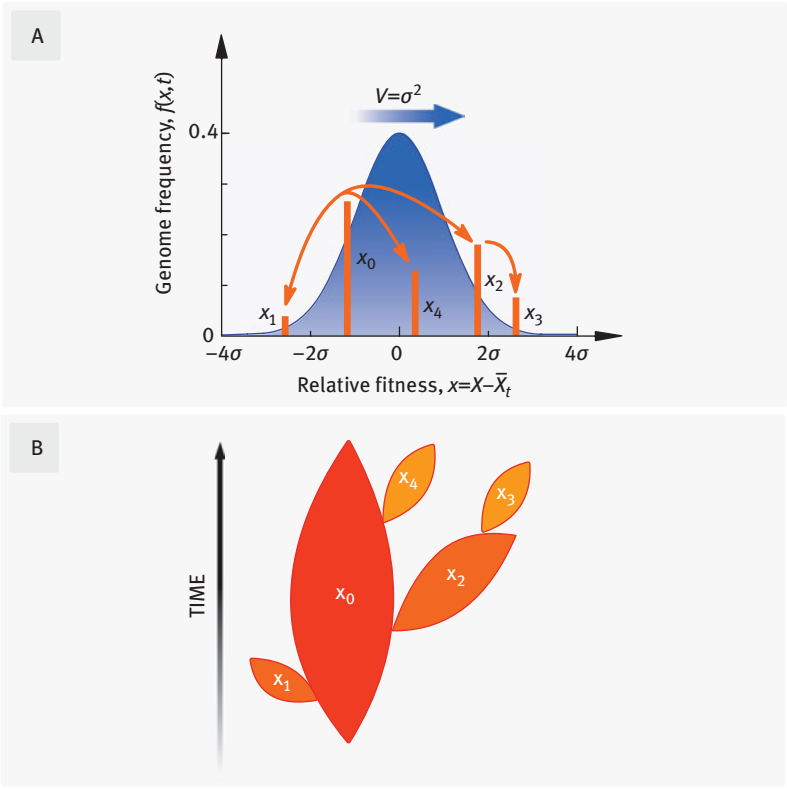

Figure 3.5: To become established, a new beneficial allele has to keep recombining with a fitter genetic background. (A) Fitness distribution of the population $f(x,t)$(blue) moves toward higher fitness values X with velocity $v=\sigma^2$. The new mutation born into a genome with fitness X_0 must jump between genomes (orange bars) to keep up with the moving fitness distribution. The mutation is established if at least one lineage survives indefinitely. (B) Diagram of clones for the case when mutation becomes extinct, that is, all lineages die out. The probability of establishment, $w(X,t)$, depends on the fitness X of the background genome and is found from eq. (3.34) (based on Neher et al. (2010)).

(the speed of fitness gain) is limited by the establishment probability of new alleles. Therefore, the speed and the establishment probability have to be determined from two self-consistent conditions (Neher et al., 2010), as described further.

3.3.1 Model and approach

As in the previous sections, we consider a population of N haploid individuals, each with its own fitness X. We assume that each mutation makes the same small change in the fitness, s. If a genome is viewed as the best-fit sequence with k deleterious

alleles, as we did in Section 3.2, we can write $X = -sk$. However, what follows does not depend on the choice of the reference sequence, hence, we will use the later notation X rather than k. As in the previous sections, we assume that a wide spectrum of genome fitness values is present, characterized by the fitness variance σ^2 of the population related to the variance of mutation load, w^2, as given by $\sigma^2 = (sw)^2$. The number of progeny per individual is random and obeys Poisson distribution with the average rate $1 + X - \bar{X}(t)$, where $\bar{X}(t)$ is the mean fitness in the population, and $X - \bar{X}(t) \ll 1$. The death rate is set to one. In addition to asexual reproduction, an individual can recombine with a probability r with another individual. As previously, we assume that the number of crossovers sites M is large, $M \gg 1$. We also assume, that the progeny fitness distribution obeys a Gaussian function centered at the average of the fitness values of the two parents and has variance $\sigma^2/2$ (Bulmer, 1980).

We must warn the reader that this model of recombination, used in the original paper (Neher et al., 2010), implicitly assumes that sequences are very strongly correlated and underestimates the typical fitness difference between progeny and parents. In the previous section, we studied an opposite case and assumed that sequences are not correlated at all. Hence, the variance of progeny fitness was much larger, $s\bar{X}/2$, where $\bar{X} = -s\bar{k}$ and $\bar{k}$ is average number of deleterious alleles at variable (segregating) loci. Our model in Section 3.2 overestimated the effect of recombination. An intermediate model describing sequence correlations more realistically than these two extreme approximations will be considered in Section 3.4 (in the absence of mutation).

We note that σ^2 is proportional to the diverse allele number. Hence, σ^2 reflects the magnitude genetic variation in the adapting population. It is not a fixed parameter of the model, but is calculated self-consistently later, as a function of model parameters.

The recombination process is characterized by outcrossing probability r and the distribution of offspring fitness Y, given that a parent with fitness X mated with a random member of the population, denoted as $K(X, Y, t)$. This function is connected to our recombination generator, $R(k, t)$ in Section 3.2, as

$$R(Y, t) = \int dX\, f(X, t) K(X, Y, t)$$

Because it is the distribution of offspring fitness, the recombination "kernel" is normalized as $1 = \int dX\, K(X, Y, t)$.

Further we assume that different loci are in linkage equilibrium, so that recombination does not change the fitness distribution $f(X, t)$, as given by

$$f(Y, t) = \int dX\, f(X, t) K(X, Y, t) \tag{3.32}$$

which is equivalent to $R(k,t) = f(k,t)$ in eq. (3.3). This corresponds to the case $p = 1$ in Section 3.2. Indeed, in the stationary case with mutation, as we show later, the speed of the wave is limited by the establishment of new mutations and not by global linkage disequilibrium, $p < 1$, as it was in the standing-variation model, eq. (3.3).

According to our recombination model, during mating, two parents are replaced with two offspring. However, there is no need to follow both offspring, because a rare allele jumps only to one. Hence, we can focus on that lineage alone and ignore the other offspring. Mating between two individuals with the same rare allele is very unlikely and can be safely neglected. Also, as in the previous model, we assume that recombination is not too infrequent, $r \gg s$.

3.3.1.1 Branching process and establishment probability

The probability that a new beneficial allele survives and establishes in the population is the limiting factor of adaptation. This rare establishment occurs, and the fate of allele is sealed, when the number of copies of the new allele is much larger than unit but much smaller than the total population size, N. In this regime, the stochastic dynamics of the new lineage can be described by a branching process that includes the factors of stochastic birth, death, and random recombination events that can move the allele between genomes.

The stochastic dynamics of the lineage is heavily influenced by the presence of average fitness constantly increasing because of beneficial alleles fixing at the other loci. The mean fitness, $\bar{X}(t)$, increases with rate $v = d\bar{X}(t)/dt = \sigma^2$, where σ^2 is the fitness variance. The trajectory of a beneficial allelic lineage between genomic backgrounds in an adapting population is depicted in Figure 3.5. To be established, the allele has to keep jumping to better-fit backgrounds (Rice, 2002).

The establishment probability at time $t - dt$ of descendants of a genome of fitness X, defined as $w(X, t - dt)$, is related to the establishment probability at later time t:

$$\begin{aligned} w(X,t-dt) = w(X,t) &- dt[1 + B(X,t) + r]w(X,t) \\ &+ dt\, B(X,t)\left[1 - (1 - w(X,t))^2\right] \\ &+ dt\, r \int dy\; K(X,Y,t)w(Y,t) \end{aligned} \tag{3.33}$$

(Barton, 1995), where $B(X) = 1 + X - \bar{X}$ is the birth rate, and we included death rate 1. After a division, each of the two offspring has probability $1 - w$ of extinction. Hence, the probability that, at least, one of these offspring will be fixed is $1 - (1 - w(X,t))^2$, see the second line in eq. (3.33).

If an allele with fitness effect s is added by mutation on a genomic background with fitness X, $B(X)$ above is replaced with

$$B(X) = 1 + X - \bar{X}(t) + s$$

The adaptation process represents a traveling wave with the velocity $v = d\bar{X}/dt$. As in Section 3.2, we assume that the profile of the fitness distribution $f(X,t)$ around mean $\bar{X}$ does not fluctuate much between realizations, and that that the distribution has a Gaussian shape, as in Section 3.2. We will not need to consider the edge of $f(X,t)$ this time, since the stochastic edge will naturally enter through establishment probability, $w(x,t)$.

In a traveling-wave solution, $w(x,t)$ depends on time only through $\bar{X}(t)$, and we center fitness at $\bar{X}(t) = vt + \text{const}$, introducing $x \equiv X - \bar{X}(t)$, $B = 1 + x + s$, and seek a solution of the form

$$w(x) = w(X - vt) = w(x,t)$$

In these terms, eq. (3.33) simplifies to

$$v\frac{dw}{dx} = (x + s - r)w(x) - (1 + x + s)w^2(x) + r\int dy\ K(x,y)w(y) \tag{3.34}$$

At small fitness differences, $s \ll 1$, selection is important only on timescales much longer than the time between generations, $t \gg 1$. In this case, term $x + s$ in the prefactor of the quadratic term is negligible, and eq. (3.34) simplifies to (Neher et al., 2010)

$$\begin{aligned}\Psi[w(x)] &\equiv \left(v\frac{d}{dx} - x + r\right)w(x) - r\int dy\ K(x,y)w(y) \\ &= sw(x) - w^2(x)\end{aligned} \tag{3.35}$$

Here we introduce linear operator Ψ which will be useful later on. Let us examine the limits of this expression at different outcrossing rates r. In the limit of very high r, we will obtain further that the left-hand side of eq. (3.35) becomes zero, and we have $w(x) = s$. This is the classical one-locus result in the absence of linkage in the genome. In the general case of intermediate r, the establishment probability of a new mutation that can arise in any individual, P_e, is the population average

$$P_e \approx \int dx\ f(x)w(x) \tag{3.36}$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-x^2/(2\sigma^2)}$$

Here $f(x)$ is the centered fitness distribution of genomes with $p = 1$ (see Section 3.2). In other words, the long-scale linkage of alleles that made the values of k in different genomes correlate and determined the dependence of speed in Nr in the

previous model, is assumed to be absent in this model, since the dependence is driven by the changes of establishment rate with Nr and is much sharper. Indeed, as will become later on, all the important changes in the speed of adaptation found below occur at $p \approx 1$.

Please note that the left-hand side of eq. (3.35) vanishes after the population averaging with the Gaussian factor $f(x)$, eq. (3.36), as is easy to verify by using $v = \sigma^2$ and eq. (3.32). This property originates from the above approximation that linkage disequilibrium is absent, eq. (3.32). Hence, after averaging eq. (3.35) over population, we get

$$\int dx\, f(x)\left[sw(x) - w^2(x)\right] = 0 \tag{3.37}$$

When combined with eq. (3.36), eq. (3.37) provides an alternative expression for the establishment probability:

$$P_e \approx \frac{1}{s}\int dx\, f(x) w^2(x) \tag{3.38}$$

Together with eq. (3.35), this equation describes the "surfing" of a beneficial allele between genomes illustrated in Figure 3.5.

The recombination kernel $K(x,y)$ depends on the recombination model (Neher et al., 2010). For the model we have chosen here, which is suitable for both viruses and dominance-free segments of genomes of higher organisms, the fitness of the progeny of two parents with fitness x and z is Gaussian distributed with the average of $(x+z)/2$ and the variance of $\sigma^2/2$. Fixing fitness x of one parent and averaging over fitness z of another, results in the recombination kernel

$$K(x,y) = \sqrt{\frac{2}{3\pi\sigma^2}} e^{-\frac{2\left[y-\frac{x}{2}\right]^2}{3\sigma^2}} \tag{3.39}$$

3.3.2 Main results

3.3.2.1 Establishment probability and the speed of adaption

In the following section, we solve eq. (3.35) and obtain the expression for the average fixation probability, P_e, of a beneficial mutation. Then, we use it to calculate σ^2 in a self-consistent manner. In the limit of interest, $r \gg s$, the final result has a form (Neher et al., 2010)

$$P_e = \frac{\sigma^3 \log(cr/s)}{sr\sqrt{2\pi}}\; e^{-\left(\sigma^2/2r^2\right)\log^2(cr/s)}, \quad s \ll r \ll \sigma \tag{3.40}$$

$$P_e = s\left(1 - \frac{4\sigma^2}{r^2} + \cdots\right), r \gg \sigma \tag{3.41}$$

where c is a numeric coefficient on the order of 1. This expression is proportional to the product of σ and a function of r/σ and $1/\sigma$. [Note that in the limit of very small s, $s \ll \exp(-cr^2/\sigma^2)$, the expression in eq. (3.40) breaks down. This is unlikely to be relevant in practice.]

In the limit of small r, the fixation probability, eq. (3.42), decreases very rapidly (exponentially) with decreasing r. This is because mutations occurring in individuals located in the high-fitness tail of the distribution have an exponentially larger probability of being fixed than mutations in the most of the fitness distribution; such rare mutations dominate the average fixation probability in the population. At large r, eq. 3.41, the initial genetic background plays only a minor role, since the allele hops quickly between genomes as compared to the speed of the wave. New alleles emerging in any fitness background have a high chance of fixing. Therefore, for large r, the result for P_e is close to the one-site result: $P_e \approx s$.

The expressions for P_e presented above depend on the variance in fitness σ^2. At a small mutation rate, according to the Fischer theorem, the variance is equal to the adaptation speed, v. The latter, in turn, is given by the product of the rate at which new beneficial mutations alleles enter the population, U_b, their fitness effect s, and the fixation probability $P_e(\sigma)$, as given by

$$v = \sigma^2 = NU_b s P_e(\sigma) \tag{3.42}$$

The adaptation rate can be obtained by solving for σ the self-consistency condition, eq. (3.42). Substituting our result for P_e from eqs. (3.40) and (3.41) into (3.42) and omitting logarithmic factors in the arguments of large logarithms, we find

$$v \approx \begin{cases} 2r^2 \dfrac{\log(NU_b)}{\log^2(r/s)}, & 1 \ll \dfrac{r}{s} \ll \sqrt{\dfrac{NU_b}{\log(NU_b)}} \\ NU_b s^2 \left(1 - \dfrac{4NU_b s^2}{r^2} + \cdots\right), & \dfrac{r}{s} \gg \sqrt{NU_b} \end{cases} \tag{3.43}$$

As in the previous model in Section 3.2, evolution rate v is not proportional to the total mutation rate in a population NU_b, at low r and sufficiently large population sizes, N. Rather, evolution rate is proportional to the logarithm of NU_b. This is the standard consequence of clonal interference of mutations arising at a large number of evolving loci. Therefore, as in the asexual case in Chapter 2, due to clonal interference, only a tiny fraction of the beneficial mutations are able to fix in a population. The tiny fixed fraction becomes larger as recombination rate increases. As a consequence, the adaptation rate increases as $r^2 \log(NU_b)$, until it crosses over to the limit of independent loci, $v = NU_b s^2$.

3.3.3 Computer simulation

In our analysis of the establishment probability of a beneficial mutation, we have made several approximations, starting from the recombination model. To compare the analytic results to simulations of a population of individual genomes, (Neher et al., 2010) used a discrete generation algorithm, as follows.

Each individual produces a Poisson-distributed number of offspring with average number $\exp(X - \bar{X})$. The population size, N', is kept approximately constant with an average of N by adjusting the overall rate of replication with factor $\alpha = (1 - N'/N)\log 2$. Each individual is represented by a binary string, where each bit represents one locus. Recombination is implemented, as follows. Each generation, offspring is randomly chosen to be asexual with probability $1 - r$ and sexual with probability r. The asexual part is passed to the next generation. The sexual individuals are matched in random pairs, and their loci are chosen randomly to produce haploid progeny. To optimize performance, whenever a locus became monomorphic because of fixation or loss of an allele, a mutation is introduced in a random individual. This technique allowed to keep as many polymorphic loci as possible. However, beneficial mutation rate per genome U_b becomes a variable quantity, which increases with L and decreases with r (Figure 3.6A, inset). The infusion rate of beneficial mutations, NU_b, is determined by measuring the average rate at which the new mutations are introduced, which is set in simulation to be the sum of the extinction and fixation rates.

The average probability of establishment, $P_e = \langle w \rangle$, is shown in Figure 3.6A as a function of the outcrossing rate, for different values of L (roughly proportional to NU_b, see the inset). It is small at small r but increases sharply at high r and saturates at $P_e = 2s$, the single-locus result for a discrete-time process. The increase of P_e starts at larger r for larger NU_b, which confirms our previous expectation that the limit of frequent recombination is reached when r much larger than s. The agreement between a numerical solution of the branching process equation, eq. (3.35), and the stochastic simulation improves at high NU_b, confirming that the approximations are valid, at least, for large populations.

The error of the analytic result is quite large at small r. A possible reason for the large error is that, as we mentioned earlier, the chosen approximation of recombination with offspring fitness variance $\sigma^2/2$ underestimates that variance. The algorithm does not use this approximation but faithfully recombines random segments of genomes. A more accurate treatment of recombination effect is considered in Section 3.4 (in the absence of mutation).

