Evolution of large populations under the joint action of deleterious and beneficial mutations

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

IN THE FACULTY OF SCIENCE

by

SONA JOHN

THEORETICAL SCIENCES UNIT
JAWAHARLAL NEHRU CENTRE FOR ADVANCED SCIENTIFIC RESEARCH
(A Deemed University)
Bangalore – 560 064

January 2017

To my Family and Friends

DECLARATION

I hereby declare that the matter embodied in the thesis entitled "**Evolution of large populations under the joint action of deleterious and beneficial mutations**" is the result of investigations carried out by me at the Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under the supervision of **Prof. Kavita Jain**, and that it has not been submitted elsewhere for the award of any degree or diploma.

In keeping with the general practice in reporting scientific observations, due acknowledgement has been made whenever the work described is based on the findings of other investigators.

Sona John

CERTIFICATE

I hereby certify that the matter embodied in this thesis entitled "**Evolution of large populations under the joint action of deleterious and beneficial mutations**" has been carried out by **Ms. Sona John** at the Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India under my supervision and that it has not been submitted elsewhere for the award of any degree or diploma.

Prof. Kavita Jain
(Research Supervisor)

Acknowledgements

First of all, I would like to thank my supervisor Prof. Kavita Jain for her earnest support and guidance throughout my Ph.D. work. She gave me my freedom and at the same time rendered enough help and motivation which enabled me to finish things on time. I really appreciate her dedication and hard working nature, and I wish to develop those qualities in me.

I would like to express my endless gratitude to my family, few teachers and close friends without whom I would never have accomplished my ambition. I am so proud of my parents, especially my mother who always supported me to pursue what I liked. I thank my sister Swapna and brother Jacob (Appu) for all the love and care extended to make our family complete.

I am thankful to all my teachers especially Dr. P. T. John, Dr. Rajan K John and Prof. Manoj Gopalakrishnan, for the motivation and guidance at the crucial times. All of them have encouraged me to aim big, and the interaction I had with them gave me the confidence to pursue my dreams. Also, they have given me the right advice at the needed occasions that made my journey ahead easier.

I would like to thank Prof. Amitabh Joshi, Prof. Shobhana Narasimhan, Prof. Subir K. Das, Prof. Vidhyadhiraja N. S., Prof. Umesh V. Waghmare, Prof. Srikanth Sastry, Prof. Swapan Pati and Dr. Meher K. Prakash for the courses offered and for the useful discussions carried out. I could learn a lot of new things from them. Being fresh to the field of evolutionary biology, AJ's course on population dynamics has helped me a lot in understanding the essential basics of the topic of my Ph.D. project.

Friends mean a lot to me, and I am very lucky to have a bunch of very good friends, who always dreamed of my success. Whenever I face difficulties, I had the freedom to seek their help and they were more than happy to help me and support me in all possible ways. I thank

all my close friends Dipu, Vybhav, Reenu, Krishnan, Jagadeesh, Ravi, Ambarish, Joji and Jithin, to name a few, for all their love and support.

During my Ph.D. period, I got the rare opportunity to work on two independent projects with my seniors Dr. Sarada Seetharaman and Dr. Himani Sachdeva. It was a very nice and valuable experience that made me learn a lot. I appreciate Himani's style of working and also her deep understanding of the subject that had helped me in having a clear picture of the problems. Also, I thank her for the effort she put in to make my writing of our unpublished results in this thesis, more clear and interesting.

Now I would like to thank my past and present lab mates Sarada, Priyanka, Ananthu, Jyoti, Tanmoy, and Archana. Lab environment was always very lively and friendly. I thank Priyanka for helping me with her awesome ability to find answers for each and every problem faced from google. Also, during our initial days of association in JNC, when we were relatively free, we had many trips together including visits to my home in Kerala that had left cheerful memories about my Ph.D. life.

I thank my seniors/friends Suman, Sutapa, and Shaista for the nice time spent together, with all the fun and enjoyment. I thank Sutapa for being a very caring friend and the best next room neighbor. Staying next to her room was a very nice experience, that gave a homely feel to my hostel life. I thank Suman for his wise tips to take life easily that had made many pestering problems look trivial in the end. Also, I thank Suman for being my breakfast partner for the initial 3 years that helped me to wake up early and to have a disciplined schedule.

Our lab always had a strong link with EMU. Our friends in EMU had often come to our help whenever we had problems with coding to install some software. I am especially thankful to a few people like Vybhav, Nandu, Deepak, Mahan, Ujjain, Siddharth and Milind for all their help with coding and also for allowing me to use their system to run codes. Also, I thank Dr. Rajesh Ranjan for this nice thesis template.

I am very thankful to my father for bearing with my careless writing and for all the corrections proposed. Also, I thank all my friends, Mahan, Ananthu, Archana, Arun, Avani and Manaswini who had helped me in checking my manuscripts and documents. Special thanks to Manaswini for the very dedicated reading of my introduction chapter and the useful suggestions.

I am very thankful to Sharanappa for making me cheerful whenever I am down, with a nice cup of coffee. He had also been a very patient Kannada teacher for me :).

Another group of people who became an important part of our lab group and claimed themselves as our secondary labmates are Meha, Vinay, Sankalp, Nikhil, Vicky, Deb, Anjali and Deepak. There is a bunch of other people like Praveen, Rafi, Chakri, Avani, Sharmila, Nisha, Shashank, Sunil, Sreelakshmi, Lakshmi, and Samrat whom Vybhav brought to my life along with him, in the last two years of Ph.D. We became very close in a short span of time. We had a lot of outing and cooking sessions together. I thank all of them for leaving with loving memories about my Ph.D. life in JNC. I also thank my other friends Simi, Priya, Ponnu, Arun, Keerthipriya, Nandini and all others whomsoever I might have inadvertently missed here.

I am also particularly thankful to all the staff in Canteen, Academic, Dhanvantari, Library, Complab and Admin. for all their help and support. Finally, I would like to thank JNC for funding and also for the lovely facilities provided in campus and hostel, including the beautiful and green surroundings, and for their proper maintenance and upkeep.

Abstract

The present state of life on earth is the outcome of millions of years of biological evolution. All organisms evolve to develop traits which make them better suited to their environment. Biological evolution is driven by several forces such as mutations (beneficial, deleterious and neutral), recombination, migration, genetic drift and natural selection. The main aim of this work is to understand the effect of beneficial mutations in the presence of deleterious mutations and other evolutionary forces. Many theoretical studies consider the effect of either one of these mutations only, but the combined effect of both beneficial and deleterious mutations is much less explored. However, it is important to take both into account because in a real biological system, they occur together.

In this thesis, we focus on two biological questions, namely, evolution of sex and recombination and dynamics of adaptation process in which beneficial mutations play a crucial role. A summary of the different models studied in this thesis is given in Table 1. The thesis is divided into six chapters as described below:

In Chapter 1, we introduce various evolutionary forces such as mutation, recombination, migration, genetic drift and natural selection. Two theoretical models (Wright-Fisher process and Moran process) used to study the role of these forces in evolution are discussed here. Further, different fitness landscapes considered in our study are also explained in this Chapter.

Recombination is very common in nature as a primary mechanism of reproduction. The reason why it is so widespread in spite of all its disadvantages is however not properly understood. Irreversible accumulation of deleterious mutations (Muller's ratchet)[1] in finite asexual populations is considered to be one of the reasons for the evolution of sex and recombination. But theoretical studies of Muller's ratchet [2] completely ignore the presence of beneficial

Beneficial and deleterious mutations	Drift	Fitness landscape	Recombination	Reference
Yes	Yes	Simple	Yes	[3]
Yes	No	Simple	No	[5]
Yes	Yes	Rugged	No	[6]
Yes	Yes	Rugged	Yes	[7]

Table 1 Summary of different models considered in this study.

mutations. In Chapter 2, we study the effect of beneficial mutations modeled as *back mutations* in halting the Muller's ratchet and attaining the equilibrium in finite populations. The main questions addressed in this study are: What is the equilibrium frequency of deleterious mutations when beneficial mutations are included? How does recombination help in reducing the mutational load in finite populations? Our results show that beneficial mutations allow the population to attain a nontrivial steady state unlike in the case of Muller's ratchet. The steady state fraction of deleterious mutations show a weak dependence on population size in the case of linked genome, which is very different from the exponential dependence predicted for an unlinked genome [3].

In Chapter 3, we explore the effect of beneficial mutations in a variant of the above mentioned model with a different mutational scheme. The assumption here is that the genome is infinitely long, and the mutation rates per genome are constant in the fitness space [4]. In this case, we obtain an exact solution for the steady state distribution of an asexual and infinitely large population. We show that this distribution is proportional to the Bessel function of the first kind [5] unlike the well known Poisson distribution [2] when beneficial mutations are neglected. We also numerically study the effect of genetic drift to find the critical population size (or, beneficial mutation rate) needed to attain a steady state in a population of finite size.

Unlike in the previous chapters, in Chapter 4, we consider a complex rugged fitness landscape which is biologically more realistic. In our previous study on a single-peak fitness landscape (discussed in Chapter 1)[3], we found an advantage of recombination in reducing the equilibrium mutational load. In this chapter, we study this effect on a class of complex fitness landscapes [8, 9]. Our study shows that recombination has a short term advantage when the population is starting from a maladapted state and the time scale over which the advantage persists is longer when the mutation rate is low and fitness landscape is maximally rugged [7].

In Chapter 5, we turn to a study of the adaptation dynamics of microbial populations to understand the details of the underlying distribution of beneficial fitness effects. Since beneficial mutations which drive adaptation are rare, they occur only in the tail of the fitness distribution which allows one to use extreme value statistics for the distribution of beneficial fitness effects (DBFEs). A previous study on adaptation dynamics in low mutation regime has identified a quantity, namely, the fitness advantage in successive adaptive steps that shows distinguishable trends for different DBFEs [10]. In Chapter 5, we study the robustness of this measure in high mutation regime and find that the qualitative behavior seen in the low mutation regime holds in the high mutation regime as well [6]. We also find that the rate of adaptation shows distinct trends for different DBFEs in both high and low mutation regimes.

Finally, in Chapter 6, we summarize our main results and discuss some interesting open problems related to our study.

List of Publications

[1] S. John and K. Jain. Effect of drift, selection and recombination on the equilibrium frequency of deleterious mutations. *J. Theor. Biol.*, 365:238–246, 2015.

[2] S. John* and S. Seetharaman*. Exploiting the adaptation dynamics to predict the distribution of beneficial fitness effects. *PLoS One*, 11(3):e0151795, 2016. (* equal contribution)

[3] K. Jain and S. John. Deterministic evolution of an asexual population under the action of beneficial and deleterious mutations on additive fitness landscapes. *Theor. Popul. Biol.*, 112:117-125, 2016.