The establishment probability $w(x)$ of a new allele born into fixed background x obtained from simulations is compared to the predictions based on eq. (3.35) in Figure 3.6B. At very large recombination rates, $w(x)$ is weakly sensitive to the background fitness, so that all new alleles have the same, nonexponentially small probability ~s, to establish. With decreasing r/σ, however, the dependence of the fixation probability on x becomes stronger and stronger, and only new alleles generated on

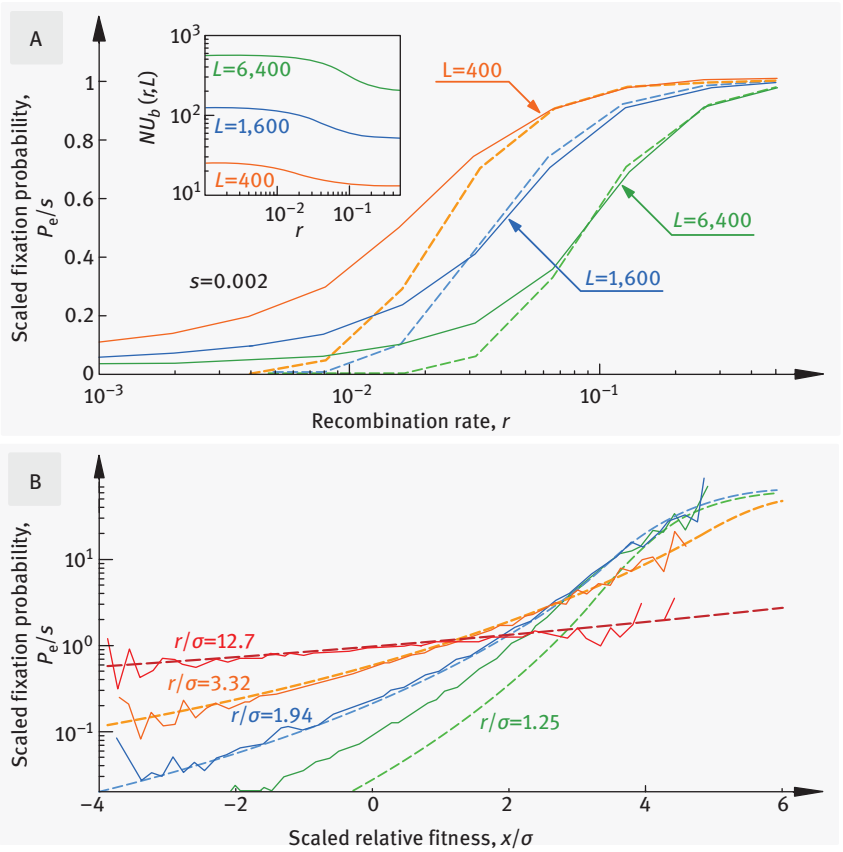

Figure 3.6: Fixation probability in a population with selection, recombination, and mutation. (A) The average fixation probability scaled to its value in the frequent-recombination limit, as a function of recombination rate per genome, r, for three different genome sizes L (shown). Parameters: $s = 0.002$, $N = 20,000$. Inset: The rate of beneficial mutations per genome per population, U_b, as a function of r (see main text). The scaled fixation probability in the simulation (solid lines) is calculated as $v/(2NU_bs^2)$ and compared to the analytic results for the scaled establishment probability P_e/s (dashed lines). (B) P_e/s as a function of the scaled relative fitness of genome, x/s. The solid lines are simulation results for $w(x)/(2s)$ for different values of r/σ (shown). Fixed parameters: $L = 6400$; $r = 0.512, 0.128, 0.064, 0.032$. The dashed lines are predictions for $w(x)/s$ obtained by numerical solution of eq. (3.35) (based on Neher et al. (2010)).

high fitness backgrounds have a chance. Note that at $r/\sigma \sim 1$, simulated $w(x)$ decays less rapidly at small x than predicted by eq. (3.35), pointing to the fact that the analytic model underestimates the effect of recombination. The likely reason for that and for the discrepancy at small N are very strong inter-sequence correlations built into the recombination model. The actual kernel is broader in progeny fitness.

3.3.4 Analysis of establishment probability

Here we derive the announced results in eqs. (3.40), (3.41), and (3.43). We have to solve eq. (3.35). Consider, first, the interval of moderate recombination rate, $s \ll r \ll \sigma$. We will analyze eq. (3.35) in different intervals x. At very large positive $x - r$, we can neglect with all terms but two, and the equation becomes $(x-r)w(x) \approx w^2(x)$ with solution $w_>(x) = x - r$ (Figure 3.7). In this regime, $w(x)$ does not need recombination jumps, but only the clonal growth reduced by r due to recombination from the clone. Establishment occurs by clonal expansion not dependent on the rest of population. In the opposite case of small x, $w(x)$ is so small that the quadratic term and $sw(x)$ can be neglected and eq. (3.35) becomes

$$\left(v\frac{d}{dx} - x + r\right) w_<(x) - r\int dy\ K(x,y)w(y) = 0$$

$$x < x_c = \sigma\theta \tag{3.44}$$

In this regime, the only way for an allele to become fixed is to recombine onto better backgrounds. We demonstrate further that the transition from $w_<(x)$to the saturated behavior at large $x, w_>(x)$, happens in a narrow interval around point $x = x_c \equiv \sigma\theta$,

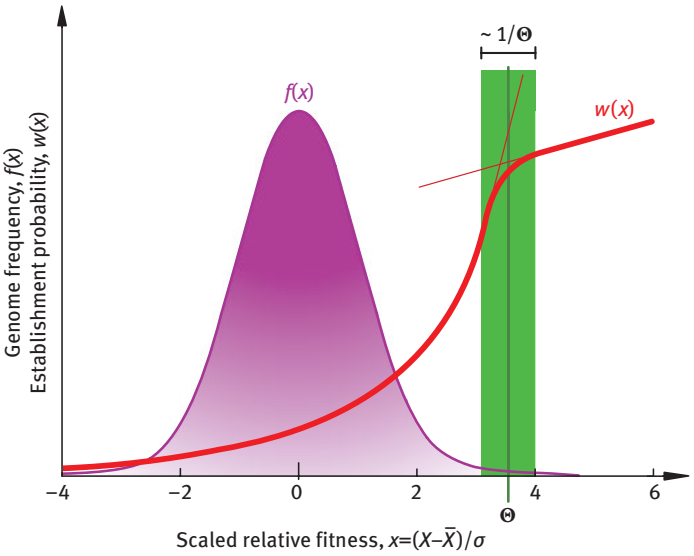

Figure 3.7: Schematic plot of the establishment probability, $w(x)$. Purple: the fitness distribution $f(x)$ of the population as a function of scaled relative fitness x/σ, where $\sigma = \sqrt{\mathrm{Var}[x]}$ is the distribution width. Red: establishment probability, $w(x)$, for $r \ll \sigma$. At small x, $w(x)$ is small and increases sharply proportionally to $e^{(x-r)^2/(2\sigma^2)}$. At larger x beyond a transition point, $\sigma\theta$, the quadratic term in eq. (3.35) becomes important, producing the one-locus result, $w(x) \approx x$. The width of the crossover region is of the order of σ/θ (based on Neher et al. (2010)).

which plays the role analogous to the leading edge of the wave discussed in the previous Sections, even though its biological meaning is different.

At intermediate x, $0 < x < \sigma\theta$, the establishment probability $w_<(x)$ increases steeply with x (but the quadratic term w^2 still stays negligibly small). Because individuals in this interval of x are much better-fit than the average individual, typically, recombination generates offspring, which are less-fit that the parents. Therefore, $w_<(x)$ is dominated by the first term in eq. (3.44), and the recombination term in eq. (3.44) is a small correction. The solution of eq. (3.44) is a Gaussian

$$w_<(x) = \phi(x)\ e^{(x-r)^2/(2\sigma^2)}$$

where $\phi(x)$ is a slowly varying function.

Note that the amplitude of $w_<(x)$is left undetermined by eq. (3.44), because it is uniform. Hence, the location $x_c = \sigma\theta$ of the crossover is not determined either. To determine these values, and to make sure that $w(x)$ satisfies the entire eq. (3.35), we will use eq. (3.37) as an additional constraint. This equation involves two first momenta of $w(x)$ averaged with fitness distribution $f(x)$. The first moment is dominated by intermediate x, because $f(x)w(x)$ decreases with x. The second is mostly contributed from a narrow peak in x at the crossover point $x = \sigma\theta$. The condition (3.37) then becomes a relation between P_e and θ

$$sP_e \approx \frac{\sigma\theta}{\sqrt{2\pi}} e^{-\theta^2/2} \tag{3.45}$$

It is convenient to rescale the variables as

$$x' = \frac{x}{\sigma},\quad r' = \frac{r}{s},\quad s' = \frac{s}{\sigma},\quad w' = \frac{w}{\sigma} \tag{3.46}$$

and use a linear transform

$$\Omega(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx' e^{-(x'-z)^2/2} w'(x') \tag{3.47}$$

In this notation, the fixation probability can be represented as $P_e = \sigma\Omega(0)$. By using the transform on eq. (3.35) we obtain an equation for $\Omega(z)$

$$\Lambda[\Omega(z)] \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx' e^{-\frac{(x'-z)^2}{2}} \Psi\left[w'(x')\right] = s'\Omega(z) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx' e^{-\frac{(x'-z)^2}{2}} \left[w'(x')\right]^2 \tag{3.48}$$

The left-hand side defines a linear operator Λ acting on $\Omega(z)$, which will be found later, and operator Ψ defined in eq. (3.35). The integral of $\left[w'(x')\right]^2$ in eq. (3.48) is

dominated by a narrow peak near $x'_c = \theta$ and can be evaluated using approximation $w' \approx \theta$:

$$\Lambda[\Omega(z)] \approx s'\Omega(z) - \frac{\theta}{\sqrt{2\pi}} e^{-(\theta - z)^2} = s'\Omega(z) - s'\Omega(0) e^{\theta z - z^2/2} \tag{3.49}$$

were we used eq. (3.45) to obtain the second equation. The continuity condition that the solution $w_<(x')$ joins smoothly to $w_>(x')$ implies that its linear transform $\Omega(z)$ does not diverge at any z and is an analytic function of z. Now we need to find $\Lambda[\Omega(z)]$.

It is useful to recall, at this point, that each offspring contains, on the average, a half of each parental genome. The parent carrying the new allele pairs with a random member of the population, which is likely to have the average fitness. Hence, after recombination, the average fitness of the offspring is at a half distance from the average population fitness compared to that of the parent. As a result of this connection between parents and offspring, operator Λ links $\Omega(z)$ to $\Omega(z)/2$, as given by

$$\Lambda[\Omega(z)] = (r' - z)\Omega(z) - r'\Omega(z/2) = s'\Omega(z) - s'\Omega(0)\, e^{\theta z - z^2/2} \tag{3.50}$$

where, we remind, $P_e = \sigma\Omega(0)$. Because we consider the case $r' \ll 1$, we can also assume $z \ll 1$. Approximating $e^{-z^2/2} \approx 1$, we can expand $\Omega(z)$ into a power series $\Omega(z) = \sum_{n=0} \Omega_n z^n$, where coefficients Ω_n satisfy an equation

$$\frac{\Omega_n}{\Omega_0} = \prod_{k=1}^{n} \frac{1}{r' - s' - r'2^{-k}} - s' \sum_{j=1}^{n} \frac{\theta^j}{j!} \prod_{k=j}^{n} \frac{1}{r' - s' - r'2^{-k}} \tag{3.51}$$

The first term on the right-hand side contains consecutive factors that approach $1/(r' - s') \gg 1$ for large n, which creates divergence in n. Therefore, it has to be canceled by the second term. By dividing the second product in eq. (3.51) by the first product, the condition for convergence is that $\Omega_n (r' - s')^n \to 0$ for $n \to \infty$, takes a form

$$1 = s' \sum_{j=1}^{\infty} \frac{\theta^j}{j!} \prod_{k=1}^{j-1} (r' - s' - r'2^{-k}) \approx \frac{cs'}{r'} e^{\theta(r' - s')} \prod_{k=1}^{\infty} (1 - 2^{-k}) \tag{3.52}$$

with $c = \prod_{k=1}^{\infty} (1 - 2^{-k}) \sim 1$. The last approximate equality in eq. (3.52) is accurate when $s' \ll r'$ and hence $(r' - s')\theta \gg 1$. From eq. (3.52), for the rescaled crossover point we obtain

$$\theta \approx \frac{\log(cr'/s')}{r' - s'} \tag{3.53}$$

From eqs. (3.45) and (3.53), we arrive at the desired final expression for the population-averaged establishment probability

$$P_e = \sigma\ \Omega(0) = \sigma \sum_{n=0} \Omega_n = \sigma \frac{\log(cr/s)}{s(r-s)\sqrt{2\pi}} e^{-\frac{\log^2(cr/s)}{2(r-s)^2}} \tag{3.54}$$

3.4 Recombination, standing variation, and inbreeding

We return to the model with natural selection and recombination in the absence of mutation considered in Section 2. Our intent is to include the effect of inbreeding between related individuals causing homology between genomes. We will show in this section that, in the long term, inbreeding has a strong, adverse effect on adaptation.

3.4.1 Inbreeding slows down adaptation

As mentioned in Sections 3.2 and 3.3, the adaptation rate depends sensitively on the model of recombination. In Section 3.2, we have considered an overly optimistic model of recombination, where individual sequences with a given fitness are not correlated (related) at all, which is correct in the short term when you start from a diverse randomized population. In Section 3.3, we considered an overly pessimistic model of recombination, in which sequences are strongly correlated, and variance in offspring progeny is limited by the half variance of the genome fitness among individuals and obtained the speed below Monte Carlo results. The truth lies in between.

In the present section, we will develop a more realistic model of recombination by calculating explicitly the dynamics of inter-sequence correlations (Rouzine and Coffin 2007; 2010). Monte Carlo simulation (Gheorghiu-Svirschevski et al., 2007) predicted that correlations due to common ancestry may be very strong for the relevant range and, worse than that, they accumulate in time (Figure 3.8). The cause of correlations is multiple mating of genomes within the limited pool of alleles, that is, the effect of inbreeding. Correlations decrease the adaptation rate and cause progressive extinction of beneficial alleles. Further, we generalize the formalism of Section 3.2 to take into account the inbreeding effect. Using the model in Section 3.2 as a starting point, we will derive the evolution rate is a function of the current average fitness and the model parameters (N, r, s, L). We will also obtain the clonal structure of fitness classes and predict the fitness distribution of remote ancestors.

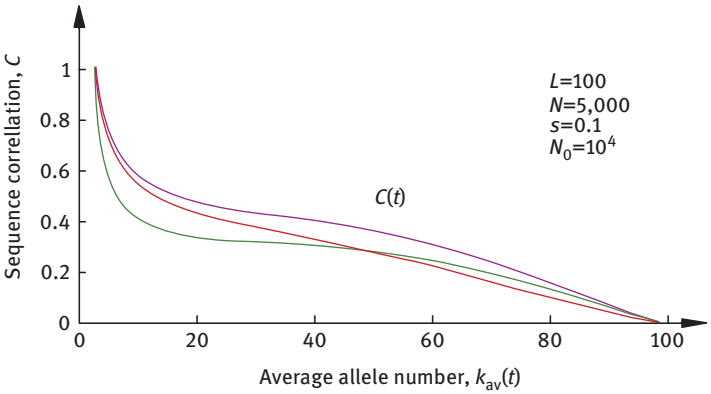

Figure 3.8: Monte-Carlo simulation with recombination of pre-existing variation and natural selection. Curves show the trajectory of the fraction of sites identical by descent C estimated in three independent ways against the average mutation load k (decreasing in time). The recombination rate per genome is assumed to scale as $r = N/N_0 = 0.5$, where $N_0 = 10^4$ is the characteristic population size. The initial population has random sequences with a small frequency of beneficial alleles at each site. Parameters are shown (based on Gheorghiu-Svirschevski et al. (2007)).

3.4.2 Model and approach

Our starting population model is similar to that in Section 3.2 (Figure 3.1). The deterministic bulk of fitness distribution is controlled by eqs. (3.3) and (3.4) or (3.7) and (3.8), which take into account the effects of selection and recombination in the absence of new mutations. These equations still predict a traveling wave with the high-fitness stochastic edge (Figure 3.2). However, here we will take into account phylogenetic relations between genomes which become important in the long-term (Figure 3.8).

First, we need to re-define the distribution of recombinant offspring in fitness. Suppose, two genomes with less-fit allele number k_1 and k_2, where k_1 and $k_2 \gg 1$, undergo recombination. In general, allele number k of a specific progeny genome is not a function of k_1 and k_2, because it depends on specific sequences and location of crossovers points. However, if k_1 and k_2 are large, $k_1 \gg 1$, $k_2 \gg 1$, and we have know something about the distribution of alleles within genomes and correlations between genomes, it is possible to make a statistical prediction about the distribution of progeny recombinants in k and, hence, describe adaptation.

In Section 3.2, we considered the simplest model where genomes are unrelated phylogenetically and alleles are distributed randomly and uniformly within each parental genome with given allele numbers, k_1 and k_2, and a random half of each

parent DNA goes to progeny. Then, the distribution of recombinant progeny over k is Gaussian with the peak at $k = \bar{k} \equiv (k_1 + k_2)/2$ and the variance $\overline{k^2 - \bar{k}^2} = d/2$, where

$$d = \bar{k}\left(1 - \bar{k}/L\right) \tag{3.55}$$

is the genetic half-distance between two random parental genomes. Equation (3.55) can be derived from the fact that, in this simple model, the contributions to k from each parents are statistically independent and obey a binomial distribution.

Now we need to generalize eq. (3.55) to account for genome relation causing allelic correlations. For this end, we align a pair of individual genomes and define a genome correlation measure, C, as the average fraction of sites that descend from the same ancestors in both genomes. By the definition, correlations are absent in the beginning of adaptation when all genomes are assumed to be unrelated, so that, $C = 0$ at $t = 0$. In the process of evolution, more pairs of homologous sites become related through common ancestors (Figure 3.8). New mutations in our new model are neglected. Therefore, all homologous sites with a common ancestor must carry alleles of the same type, the same as their ancestor. Therefore, such pairs do not contribute the genetic distance and we have

$$d = (1 - C)\bar{k}\left(1 - \bar{k}/L\right) \tag{3.56}$$

Such related pairs do not contribute to the effect of recombination on fitness variation of offspring.

When C becomes close 1 (which, as we show later, happens at small recombination rates in the end of adaptation), a significant fraction of loci will be related not only for pairs of genomes (sample size $n = 2$), but also across larger samples ($n > 2$) and, eventually, across the entire population ($n = N$). We denote the frequency of these completely correlated sites, at which the full population has a single ancestor, as C_{loss}. Such a all-correlated site is uniform in either the better-fit allele because the less-fit allele became extinct, or in the less-fit allele because the better-fit allele became lost. Monte-Carlo simulation (Gheorghiu-Svirschevski et al., 2007) in the parameter range relevant for viruses shows that the loss of beneficial alleles is much important in the traveling wave regime. The loss of less-fit alleles occurs only in the end when the traveling wave has already arrived at its final destination (Section 3.4.7). Therefore, we will neglect the loss of less-fit alleles.

Because the sites that lost beneficial alleles are uniformly and permanently less fit (we assume no new mutations), do not evolve, and do not contribute to recombination, we need to exclude them from consideration by replacing $L \to L - LC_{\text{loss}}$, $\bar{k} \to \bar{k} - LC_{\text{loss}}$. Then, the genetic half-distance in eq. (3.56) is generalized (Rouzine and Coffin 2007, 2010)

$$d = L(1 - C)q \tag{3.57}$$

$$q \equiv \frac{(f_1 - C_{\text{loss}})(1-f_1)}{1-C_{\text{loss}}} \tag{3.58}$$

where $f_1 = \bar{k}/L$ is the less-fit allele frequency. As it is easy to show, q cannot exceed $f_1(1-f_1)$. Thus, genetic distance d is decreased by the factor of $1-C$ due to pairwise correlations and by a factor of $q/[f_1(1-f_1)]$ due to N-wise correlations. We emphasize that in our model, inter-genome correlations affect derivation only through these two factors: in this model, correlations for samples of size larger than 2 but less than N are not important.

3.4.2.1 Including genomic correlations

By the definition, $C(t)$ is the chance that the most recent common ancestor for two homologous sites exist within the time of evolution t, as given by $T_{\text{MRCA}} < t$. Therefore, $C(t)$ monotonously increases in time as more pairs of genomes become related at an average site. We remind that initial sequences are not correlated, $C(0) = 0$. To describe dynamics of $C(t)$, it is convenient to introduce the density of coalescent events in time. Term "coalescent" refers to two lineages fusing to the common ancestor when moving back in time. We define the effective population size for genealogy, $N_{\text{anc}}(t)$, so that $1/N_{\text{anc}}(t)$ is the probability that two randomly sampled sequences at a random locus descend in the previous generation from a common ancestor. In the selectively neutral model, we would simply have $N_{\text{anc}}(t)$ equal to the population size, N (Kingman, 1982a, b). The factor of directional selection makes $N_{\text{anc}}(t)$ smaller. In these terms, the master equation for the dynamics of correlations $C(t)$ has a form

$$\frac{dC}{dt} = \frac{1-C(t)}{N_{\text{anc}}(t)} \tag{3.59}$$

In traveling wave regime, as we will show later, the coalescent density $1/N_{\text{anc}}(t)$ is a function of the current state of population only, that is, it can be expressed in terms of state variables and model parameters, but not on the initial conditions explicitly (Section 3.4.7). Specifically, $N_{\text{anc}}(t)$ will be expressed in terms of the current half-distance, $d(t)$, and the model parameters ($N, s, r,$ and L).

The average substitution rate $V = -d\bar{k}/dt$ is shown below to be approximately

$$V = psd \approx sd.$$

which is eq. (3.5) with $p \approx 1$ and $\bar{k}$ replaced with d. Therefore, eq. (3.59) can be rewritten as

$$\frac{dC}{df_1} \approx -\frac{L(1-C)}{sdN_{\text{anc}}} \tag{3.60}$$

We will show later that N_{anc} depends on time only through $d(t)$, which, in turn, can be expressed in terms of C, C_{loss}, and f_1, as given by eqs. (3.57) and (3.58). Therefore,

the right-hand side of eq. (3.60) can be expressed in terms of $C(f_1)$, $C_{\text{loss}}(f_1)$, f_1 and the constant model parameters.

We still lack an equation for $C_{\text{loss}}(t)$. The proper treatment would be difficult, and we take a shortcut. We will assume that C_{loss} can be expressed directly in terms of C alone using a relation following from a stationary neutral model. This nontrivial approximation is justified analytically in Section 3.4.7 and tested below in simulation. The dependence of C_{loss} on C can be conveniently represented by an interpolation formula (Figure 3.9)

$$C_{\text{loss}} \approx \exp[-2.53(1/C-1)] \tag{3.61}$$

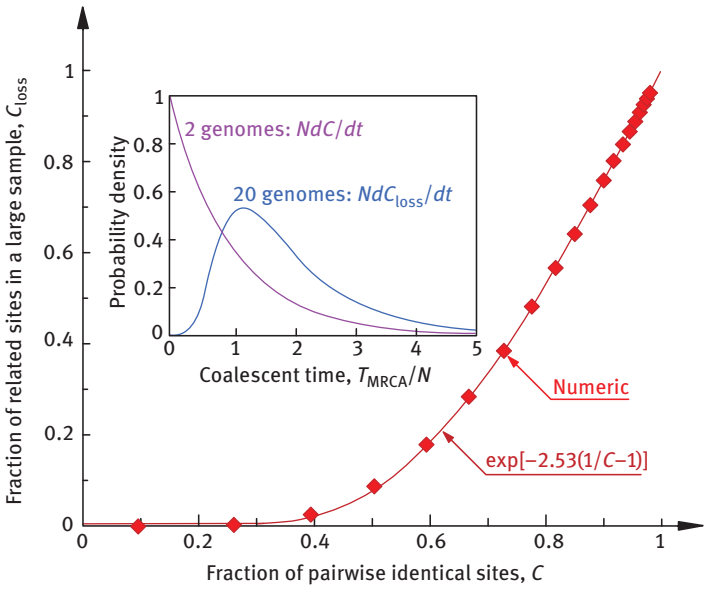

Figure 3.9: The average fractions of homologous sites with a common ancestor for a pair of individuals, C, and for a large sample, C_{loss}, as predicted by a selectively neutral model (Kingman, 1982b). Diamonds: The relation calculated numerically from the cumulative probability distribution of the time to the common recent ancestor for samples of 2 and 20 individuals. Smooth red line: Interpolation formula, eq. (3.61). Inset: Rescaled probability density of the coalescent time for samples of 2 (purple) and 20 (blue) sites, dC/dt and dC_{loss}/dt (based on Rouzine and Coffin (2010)).

Further, C_{loss}, C, N_{anc} and d will be found self-consistently, as a function of average allele frequency f_1.

3.4.3 Main results

The results below are valid in a range of model parameters, as follows. This conditions are explained as we go and summarized in Section 3.4.7: (i) To ensure the traveling wave being far from the best-fit possible genome, $k = 0$, or $|x_0| \ll k_{av}$, the total number of loci L and the average allele number k_{av} should be both much larger than $\log(Nr)$. (ii) At the same time, for a traveling wave to exist, the lead must be longer than the distribution width, $|x_0| \gg \sqrt{d}$, which takes place at large population sizes, $Ns \gg 1$ and $Nr \gg 1$. The first inequality is the condition of that selection is important in the one-site model (Chapter 1). (iii) Triple inequality $s \ll r \ll sL^{1/2} \ll \log^{1/2}(Nr)$.

The first inequality ensures that, in the main region of interest $r \sim [L/\log(Nr)]^{1/2}$, the total number of recombination events per population Nr is large. Please note that L in animals and plants is in millions, so that if s is not extremely small, the second inequality $r \ll sL^{1/2}$ is met even for $r = 1$, that is, a fully sexual population. Therefore, our analysis, although inspired by virus evolution, is relevant for organisms as well. The third inequality implies that the log-fitness difference between the best-fit and the average genome, $s|x_0|$, is smaller than 1.

As in Section 3.2, fitness distribution satisfies the wave equation (3.7), with $\overline{k}$ in eq. (3.8) replaced by d, and the solution has a form of traveling wave with speed V and shape $\phi(x)$. We will obtain asymptotically accurate solutions for two important overlapping intervals of the recombination rate (Rouzine and Coffin, 2007, 2010).

3.4.3.1 Small recombination rates

Analysis is especially simple if recombination rate is small, as given by $r \ll s[L/\log(Nr)]^{1/2}$. Then, the probability density of genomes with the less-fit allele number k has the form of traveling wave that represents a Gaussian

$$\phi(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi p d}} e^{-\frac{x^2}{2pd}}, & x > x_0 \\ 0, & x < x_0 \end{cases} \tag{3.62}$$

$$V = psd \tag{3.63}$$

where $p = V/sd$, and d given by eq. (3.57). Note that eq. (3.62) is a straightforward generalization of eq. (3.15) with variance $w^2 = pd$. As in Section 3.2, the wave has a high-fitness edge, $x_0 < 0$. Here parameter p and the edge location, $x_0 < 0$, are to be determined from the same stochastic edge approximation as in Section 3.2. The result is the generalization of eqs. (3.5) and (3.24) for the parameter of large-scale linkage disequilibrium, p, and lead, x_0, to

$$p = \Lambda_1/(\Lambda_1 + 2\Lambda_2) \tag{3.64}$$

$$x_0^2 = d\frac{2\Lambda_1(\Lambda_1+\Lambda_2)}{\Lambda_1+2\Lambda_2} \tag{3.65}$$

Here we introduced two large logarithms by recursive relations

$$\Lambda_1 \equiv \log\frac{Nr(1+p)}{4p\sqrt{\pi\Lambda_1}} \gg 1 \tag{3.66}$$

$$\Lambda_2 \equiv \log\left(\frac{\Lambda_2}{\beta}\sqrt{\frac{2+2p}{p}}\right) \gg 1 \tag{3.67}$$

and an important dimensionless parameter β which we will refer to as "the clone decay parameter" given by

$$\beta \equiv \frac{r|x_0|}{V} = \frac{r|x_0|}{psd} \tag{3.68}$$

The composite parameter β, which features extensively in the story, represents the total number of recombination events per genome during the time in which the traveling wave moves by its lead, $|x_0|$.

These results are valid if the recombination rate is sufficiently small, $r \ll sw^2/|x_0|$, which is equivalent to condition $1-p \gg 1/\Lambda_1$ or to condition $\beta \ll 1$, which ensures that the logarithm Λ_2 is much larger than 1. The case of arbitrary β and larger r will be considered later as well. As we mentioned, we assume a large total number of recombination events per population per generation, $Nr \gg 1$, hence $\Lambda_1 \gg 1$. Note that eqs. (3.66) and (3.67) are equations for Λ_1 and Λ_2 which can be solved iteratively. Because the dependence of the right-hand sides on Λ_1 and Λ_2 is logarithmically slow, one or two iterations give a good accuracy.

The recombinant generator $\rho(x)$ function introduced in eq. (3.7) is also easily generalized from (3.17)

$$\rho(x) = \frac{1}{\sqrt{\pi(1+p)d}}e^{-\frac{x^2}{(1+p)d}} \tag{3.69}$$

Note that its width is large than that of the fitness distribution given by eq. (3.62) due to the condition $p<1$ and, unlike that distribution, does not have a cutoff at high finesses, since it is a continuous noninteger function (Figure 3.2).

Based on eq. (3.64), parameter p slowly monotonously increases with the population size, N, and the recombination rate, r, combined together into $\Lambda_1 \approx \log(Nr)$. When $\log(Nr)$ reaches $d(s/r)^2 \sim L(s/r)^2$, which value is much larger than 1 according to condition (iii), we have $\Lambda_2 \sim 1$ and p is close to 1, as given by $1-p \sim 1/\log(Nr)$. In this case, if inter-sequence correlations were absent, the substitution rate V would saturate at one-site result

$$V = V_{1\,\mathrm{site}} = s\,k_{\mathrm{av}}(1-k_{\mathrm{av}}/L), \quad \text{if} \quad C = C_{\mathrm{loss}} = 0 \tag{3.70}$$

At smaller $\ln(Nr)$, we have $p < 1$ and $V < V_{1\text{site}}$. Thus, factor p represents the adverse effect of global linkage, which is partly offset by recombination, in the case when site-site correlations are absent. Adaptation is slowed down compared to the deterministic limit due to a combined effect of limited population size and limited recombination rate, as follows. Firstly, due to finite population, the fitness distribution has a high-fitness cutoff. Indeed, the lead $|x_0|$ diverges with $\Lambda_1 \to \infty$, eq. (3.65). Secondly, finite generation rate of new recombinants at the edge limits the speed of the wave.

The relative roles of global linkage p and correlations C are easy to understand from Fisher Theorem, $V = s\mathrm{Var}[k]$, we already mentioned in previous sections. Our problem is a particular case of it. Indeed, from eqs. (3.62) and (3.63), we have

$$\mathrm{Var}[k] = pd, \quad V = spd$$

so the theorem is met. If we forget about the loss of alleles for a moment, $C_{\text{loss}} = 0$, we can write

$$\mathrm{Var}[k] = pd = p(1-C)[k_{\text{av}}(1-k_{\text{av}}/L)] \tag{3.71}$$

Factor $k_{\text{av}}(1-k_{\text{av}}/L)$ in eq. (3.71) corresponds to the binomial distribution in the independent-site limit, that is, for a very large recombination rate or a very large population size. Factors p and $1-C$ represent the compression of the traveling wave due to two different types of inter-genome correlations: factor p describes correlation in genome fitness $-sk$ arising due to genome competition and finite population. Because the fitness distribution has an edge with limited extension speed, selection squeezes the distribution of k against the edge. As a result, $\mathrm{Var}[k]$ connected to the adaptation speed is decreased even if the genetic distance d is not affected. In contrast, correlations of individual loci, represented by factor C, decrease $\mathrm{Var}[k]$ by decreasing the genetic distance. The loss of diverse sites, $C_{\text{loss}} > 0$, aggravates this effect. As we will show, in the long term, site-site correlations have a much stronger adverse effect on adaptation than global linkage has.

3.4.3.2 Large recombination rates

As we show later, all the interesting changes in correlation factor C and substitution rate V occur at the characteristic value of recombination rate, $r \sim sd/|x_0|$. At this point, we have $\beta \sim 1$, and parameter p is almost 1. In this region, the fitness distribution substantially deviates from the Gaussian in eq. (3.62). Hence, here we need to use another approximation, which does not assume $r \ll sd/|x_0|$, but instead treats $1-p \ll 1$ as a small parameter. In this approximation, the fitness distribution $\phi(x)$ has a form (see the derivation in Section 3.4.6)

$$\phi(x) = \frac{1}{\sqrt{2\pi d}} e^{-\frac{x^2}{2d} - \varepsilon_\beta h_\beta(u)}, \quad x > x_0 \tag{3.72}$$

$$\rho(x) = \frac{1}{\sqrt{2\pi d}} e^{-\frac{x^2}{2d} - 2\varepsilon_\beta h_\beta(u/2)} \tag{3.73}$$
$$\varepsilon_\beta = \left(x_0^2/d\right)(1-p)$$
$$u = x/|x_0|$$

In the exponential in eq. (3.72), the term $x^2/(2d)$ corresponds to the infinite population (recombination rate) limit, $p=1$. The term $\varepsilon_\beta h_\beta(u)$ is not quadratic in u and, hence, creates a non-Gaussian correction to the traveling wave. The values of $\varepsilon_\beta, h_\beta(u)$ are determined by parameter β and will be calculated in Section 3.4.6. The lead of the distribution $|x_0|$ is given by

$$x_0^2 = 2d \cdot \left[\Lambda_1 - 2\varepsilon_\beta h_\beta(-1/2)\right]$$
$$\Lambda_1 = \log \frac{Nr}{2\sqrt{\pi\Lambda_1}} \gg 1 \tag{3.74}$$

The adaptation rate is given by

$$V = sd\left(1 - d\varepsilon_\beta/x_0^2\right)$$

where the second term in parentheses is relatively small, $O(1-p)$. Thus, the decrease of adaptation rate V below the single-site model value, eq. (3.70), is mostly due to the decrease in genetic distance d caused by correlations calculated further.

In the limit of small recombination rates considered before, $\beta \ll 1$, we have

$$h_\beta(u) = u^2/2 + (\text{small term diverging at } u = -1)$$
$$\varepsilon_\beta = 4\Lambda_2 \tag{3.75}$$
$$\Lambda_2 = \log(2\Lambda_2/\beta), \quad \beta \ll 1$$

which confirms eq. (3.62) obtained in that limit. In the opposite limit $\beta \gg 1$, ε_β is exponentially small.