[4] H. Sachdeva and S. John. (Dis-)Advantage of recombination on rugged fitness landscapes. *(in preparation)*.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Modeling biological evolution

Theoretical population genetics aims to understand the action of different evolutionary forces such as mutation, selection, recombination, migration and genetic drift in driving the process of evolution [14]. But, as the incorporation of all the complexity of a real biological system to a mathematical model will make it impossible to make any progress, it is very important to make appropriate simplifications that preserve the intricacies of the real system and at the same time simplifies the problem to a level where one can handle it mathematically or numerically. In this thesis, we assume that the evolutionary forces remain constant over time and the relevant evolutionary forces such as mutation and selection are modeled as simple one parameter process [14]. We will see in the coming Chapters that, even these simplified models are mathematically nontrivial but give important insights into the process of evolution. In this Chapter, we explain how we model the process of evolution under the action of different evolutionary forces.

1.2 Sequence space and fitness landscapes

We model a genome as a binary string with two allelic state 0 or 1. Then a genome of length L has 2^L possible configurations. The Hamming distance between two sequences is defined as the number of positions where these two genotypes differ. For $L = 2$, the four sequences can be represented as the vertices of a square. When $L = 3$, there are 8 possible configurations which form a cube with sequences of one Hamming distance away arranged as the first nearest neighbors and two Hamming distance as the second nearest neighbors and so on. For large values of L, the sequence space is a higher-dimensional hypercube.

Once the sequence space is generated, each genetic sequence can be assigned a particular fitness value, by measuring a fitness proxy such as cell size, drug resistance, or reproductive ability, to create the fitness landscape. In our study, we define fitness as the reproductive ability.

1.2.1 Single peak fitness landscape

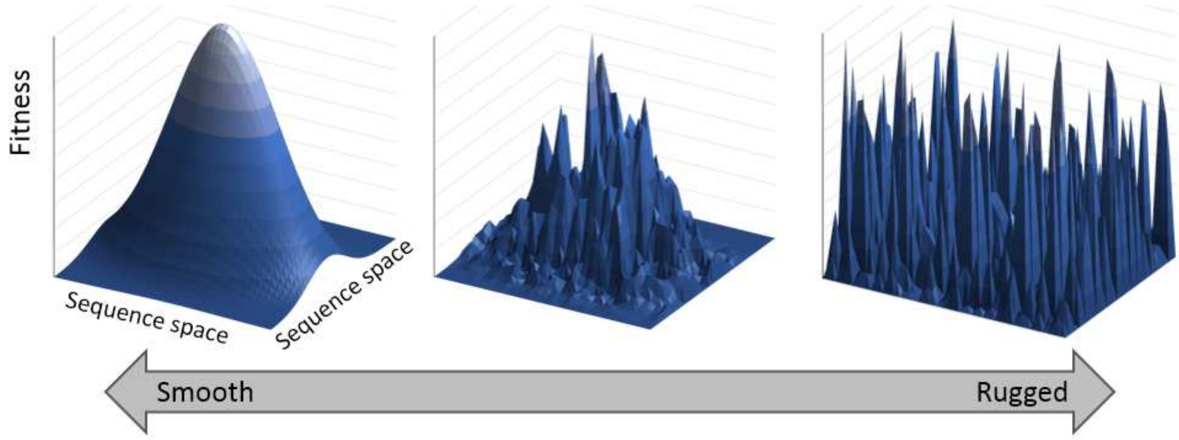
A single peak, non-epistatic fitness landscape can be created by assuming the fitness function to be

$$W(j) = (1-s)^j, \tag{1.1}$$

where j represents the number of deleterious alleles (ones) the genome carries. Such fitness landscape is relevant to a population growing in a medium which contains only one type of food source [15]. Even though such fitness landscapes are biologically unrealistic, mathematical model considers such simplified functions to get an insight into complex questions. We consider this fitness landscape in our study and the results obtained in Chapters 2 and 3 are based on this fitness landscape.

1.2.2 Maximally rugged fitness landscape

Experiments suggest that the fitness landscapes are quite complex and have many local fitness peaks [16–19]. Maximally rugged fitness landscapes can be generated in mathematical models by assigning fitness values randomly for each configuration from a fixed distribution. This distribution is chosen with reference to the experimental data obtained from the measurements of fitness landscapes. This type of fitness landscape is considered in Chapter 5 of this thesis.



Ref: commons.wikimedia.org/wiki/File:Epistasis_and_landscapes.png

Fig. 1.1 Schematic diagram of different types of fitness landscapes.

1.2.3 Tunable rugged fitness landscape: Rough Mount Fuji (RMF)

The maximally rugged and single peak fitness landscapes are two extreme cases, which are comparatively easy to handle mathematically. But, in more realistic scenario, the fitness landscapes show an intermediate level of ruggedness [9]. One example of such fitness landscape is the Rough Mount Fuji (RMF) model [20], which consists of a smooth component with a selection gradient c, along with a random component with amplitude B, which is drawn from the unit normal distribution. The fitness function for this fitness landscape is defined as:

$$W(\{\sigma\}) = cd_{\sigma,\sigma_*} + B\eta(\sigma). \tag{1.2}$$

Here, d_{σ,σ_*} is the Hamming distance between σ and the reference sequence $\sigma_* = (1,1,1,1...1)$, which is equal to the number of zeros in the sequence and η is a random variable drawn independently for each sequence from a normal distribution with mean zero and variance one. Thus, the RMF fitness landscape is parameterized by the ratio $\theta = B/c$. This type of fitness landscape is considered in Chapter 4 of this thesis. A two dimensional schematic presentation of all three types of fitness landscapes is shown in Fig. 1.1

1.3 Evolutionary forces

The process of evolution is driven by several evolutionary forces; we describe below the forces that are relevant to the discussion in this thesis.

1.3.1 Natural selection

Natural selection acts on the variation in the population and gives an advantage to the one with favourable traits by giving them a higher probability of reproducing. This will result in an increase in fraction of population with favourable traits and that in turn will result in an increase of high average fitness of the population. This helps the population to attain the fitness peak in the fitness landscape. This process is reasonably simple to understand and easy to implement in models by assigning the probability of reproduction as a function of relative fitness of the genotype. In this study, we consider only one form of selection, namely, directional selection that reduces the variation in a population. Other forces- mutation, recombination, and migration described below generate new genetic variation.

1.3.2 Mutation

Mutations are random changes that result in a new genotype. To understand the process, let us consider the simplest case of asexual reproduction by binary fission. During asexual reproduction, cell first makes a copy of its genome and then divides by giving each daughter cell one copy. In this simple process of reproduction, the only way to generate variation is through mutations. Three major sources of mutation are, errors in replication, errors in segregation of the replicated genome and genome modification by DNA damage [21].

We consider point mutations where the change occurs at one base pair in a single position in the DNA. If the change is from one purine (G and A) to another or one pyrimidine (C and T) to another it is called transition. If a purine changes to pyrimidine or vice versa it is called transversion [21]. Transversion can have a strong effect on phenotype, and results in serious diseases. Mutations based on their fitness effect on the phenotype are classified into two, synonymous mutations and nonsynonymous mutations. Synonymous mutations are

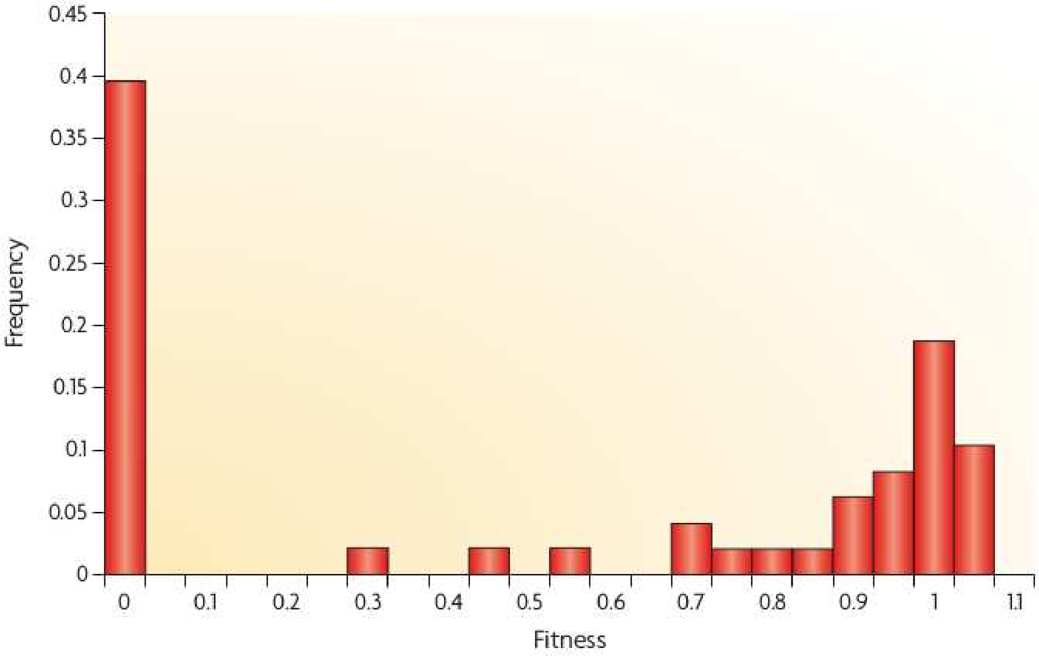


Fig. 1.2 Figure shows the distribution of fitness effects of mutations in *vesicular stomatitis virus*. In this experiment, random mutations were introduced into the virus and the fitness of each mutant was compared with the ancestral type. A fitness of zero, less than one, one, more than one, respectively, indicates that mutations are lethal, deleterious, neutral, and beneficial [11, 12].

nearly neutral and have a mild effect on fitness. One example of such mutations is the mutation between preferred and unpreferred codon. Nonsynonymous mutations are the ones which have larger fitness effects. These mutations are important in the study of adapting populations.

1.3.2.1 Frequency of mutations

Different types of mutations occur at different rates. For example, in point mutations, transitions occur at a much higher rate compared to transversions. Moreover different regions of the genome will also have different mutation rates. It is difficult to handle such complex scenario in a mathematical model. Here, we work with the average mutation rate per locus per generation for a particular genome for simplicity. Mutation rate varies for different species having a range from 10^{-10} to 10^{-3} [21] per bp per generation from humans to viruses. This includes all types of mutations with different fitness effects: deleterious ones that reduce fitness, beneficial mutations which increase fitness and neutral ones that have no fitness effect.

The mutation rates of nonsynonymous mutations which carry a fitness effect are highly skewed with more than 90% of them being deleterious [11], which can be seen from Fig. 1.2. This uneven distribution of mutations results in ignoring the presence of beneficial mutations in most of the theoretical studies as a first approximation. Even though the fraction of beneficial mutations is small, they play an important role in evolutionary dynamics. More evidence for the importance of this small fraction of mutations is reported in recent times [22]. As the title of this thesis suggests, our aim is to study the combined effect of both deleterious and beneficial mutations in evolution.

In Chapter 5, we study the distribution of beneficial fitness effects (DBFEs), that only include mutations which have fitness relative to the wildtype more than one in Fig. 1.2, to get a better understanding of adaptation dynamics of asexual populations. We discuss this point in detail in Section 1.5.2.

We have also considered nearly neutral or synonymous mutations, to understand the problem of codon usage bias. Here the rate of beneficial and deleterious mutations are of the same order. The results obtained in this study are discussed in Chapter 2.

1.3.2.2 Rates of mutations

When an error occurs in the proofreading mechanism of DNA, the mutation rate of the strain increases and can be as high as 1000 times that of the wildtype [23, 24]. During adaptation, such mutators hitchhike with the beneficial mutations they produce, and increase in frequency, which raises the mutation rate of the entire population. On the other hand, for adapted populations, higher mutation rate means a higher load of deleterious mutations, and hence, selection favours lower mutation rates, which is in agreement with many experimental [25] as well as theoretical [26, 27] findings.

1.3.3 Recombination

Genetic variation can also arise through recombination. In eukaryotes, this results from sexual reproduction and the formation of gametes which are genetically different from those united to reproduce. This could happen by independent segregation of nonhomologous chromosomes, or by crossing over between homologous chromosomes. Recombination or lateral gene transfer in bacteria also leads to mixing of genetic sequence from different lineages. Another example is the genetic exchange happening in virus genome when more than one virus strain infects the same host. The new viral particles produced will contain a mixture of chromosome from the two original strains. All these processes come under the class of recombination without sexual reproduction, and this process is equally effective in generating genetic variation [21].

Recombination is a much more complex process, in which there is mixing of two parent genomes in each generation to produce the offspring genome. This process is very efficient in creating large variation in a short time span of the order of one generation. It is evident from all higher organisms that has sex or recombination as the mechanism of reproduction, that they carry large amount of genetic variation compared to the asexual populations. This in turn helps selection to act more effectively and have faster adaptation. But, recombination is expensive and there are many 'costs' associated with it, which makes the evolution and widespread of recombination a puzzle. We discuss this point in detail in Section. 1.5.1.

1.3.4 Random genetic drift

Real populations have a finite size and this finiteness is responsible for stochastic evolution of a population. Random genetic drift is a very important force in evolution that decides the fate of a new genotype appearing in the population due to mutation or recombination. When a new genotype appears first in the population it will be in a small fraction, and the initial evolution of this genotype is mostly driven by drift. Even the beneficial (deleterious) mutant has a chance of getting lost (fixed) in the population because of the stochastic fluctuations in number.

The probability of fixation of a mutant under drift is given as [4]

(i) Deleterious mutations:

$$P_d = \frac{e^{2s}-1}{e^{2Ns}-1} \tag{1.3}$$

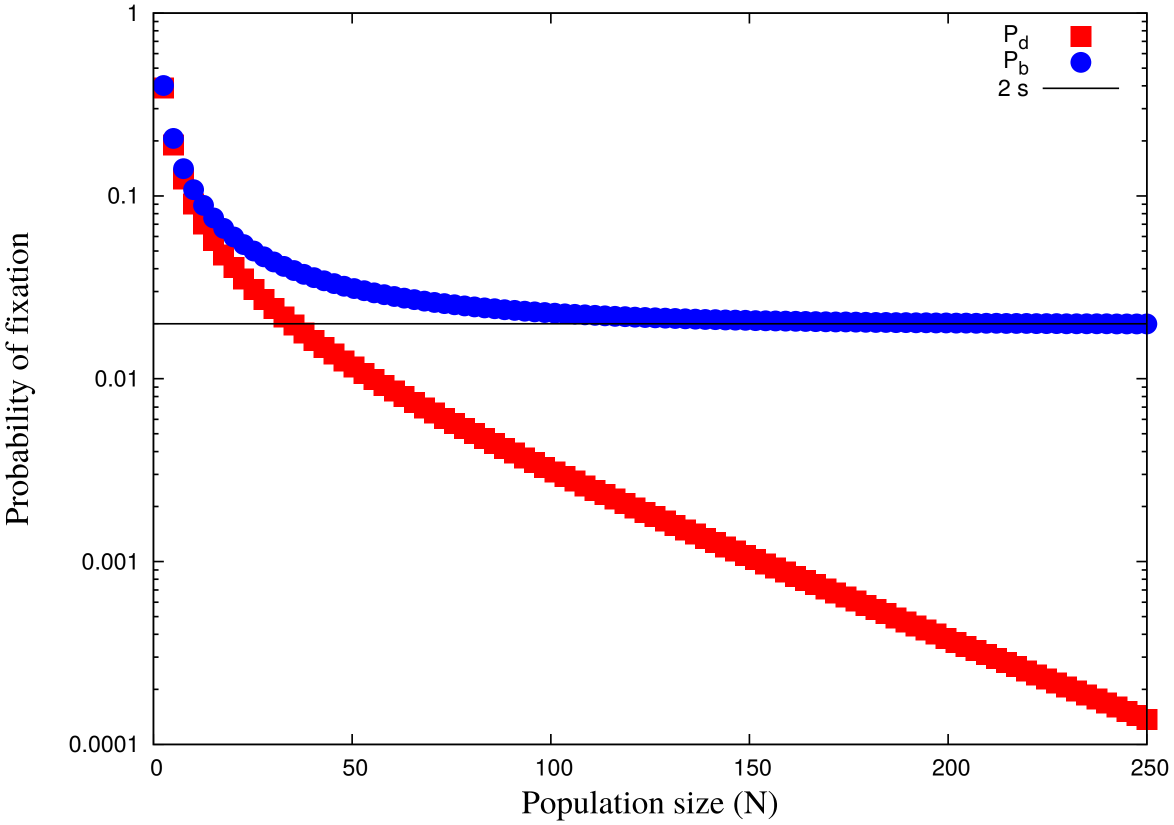


Fig. 1.3 Figure shows the fixation probability of beneficial (P_b) and deleterious (P_d) mutations with population size. Selection coefficient $s = 0.01$ is kept fixed.

(ii) Beneficial mutations:

$$P_b = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}. \tag{1.4}$$

The probability of fixation of the deleterious and beneficial mutations with population size is shown in Fig. 1.3. When N is small, deleterious mutants have a finite probability to get fixed in the population that goes to zero as N increases. Similarly, the fixation probability of a beneficial mutant will saturate to a constant equal to $2s$ as N increases.

1.4 Mathematical Models

Here we aim to include the evolutionary forces discussed above, such as selection, mutation, recombination, and random genetic drift into mathematical models, to study combined effect of these forces in driving the process of evolution.

1.4.1 Deterministic models

The assumption of $N \to \infty$, results in a deterministic model, that ignores the effect of random genetic drift. This is a major simplification, but it is useful because in some cases such models are exactly solvable and the answers obtained in this limit are useful in developing approximation for the finite-sized populations.

In this limit, we can write an equation for the change in the population fraction of a genotype σ at generation t to σ' at generation $t+1$, under the action of mutation and selection as [28],

$$X(\sigma', t+1) = \frac{\sum_\sigma M(\sigma' \leftarrow \sigma) W(\sigma) X(\sigma, t)}{\sum_\sigma W(\sigma) X(\sigma, t)}. \tag{1.5}$$

Here, $M(\sigma' \leftarrow \sigma)$ is the probability that a genotype σ mutate to genotype σ' in the next generation, and $W(\sigma)$ is the fitness of the genotype σ. The solution for the deterministic equation for two different mutation schemes and fitness function given in 1.1, corresponding to a single peak fitness landscape is discussed in Chapters 2 and 3.

1.4.2 Stochastic models

Here we consider two well known models, Wright-Fisher model, and the Moran model to study the evolution of a finite size population under the action of other evolutionary forces.

1.4.2.1 Wright-Fisher Model

In this model, generations are assumed to be discrete and nonoverlapping, i.e., all the individuals in the tth generation reproduce at the same time. In every generation, one individual is chosen with probability proportional to its relative fitness to reproduce. While reproducing it

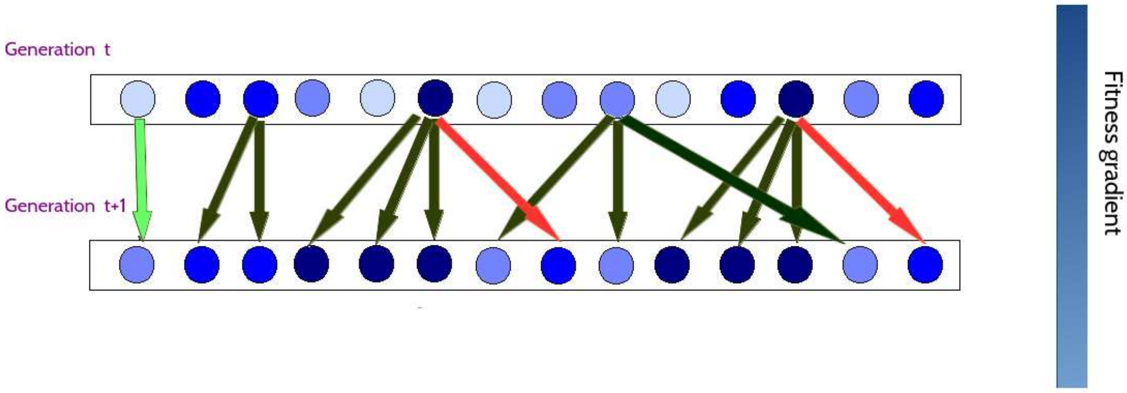


Fig. 1.4 Schematic diagram for Wright-Fisher process. Relative fitness of an individual is represented by the gradient of blue, with dark blue representing the fittest. Reproduction with mutations are shown by red and light green arrows, where the colours represent deleterious and beneficial mutations respectively. Reproduction without mutations is shown by dark green arrows.

can undergo mutations with probability specified by the parameters of the model. This process is repeated until the $(t+1)$th generation has N individuals. A schematic representation of this process is shown in Fig.1.4.

When the genome length is finite we can include recombination also. To implement recombination, we choose two parents at random, and with probability r, the genome of these two parents break and recombine to produce the offspring genome. The number of break points is chosen from a binomial distribution with mean $L\,r$, where L is the genome length. We use this protocol to study the effect of recombination, and the results obtained are discussed in Chapters 2 and 4.

1.4.2.2 Moran Model

The main difference between this model and the Wright-Fisher model is that- this model considers overlapping generations. Here an individual is chosen with probability proportional to its relative fitness to reproduce, and after reproduction to keep the population size constant another individual chosen at random is killed (replaced by the offspring). At the time of reproduction, the individual can be allowed to mutate as well. This model is described in a schematic diagram shown in Fig. 1.5.