3.4.3.3 Genealogical properties

In order to calculate the dynamics of correlations due to phylogenetic relations, we need to consider genealogy. We present an expression for the effective population size, N_{anc}, which determines the density of coalescent events in time, $1/N_{\text{anc}}$. Consider two homologous sites in two randomly sampled genomes in current generation t. We introduce $1/N_{\text{anc}}$ as the probability of their two respective lineages in the past cross, by an accident, in a selected generation far ago. It is given by

$$1/N_{\text{anc}} = \int_{x_0}^{\infty} dx\, \varphi^2(x) P_{cl}(x) \tag{3.76}$$

The factor $\varphi^2(x)$ in eq. (3.76) is the probability density that the two ancient ancestors of two sampled sites belong to the same fitness class x. The term $P_{\text{cl}}(x)$ is defined as the probability that two genomes in fitness class x also belong to the same clone of identical sequences. The accuracy of eq. (3.76), which implies that $1/N_{\text{anc}}$ depends locally on time on relevant timescales, is discussed in Section 3.4.7.

The fitness distribution of a remote ancestor of a site will be obtained in Section 3.4.6. Because that distribution is conditioned on leaving surviving progeny in the far future, it differs strongly from the current fitness distribution, $\phi(x)$. We will show that the ancestor distribution function can be rescaled as

$$\varphi(x) \equiv (1/|x_0|) y_\beta(x/|x_0|)$$

where function $y_\beta(x/|x_0|)$ depends on a single external parameter, β (see Figure 3.10 and Section 3.4.6).

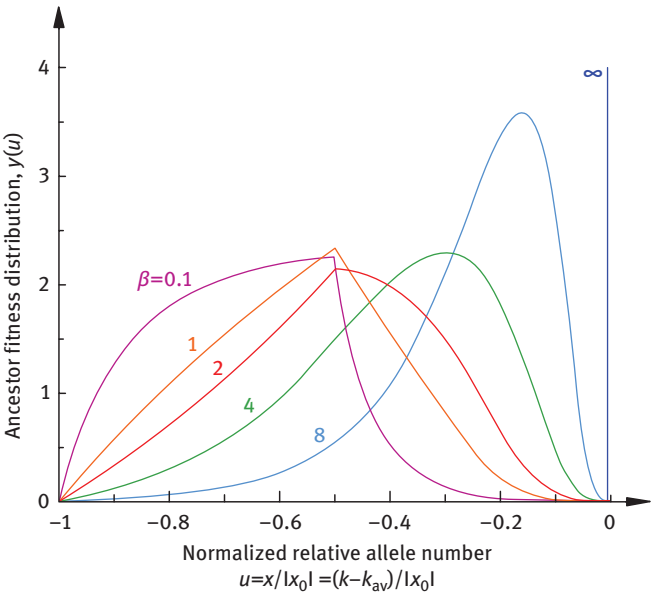

Figure 3.10: Fitness distribution of remote ancestors. Solid lines: Rescaled probability density $y_\beta(u)$ as a function of the scaled relative fitness of a remote ancestor, u. Values of β are shown on the curves. The results are obtained numerically from eqs. (3.110) and (3.112) (based on Rouzine and Coffin (2007)).

To derive $P_{\text{cl}}(x)$, it is not enough to know the fitness distribution. We must learn the fine clone structure of fitness classes, as it is done later on (Section 3.4.5). The result has a form

$$
\begin{aligned}
&P_{cl}(x) = \frac{1}{2\Lambda'_1} F_\beta(u) \\
&F_\beta(u) \equiv \beta \exp\left\{-\beta(1+u) + \varepsilon_\beta\left[(1-u^2)/2 + h_\beta(u) - 2h_\beta(-1/2)\right]\right\} \\
&\Lambda'_1 \equiv \log\left(Ns\sqrt{\Lambda'_1}\right) \gg 1 \\
&u \equiv x/|x_0|,\ -1 < u < 0
\end{aligned}
\tag{3.77}
$$

Substituting $P_{\mathrm{cl}}(x)$ from these equations into eq. (3.76) and using the rescaled form for the ancestor fitness distribution, $\varphi(x) = (1/|x_0|)y_\beta(x/|x_0|)$, we get

$$
\frac{1}{N_{\mathrm{anc}}} = \frac{1}{d^{1/2}(2\Lambda'_1)^{3/2}} \int_{-1}^{0} du\, y_\beta^2(u) F_\beta(u) \tag{3.78}
$$

Equation (3.78) represents a central result of this Section. It shows that the time to the most recent common ancestor of a site pair is a product of $d^{1/2}(2\Lambda'_1)^{3/2}$ and a universal function of the clone decay parameter, β.

In the limit of small or large β, we can obtain the asymptotic expressions for the effective population size in eq. (3.78) in the general form (Section 3.4.7). At $\beta \sim 1$, we calculated the integral in eq. (3.78) numerically based on eq. (3.77) for $F_\beta(u)$ and results for ε_β, $h_\beta(u)$, and $y_\beta(u)$ given below in Section 3.4.6. Between the two asymptotic limits, it can be approximated by an interpolation formula (Figure 3.11)

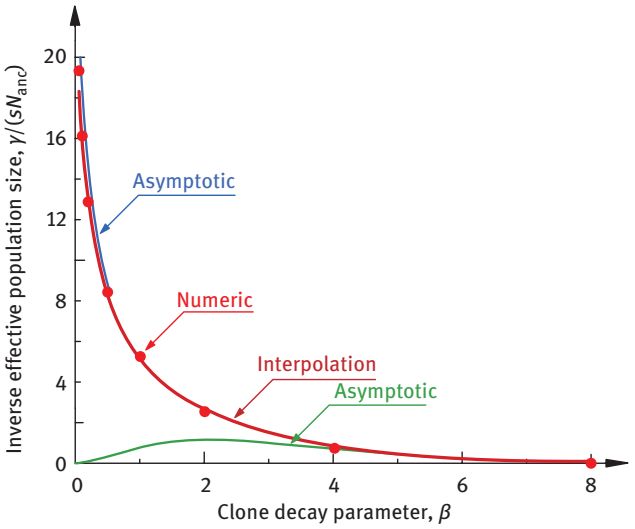

Figure 3.11: Dependence of the effective population size on the clone decay parameter β. Dots: Numerical results for $d^{1/2}(2\Lambda'_1)^{3/2}$ from eq. (3.78). Blue and green curves: the asymptotic at small and large β, eq. (3.127). Fat red line: interpolation formula, eq. (3.79) (based on Rouzine and Coffin (2010)).

$$\begin{aligned} \frac{1}{N_{\text{anc}}} &= \frac{\sqrt{2}}{d^{1/2}\Lambda_1'^{3/2}}\,\Lambda(\beta)e^{-\beta} \\ \Lambda(\beta) &\equiv \log\left(\frac{2\Lambda(\beta)}{\beta} + 15.0e^{0.53\beta^2}\right) \end{aligned} \tag{3.79}$$

This interpolation is asymptotically exact at very small and very large β and has the accuracy of 1% in the entire interval of β, tested by the numeric calculation of eq. (3.78) (Rouzine and Coffin, 2010). Thus, the density of coalescent events is a function of only two parameters: $d\Lambda_1^3$ and β.

Note that the coalescent timescale N_{anc} decreases monotonously with clone decay parameter β (Figure 3.11). This result is easy to understand intuitively, because β controls the clonal structure of a population studied in Section 3.4.5. At small β, a typical fitness class comprises one or a few large clones born at the high-fitness edge of a population, P_{cl} is relative large, so that the time to common ancestor N_{anc} is short. At relative large outcrossing rates, such that $\beta \gg 1$, a fitness class is broken into many small clones, P_{cl} is getting small, and the time to common ancestor becomes large.

3.4.4 Dynamics of inbreeding

Using eqs. (3.60) and (3.79), we can calculate the dynamics of the correlation parameter C. Quantities d, β, and N_{anc}, defined in eqs. (3.57), (3.68) and (3.79) can be expressed in terms of C, C_{loss} and f_1 as

$$d = L(1-C)q$$

$$\beta = \frac{\beta'}{\sqrt{(1-C)q}} \tag{3.80}$$

$$\frac{1}{N_{\text{anc}}} = \frac{s}{\gamma}\,\frac{\Lambda(\beta)e^{-\beta}}{\sqrt{(1-C)q}} \tag{3.81}$$

where C_{loss} is related to C by eq. (3.61), and the factor q depends on f_1 and C_{loss} as given by eq. (3.58).

We introduced in eqs. (3.80) and (3.81) two composite, constant model parameters

$$\beta' \equiv \frac{r\sqrt{2\Lambda_1}}{s\sqrt{L}} \tag{3.82}$$

$$\gamma \equiv s\sqrt{L\Lambda_1'^3/2} \tag{3.83}$$

where large logarithms Λ_1 and Λ_1' are defined in eqs. (3.66) and (3.77), respectively. Equations (3.80), (3.58) show that clone decay parameter β is a product of a rescaled recombination rate, β', and a universal function of correlations and allelic frequency. At the intermediate level of correlations in the middle of adaptation, $C \sim f_1 \sim 0.5$, we have $q \sim 1$, and parameters β and β' are of the same order of magnitude.

The composite model parameter γ characterizes the effective strength of selection for L sites, given the population size. It is analogous to product Ns in a one-site model (Chapter 1).

Substituting N_{anc} from eqs. (3.80) and (3.81) into eq. (3.60), we obtain the master equation for correlations caused by inbreeding (Rouzine and Coffin, 2010)

$$\frac{dC}{df_1} = -\frac{1}{N_{\text{anc}} sq} = -\frac{\Lambda[\beta(\mathrm{C}, f_1)] e^{-\beta(C, f_1)}}{\gamma\sqrt{(1-\mathrm{C})q^3(C, f_1)}}. \tag{3.84}$$

Note that the right-hand side of eq. (3.84) depends on C and f_1 and two constant parameters β' and γ. We assume the initial condition of a form $C(f_1 = f_0) = 0$, where $1 - f_0 \ll 1$ is the initial frequency of beneficial alleles. Specific choice of $1 - f_0$ has a minor effect on our results.

Equation (3.84) is a first-order nonlinear ordinary differential equation for $C(f_1)$, which is difficult to solve analytically. Numeric solution, at various scaled recombination rates β' and $\gamma = 10$, is shown in Figure 3.12. We observe that at (relatively) large recombination rates, correlations are weak, and the loss of beneficial alleles is very small as well. At small β', the effect of inbreeding is strong, and correlations accumulate to high levels (Figure 3.12A). Eventually, adaptation fails when deleterious allele frequency reaches the end-point $f_1 = f_{\text{end}}$ given by an equation

$$f_{\text{end}} = C_{\text{loss}}[f_1 = f_{\text{end}}] \tag{3.85}$$

The point $f_1 = f_{\text{end}}$ represents the final, minimum frequency of less-fit alleles, and the best recombination can do without new mutations. At this point, the population becomes so completely inbred that recombination becomes useless without adding new alleles. Hence we need to include new mutation, and we are back at the mutation-recombination-selection regime described in Section 3.3.

As recombination and hence β' decrease, the normalized substitution rate

$$V/(sL) \approx d/L = (1 - C)q \tag{3.86}$$

decreases and finally vanishes at the end-point. In addition, the dependence of the average adaptation rate V on f_1 deviates from the elliptical shape predicted by the single-site model (Figure 3.12B).

The clone decay parameter β given by eq. (3.80) has a flat minimum in both f_1 and β' and diverges at the ends of the interval, $f_1 = 1$ and $f_1 = f_{\text{end}}$ (Figure 3.12C). The divergence of β is due to a slowing speed of the traveling wave (formally, due to small values of q) in the beginning and the end of evolution. At a slow speed,

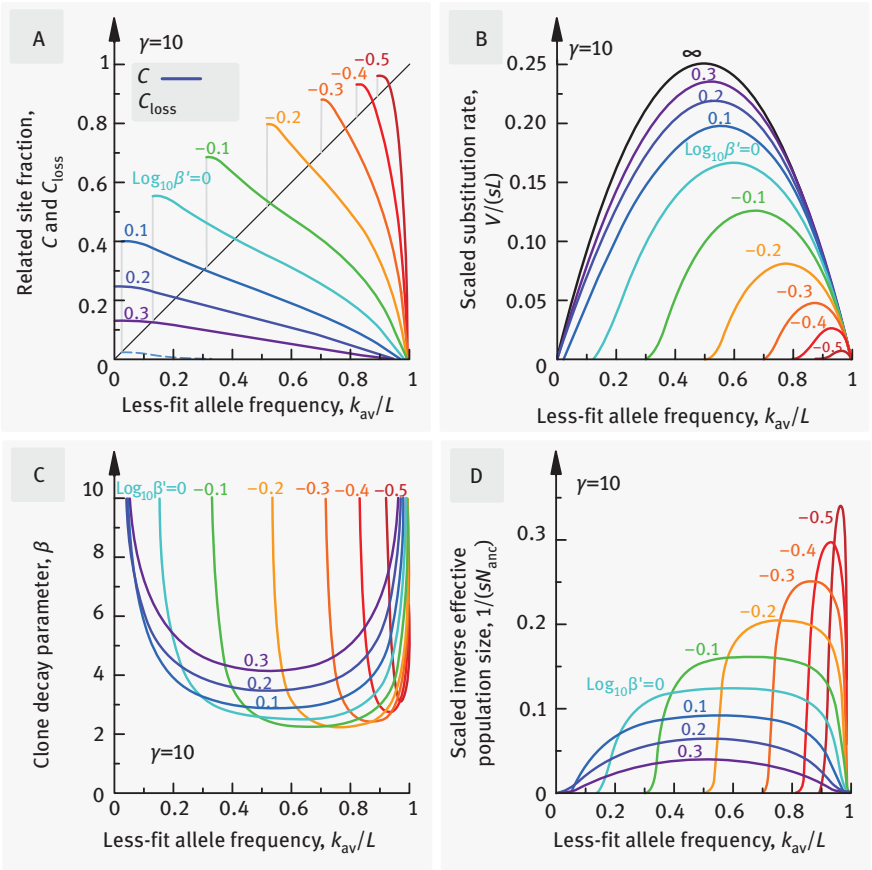

Figure 3.12: Inter-genome correlations, C, C_{loss}, the substitution rate, V, the clone decay parameter, β, and the effective population size, N, as the function of the fraction of less-fit alleles k_{av}/L. (A) Solid curves: Fraction of homologous sites related for an average genome pair, C, at different values of scaled recombination rate, eq. (3.82), $\log_{10}\beta'$ is shown on the curves. Dashed curves: Fraction of sites related across the population, C_{loss}. Vertical gray lines: End point of evolution. (B) Solid lines: Scaled substitution rate $V/(sL)$ as a function of k_{av}/L. Black line: The same value in the limit of infinite N or r. (C) Clone decay parameter β, eq. (3.80). (D) Scaled inverse effective population size for coalescent, $1/(sN_{\text{anc}})$. Scaled selection coefficient $\gamma \equiv s\left(L\Lambda_1'^3/2\right)^{1/2} = 10$. The results are obtained by solving eq. (3.84) numerically with the use of eqs. (3.58), (3.61), (3.80), and (3.79) (based on Rouzine and Coffin (2010)).

sequences have more time for recombination events. Consequently, the density of coalescent events $1/N_{anc}$, which is a function of β, has a flat maximum at intermediate f_1 and sharply declines toward the beginning and the end of adaptation, eq. (3.81) (Figure 3.12D). The flatness of the maximum is quite remarkable and supports our assumption about the quasi-neutral shape of the tree, which we used to connect C and C_{loss}.

These results are meaningful only at those values of f_1 and $N_{\text{anc}} \ll N$. The opposite inequality would imply that an ancestral clone consist of less than one individual. Practically, for correct evaluation of dynamics of correlations, the average $1/N_{\text{anc}}$ over time has to be much larger than $1/N$.

After the wave stops at the endpoint, it quickly collapses to a uniform population, and the value of C abruptly (compared to our timescales) jumps to 1. The rapid collapse to fully inbred population, evident in simulations, is beyond the scope of our theory.

3.4.4.1 Averaging adaptation time, end point, and the timescale of genealogy

Thus, both evolution speed V and genealogical population size N_{anc} depend on allele frequency f_1 (Figure 3.12B and D), which itself depends on time from near 1 to 0: $df_1/dt = -V/L$. Now we average both over the period of the adaptation process. A convenient measure is the normalized total time of adaptation

$$\frac{T}{T_{1\text{site}}} = \frac{T}{T_{1\text{site}}} \int_{f_{\text{end}}+1-f_0}^{f_0} df_1 \frac{L}{V} = \frac{1}{2|\log(1-f_0)|} \int_{f_{\text{end}}+1-f_0}^{f_0} \frac{df_1}{(1-C)q} \tag{3.87}$$

The upper limit of the integral f_0 is the initial frequency of less-fit alleles, $1-f_0 \ll 1$ being that of beneficial alleles. The low limit of the integral implies that beneficial alleles are fixed when their frequency averaged over the sites that did not lose alleles and completed adaptation is equal to $1-f_0$. Here

$$T_{1\text{site}} = T(C=0) = \frac{2}{s}\log\frac{1}{1-f_0} \tag{3.88}$$

is the value of T in the independent-site case, that is, in the limit of strong recombination. We notice that the integral in eq. (3.87) is mostly contributed from the two divergence regions near the ends of integration interval, for the relative increase of adaptation time due to linkage we obtain

$$\frac{T}{T_{1\text{site}}} \approx \frac{1-C(f_{\text{end}})/2}{1-C(f_{\text{end}})}, \tag{3.89}$$

where $C(f_{\text{end}})$ is the final value of C. Here we assumed $|\log(1-f_{\text{end}})| \ll |\log(1-f_0)|\,\beta'$, which condition is met when clonal decay parameter β' is not extremely small.