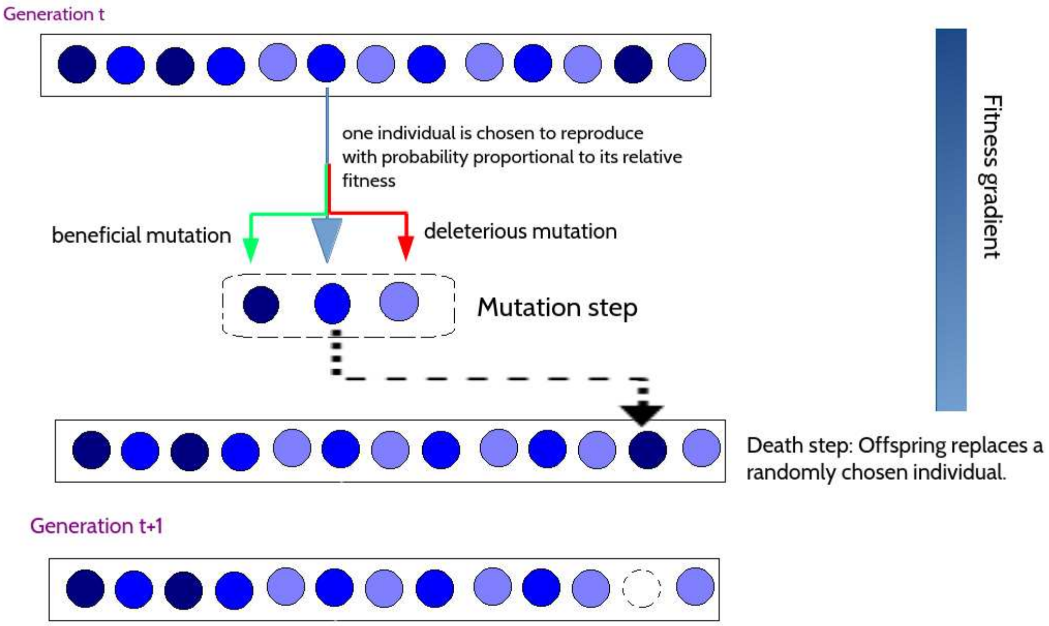

Generation t
one individual is chosen to reproduce with probability proportional to its relative fitness
beneficial mutation
deleterious mutation
Mutation step
Death step: Offspring replaces a randomly chosen individual.
Generation t+1
Fitness gradient

Fig. 1.5 Schematic diagram for Moran process. Relative fitness of an individual is represented by the gradient of blue with dark blue representing the fittest. At generation t, one individual chosen with probability proportional to its relative fitness to reproduce with mutations. Once the offspring is produced, it replaces a randomly chosen individual from the population to keep the population size constant. Offspring genotype will be one of the three possible types at the mutation step.

1.5 Questions we want to address

The main focus of this work is to understand the effect of beneficial mutations in the presence of other evolutionary forces. As already discussed in Section 1.3.2.1, beneficial mutations are rare and their effect is not completely understood. Here we want to study the combined effect of beneficial and deleterious mutations for two biological questions, namely the evolution of sex and recombination and the dynamics of adaptation.

1.5.1 Role of beneficial mutations in the evolution of sex and recombination

The ubiquity of sex and recombination is one of the major questions in evolutionary biology, which has attracted the attention of biologists over the past many decades [29–32]. Even though recombination is effective in creating variation, it has many costs associated with it, which include the time and energy required to find a partner, and the risks involved in sexual reproduction such as the transmission of disease and predation. Moreover, when only one sex is capable of reproduction, the reproductive output of the whole population is halved as compared to asexual populations [21]. Despite all these 'costs', why recombination is so common in nature has not been fully understood so far.

One major reason for the evolution of sex and recombination is believed to be an irreversible accumulation of deleterious mutations (also known as Muller's ratchet) in asexual populations [1]. However, Muller's ratchet ignores beneficial mutations and it is important to take them into account. In this thesis, we study the effect of beneficial mutations on Muller's ratchet, and ask if beneficial mutations can halt the ratchet. Also, we study the effect of recombination in a population evolving on single peak as well as rugged fitness landscapes. Chapters 2, 3 and 4 of this thesis aim to address this question, and a general conclusion of all the results obtained is discussed in Chapter 6.

1.5.2 Relation between adaptation dynamics and the distribution of beneficial fitness effects

Adaptation is the process in which organisms develop new traits which make them more favourable to survive in their environment. One example of evolution of such traits is the antibiotic resistance developed by bacteria. Since this process aims to improve the survival probability of an individual, it is obvious that it will be driven by beneficial mutations.

As we discussed in Section 1.3.2, the fractions of mutations that are beneficial are rare and occur at the extreme tail of the fitness distribution. This suggests that we can use the extreme value statistics to study the distribution of beneficial fitness effects (DBFEs). In Chapter 5, we find experimentally measurable quantities that can be used to predict the underlying beneficial fitness distribution. We found two quantities which show distinguishable trends for different DBFEs, and these results are discussed in Chapter 5.

Chapter 2

Effect of drift, selection and recombination on the equilibrium frequency of deleterious mutations

2.1 Introduction

In this Chapter, we study the effect of linkage in a finite size population evolving under the action of both ways mutation, selection, recombination and random genetic drift. This model is useful in addressing the question of evolution of sex and recombination, and also the linkage effect on codon usage bias [3].

A large number of population genetic studies assume one-way mutation- in some situations, beneficial mutations are neglected as they occur rarely [1, 33, 2, 34] while in adaptation studies, deleterious mutations are ignored as they are unlikely to fix under strong selection conditions [35, 36, 10]. The assumption of one-way mutation has an important effect on the nature of the state at large times. If the population size is infinite, a time-independent stationary state can be reached due to a balance between mutation and selection even if the mutational forces are unidirectional [2]. However in a finite population, when mutations are completely neglected or only unidirectional mutations are allowed, a population evolving under the influence of other evolutionary forces either does not reach an equilibrium state [2], or achieves a trivial one in which one of the variants gets fixed at large times [37]. It is when both beneficial and deleterious mutations are taken into account, a finite population reaches a nontrivial stationary state [38].

An example of such a steady state is seen in the context of synonymous codons that represent the same amino acid but do not occur in equal frequencies [39, 40]. In a gene coding for a two-fold degenerate amino acid, while selection favors the preferred codon, reversible mutations between preferred and unpreferred codons and random genetic drift maintain the unpreferred one [41, 42]. Assuming that the sites in the sequence evolve independently, analytical results for the equilibrium frequency of unpreferred codons have been obtained [41–43]. However as the evolutionary dynamics at a genetic locus are affected by other loci [44], a proper theory of codon usage bias must account for the Hill-Robertson interference between sequence loci [45–48].

Reverse and compensatory mutations have been proposed as a possible mechanism to stop the degeneration of asexual populations [49–51]. In a finite nonrecombining population, if beneficial mutations are completely ignored, deleterious mutations accumulate irreversibly due to stochastic fluctuations by a process known as Muller's ratchet [1, 52]. But when rare beneficial mutations are taken into account, the population reaches an equilibrium [53, 22, 52]. The model studied here accounts for the reversibility of nucleotide substitutions. Recently [51] calculated the amount of beneficial mutations required to achieve a stationary state. But these authors assumed the mutation rates to be independent of the fitness, contrary to experimental evidence [22]. Moreover their solution for the equilibrium frequency can become negative in some parameter range.

In this Chapter, we are interested in understanding the stationary state of a multilocus model, which is described in detail in the following section. We consider a class of non-epistatic fitness landscapes where the fitness depends only on the number of deleterious mutations in a sequence (*fitness class*). As in previous works [41, 45, 46], we assume that the beneficial mutations are back mutations, the probability of whose occurrence depends on the fitness class. More precisely, if the mutation probability per site is small, the total probability of a beneficial (deleterious) mutation increases (decreases) linearly with the fitness class. We consider the evolution of both infinitely large and finite populations, and to analyse the effect of linkage amongst the loci, we allow recombination to occur. We are primarily interested in the population size dependence of the average number of disadvantageous mutations at equilibrium. We obtain analytical results when the sites are completely linked, and compare them with the known results for a freely recombining population. For intermediate recombination rates, we obtain numerical results.

We find that the number of deleterious mutations decreases in a reverse sigmoidal fashion, as the population size is increased. For small populations, the fraction of disadvantageous mutations is seen to be roughly independent of population size and recombination rate. An understanding of this behavior is obtained from an exact solution and numerical simulations for a neutral finite population. For very large populations that can be described by a deterministic model, we find the stationary state exactly which is also unaffected by recombination. However for moderately large populations, recombination is found to alleviate the effect of deleterious mutations [44, 33, 54, 47], and the extent to which it does so depends on the beneficial mutation rate relative to the deleterious one. We find that when beneficial mutations are rare, the equilibrium frequency of disadvantageous mutations decreases logarithmically with population size when the loci are completely linked, but exponentially fast when linkage is absent. On the other hand, when disadvantageous mutations are rare, the deleterious mutation fraction drops exponentially fast, irrespective of the recombination rate. Thus we expect that the linkage has a weak effect on codon bias where the rates at which mutations between preferred and unpreferred codons occur are of the same order [55, 56]. But in adapting microbial populations where beneficial mutations are rare [57], recombination may be expected to reduce the frequency of disadvantageous mutations significantly.

2.2 Models

We consider a haploid population of size N in which each individual carries a diallelic (either zero or one) sequence of finite length L, where zero represents the wildtype allele and one denotes the deleterious mutation. The population is evolved in computer simulations using a

Wright-Fisher process which is described in Chapter 1, with recombination followed by mutation and selection occurs in discrete, non-overlapping generations. To create an offspring, two parent individuals are chosen at random with replacement. With probability $r \leq 1/2$, a single crossover event occurs in the parent sequences at one of the $L-1$ equally likely break points to form two recombinant sequences, while with probability $1-r$, the parent sequences are copied to the offspring sequences. In either case, one of the offspring is chosen with probability half to undergo mutations and selection, and the other one is discarded. In the offspring sequence, a deleterious mutation occurs at a locus with a wildtype allele with probability μ and a reverse beneficial mutation on mutant allele with probability ν. The resulting sequence is allowed to survive with a probability equal to its fitness, where the fitness of a sequence with j deleterious mutations is assumed to be a nonepistatic, and given by $w(j) = (1-s)^j$, $0 \leq s < 1$. This corresponds to the single peak fitness landscape discussed in the Chapter 1.

We have been able to implement the procedure described above for sequences of length up to 500 and population sizes of the order 10^3. For larger populations with long nonrecombining sequence, the computational difficulties were overcome by tracking only the number of deleterious mutations (fitness class) carried by the individual since the fitness of a sequence depends only on the number of deleterious mutations in the sequence. Here a parent chosen at random produces a clone of itself, and the offspring may undergo mutations with a probability that depends on its fitness class. In a sequence with j deleterious mutations, as a deleterious (beneficial) mutation can happen at any one of the $L-j$ (j) sites, the rate of deleterious and beneficial mutations is given by $(L-j)\mu$ and $j\nu$ respectively. To find the number of beneficial (b) and deleterious (d) mutations acquired by the offspring, random variables were drawn from Poisson distribution with mean $j\nu$ and $(L-j)\mu$ respectively. The total number of deleterious mutations in the offspring is then given by $j' = j+d-b$. If j' turns out to be greater than L or less than zero, the offspring individual is produced with $j' = j$ mutations. As before, the offspring is allowed to survive with probability $w(j')$, and the process is repeated until N individuals in the next generation are obtained.

All the numerical results presented here are obtained with an initial condition in which none of the individuals carry deleterious mutations. In each stochastic run, the Wright-Fisher process was implemented for about 10^4 generations and it was ensured that the stationary state is reached. In the equilibrium state of each run, we measured the number of deleterious mutations present in the population and averaged them over another 10^4 generations. The data were also averaged over 100 independent stochastic runs. Although all the simulation results presented here are obtained using the Wright-Fisher process, we will also use a continuous time Moran model for some analytical calculations which is described in a later section. If the population is infinitely large, the dynamics and equilibrium state of the population fraction can be described by a deterministic equation, which we discuss next.

2.3 Infinite Population

2.3.1 Nonrecombining population

For small selection coefficient and mutation rates, the population fraction $X(j,t)$ in the jth fitness class at time t evolves in continuous time according to

$$\begin{aligned} \frac{\partial X(j,t)}{\partial t} &= -(sj+\overline{w}(t))X(j,t)-[(L-j)\mu+j\nu]X(j,t) \\ &+ (L-j+1)\mu X(j-1,t)+(j+1)\nu X(j+1,t)\ ,\ 0\leq j\leq L \end{aligned} \tag{2.1}$$

where $\overline{w}(t)=\sum_{k=0}^{L}\ln w(k)\ X(k,t)\approx -s\sum_{k=0}^{L}k\ X(k,t)$ is the average Malthusian fitness and $X(-1,t)=X(L+1,t)=0$ at all times. In the above equation, the first term on the right hand side (RHS) represents the contribution to the change in $X(j,t)$ due to reproduction and the second term gives the loss in the population fraction due to mutations. The last two terms are the gain terms due to deleterious and beneficial mutations respectively. The dynamics and the steady state solution of the deterministic model defined by (2.1) can be found exactly. Below we discuss the stationary state and refer the reader to Appendix A.1 for the time-dependent solution.

In the steady state, the left hand side (LHS) of (2.1) equals zero and the population fraction carrying j deleterious mutations is of the following product form [58]:

$$X(j) = \binom{L}{j} x^j \, (1-x)^{L-j} \tag{2.2}$$

On using the above ansatz in (2.1) for $j = 0$ and L, we find that the average fitness $\bar{w} = L(\nu\tilde{x} - \mu)$ where $\tilde{x} = x/(1-x)$ is a solution of the following quadratic equation:

$$\nu\tilde{x}^2 + (s + \nu - \mu)\tilde{x} - \mu = 0 \tag{2.3}$$

Plugging the ansatz (2.2) in the bulk equations corresponding to $j = 1, \ldots, L-1$ and rearranging the terms, we get

$$j\left(\mu - s - \nu + \mu\tilde{x}^{-1} - \nu\tilde{x}\right) - \bar{w} + L(\nu\tilde{x} - \mu) = 0 \tag{2.4}$$

which, by virtue of the results obtained above, shows that the ansatz (2.2) is consistent with the bulk equations. Since the population fraction must be positive, the allowed solution of (2.3) gives the fraction x to be

$$x = \frac{2\mu}{\mu + \nu + s + \sqrt{(s + \nu - \mu)^2 + 4\mu\nu}} \tag{2.5}$$

Furthermore, as the RHS of (2.2) is a binomial distribution, the average fraction of deleterious mutations defined as $q = \bar{j}/L = \sum_{j=0}^{L} jX(j)/L$ equals x.

To get some insight in the solution obtained above, we first consider some special cases by setting one of the parameters equal to zero.

(i) In the absence of selection ($s = 0$), we get

$$\begin{aligned} X(j) &= \binom{L}{j}\left(\frac{\mu}{\mu+\nu}\right)^j \left(\frac{\nu}{\mu+\nu}\right)^{L-j} && (2.6)\\ q &= \frac{\mu}{\mu+\nu} && (2.7) \end{aligned}$$

(ii) When the reverse mutation probability ν equals zero, the fraction $x = \mu/s$, $\mu < s$ and therefore

$$X(j) = \binom{L}{j} \left(\frac{\mu}{s}\right)^j \left(1 - \frac{\mu}{s}\right)^{L-j} \; , \; \mu < s \tag{2.8}$$

while for $\mu > s$, the fraction $X(j) = \delta_{j,L}$, thus signaling the well known error threshold transition [59]. On the other hand, if the probability μ is zero, we have the trivial solution that the fitness class with zero deleterious mutations has frequency one, for all ν.

When all the three parameters are nonzero and the sequence length is large, the following cases may be considered [60]:

1. If μ, ν, s are kept fixed but the sequence length is increased, we find that the population fraction of deleterious mutations is a Gaussian centred about the average number Lx.

2. If the deleterious mutation rate per genome $U_d = L\mu$ is held fixed while $\mu \to 0, L \to \infty$, the fraction $x \approx \mu/(s+\nu)$ approaches zero for finite ν and s. In this limit, the population fraction is a Poisson distribution given by [61]

$$X(j) = e^{-\frac{U_d}{s+\nu}} \frac{1}{j!} \left(\frac{U_d}{s+\nu}\right)^j \tag{2.9}$$

3. However when both $\mu, \nu \to 0$ and $L \to \infty$ such that the product $U_d = L\mu, U_b = L\nu$ remains finite, taking $\nu \to 0$ in (2.9), we immediately find that the population fraction is *independent* of the beneficial mutation rate. To understand this rather surprising result, we first note that when beneficial mutations are completely absent, due to (2.8), the mean number of deleterious mutations is of order unity *i.e.* it does not increase with L. However when beneficial mutations are present, the average number of advantageous mutations that can occur is $\sim \bar{j}\nu$ which approaches zero as $\nu \to 0$, and thus the population remains unaffected by beneficial mutations.

2.3.2 Recombining population

So far, we discussed the stationary state of the deterministic model when recombination is absent. But in an infinitely large population, if epistasis is absent (as is the case here), the linkage disequilibrium (LD) stays at its initial value [62]. Since we start with an initially monomorphic population with zero LD, the results obtained above are expected to hold in a recombining population as well. In fact, when the sequence loci are completely unlinked ($r = 1/2$) [9], [42] has shown that the average fraction of deleterious mutations is given by (2.5).

2.4 Finite population without selection

2.4.1 Nonrecombining population

We consider a neutral Moran process for an asexual population of finite size with a mutation scheme which is more general than that described in Section 2.2. In this model, a parent is randomly chosen with replacement to replicate. If the offspring has j mutations relative to the wildtype, the number of mutations increases (decreases) by one with probability μ_j (ν_j) and remains unchanged with probability $1-\mu_j-\nu_j$. It is obvious that $\mu_L = \nu_0 = 0$. An individual in the parent population is then randomly chosen to die and is replaced by the possibly mutated offspring. As explained in the Appendix A.2, the average number $\bar{n}(j)$ of individuals carrying j mutations evolves according to (A.14). In the stationary state, we obtain

$$\frac{\bar{n}(j)}{N} = \frac{1}{1+\sum_{k=1}^{L}\prod_{i=0}^{k-1}\frac{\mu_i}{\nu_{i+1}}}\prod_{i=0}^{j-1}\frac{\mu_i}{\nu_{i+1}} \tag{2.10}$$

which is *independent* of the population size.

For the model with back mutations, as explained in Section 2.2, the probability $\mu_j = (L-j)\mu$ and $\nu_j = j\nu$. Using this in the above equation, we find that the average population fraction carrying j mutations is given by the deterministic solution (2.6) and the average fraction $q = \bar{j}/L$ by (2.7), where $\bar{j} = N^{-1}\sum_{j=0}^{L} j\bar{n}(j)$. These results are verified in numerical simulations of the Wright-Fisher process and are shown in Fig. 2.1.

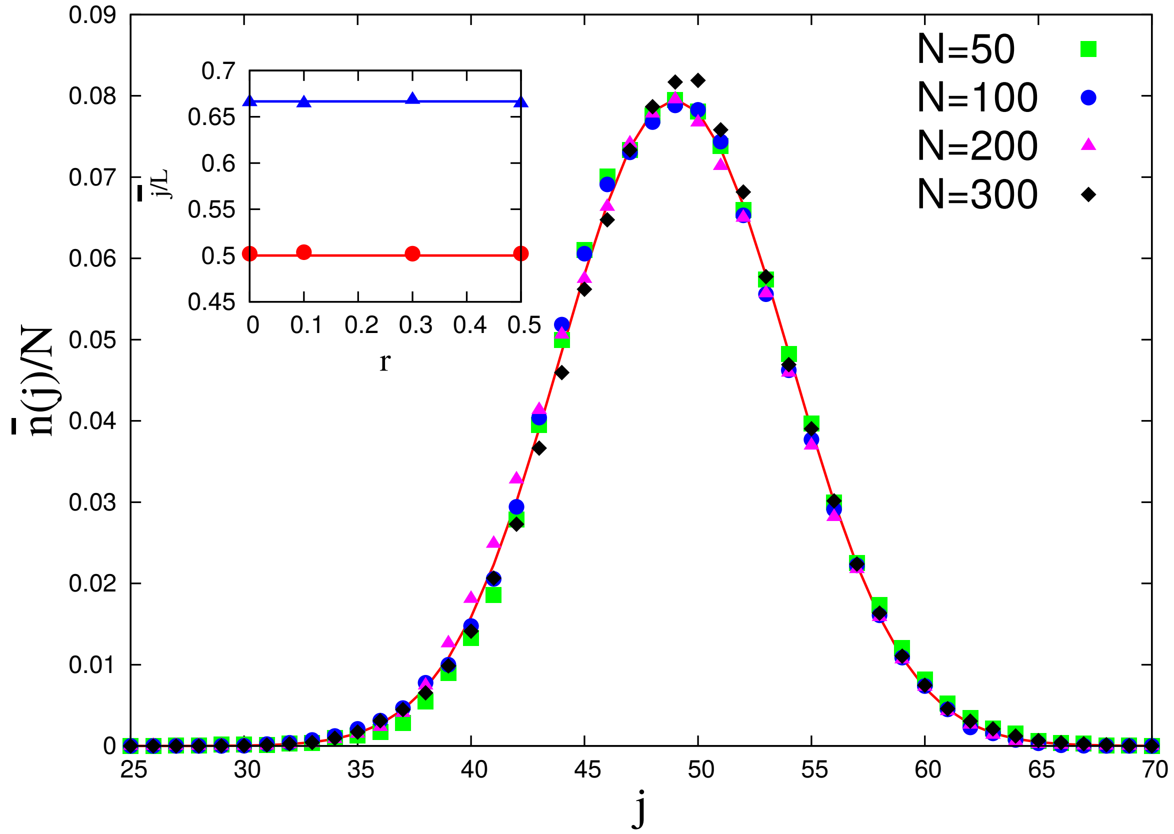


Fig. 2.1 Neutral case: Main figure shows the steady state fraction $\bar{n}(j)/N$ in the mutant class j with $\mu = 4.9 \times 10^{-5}$, $\nu = 5.1 \times 10^{-5}$ and $L = 100$. The distribution is independent of the population size N, and matches with the deterministic solution (2.6) shown by solid line. The inset shows the average fraction of mutations relative to the wildtype as a function of recombination probability r for $L = 300$, $N = 300$ and $\mu = 10^{-4}$ when $\nu = \mu$ (●) and 0.5μ (▲). The solid lines give the theoretical prediction (2.7).