As recombination rate and hence parameter β' increase, the time of adaptation T decreases and eventually saturates at its independent-locus limit $T_{1\text{site}}$ (Figure 3.13A). However, at small recombination rates, adaptation is much slower, $T/T_{1\text{site}} \gg 1$. The half-point in β' depends on parameter γ; at $\gamma = 1$, it is nearly $\beta' = 1$, which corresponds to

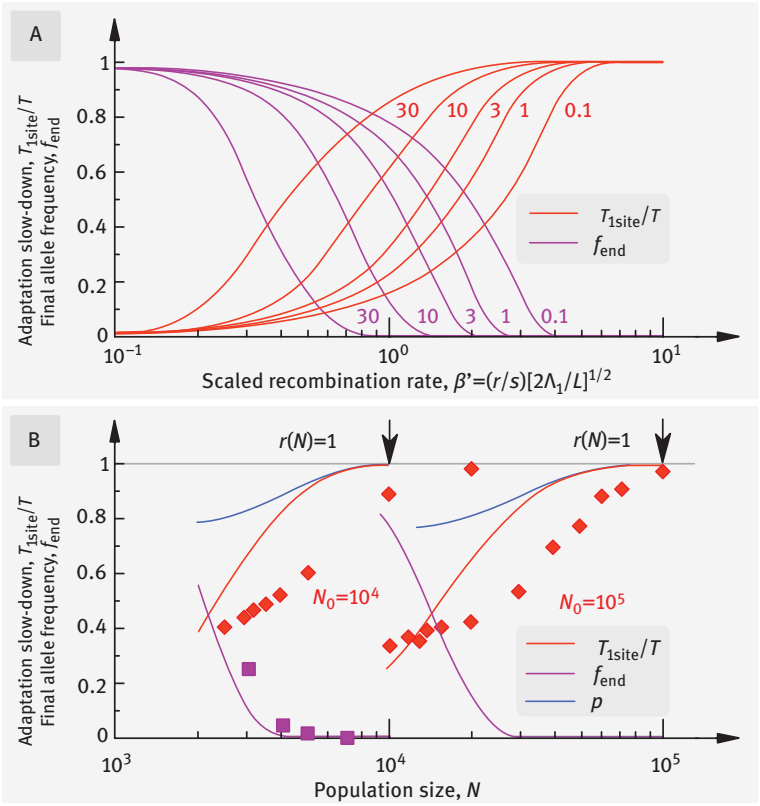

Figure 3.13: Adaptation slow-down and the fraction of sites that fail adaptation, f_{end} . (A) Inverse scaled adaptation time T_{1site}/T (red) and f_{end} (purple) as a function of scaled recombination rate β'. Values of scaled selection coefficient γ, eq. (3.83), are on the curves. (B) Purple curves: f_{end} as a function of the population size N for the case $r(N) = N/N_0$. Red curves: T_{1site}/T. Blue curves: values of $\langle p \rangle_{f_1}$ calculated from eqs. (3.92) and (3.111), numeric results for ε_β in Figure 3.17, and the dependence $\beta(f_1)$ in Figure 3.12. Open symbols: simulation results from Gheorghiu-Svirschevski et al. (2007). Parameters: $L = 100$, $s = 0.1$, two values of N_0 are shown on the curves (based on Rouzine and Coffin (2010)).

$$r = r_c = s\sqrt{\frac{L}{2\log(Nr)}}$$

This characteristic value of outcrossing rate r_c may be much larger or much smaller than the full sexual reproduction point, $r = 1$, depending on the relevant number of loci and the selection coefficient range.

For example, 740 million people living in Europe differ roughly in 0.1% of their DNA which corresponds to $3 \cdot 10^6$ polymorphic nucleotides. Assuming that 1% of these diverse loci, $L = 3 \cdot 10^4$, are neither selectively neutral nor under balancing selection but under directed selection, and that they can produce beneficial mutations with $s = 3\%$ or larger, we obtain that $r_c \gg 1$. Then the European population

with $r = 1$ is in the rare recombination regime ($\beta' < 1$). Conversely, if much fewer, only 0.01% of the diverse loci are under directed selection, then we have $r_c \approx 0.1$, so that the population with $r = 1$ is in the high-recombination limit ($\beta' \gg 1$). The exact estimate of L depends on selection conditions.

The total maximal gain in beneficial alleles at the end-point, $1 - f_{\text{end}}$, plotted as a function of β' roughly resembles the curve $T/T_{1\text{site}}$ (Figure 3.13A). As already mentioned, after time T has passed, the recombination-driven phase of evolution is over, and the evolution rate is limited by new mutations. The role of recombination is then reduced to assisting their fixation (Section 3.3.).

Now we average out over time the genealogy population size, N_{anc}. Integrating eq. (3.59) in time, we can express correlation parameter at the end of adaptation, $C(f_{\text{end}})$, in terms of the harmonic average in time, $\bar{N}_{\text{anc}}$, as given by

$$\begin{aligned} C(f_{\text{end}}) &= 1 - \exp\left[-T/\bar{N}_{\text{anc}}\right] \\ 1/\bar{N}_{\text{anc}} &\equiv \frac{1}{T}\int_0^T \frac{dt}{N_{\text{anc}(t)}} \end{aligned} \tag{3.90}$$

Note that eq. (3.90) has the form of the cumulative distribution function of the coalescent time, with $\bar{N}_{\text{anc}}$ analogous to the average time to the most recent common ancestor, $\langle T_{\text{MRCA}} \rangle$. In fact, the average coalescent time is not well-defined in our case, because a population is not described at $t < 0$, and some homologous site pairs never get common ancestors, as given by $C(f_{\text{end}}) < 1$. Yet, in the absence of better options, we can interpret the harmonic average in time, $\bar{N}_{\text{anc}}$, as an analogue of $\langle T_{\text{MRCA}} \rangle$.

It is informative to compare genealogical time $\bar{N}_{\text{anc}}$ to the total adaptation time, T, and to its independent-locus analogue. Remarkably, making use of eqs. (3.89) and (3.90), both are expressed in terms of final correlation parameter $C(f_{\text{end}})$ only

$$\begin{aligned} \bar{N}_{\text{anc}}/T &= \log^{-1}\frac{1}{1 - C(f_{\text{end}})} \\ \bar{N}_{\text{anc}}/T_{1\text{site}} &= \frac{1 - C(f_{\text{end}})/2}{1 - C(f_{\text{end}})}\log^{-1}\frac{1}{1 - C(f_{\text{end}})} \end{aligned} \tag{3.91}$$

We can see that ratio $\bar{N}_{\text{anc}}/T$ monotonously decreases with $C(f_{\text{end}})$ and, therefore, monotonously increases with both parameters β' and γ (Figure 3.14A). For rare recombination, $\beta' \ll 1$, the inbreeding effect is strong, and $\bar{N}_{\text{anc}}/T$ is small. For frequent recombination, the inbreeding effect is weak, and $\bar{N}_{\text{anc}}/T$ is large.

The ratio $\bar{N}_{\text{anc}}/T_{1\text{site}}$ (and, hence, the genealogical time $\bar{N}_{\text{anc}}$ itself) has a more complex dependence on $C(f_{\text{end}})$. It has a minimum at $C(f_{\text{end}}) \approx 0.72$, eq. (3.91). As a result, time scale $\bar{N}_{\text{anc}}$ has an absolute minimum in β' and γ' where $\bar{N}_{\text{anc}} \approx 1.8\ T_{1\text{site}}$ and $T \approx 2.3\ T_{1\text{site}}$ (Figure 3.14A). The increase of the effective population size at smaller β' is caused by the increasing loss of variable sites, C_{loss}, which decreases the genetic distance ($q \ll 1$) and hence increases the clone decay parameter β (see Figure 3.12C).

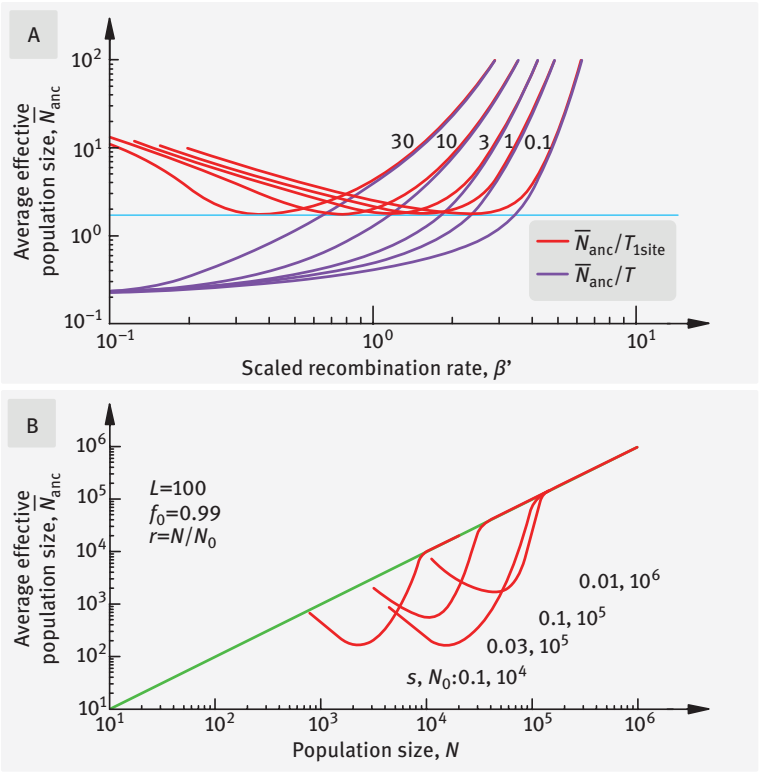

Figure 3.14: Average effective population size N_{eff} for genealogy. (a) Blue curves: Harmonic average of $\bar{N}_{anc}$, normalized to the total adaptation time in the deterministic limit, T_{1site}, as a function of scaled recombination rate β' at different values of scaled selection strength γ (on the curves). Green curves: $\bar{N}_{anc}$, normalized to the total adaptation time, T. Brown line: Minimum value of $\bar{N}_{anc}/T_{1site}$. (b) $\bar{N}_{anc}$, as a function of population size N in the "dilute virus" case, $r = N/N_0$. Parameters are shown. (a), (b) Results are obtained from eq. (3.91) (based on Rouzine and Coffin (2010)).

The value of $\bar{N}_{anc}$ as a function of population size N is shown for the "dilute virus model", $r = N/N_0$ and parameter values relevant for an HIV population an infected untreated patient (Rouzine and Coffin, 2010) (Figure 3.14B). The reason why recombination rate is assumed to be proportional to population density is that only a small fraction of cells coinfected with two different virus genomes can produce recombinants, the rest produce simple copies, that is, reproduce virus asexually. (Such a density-dependent scaling of r is not appropriate for organisms, but the comparison is still useful in this case, as long as one keeps in mind the rescaling.) In a window of N values, time $\bar{N}_{anc}$ can be much less than the prediction of the selectively neutral model, $\bar{N}_{anc} = N$ (Kingman, 1982a, b), provided N_0, s, or L is sufficiently large. At the minimum of $\bar{N}_{anc}$, the condition is $N \gg 1.8T_{1site}$, which is equivalent to $Ns \gg 4|\log(1-f_0)|$. In the opposite neutral case, where $\bar{N}_{anc} > N$, our derivation ceases to apply.

We have showed analytically, that the global linkage parameter $p = V/(sd)$ is approximately equal to 1, based on the inequality $\log(Nr) \gg 1$. To test the accuracy of this approximation numerically for parameter values representative for a virus, p was averaged out over f_1, as given by

$$\langle p \rangle_{f_1} = \frac{1}{1 - f_{\text{end}}} \int_{f_{\text{end}}}^{1} p(f_1) df_1 \tag{3.92}$$

In the region plotted in Figure 3.13B, we have $1 - \langle p \rangle < 0.2$, so that the approximation works.

3.4.4.2 Summary of Section 3.4

To summarize this section, the dynamics of inter-sequence correlations caused by phylogenetic relation, the adaptation time, T, and the genealogical time $\bar{N}_{\text{anc}}$ normalized to T depend monotonously on only two composite model parameters: the rescaled recombination rate, $\beta' \approx (r/s)\left[2\log(Nr)/L\right]^{1/2}$ and the rescaled selection coefficient, $\gamma \approx s\left[L\log^3(Ns)/2\right]^{1/2}$. These two scaled parameters play the same role as the two parameters of a two-locus model, Nr_2 and Ns, respectively, where r_2 is the crossover rate between a locus pair. In the limit of frequent recombination or large population size, correlations become small, and the adaptation speed saturates at its single-site model limit.

For rare recombinations, correlations are strong, adaptation fails at many sites that lose beneficial alleles, and is slow at the sites which complete adaptation. In the general case, the population becomes completely inbred far before adaptation is complete. Further adaptation can be continued only with the help of new mutations, as described in Section 3.3.

As we show in the next subsection, the sensitivity of the outcome to the value of r reflects rapid changes in the clone structure of a population affecting, in their turn, genealogy time $\bar{N}_{\text{anc}}$. We will show that, at small β', each fitness class is a single clone of identical sequences, while at large β', fitness class comprises many small clones.

3.4.5 Clone structure of fitness classes

To derive the coalescent probability given above in eq. (3.76), we need to understand the clonal structure of population under selection and recombination (Rouzine and Coffin, 2007). As we show now, each fitness class consists of clones, that is, groups of identical sequences that are recombinants of different age and different size. A recombinant sequence generated near the high-fitness end of distribution, expands due to natural selection and forms a clone, while also decaying due to recombination. The relative size of a clone within a fitness class is determined by its age, that is, the

time passed since its birth. We have competition between two factors: size of clones and the number of clones of a given size. Older clones are much larger, because of their exponential growth in time, but also much more rare than younger clones. Indeed, to be old, a clone must be generated in the tail of generator function, $\rho(x)$, with a high starting fitness $-sx \approx -sx_c$. Hence, their generation rate is much smaller than for younger and smaller clones (Figure 3.15). The relative contribution of clones born at earlier and later times into a fitness class depends on parameter r. Further, we show that, at sufficiently small r, each fitness group is dominated by lineages born near the high-fitness edge.

3.4.5.1 Fitness of most likely parents

The fitness of a recombinant, $-sx$, fluctuates around the average fitness of its parents, because locations of alleles in two parents are random and uncorrelated, except for the loci that have a common ancestor (fraction C). New recombinants with different fitness values $-sx$ are generated at rates determined by the recombinant generation profile $\rho(x)$, eq. (3.69). As we will show, for a highly-fit recombinant $-sx$, where $\sqrt{d} \ll x$, both parents have fitness values near an optimal point $-sxp/(p+1)$, that is, at $p \approx 1$, half-way between the offspring and the traveling wave maximum (Figure 3.15). Indeed, parents with a higher fitness value are too few, and less-fit parents are unlikely to produce such a highly-fit offspring. Indeed, The general expression for the recombinant generation rate is given by eq. (3.8) where $\bar{k}$ is replaced with eq. (3.57) for the genetic half-distance, d, taking into account genome correlations:

$$\rho(x) = \frac{1}{\sqrt{\pi \bar{k}}} \int dx_1 \int dx_2 \; \phi(x_1)\phi(x_2) e^{-[x-(x_1+x_2)/2]^2/d} \tag{3.93}$$

Substituting eq. (3.62) for $\phi(x)$, we observe, that the integrand, which has a Gaussian form, peaks at $x_1 = x_2 = xp/(p+1)$, and that the peak has a small width $\sqrt{d} \ll x$ Integrating in x_1 and x_2, we obtain, again, the recombinant generator in eq. (3.69). Thus, for a recombinant genome with $x < 0$, the most likely parents are $xp/(p+1)$ (Figure 3.15). Thus, the parents of any offspring that is far above average are likely to be "not at the level" with their child.

3.4.5.2 Analysis of clone structure

As already mentioned, a fitness class consists of lineages of different age. Each lineage starts from a single recombinant individual who was fortunate to escape extinction by random drift and be established. In the traveling wave framework, it is most convenient to label clones by their birth fitness, $-sx'$, (Figure 3.15). The number of clones born and established within the interval $[x', x'+dx]$ is denoted as $m(x')dx'$, where

$$m(x') \equiv \left[Nr\rho(x')\right]\left(s|x'|\right)(1/V), \quad x' < 0 \tag{3.94}$$

where recombination generator $\rho(x)$ is given by eq. (3.69). The first factor in the right-hand side is the recombinant birth rate per generation per unit x, the second factor is the establishment probability for the new recombinant, and factor $1/V$ is the time interval in which the wave moves by unit in x.

A new sequence created at location $x' < 0$ has a fitness advantage $-sx'$ with respect to an average genome. The lineage is established in the population if its size exceeds the stochastic threshold, $1/(s|x'|)$, and it is fixed with certainty due to natural selection. Then, as the wave peak moves toward it, the clone is grows deterministically (Figure 3.15) and, simultaneously, loses sequences due to recombination with genomes from other clones. The size $n(x', x)$ of a clone born at x' and sampled later, when it has relative mutation load $x > x'$, can be obtained from eq. (3.7) with $n(x', x)$ instead of $\phi(x)$ and without the term $r\,\rho\,(x)$ responsible for generation of new recombinants. The initial condition $n(x', x') = 1/(s|x'|)$ corresponds to the stochastic threshold of establishment. (Although it is defined up to a numeric factor ~1, below we verify that the coefficient 1 is correct.) Solving eq. (3.7), we obtain

$$n(x', x) \equiv \frac{1}{s|x'|}\, e^{\frac{s\left(x'^2 - x^2\right)}{2V} + \frac{r(x' - x)}{V}} = \frac{1}{s|x'|}\, e^{\frac{1}{pd}\left[\frac{x'^2 - x^2}{2} + \frac{r(x' - x)}{s}\right]} \tag{3.95}$$

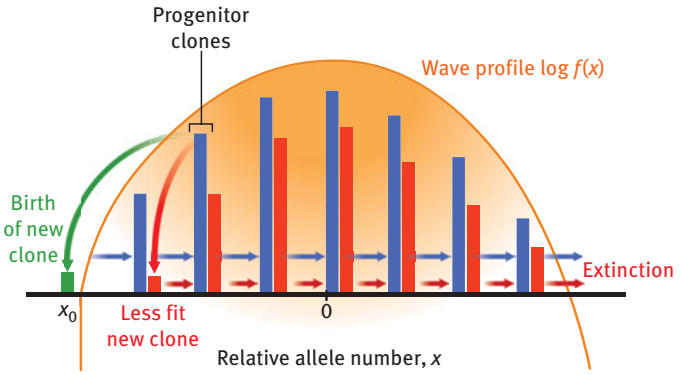

Figure 3.15: Each fitness class comprises many clones of different sizes born with different relative fitness values x. The sketch shows the life cycles of two clones (blue and red bars, respectively) from the same parents undergoing birth, growth, parenthood, further growth, contraction, and extinction. A blue (or red) bar shows locations of the respective clone in x at consecutive time intervals due to motion of the reference frame $k_{av}(t)$ to the left. In the middle of the high-fitness tail of the distribution, clones recombine and create new offspring. Green: A clone is born at the edge, $x=x_0$, and later becomes as tall as its parents. Red: A clone is born short of the edge and never grows enough to contribute to evolution (based on Rouzine and Coffin (2007)).

The probability density of genomes in x, $\phi(x)$, can be written as an integral over clones generated in different locations (clones with different age), as given by

$$\phi(x) = \frac{1}{N}\int_{x_0}^{x} dx' m(x') n(x', x) \tag{3.96}$$

where $m(x)$ is given by eq. (3.94).

Note that factors $m(x')$ and $n(x', x)$ in the integrand of eq. (3.96) both change exponentially with x', but in the opposite directions: at negative x', $m(x')$ is increasing, and $n(x', x)$ is decreasing with x'. In the case of rare recombination, $\beta \ll 1$, we have $1-p \gg d/x_0^2$, so that the second factor wins, and the integrand in eq. (3.96) is maximal at the lower limit, $x = x_0$, with a sharp decrease toward smaller $|x|$. Also, at $\beta \ll 1$, the term with r in eq. (3.95) describing the loss of sequences due to recombination can be neglected. We conclude that old clones born near the high-fitness edge of the distribution x_0 dominate fitness class x is dominated by (Figure 3.15). Using this fact and integrating in eq. (3.96) over x', for fitness classes not too close to the edge, we arrive at a non-normalized Gaussian form of $\phi(x)$, eq. (3.62). The normalization factor follows from condition $\int dx\, \phi(x) = 1$, and we also obtain a relationship between parameter p and the cutoff point, x_0

$$x_0^2 \approx d\frac{2p(1+p)}{1-p}\log\left[\frac{s\sqrt{d(1-p)}}{r}\right], \quad r \ll s\sqrt{d(1-p)} \tag{3.97}$$

which is a direct generalization of eq. (3.18) with $\bar{k} \to d$. The validity condition of eq. (3.97) can be written as $r|x_0| \ll V$, which implies that most sequences do not have recombination before they become extinct. This condition confirms our above approximation neglecting the recombination term in eq. (3.97). Using the generalized form of eq. (3.24) with $\bar{k} \to d$ obtained from the stochastic edge consideration

$$x_0^2 \approx d(1+p)\log\frac{Nr}{p} \tag{3.98}$$

we again arrive at eqs. (3.64) and (3.65) for parameter p and lead x_0. Thus, we reobtained the results for the fitness distribution obtained earlier in this section, using an independent argument based on clone structure.

3.4.5.3 Life cycle of a clone

It is very informative for what follows to understand the fitness trajectory of a representative large lineage dominating a fitness class (Figure 3.15). Suppose, a new lineage is generated by recombination and established near the edge, $x \approx x_0$. As the fitness distribution advances, the value of x decreases, and the clone grows in size with a decreasing exponential speed. After the clone reaches the point $x = 0$, which

corresponds to the average fitness, it starts to contract due to negative selection until, eventually, it becomes extinct. Before that, when the clone is at the most likely parenthood point $x = x_0 p/(1+p)$(see earlier), its individuals mate with other individuals of similar fitness. The offspring is broadly scattered in fitness, but one of its genomes is lucky to create one offspring far at the high-fitness edge of the distribution. This offspring starts a new lineage, grows up, reaches the parenthood point, and the reproduction cycle repeats again.

We now choose an individual genome randomly and choose any locus in this genome. Our task is trace the fitness of its ancestor lineage, $-sk_{\text{anc}}(t)$, back in time based on the reproduction cycle shown in Figure 3.15. For that we need to recall that this figure is in reference frame of the average population fitness, and that the traveling wave is moving, while the fitness of each clone is fixed. A representative trajectory $k_{\text{anc}}(t)$ is a broken line (Figure 3.16A) consisting from vertical segments corresponding to the clonal expansion during asexual reproduction and horizontal segments due to recombination with another genome. The length between mating events fluctuates around $\langle \Delta k \rangle = |x_0|/(1+p)$, which is the distance between the edge and the optimal parenthood point in k. The average time interval between mating events is $\langle \Delta t \rangle = \langle \Delta k \rangle / V = |x_0|/\langle V(1+p) \rangle$. Because the birth point of an edge clone and the point of parenthood both fluctuate between mating events, the jump length Δk wobbles as well, according to a narrow distribution with a variance, $(\text{Var}[k])^{1/2} \sim d \ll |x_0|$. Therefore, the long-range periodicity is absent from this broken-line trajectory. Therefore, sufficiently far back in time, independently on the initial fitness today, the ancestor's relative fitness can be found, with an almost uniform probability density, anywhere in the interval between the birth point and the optimal parenthood point, as given by $x_0 < x < x_0 p/(1+p)$ (Figure 3.16B). Thus, only genomes with a very high fitness above the parenthood point are likely to establish an uninterrupted lineage in the future.

Speaking in terms of the age of clones, the time interval $\Delta t = |x_0|/[V(1+p)]$ is the optimal reproduction age. In other words, only sufficiently young clones are likely to leave descendants in the future.

3.4.5.4 Probability of finding two individuals in the same clone

In the previous sections, we expressed the density of coalescent events in terms of the probability of two individuals with the same fitness to be accidentally found within the same clone, eq. (3.77). Now we will derive it. If we use the continuous approximation in x', we obtain

$$P_d(x) = \frac{\int_{x_0}^{x} dx'\, m(x') n^2(x', x)}{[N\phi(x)]^2} \tag{3.99}$$

where $m(x)$ and $n(x', x)$ are given by eqs. (3.94) and (3.95), respectively.

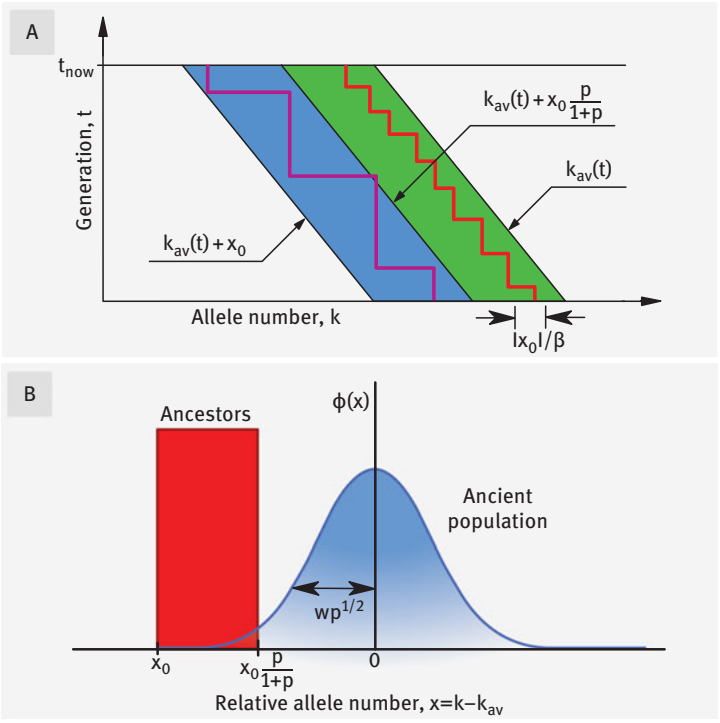

Figure 3.16: Fitness history in reverse and the fitness distribution of ancestors. (A) Thick broken line in the blue area: a representative trajectory of the relative less-fit allele number, $x(t)$, of ancestral lineage of a site in a genome at low recombination rates, $r \ll sd/|x_0|$, that is $\beta \ll 1$. Thick broken line in the green area: the trajectory at $\beta \gg 1$. Thin black lines: trajectories of the high-fitness edge, the most likely parenthood point at $\beta \ll 1$, and the $x = 0$, respectively. (B) Fitness distribution for an ancestor at $\beta \ll 1$ compared to its contemporary population (based on Rouzine and Coffin (2007)).

However, there are two reasons why the continuous-in-x' approximation is inaccurate in this case, both following from the fact that $n(x', x)$ enters the integrand of eq. (3.99) in the second power. In the integrand of eq. (3.99), at large negative x', $m(x')$ decreases as $\exp(-x'^2/2d)$, and $n^2(x', x)$ increases as $\exp(x'^2/d)$. Therefore, the integrand in eq. (3.99) decreases sharply from its lower limit. The rapid increase has two effects, as follows. First, expanding the net exponential linearly near the leading edge in $x' - x_0$, we observe that the integral in x' is mostly contributed from a narrow interval near the edge

$$x' - x_0 \sim [d \ln \rho / dx]^{-1}_{x_0} \sim \omega^2 / x_0$$

The region is of the same order as the typical distance between the birth locations in x of adjacent clones born at the leading edge, Δx, given by

$$\Delta x = \left[\frac{d\log\rho}{dx}\right]^{-1}_{x_0} \tag{3.100}$$

Thus, two individuals (representing two ancestral lineages) are most likely to meet in a small number of edge-born clones, and the contribution to $P_{\rm cl}$ of two adjacent clones may differ several-fold. Therefore, a discrete sum instead of an integral is due, with averaging over their birth locations.

The second issue is fluctuations between realizations. In eq. (3.99), we assume that the lower limit in x' is given by the average cutoff at x_0, eq. (3.98). However, $P_{\rm cl}$ is obviously quite sensitive to the birth place of the largest clone in a fitness class, which we denote $x_0{}'$, because it is in the argument of an exponential. The clones that are, at rare times, created far ahead of the average edge, are much larger and thus contribute much more to $P_{\rm cl}$ than representative edge clones near x_0. In stands to reason that $P_{\rm cl}(x)$ is mostly contributed from rare generations in a single large clone born far ahead the average edge at x' makes most of the entire fitness class, x, as given by

$$n\,(x_0', x) = N\phi(x) \tag{3.101}$$

which implies that x_0' depends on x. Hence the probability of having two individuals meeting in the same clone is given by the probability to have the largest clone born at $x' < x_0'$ given by

$$P_{\rm cl}(x) = \int_{-\infty}^{x_0{}'} dx' m(x') \tag{3.102}$$

which replaces eq. (3.99). Substituting $\phi(x)$ and $n(x', x)$ from eqs. (3.72) and (3.95) into eq. (3.101), for the birth location $x_0{}'$ we obtain

$$\frac{\sqrt{2\pi d}}{Ns\left|x_0'\right|}\exp\left\{\frac{x_0'^2}{2d} + \frac{\varepsilon_\beta}{2}\left[1 - u^2 + 2h_\beta(u)\right] - \beta(1+u) + O\left[\ln^{-1}(Nr)\right]\right\} = 1 \tag{3.103}$$

Next, substituting $m(x)$ from eq. (3.94) and $\rho(x)$ from eq. (3.72) into eq. (3.102), we get

$$P_{\rm cl} = \frac{d}{\left|x_0'\right|} m(x_0') = \frac{Nr}{\sqrt{2\pi d}}\exp\left\{-\frac{x_0'^2}{2d} - 2\varepsilon_\beta h_\beta(-1/2) + O\left[\ln^{-1}(Nr)\right]\right\} \tag{3.104}$$

Finally, solving eq. (3.103) for $x_0{}' \gg d^{1/2}$ iteratively in the first approximation and substituting it into eq. (3.104), we arrive at eq. (3.77) of the main text.

3.4.6 Fitness distribution of remote ancestors

Another important element used in Section 3.4.3 to derive the density of coalescent events $1/N_{\text{anc}}$ is the distribution of ancestor fitness. As it is clear from lineage trajectories discussed in Section 3.4.5, the fitness distribution of ancestors, $y(x)$, is very different from the fitness distribution of a population, $\phi(x)$ (Figure 3.16B), and requires a separate analysis.

3.4.6.1 Small recombination rates ($\beta \ll 1$)

We start with the case when the clone decay parameter β defined in eq. (3.68) is small. We choose a site on an individual genome and address the fitness of a distant ancestor. While the ancient population distribution is localized symmetrically near its peak, the ancestor distribution, $\varphi(x)$, is nearly uniform with fitness above the optimal parenthood point

$$\varphi(x) = \begin{cases} |x_0|(1+p), & x_0 < x < x_0 p/(1+p) \\ 0, & \text{otherwise} \end{cases} \tag{3.105}$$

3.4.6.2 Population distribution in fitness at any recombination rate

Our next task is to calculate the modern fitness distribution in the general case of β. Now we consider the general case when the clone decay parameter

$$\beta = r|x_0|/V = r|x_0|/(psd) \tag{3.106}$$

is not necessarily small. Here x_0 and p are given by eqs. (3.64) and (3.65).

In case $\beta \ll 1$, we could neglect with the clone decay due to recombination, and the integrand in eq. (3.96) peaked at the low limit of integration. At larger recombination rates, such that $\beta \sim 1$ or $\beta \gg 1$, we have $1-p$ less or on the order of d/x_0^2, so that the integrand in eq. (3.96) no longer has a sharp peak at the lower limit $x = x_0$. This implies that clones born far from the edge inside of the wave, $x > x_0$, now give important contribution to fitness classes (and also decay significantly), affecting our conclusions regarding the Gaussian form of $\phi(x)$, as well the ancestor history. For this case, we have to use an alternative formalism.

We start by observing that the relative difference between logarithm of fitness distribution $\log\phi(x)$ and its deterministic limit is on the order of $1-p \sim 1/\log(Nr)$, eq. (3.62). Therefore, we can calculate the correction to the Gaussian as a first-order correction in small parameter $1/\log(Nr) \ll 1$. We seek $\phi(x)$ in the form

$$\phi(x) = \frac{1}{\sqrt{2\pi d}} e^{-\frac{x^2}{2d} - \varepsilon h(u)} \tag{3.107}$$

$$\begin{aligned} \varepsilon &\equiv \left(x_0^2/d\right)(1-p) \approx 2\lambda_1(1-p) \\ u &\equiv x/|x_0| \end{aligned} \tag{3.108}$$

For the sake of convenience, to decrease the number of parameters, we replaced $1-p$ and x with scaled variables, ε and u, which are both on the order of 1 for $\beta \sim 1$. In this notation, the recombinant generator in eq. (3.93) takes a form

$$\rho(x) = \frac{1}{\sqrt{2\pi d}} e^{-\frac{x^2}{2d} - 2\varepsilon h(u/2)} \tag{3.109}$$

An equation for non-Gaussian correction $h(u)$ that follows from eqs. (3.7), (3.107), and (3.109) has a form

$$\varepsilon h(u) = \frac{\varepsilon u^2}{2} + \beta\left[u - \int_0^u du'\, e^{\varepsilon h(u') - 2\varepsilon h(u'/2)}\right] \tag{3.110}$$

were β is the only fixed parameter. The value of ε is not an independent parameter, it needs to be determined from the boundary condition at the edge. From the clonal representation of the traveling wave, eq. (3.96), the fitness density is zero at the edge, $\phi(x_0) = 0$, which implies that $h(u)$ must diverge at $u = -1$, $h(u) \to \infty$. eq. (3.110) can be solved numerically with respect to $h(u)$ and ε for different values of fixed parameter β (Figure 3.17A).

In the limit $\beta \ll 1$, function $h(u)$ is quadratic except near the edge, which agrees with eq. (3.62). For the value of ε, in this range we have

$$\varepsilon \approx 4\ln[(2/\beta)\log(2/\beta)] \approx 4\Lambda_2, \quad \beta \ll 1$$

Hence, at small β, we have $1-p \approx 2\Lambda_2/\Lambda_1$, which overlaps with expression for p in the limit $1-p \ll 1$, eq. (3.64). The two asymptotic expressions for p are valid in two overlapping intervals of r and hence can be combined into an accurate interpolation formula

$$p = \frac{\Lambda_1}{\Lambda_1 + \varepsilon(\beta)/2} \tag{3.111}$$

where $\varepsilon(\beta)$ is plotted in inset in Figure 3.17A.

All these results have been obtained for r and N as independent parameters, which is reasonable in most real-life scenarios. For the specific example of a dilute viral infection, recombination is proportional to coinfection, and we have scaling $r = N/N_0$, where N_0 is a fixed parameter. Examples of the dependence of p on N for such scaling are plotted in Figure 3.17B. If inter-genomic correlations were absent ($C = 0$), $p(N)$ would represent the speed of evolution scaled to the deterministic limit for independentsites.