2.4.2 Recombining population

When the recombination probability is equal to half, as the sequence loci evolve independently, the results from single locus theory are expected to hold. In this case, the frequency of mutations is given exactly by [38, 63]

$$\bar{j}_1 = \frac{\mu}{\mu + \nu} \tag{2.11}$$

Thus the average number of mutations in the two limiting cases, namely for a nonrecombining population ($r = 0$) and a freely recombining one ($r = 1/2$), is same. Furthermore, the results of our numerical simulations displayed in the inset of Fig. 2.1 for $0 \leq r \leq 1/2$ show that the average fraction q is independent of the recombination probability.

2.5 Finite population under selection

2.5.1 Effect of sequence length

Our numerical simulations show that, unlike in the deterministic case, the fraction of deleterious mutations initially varies with the sequence length and approaches a constant value for long enough sequences. Motivated by the discussion for the deterministic model, we consider the three cases when the sequence length is large.

1. The limit in which μ, ν, s are kept fixed but the sequence length is increased has been studied in previous works to gauge the effect of Hill-Robertson interference on the fraction of deleterious mutations [45, 48] and to understand the effect of nonrecombining regions of different lengths in the genome of various species [45, 64]. Here for a given Ns, the average fraction of deleterious mutations is found to increase with increasing sequence length, but saturates to a finite constant smaller than unity for long sequences. Our simulation data for minimum number of deleterious mutations shown in Fig. 2.2 is also consistent with this observation.

2. When U_d and ν are kept finite and sequence length is increased, our simulations show that for long enough sequences, the average number of deleterious mutations $\bar{j}$ is a constant, as in the deterministic model.

3. In the rest of the Chapter, we will consider the biologically relevant limit in which the genome mutation rates U_b and U_d remain finite, as the number of loci in the sequence is increased [65]. We find that unlike in the deterministic case, here the average *fraction* of deleterious mutations is finite and sensitive to the beneficial mutation rate. Figure 2.3 shows that the fraction $\bar{j}$ decreases to a constant value, as the sequence length is increased. The data shown in the other figures of this Chapter refers to this large-L limit.

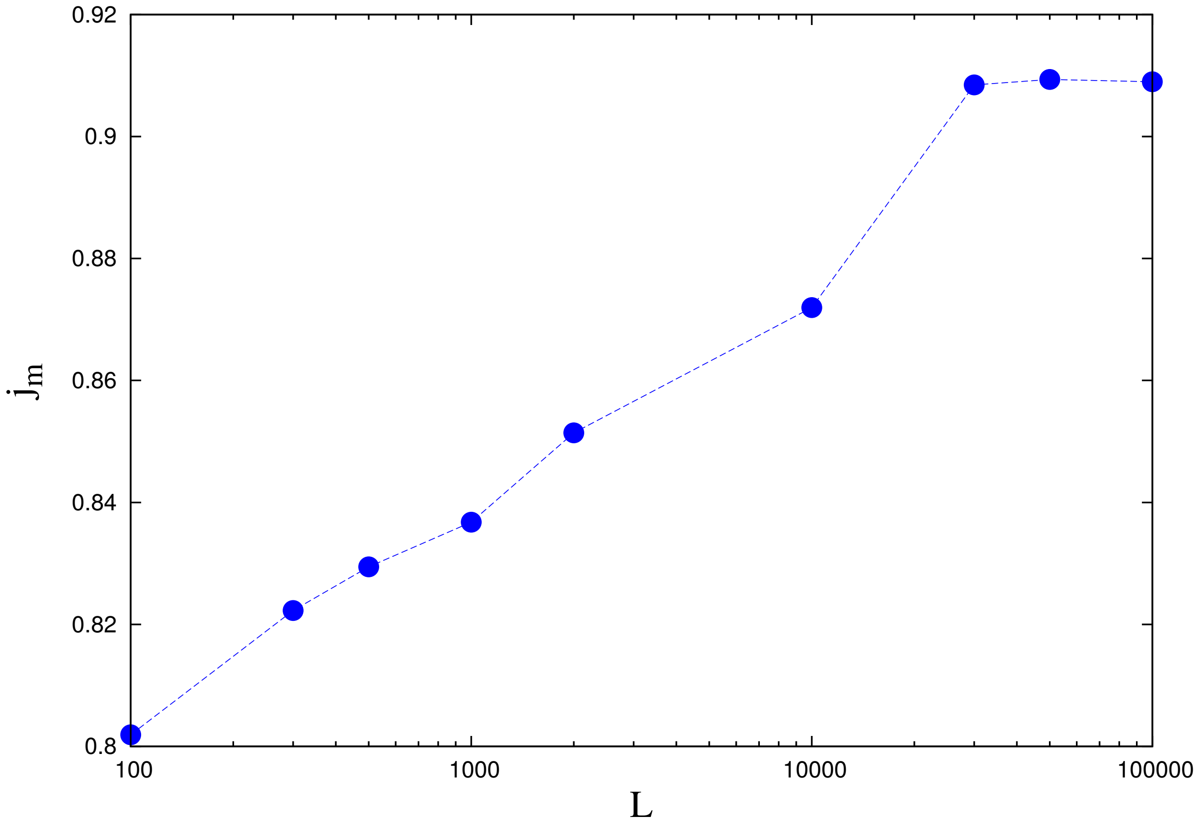


Fig. 2.2 Figure shows the minimum fraction of deleterious mutations with genome length L. Other parameters are $N = 100, \mu = 0.008, \nu = 0.00008$ and $s = 0.05$.

2.5.2 Nonrecombining population

The neutral Moran model described in the last section can be straightforwardly generalised to include selection, but we find that the evolution equation for the average number distribution $\bar{n}(j)$ does not close in the presence of selection *i.e.* it involves quantities that can not be expressed in terms of $\bar{n}(j)$. Therefore to understand the population size dependence of the average frequency q of deleterious mutations, we use the results obtained in the last two sections, and employ an analytical argument which is described below.

2.5.2.1 Small and very large populations

Figures 2.4 and 2.5 show that the fraction of disadvantageous mutations decreases monotonically with the population size N. When the selection is weak ($Ns \ll 1$), the fraction q is

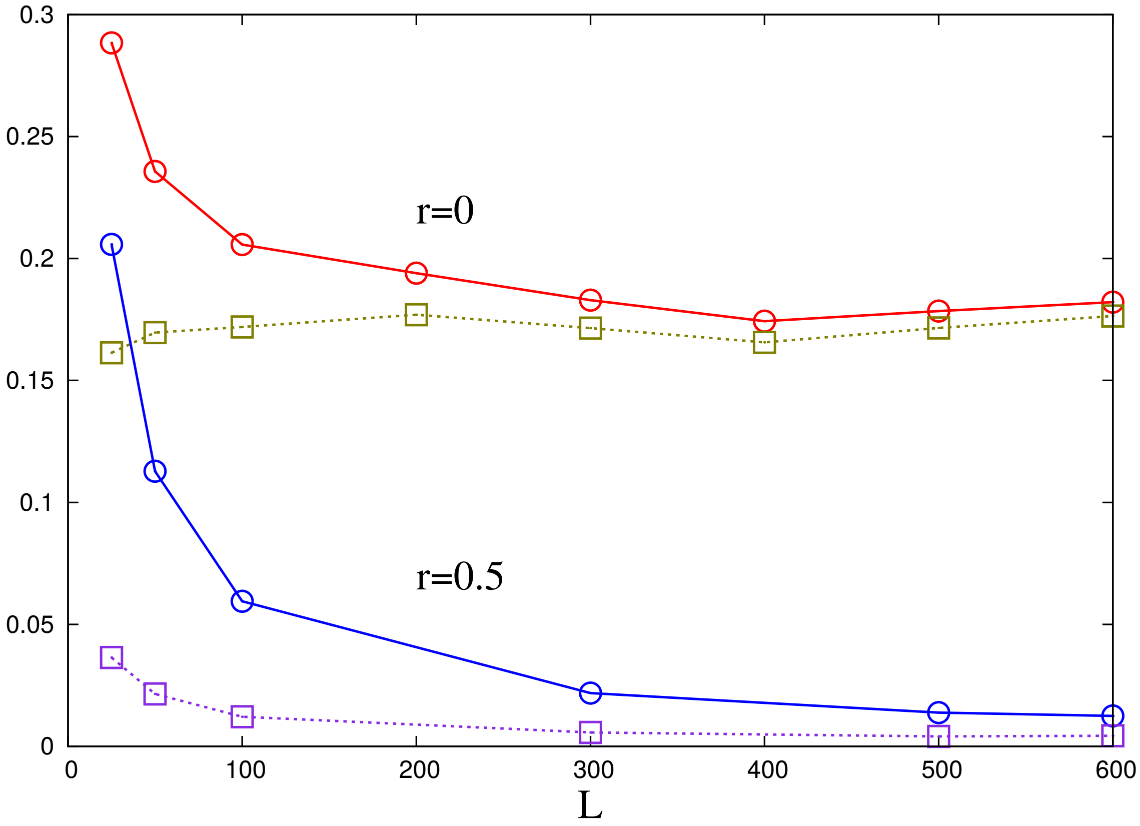


Fig. 2.3 Variation of minimum number j_m (broken line with □) and average number $\bar{j}$ (solid line with ○) of deleterious mutations with sequence length L for $N = 200$, $s = 2 \times 10^{-2}$, $U_d = 10^{-1}$ and $U_b = 5 \times 10^{-2}$.

expected to be close to the neutral value (2.7), in agreement with the data in Figs. 2.4 and 2.5. For very large populations, the deterministic solution (2.5) is expected to hold, and Fig. 2.5 clearly shows that this expectation is borne out by numerical simulations.