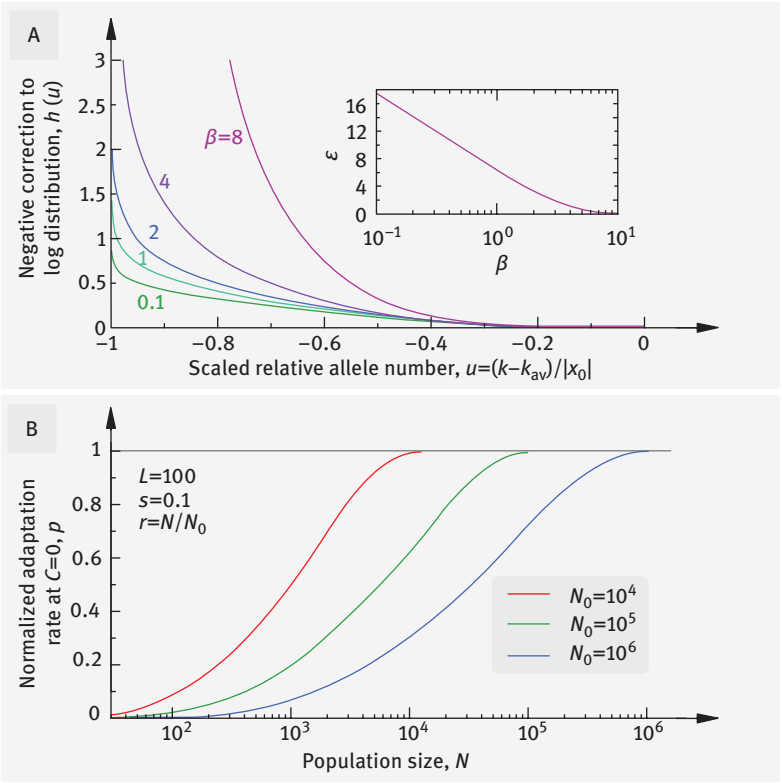

Figure 3.17: Non-Gaussian correction to the log fitness distribution profile and the scaled adaptation rate at $C = 0$, p, in the general case of non-small recombination rates (β can be ~1 or large). (A) Solid lines: normalized negative correction $h(u)$ to log $\phi(x)$, eq. (3.107), plotted at different values of the parameter β (shown), as a function of the scaled relative number of less-fit alleles, $u = x/|x_0|$. Results are obtained by solving eq. (3.110) numerically. Inset: scaled negative correction to the evolution rate ε, eq. (3.108), as a function of β. (B) The dependence of p on N for $C = 0$, for the case $r = N/N_0$, at different N_0 (shown), based on eq. (3.111). Parameters are shown (based on Rouzine and Coffin (2007)).

3.4.6.3 Ancestor fitness distribution at any recombination rate

The ancestor fitness distribution $\varphi(x)$ has the rescaled form where rescaled distribution $\varphi(x) \equiv (1/x_0)\ y(u)$ satisfies an equation

$$y(u) = \beta \int_{\max(2u,\,-1)}^{u} du' e^{\varepsilon\left[h(u') - 2h(u'/2)\right]} y(u') \tag{3.112}$$

where β is the only external parameter, $u < 0$, and $h(u)$ and ε are determined from eq. (3.110). Numeric solution of eq. (3.112) for function $y(u)$ is shown at different β in Figure 3.10. In the limit $\beta \to 0$, the distribution is almost uniform within the

interval $-1 < u < -1/2$, just as we obtained earlier (Figure 3.16B). At $\beta \sim 1$, the distribution shifts toward lower fitness values and, at $\beta \gg 1$, assumes a universal asymmetric shape with a scale $|u| \sim 1/\beta$. Thus, unless β is extremely large, $\beta \sim \sqrt{d} \sim \sqrt{L}$, the conclusion that all ancestors are exceptionally well-fit for their time still holds, although they no longer occupy the upper half of the fitness distribution tail as they do in the rare recombination case $\beta \ll 1$.

To start the derivation of eq. (3.112), we note that each genomic site in a genome has an ancestor genome in each past generation. Consider a site in a genome at time t_{now} and its ancestral genome at time $t < t_{\text{now}}$. The probability density $\varphi(x, t)$ of ancestor allele number $x = k - k_{\text{av}}(t)$ obeys a Markovian equation

$$\varphi(x,t) = \int_{x_0}^{\infty} dx' P(x|x') \varphi(x', t+1), \; x > x_0 \tag{3.113}$$

where $P(x|x')$ is the probability density of the relative less-fit allele number of a parent x, given the number x' of a progeny genome. Then eq. (3.113) is solved backward in time with the initial condition, $\varphi(x, t_{\text{now}}) = \delta[x - x(t_{\text{now}})]$. The kernel $P(x|x')$ can be written in the form

$$P(x|x') = A(x') \left[\phi(x' - V) \delta(x' - x - V) + r\rho(x' - V) P_{\text{sex}}(x|x') \right] \tag{3.114}$$

The first and the second terms in the brackets represent asexual reproduction by clonal expansion and sexual reproduction by recombination, respectively. Population density in fitness $\phi(x)$ and recombination generation rate $\rho(x)$ are given by eqs. (3.107) and (3.109), respectively. Kernel $P_{\text{sex}}(x|x')$ is the normalized probability density of parental fitness x assuming sexual reproduction and given progeny genome fitness x'. The term $-V$ appears due to the shift of the population fitness distribution between the adjacent generations. Prefactor $A(x)$ensures normalization with respect to parental fitness. Recollecting that $\phi(x)$ must be zero at $x \le x_0$, the normalization factor $A(x)$ takes a form

$$A(x') = \begin{cases} 1/[\phi(x' - V) + r\rho(x' - V)], & x' > x_0 + V \\ 1/[r\rho(x' - V)], & x_0 < x' < x_0 + V \end{cases}. \tag{3.115}$$

As we have mentioned several times, $P_{\text{sex}}(x|x')$ is maximal at the likely parenthood point, $x \approx x'/2$, and its characteristic width in x is $\sim\sqrt{d}$ (Figure 3.15). On the other hand, the characteristic scale of the ancestor distribution $\varphi(x, t)$ in x is on the order of $|x_0|$, where $|x_0| \gg \sqrt{d}$. Therefore, in eq. (3.113), we can approximate $P_{\text{sex}}(x|x')$ with a delta-function, as given by

$$P_{\text{sex}}(x|x') \approx \delta(x - x'/2) \tag{3.116}$$

Substituting eqs. (3.115) and (3.116) into eq. (3.114), we obtain

$$P(x|x') = [1-\bar{r}(x')]\delta(x'-x-V) + \{\bar{r}(x') + [1-\bar{r}(x')]\theta(x_0+V-x')\}\delta(x-x'/2) \tag{3.117}$$

$$\bar{r}(x') \equiv \frac{r\rho(x'-V)}{\phi(x'-V)+r\rho(x'-V)} \tag{3.118}$$

Due to our basic assumption that evolution is slow, $|x_0|/V \gg 1$, since the fitness difference $s|x_0|$ between the average and the best-fit genome is smaller than 1, we can formally and linearly expand the first delta function and the theta function in eq. (3.117) in V, which yields

$$P(x|x') = \delta(x'-x) + V\Big[-\delta'(x'-x) + \delta(x_0-x')\delta(x-x'/2)\Big] + \bar{r}(x')\big[\delta(x-x'/2) - \delta(x'-x)\big] \tag{3.119}$$

Substituting eqs. (3.119) into eq. (3.113), approximating as usual

$$\varphi(x,t+1) - \varphi(x,t) \approx \partial\phi/\partial t$$

and integrating eq. (3.113) in x, we obtain

$$\begin{aligned} -\frac{\partial\varphi(x,t)}{\partial t} = V\frac{-\partial\varphi(x,t)}{\partial x} + V\varphi(x_0+0,t)[\delta(x-x_0/2) - \delta(x-x_0)] \\ + 2\bar{r}(2x)\varphi(2x,t)\theta(x-x_0/2) - \bar{r}(x)\varphi(x,t) \end{aligned} \tag{3.120}$$

We can decrease the number of independent parameters, by rescaling the variables as given by

$$\tau \equiv tr/\beta, \quad u \equiv x/|x_0|, \quad y(u,\tau) \equiv |x_0|\varphi(x,t) \tag{3.121}$$

In the rescaled notation, eq. (3.120) writes

$$\begin{aligned} -\partial y(u,t)/\partial\tau = \partial y(u,t)/\partial u + y(-1+0,\tau)[\delta(u+1/2) - \delta(u+1)] \\ + \beta\ [2\eta(2u)y(2u,t)\theta(u+1/2) - \eta(u)y(u,t)] \end{aligned} \tag{3.122}$$

$$\eta(u) = \exp\{\varepsilon[h(u) - 2h(u/2)]\} \tag{3.123}$$

where eqs. (3.107), (3.109) and (3.118) and strong inequality $r \ll 1$ are used.

The characteristic timescale in eq. (3.122) for is $\tau \sim 1$, or $t \sim \beta/r$. If we go farther back in time, the ancestor fitness distribution takes a stationary form that satisfies an equation

$$\frac{dy(u)}{du} = y(-1+0)[\delta(u+1) - \delta(u+1/2)] + \beta[\eta(u)y(u) - 2\eta(2u)y(2u)\theta(u+1/2)] \tag{3.124}$$

Note that $\eta(u)$ diverges at the leading edge $u \to -1$ due to the divergence of non-Gaussian correction $h(u)$, eq. (3.124). To keep derivative dy/du finite, $y(u)$ has to

vanish at $u \to -1$. Therefore, in eq. (3.12), the term with delta functions must be zero. Integrating eq. (3.124) in u, we arrive at the promised eq. (3.112) we have used in the previous sections.

We conclude that a single composite model parameter β proportional to the recombination (outcrossing) rate, r, and inversely proportional to the selection coefficient, s, decides both the history of dominant lineages and the rescaled fitness distribution of distant ancestors. The average substitution rate, the fitness lead, and the width of the fitness distribution of the entire population are all expressed in terms of β and the genetic distance d (addressed in Sections 3.4.3 and 3.4.4).

3.4.7 Main approximations

Let us check approximations we have used in Section 3.4.

3.4.7.1 Neglecting the loss of deleterious alleles

Based on simulation results (Gheorghiu-Svirschevski et al., 2007), we assumed in model formulation in Section 3.4.2 that common ancestor sites have, for large samples, only less-fit alleles. The rationale for neglecting the loss of deleterious alleles is that, in the initial population ($1 - f_0 \ll 1$), less-fit alleles are much more abundant.

3.4.7.2 Time locality for the coalescent density

The present work considers a nonstationary process of adaptation. Therefore, the coalescent event density $1/N_{\text{anc}}$ expressed in terms of the effective population size N_{anc} in eq. (3.59) varies in time, specifically, through the genetic half-distance d and clone decay parameter β. Please note that eq. (3.59) does not include a time delay treating N_{anc} as a variable depending on the current state variables. This treatment is in apparent contradiction with eq. (3.76), which links N_{anc} at current time t to the fitness distribution of remote ancestors $\varphi(x)$. Also, $P_{\text{cl}}(x)$, which also depend on time through d, must refer to the ancestral, earlier population as well. In general, that could create a time delay in the original correlation dynamics.

Fortunately, as it has been shown in Section 3.4.6, the ancestor fitness distribution going back in time becomes independent on the initial condition and assumes the calculated ancient value at $t \gg \beta/r = |x_0|/V$. Coalescent events in this time interval can be neglected provided the time interval of the wave shift by its lead, $|x_0|/V$, is much less than the coalescent time, N_{anc}. The resulting validity condition of eqs. (3.59) and (3.76) under which they are asymptotically accurate, is $|x_0| \ll L$ or $\log(Nr) \ll L$. This strong inequality is equivalent to condition (i) stated in beginning of Section 3.4.3 which ensures the existence of the traveling wave regime. Thus, unless populations are astronomically large for large L, "remote ancestors" are not that far back in time.

We also note that eq. (3.76) does not include the term of the probability that the two individuals found within the same clone are also the same individuals. This is because, once the two ancestral lineages converge, backward in time, into the same clone, they will coalesce to the same individual in less than $|x_0|/V$ generations backward with a probability equal to 1. As we just showed, such a short time is negligible compared to the average coalescent time N_{anc}.

3.4.7.3 Neutral model relation between C_{loss} and C

We have assumed earlier that the fractions of related sites in a pair of genome and a large sample of genomes are connected by the same relation as in a model with selection. This is a nontrivial assumption, since it implies the same shape of the genealogical tree, in the statistical sense, as in the neutral theory (Kingman, 1982a, b). The justification is, as follows. At a large rescaled selection coefficient, $\gamma \sim sL^{1/2} > 1$, the coalescent event density in time, $1/N_{\text{anc}}$, does not change much through most of the interval, $f_1 = [0;1]$. The coalescent density plummets rapidly only at the ends of the interval, that is, near the beginning and the end of adaptation (Figure 3.12D). Therefore, coalescent events are present mostly where their frequency is almost constant. Therefore, the statistical shape of the phylogenetic tree resembles the tree predicted by the stationary neutral model where that the coalescent frequency is given by $1/N$ (Kingman, 1982a, b), except that the timescale is now N_{anc} rather than the total population size, N. Therefore, the relation between C_{loss} and C, which are the cumulative distributions of the coalescent time for an infinite sample and a pair of genomes, respectively, can be approximated by the neutral relation [Figure 3.9, eq. (3.61)]. This result in striking contrast to a purely asexual regime where not only the timescale, but the shape of the tree is very different from Kingman's coalescent (Brunet et al., 2007; Desai et al., 2013; Neher and Hallatschek, 2013).

3.4.7.4 Asymptotics of coalescent density

Let us derive the asymptotic expressions of $F_\beta(u)$ at small and large β in eq. (3.77). At $\beta \ll 1$, asymptotic expressions for that $h_\beta(u)$ and ε_β are determined by eq. (3.75). At $\beta \gg 1$, the value of ε_β is exponentially small (cf. Figure 3.17A), and the second term in the exponential in eq. (3.77) for $F_\beta(u)$ can be neglected. Based on this information, we obtain two limiting cases

$$F_\beta(u) = \begin{cases} 2\Lambda_2,\ \Lambda_2 \equiv \log(2\Lambda_2/\beta), & \beta \ll 1 \\ \beta e^{-\beta(1+u)}, & \beta \gg 1 \end{cases} \tag{3.125}$$

Asymptotics of $y_\beta(u)$, $u < 0$, can also be derived analytically from eq. (3.112) to obtain

$$y_\beta(u) = \begin{cases} 2\theta(-1/2-u)\theta(u+1) & \beta \ll 1, \quad u+1 \gg 1/\beta \\ \beta\sigma(\beta u),\ \sigma(v) = \int\limits_{2v}^{v} dv'\sigma(v') & \beta \gg 1 \end{cases} \tag{3.126}$$

Substituting eqs. (3.125) and (3.126) into eq. (3.78), we obtain asymptotic expressions for genealogical timescale N_{anc}

$$\frac{1}{N_{\text{anc}}} = \frac{\sqrt{2}}{d^{1/2}\Lambda_1'^{3/2}} \times \begin{cases} \Lambda_2 & \beta \ll 1 \\ (\beta^2/4)e^{-\beta} \int\limits_{-\infty}^{0} dv\,\sigma^2(v)e^{-v} \approx 0.53\,\beta^2 e^{-\beta} & \beta \gg 1 \end{cases} \tag{3.127}$$

where numeric coefficient 0.53 is obtained by solving numerically eq. (3.126) for normalized function $\sigma(v)$ and calculating the integral in v in eq. (3.127). These limits can be interpolated using $\Lambda(\beta)$ in eq. (3.79), which has 1% accuracy.

References

Abramowitz, M., and Stegun, I., eds. (1964). Handbook of mathematical functions (New York: National Bureau of Standards).

Acevedo, A., Brodsky, L., and Andino, R., (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature, *505*, 686–690.

Arjan, J.A., Visser, M., Zeyl, C.W., Gerrish, P.J., Blanchard, J.L., and Lenski, R.E., (1999). Diminishing returns from mutation supply rate in asexual populations. Science, *283*, 404–406.

Balfe, P., Simmonds, P., Ludlam, C.A., Bishop, J.O., and Leigh Brown, A.J., (1990). Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J Virol, *64*, 6221–6233.

Barton, N.H., (1995). Linkage and the limits to natural selection. Genetics, *140*, 821–841.

Barton, N.H., and Rouhani, S., (1991). The Probability of Fixation of a New Karyotype in a Continuous Population. Evolution, *45*, 499–517.

Batorsky, R., Kearney, M.F., Palmer, S.E., Maldarelli, F., Rouzine, I.M., and Coffin, J.M., (2011). Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc Natl Acad Sci U.S.A., *108*, 5661–5666.

Bedford, T., Riley, S., Barr, I.G., Broor, S., Chadha, M., Cox, N.J., Daniels, R.S., Gunasekaran, C.P., Hurt, A.C., Kelso, A., *et al*. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature, *523*, 217–220.

Brunet, E., Derrida, B., Mueller, A.H., and Munier, S., (2007). Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization. Phys Rev E, *76*, 041104–041101.

Brunet, E., Rouzine, I.M., and Wilke, C.O., (2008). The stochastic edge in adaptive evolution. Genetics, *179*, 603–620.

Bulmer, M.G., (1980). The mathematical theory of quantitative genetics, (Oxford New York: Clarendon Press; Oxford University Press).