2.5.2.2 Moderately large populations

We now discuss a rate matching argument that allows us to find the *minimum* number j_m of deleterious mutations in the population. The basic idea is that if beneficial mutations are neglected, due to stochastic fluctuations, all the individuals in the least-loaded fitness class j_m will acquire deleterious mutations and it will get lost from the population at a degeneration rate $r_{j_m}^{-}$ [1, 2]. However due to beneficial back mutations, this process can be reversed and the population in the fitness class j_m will be regenerated at a rate $r_{j_m}^{+}$. In the stationary state,

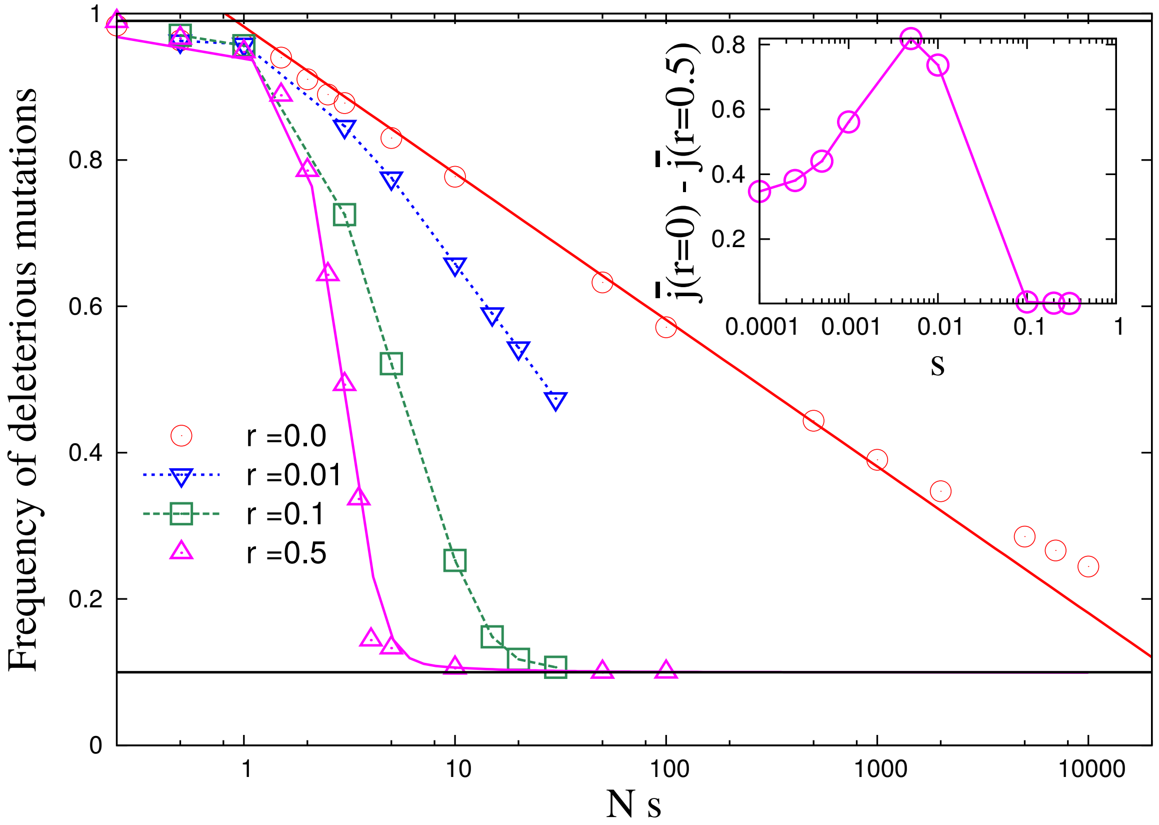


Fig. 2.4 Directional selection and rare beneficial mutations: Figure shows the average frequency of disadvantageous mutations as a function of population size when $U_d = 10^{-1}$, $U_b = 10^{-3}$, $s = 10^{-2}$ and $L = 100$. For nonrecombining population, the line shows the best fit curve $0.087 \ln(Ns) + 0.98$ to the numerical data and for freely recombining population, (2.17) is shown. The broken lines joining the numerical data for $r = 10^{-2}$ and $r = 10^{-1}$ are a guide to the eye. The solid line at the bottom is the deterministic expression (2.5) and the one at the top shows the prediction (2.7) from the neutral theory. The inset shows the nonmonotonic behavior of the difference between the deleterious mutations in a nonrecombining and freely recombining population for $N = 1000, L = 100, U_d = 10^{-1}$ and $U_b = 10^{-3}$.

on equating these two rates, the least-loaded fitness class j_m can be found [51]. The variation of these rates with the fitness class is shown in the inset of Fig. 2.6, and we observe that with increasing number of deleterious mutations, the degeneration rate decreases while the regeneration rate increases. This is a direct consequence of the fact that for the fitness-dependent mutation scheme considered here (refer Section 2.2), the total deleterious mutation rate $(L-j)\mu$ decreases with increasing j, but the beneficial mutation rate $j\nu$ decreases with decreasing j.

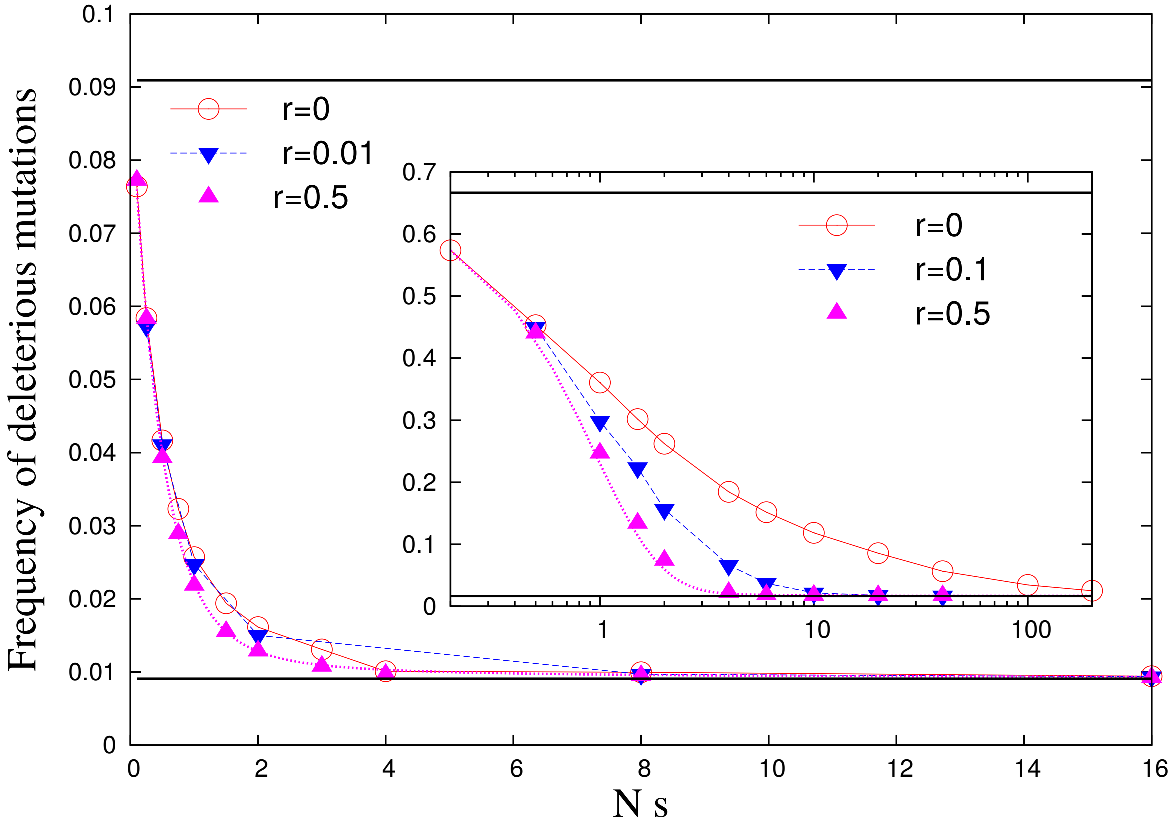


Fig. 2.5 Directional selection and frequent beneficial mutations: Figure shows the average frequency of disadvantageous mutations as a function of population size when $L = 100, s = 10^{-2}, U_d = 10^{-2}, U_b = 10U_d$ (main) and $L = 300, s = 2\times 10^{-2}, U_d = 10^{-1}, U_b = 0.5U_d$ (inset). For freely recombining population, (2.17) is shown while rest of the curves are a guide to the eye. The solid line at the bottom is the deterministic expression (2.5) and the one at the top shows the prediction (2.7) from the neutral theory.

In the absence of beneficial mutations, as shown in Appendix A.3, the average number of individuals in the least-loaded fitness class J is given by $n_J = NX_J^{(0)}(J) = N(1-\mu/s)^{L-J}$ which grows exponentially with J. As a result, an initially fast-clicking ratchet with $n_J s \ll 1$ crosses over to a slow-clicking ratchet with $n_J s \gg 1$, when $n_J s$ is of order unity [2, 66]. Using a diffusion theory for the slow ratchet [67, 34, 66], we find that when $n_J \gg 1$, the degeneration rate is given by

$$r_J^- = \sqrt{\frac{NX_J^{(0)}(J)c^3s^3}{\pi}}\, e^{-csNX_J^{(0)}(J)} \tag{2.12}$$

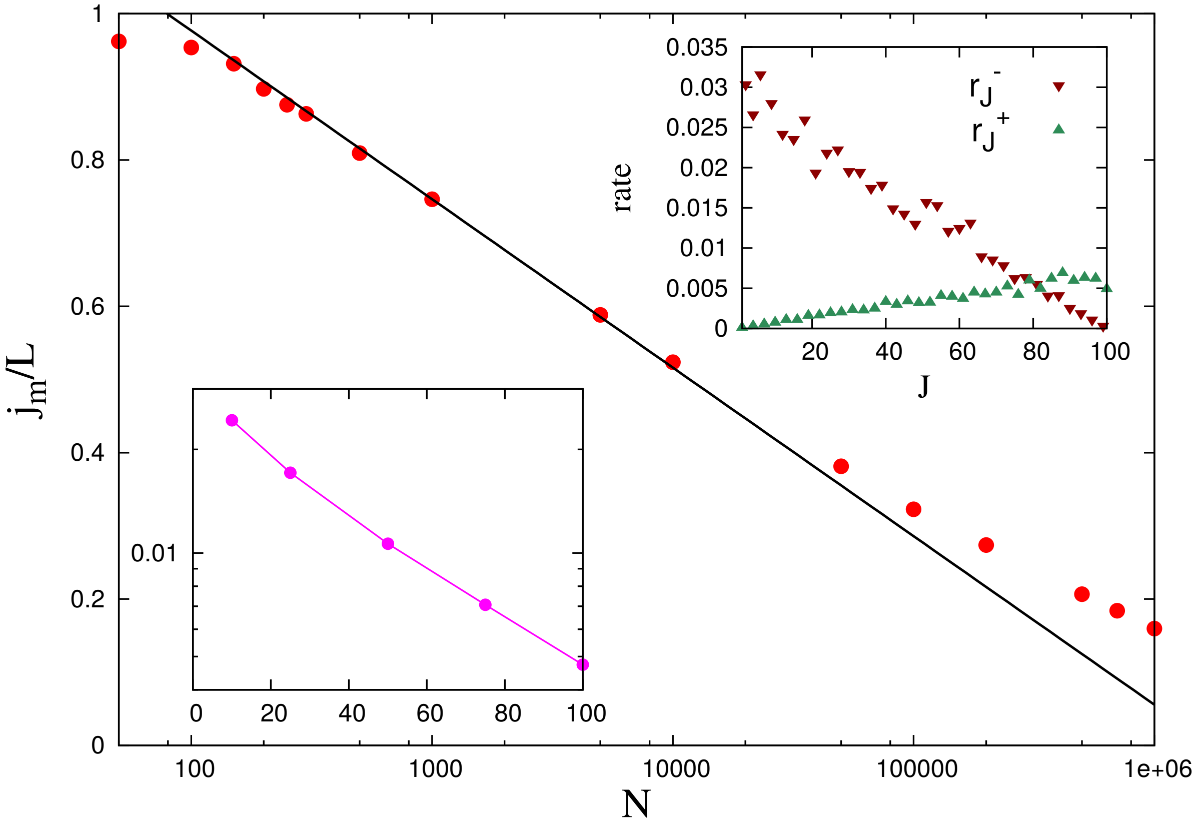


Fig. 2.6 Behavior of least-loaded class: Main figure shows the logarithmic decrease of the deleterious mutation fraction j_m/L with population size N when beneficial mutations are rare. The theoretical prediction (2.15) is shown with a best fit for the intercept as 1.437. The parameters are $s = 10^{-2}, U_d = 10^{-1}, U_b = 10^{-3}$ and $L = 100$. Bottom, left inset: Plot to show the exponential decay of j_m/L as predicted from (2.16) when beneficial mutations occur frequently. The line joining the points is a guide to the eye. Here $U_b = 10^{-1}$ and $U_d = 10^{-2}$, and the other parameters are same as those in the main figure. Top, right inset: Degeneration and regeneration rates calculated in numerical simulations starting from all the individuals in the best and worst fitness class respectively. Parameters: $L = 100, N = 50, U_d = 5 \times 10^{-2}, U_b = 5 \times 10^{-3}$ and $s = 10^{-2}$.