Burns, D.P., and Desrosiers, R.C., (1994). Envelope sequence variation, neutralizing antibodies, and primate lentivirus persistence [review]. Curr Top Microbiol Immunol, *188*, 185–219.

Chavda, S.C., Griffin, P., Zhen, H.L., Keys, B., Vekony, M.A., and Cann, A.J., (1994). Molecular determinants of the V3 loop of human immunodeficiency virus type 1 glycoprotein gp120 responsible for controlling cell tropism. J Gen Virol, *75*, 3249–3253.

Cleland, A., Watson, H.G., Robertson, P., Ludlam, C.A., and Brown, A.J.L., (1996). Evolution of zidovudine resistance-associated genotypes in human immunodeficiency virus type 1-infected patients. J Acquir Immune Dedic Syndr Hum Retrovirol, *12*, 6–18.

Coffin, J.M., (1995). HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science, *267*, 483–488.

Desai, M.M., and Fisher, D.S., (2007). Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics, *176*, 1759–1798.

Desai, M.M., Walczak, A.M., and Fisher, D.S., (2013). Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics, *193*, 565–585.

Eigen, M., and Biebricher, C.K., (1988). Sequence space and quasispecies distribution, In RNA Genetics: Volume III, E. Domingo, J.J. Holland, and P. Ahlquist, eds. (Boca Raton, Fla.: CRC Press), 3–22.

Excoffier, L., and Ray, N., (2008). Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol Evol (Amst.), *23*, 347–351.

Felsenstein, J., (1974). The evolutionary advantage of recombination. Genetics, *78*, 737–756.

Fisher, R.A., (1922). On the dominance ratio. Proc Roy Soc Edinburgh, *42*, 321–341.

Fisher, R.A., (1930). The genetical theory of natural selection, (Oxford, United Kingdom: Clarendon Press), 1958.

https://doi.org/10.1515/9783110615456-004

Fogle, C.A., Nagle, J.L., and Desai, M.M., (2008). Clonal interference, multiple mutations and adaptation in large asexual populations. Genetics, *180*, 2163–2173.

Gerrish, P.J., and Lenski, R.E., (1998). The fate of competing beneficial mutations in an asexual population. Genetica, *102/103*, 127–144.

Gheorghiu-Svirschevski, S., Rouzine, I.M., and Coffin, J.M., (2007). Increasing sequence correlation limits the efficiency of recombination in a multisite evolution model. Mol Biol Evol, *24*, 574–586.

Good, B.H., Rouzine, I.M., Balick, D.J., Hallatschek, O., and Desai, M.M., (2012). Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci U.S.A., *109*, 4950–4955.

Gordo, I., and Charlesworth, B., (2000). The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics, *154*, 1379–1387.

Goyal, S., Balick, D.J., Jerison, E.R., Neher, R.A., Shraiman, B.I., and Desai, M.M., (2012). Dynamic mutation-selection balance as an evolutionary attractor. Genetics, *191*, 1309–1319.

Groenink, M., Andeweg, A.C., Fouchier, R.A.M., Broersen, S., van der Jagt, R.C.M., Schuitemaker, H., de Goede, R.E.Y., Bosch, M.L., Huisman, H.G., and Tersmette, M., (1992). Phenotype-associated env gene variation among eight related human immunodeficiency virus type 1 clones: evidence for in vivo recombination and determinants of cytotropism outside the V3 domain. J Virol, *66*, 6175–6180.

Haase, A.T., (1999). Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics in lymphatic tissues. Annu Rev Immunol, *17*, 625–656.

Haigh, J., (1978). The accumulation of deleterious genes in a population – Muller's ratchet. Theor Popul Biol, *14*, 251–267.

Haldane, J.B.S., (1924). A mathematical theory of natural and artificial selection. Part I. Trans Camb Phil Soc, *23*, 19–41.

Haldane, J.B.S., (1927). A mathematical theory of natural and artificial selection. Part V: Selection and mutation. Proc Camb Phil Soc, *23*, 838–844.

Hallatschek, O., (2010). The noisy edge of traveling waves. Proc Natl Acad Sci U.S.A., *108*, 1783–1787.

Hallatschek, O., and Nelson, D.R., (2008). Gene surfing in expanding populations. Theor Popul Biol, *73*, 158–170.

Hegreness, M., Shoresh, N., Hartl, D., and Kishony, R., (2006). An equivalence principle for the incorporation of favorable mutations in asexual populations. Science, *311*, 1615.

Hermisson, J., and Pennings, P.S., (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics, *169*, 2335–2352.

Hill, W.G., and Robertson, A., (1966). The effect of linkage on limits to artificial selection. Genet Res, *8*, 269–294.

Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M., and Markowitz, M., (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV infection. Nature, *373*, 123–126.

Holland, S.M., Chavez, M., Gerstberger, S., and Venkatesan, S., (1992). A specific sequence with a bulged guanosine residue(s) in a stem-bulge-stem structure of rev-responsive element RNA is required for trans activation by human immunodeficiency virus type 1 rev. J Virol, *66*, 3699–3700.

Imhof, M., and Schlotterer, C., (2001). Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U.S.A., *98*, 1113–1117.

Karlin, S., and McGregor, J., (1964). On some stochastic models in genetics, Stochastic models in medicine and biology, M., ed. (Madison: University of Wisconsin Press), 245–271.

Kassen, R., and Bataillon, T., (2006). Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet, *38*, 484–488.

Kessler, D.A., Levine, H., Ridgway, D., and Tsimring, L., (1997). Evolution on a smooth landscape. J Stat Phys, *87*, 519–544.

Keys, B., Karis, J., Fadeel, B., Valentin, A., Norkrans, G., Hagberg, L., and Chiodi, F., (1993). V3 sequences of paired HIV-1 isolates from blood and cerebrospinal fluid cluster according to host and show variation related to m clinical stage of disease. Virology, *196*, 475–483.

Kimura, M., (1954). Process leading to quasi-fixation of genes in natural populations due to random fluctuations of selection intensities. Genetics, *39*, 280–295.

Kimura, M., (1955a). Solution of a process of random genetic drift with a continuous model. Proc Nat Acad Sci USA, *41*, 144–150.

Kimura, M., (1955b). Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harb Symp Quant Biol, *20*, 33–53.

Kimura, M., (1962). On the probability of fixation of mutant genes in a population. Genetics, *47*, 713–719.

Kimura, M., (1964). Diffusion models in population genetics. J Appl Probab, *1*, 177–232.

Kimura, M., (1989). The neutral theory of molecular evolution and the world view of the neutralists. Genome, *31*, 24–31.

Kimura, M., (1994). Population genetics, molecular evolution, and the neutral theory, Selected papers. (Chicago: The University of Chicago Press).

Kimura, M., and Ohta, T., (1969). The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. Genetics, *61*, 763–771.

Kingman, J.F.C., (1982a). The coalescent. Stochastic Processes Appl, *13*, 235–248.

Kingman, J.F.C., (1982b). On the genealogy of large populations. J Appl Probab, *19A*, 27–43.

Kolmogorov, A., (1931). Ueber die analitischen Methoden in der Wahrscheinlichkeitrechnung. Math Ann, *104*, 415–458.

Kondrashov, A.S., (1993). Classification of hypotheses on the advantage of amphimixis. J Hered, *84*, 372–387.

Lamers, S.L., Sleasman, J.W., She, J.X., Barrie, K.A., Pomeroy, S.M., Barrett, D.J., and Goodenow, M.M., (1993). Independent variation and positive selection in env V1 and V2 domains within maternal-infant strains of human immunodeficiency virus type 1 in vivo. J Virol, *67*, 3951–3960.

Lande, R., (1998). Risk of population extinction from fixation of delterious and reverse mutations. Genetica, *102/103*, 21–27.

Lech, W.J., Wang, G., Yang, Y.L., Chee, Y., Dorman, K., McCrae, D., Lazzeroni, L.C., Erickson, J.W., Sinsheimer, J.S., and Kaplan, A.H., (1996). In vivo sequence diversity of the protease of human immunodeficiency virus type 1: presence of protease inhibitor-resistant variants in untreated subjects. J Virol, *70*, 2038–2043.

Levy, D.N., Aldrovandi, G.M., Kutsch, O., and Shaw, G.M., (2004). Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci U.S.A., *101*, 4204–4209.

Lopez-Galindez, C., Rojas, J.M., Najera, R., Richman, D.D., and Perucho, M., (1991). Characterization of genetic variation and 3′-azido-3′-deoxythymidine- resistance mutations of human immunodeficiency virus by the RNase A mismatch cleavage method. Proc Natl Acad Sci U.S.A., *88*, 4280–4284.

Lukashov, V.V., Kuiken, C.L., and Goudsmit, J., (1995). Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J Virol, *69*, 6911–6916.

Luksza, M., and Lassig, M., (2014). A predictive fitness model for influenza. Nature, *507*, 57–61.

Mansky, L.M., and Temin, H.M., (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol, *69*, 5087–5094.

Maynard Smith, J.M., (1971). What use is sex?. J Theor Biol, *30*, 319–335.
Moran, P.A.P., (1958). A general theory of the distribution of gene frequencies. I. Overlapping generations. II. Non-overlapping generations. Proc Roy Soc London, *B 149*, 102–116.
Muller, H.J., (1932). Some genetic aspects of sex. Am Nat, *66*, 118–128.
Muller, H.J., (1964). The Relation of Recombination to Mutational Advance. Mutat Res, *106*, 2–9.
Neher, R.A., and Hallatschek, O., (2013). Genealogies of rapidly adapting populations. Proc Natl Acad Sci U.S.A., *110*, 437–442.
Neher, R.A., and Leitner, T., (2010). Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput Biol, *6*, e1000660.
Neher, R.A., Shraiman, B.I., and Fisher, D.S., (2010). Rate of adaptation in large sexual populations. Genetics, *184*, 467–481.
Nei, M., (1972). Genetic distance between populations. Am Nat, *106*, 283–292.
Nietfield, W., Bauer, M., Fevrier, M., Maier, R., Holzwarth, B., Frank, R., Maier, B., Riviere, Y., and Meyerhans, A., (1995). Sequence constraints and recognition by CTL of an HLA-B27-restricted HIV-1 gag epitope. J Immunol, *154*, 2189–2197.
Pennings, P.S., Kryazhimskiy, S., and Wakeley, J., (2014). Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet, *10*, e1004000.
Rice, W.R., (2002). Experimental tests of the adaptive significance of sexual recombination. Nat Rev, *3*, 241–251.
Rodrigo, A.G., and Felsenstein, J., (1999). Coalescent approaches to HIV population genetics, In Molecular evolution of HIV, K. Crandall, ed. (Baltimore: John Hopkins University Press).
Rodrigo, A.G., Shpaer, E.G., Delwart, E.L., Iversen, A.K., Gallo, M.V., Brojatsch, J., Hirsch, M.S., Walker, B.D., and Mullins, J.I., (1999). Coalescent estimates of HIV-1 generation time in vivo. Proc Natl Acad Sci U.S.A., *96*, 2187–2191.
Rouzine, I.M., Brunet, E., and Wilke, C.O., (2008). The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. Theor Popul Biol, *73*, 24–46.
Rouzine, I.M., and Coffin, J.M., (1999a). Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc Natl Acad Sci U.S.A., *96*, 10758–10763.
Rouzine, I.M., and Coffin, J.M., (1999b). Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus. J Virol, *73*, 8167–8178.
Rouzine, I.M., and Coffin, J.M., (2005). Evolution of human immunodeficiency virus under selection and weak recombination. Genetics, *170*, 7–18.
Rouzine, I.M., and Coffin, J.M., (2007). Highly fit ancestors of a partly sexual haploid population. Theor Popul Biol, *71*, 239–250.
Rouzine, I.M., and Coffin, J.M., (2010). Multi-site adaptation in the presence of infrequent recombination. Theor Popul Biol, *77*, 189–204.
Rouzine, I.M., Coffin, J.M., and Weinberger, L.S., (2014). Fifteen Years Later: Hard and Soft Selection Sweeps Confirm a Large Population Number for HIV In Vivo. PLoS Genet, *10*, e1004179.
Rouzine, I.M., Rodrigo, A., and Coffin, J.M., (2001). Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. Microbiol Mol Biol Rev, *65*, 151–185.
Rouzine, I.M., and Rozhnova, G., (2018). Antigenic evolution of viruses in host populations. PLoS Pathog, *14*, e1007291.
Rouzine, I.M., Wakeley, J., and Coffin, J.M., (2003). The solitary wave of asexual evolution. Proc Natl Acad Sci U.S.A., *100*, 587–592.
Rouzine, I.M., and Weinberger, L., (2013). The quantitative theory of within-host viral evolution [review]. J Stat Mech: Theory Exp, P01009.

Schiffels, S., Szollosi, G.J., Mustonen, V., and Lassig, M., (2011). Emergent neutrality in adaptive asexual evolution. Genetics, *189*, 1361–1375.

Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., *et al*. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol, *73*, 10489–10502.

Simmonds, P., (2004). Genetic diversity and evolution of hepatitis C virus–15 years on. J Gen Virol, *85*, 3173–3188.

Smith, D.J., Lapedes, A.S., de Jong, J.C., Bestebroer, T.M., Rimmelzwaan, G.F., Osterhaus, A.D., and Fouchier, R.A., (2004). Mapping the antigenic and genetic evolution of influenza virus. Science, *305*, 371–376.

Stephan, W., Chao, L., and Smale, J.G., (1993). The advance of Muller's ratchet in a haploid asexual population: approximate solutions based on diffusion theory. Genet Res, *61*, 225–231.

Stern, A., Bianco, S., Yeh, M.T., Wright, C., Butcher, K., Tang, C., Nielsen, R., and Andino, R., (2014). Costs and benefits of mutational robustness in RNA viruses. Cell Rep, *8*, 1026–1036.

Takahashi, H., Houghten, R., Putney, S.D., Margulies, D.H., Moss, B., Germain, R.N., and Berzofsky, J.A., (1989). Structural requirements for class I MHC molecule-mediated antigen presentation and cytotoxic T cell recognition of an immunodominant determinant of the human immunodeficiency virus envelope protein. J Exp Med, *170*, 2023–2035.

Tsimring, L.S., Levine, H., and Kessler, D., (1996). RNA virus evolution via a fitness-space model. Phys Rev Lett, *76*, 4440–4443.

Watterson, G.A., (1975). On the number of segregating sites in genetical models without recombination. Theor Popul Bio, *7*, 256–276.

Wei, X., Ghosh, S., Taylor, M.E., Johnson, V.A., Emini, E.A., Deutsch, P., Lifson, J.D., Bonhoeffer, S., Nowak, M.A., Hahn, B.H., *et al*. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. Nature, *373*, 117–122.

Wolfs, T.F., de Jong, J.J., Van den Berg, H., Tijnagel, J.M., Krone, W.J., and Goudsmit, J., (1990). Evolution of sequences encoding the principal neutralization epitope of human immunodeficiency virus 1 is host dependent, rapid, and continuous. Proc Natl Acad Sci U.S.A., *87*, 9938–9942.

Wolfs, T.F., Zwart, G., Bakker, M., Valk, M., Kuiken, C.L., and Goudsmit, J., (1991). Naturally occurring mutations within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology, *185*, 195–205.

Wrenbeck, E.E., Azouz, L.R., and Whitehead, T.A., (2017). Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nat Commun, *8*, 15695.

Wright, S., (1931). Evolution in Mendelian Populations. Genetics, *16*, 97–159.

Wright, S., (1945). The differential equation of the distribution of gene frequencies. Proc Nat Acad Sci, *31*, 382–389.

Xiao, Y., Rouzine, I.M., Bianco, S., Acevedo, A., Goldstein, E.F., Farkov, M., Brodsky, L., and Andino, R., (2017). RNA Recombination Enhances Adaptability and Is Required for Virus Spread and Virulence. Cell Host Microbe, *22*, 420.

De Gruyter Series in Mathematics and Life Sciences

Volume 7
George Dassios, Athanassios S. Fokas
Electroencephalography and Magnetoencephalography, 2020
ISBN 978-3-11-054583-8, ISBN (PDF) 978-3-11-054753-5, ISBN (EPUB) 978-3-11-054578-4

Volume 6
Piotr Biler
Singularities of Solutions to Chemotaxis Systems, 2019
ISBN 978-3-11-059789-9, ISBN (PDF) 978-3-11-059953-4, ISBN (EPUB) 978-3-11-059862-9

Volume 5
S. V. Masiuk, A. G. Kukush, S. V. Shklyar, M. I. Chepurny, I. A. Likhtarov
Radiation Risk Estimation. Based on Measurement Error Models, 2016
ISBN 978-3-11-044180-2, ISBN (PDF) 978-3-11-043366-1, ISBN (EPUB) 978-3-11-043347-0

Volume 4
Sergey Vakulenko
Complexity and Evolution of Dissipative Systems: An Analytical Approach, 2013
ISBN 978-3-11-026648-1, e-ISBN 978-3-11-026828-7

Volume 3
Zoran Nikoloski, Sergio Grimbs
Network-based Molecular Biology: Data-driven Modeling and Analysis, 2013
ISBN 978-3-11-026256-8, e-ISBN 978-3-11-026266-7

Volume 2
Shair Ahmad, Ivanka M. Stamova (Eds.)
Lotka-Volterra and Related Systems: Recent Developments in Population Dynamics, 2013
ISBN 978-3-11-026951-2, e-ISBN 978-3-11-026984-0

Volume 1
Alexandra V. Antoniouk, Roderick V. N. Melnik (Eds.)
Mathematics and Life Sciences, 2012
ISBN 978-3-11-027372-4, e-ISBN 978-3-11-028853-7

www.degruyter.com