where

$$X_J^{(0)}(J) \approx e^{-\frac{U_d}{s}\left(1-\frac{J}{L}\right)} \tag{2.13}$$

and c is a number of order unity [68, 69]. When deleterious mutations are absent, a maladapted population adapts at a rate that depends on the number NU_b of beneficial mutants produced per generation. For $NU_b \ll 1$, the beneficial mutants arise one at a time and go to fixation

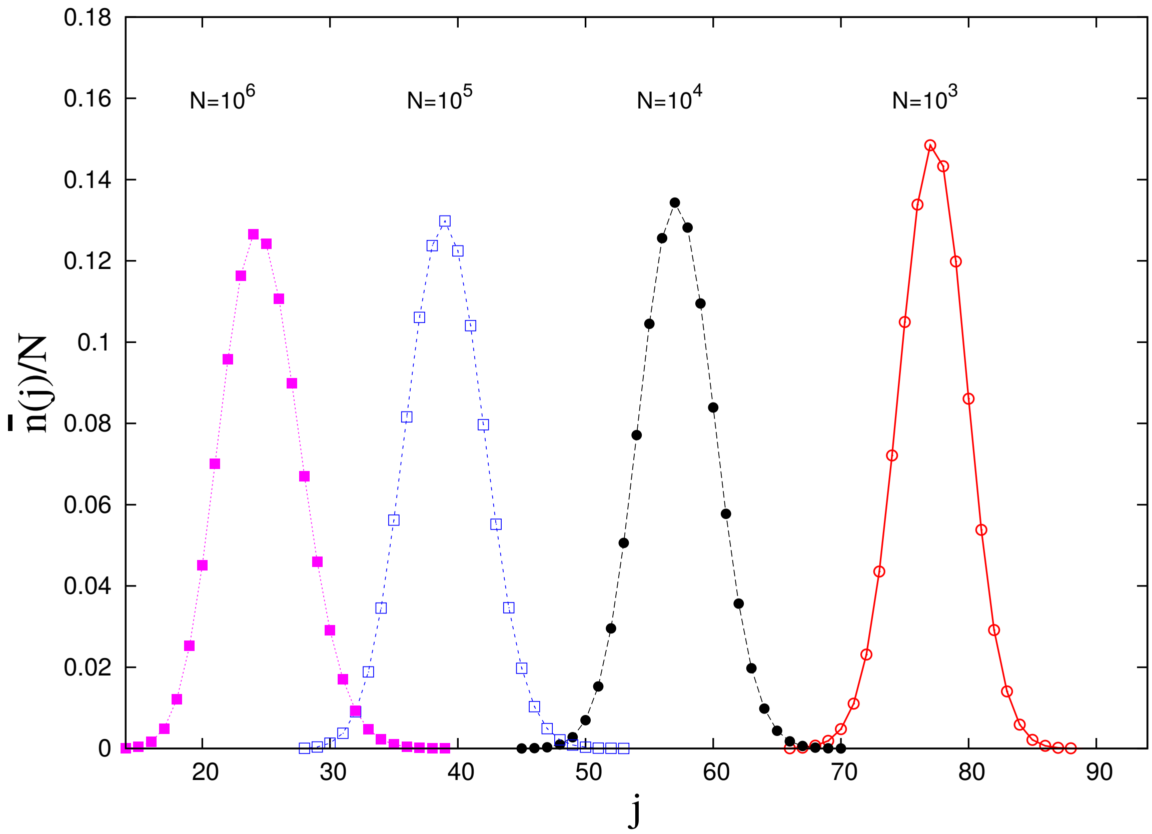


Fig. 2.7 Distribution of the average fraction of individuals in each mutant class for a completely linked sequence for various population sizes, and $L = 100$, $U_d = 10^{-1}$, $U_b = 10^{-3}$ and $s = 10^{-2}$.

sequentially, while they interfere with each other for $NU_b \gg 1$ [35]. The regeneration rate in these two parameter regimes is given by [70, 51]

$$r_J^+ \sim \begin{cases} 2sNU_b\,(J/L)\;,\; NU_b \ll 1 \\ \dfrac{s\ln N}{\ln^2 U_b}\,\dfrac{f(J)}{L}\;,\; NU_b \gg 1 \end{cases} \tag{2.14}$$

where, our numerical simulations for large populations indicate that $f(J)$ is of the form $\delta_1\sqrt{J}+\delta_2 J$. The above equation shows that the rate r_J^+ depends weakly on N, and increases linearly with J for large J.

i. Rare beneficial mutations ($U_b \ll U_d$): An expression for j_m can be obtained by matching the rates (2.12) and (2.14). But as the degeneration rate decays fast with N whereas regeneration rate depends weakly on population size, we may treat the rate $r^+_{j_m}$ as a constant in N. This simplification implies that $r^-_{j_m} \sim e^{-csNX^{(0)}_{j_m}(j_m)} \sim 1$ which immediately leads to

$$\frac{j_m}{L} \sim -\frac{s}{U_d} \ln(Ns) \tag{2.15}$$

Our analytical result (2.15) is compared with the results of numerical simulations in Fig. 2.6 and for a wide range of population sizes, we see a good agreement. Figure 2.7 shows that the average population fraction is distributed over a narrow range of fitness classes [41], and therefore we may expect $\bar{j}$ to behave in a manner similar to j_m. Indeed as shown in Fig. 2.4, the average fraction of disadvantageous mutations also decreases logarithmically with population size, albeit with a prefactor smaller than s/U_d.

ii. Frequent beneficial mutations ($U_b \gg U_d$): When $U_b \gg U_d$, the average frequency of deleterious mutations lies between the neutral value μ/ν (refer (2.7)) and the deterministic value $\mu/(s+\nu)$ (refer (2.5)), and thus $q \ll 1$ for a wide range of population sizes. This implies that j_m/L is also small compared to unity. Using this in (2.13), and that the degeneration rate $r^+_{j_m}$ is linear in j_m, we have

$$\frac{j_m}{L} \sim e^{-csNe^{-U_d/s}} \tag{2.16}$$

which decreases exponentially fast with population size and is consistent with our numerical observations shown in the inset of Fig. 2.6. Same behaviour is observed for the average fraction of deleterious mutations $\bar{j}/L$, refer Fig. 2.5.

2.5.3 Recombining population

Having discussed the case of complete linkage ($r = 0$), we now turn to the limit of completely unlinked loci ($r = 1/2$) where single locus theory applies. When selection is present, a diffusion theory calculation [38] gives the frequency of deleterious mutations for a haploid population to be [71]

$$\bar{j}_1 = \frac{\mu}{\mu+\nu} \frac{{}_1F_1(2N\mu+1, 2N(\mu+\nu)+1, -2Ns)}{{}_1F_1(2N\mu, 2N(\mu+\nu), -2Ns)} \tag{2.17}$$

where ${}_1F_1(a,b,z)$ is the confluent hypergeometric function. For $s=0$, the above expression reduces to (2.11). When Ns is small, we have

$$\bar{j}_1 = \left(1+\frac{\nu}{\mu}e^{2Ns}\right)^{-1} \tag{2.18}$$

which may be obtained either from (2.17) [41, 42, 72, 49] or a rate matching argument [42, 49]. When Ns is large, (2.17) approaches μ/s [71] as one would also expect from the deterministic solution (2.5). Thus as Figs. 2.4 and 2.5 show, the fraction q decreases exponentially fast in a reverse sigmoidal fashion, as the population size N is increased when there is no linkage between loci.

As in the two extreme cases of complete linkage and no linkage, for $0<r<1/2$, we discern three distinct regimes in the behavior of the fraction q of disadvantageous mutations. Our numerical data in Figs. 2.4 and 2.5 shows that the fraction q is roughly constant in population size and recombination rate when the population is small or very large. But for moderately large population, q decreases with increasing population size and the general effect of recombination is to decrease the equilibrium frequency of the deleterious mutations.

2.6 Discussion

In this Chapter, we examined the stationary state of a model in which both beneficial and deleterious mutations can occur. The multilocus model studied here differs from that in previous works [73, 51] where these mutation rates are assumed to be independent of the fitness. Here we considered a biologically realistic situation of forward and backward mutations where the rates depend linearly on the logarithmic fitness. In the general scenario where compensatory mutations can occur, nonlinear relationship between the mutation rates and logarithmic fitness has been experimentally observed [22]. Here we are mainly concerned with the variation of the average number $\bar{j}$ of deleterious mutations with the population size.

2.6.1 Exact bounds on the number of deleterious mutations

For an infinitely large and nonrecombining population, exact results for the population frequency have been obtained for special choice of parameters [58, 74, 75, 61], and here these results were generalised to obtain exact stationary state and dynamics. Since we consider non-epistatic fitnesses, the stationary state solution does not depend on the recombination rate [62]. Moreover as the deterministic limit corresponds to very strong selection which is not favorable for disadvantageous mutations, this analysis provides a lower bound on the average number $\bar{j}$ of deleterious mutations.

The upper bound on $\bar{j}$ can be found by considering the neutral limit for a finite population. For completely linked loci, we calculated the average frequency of mutations (relative to the wildtype) exactly, and found it to be independent of the population size. Although the latter result is known from previous studies on one locus models [63], to our knowledge, such a result has not been obtained using a multilocus model. Using numerical simulations and the known results for freely recombining population [38, 63], we found that the number $\bar{j}$ is independent of the recombination rate in the neutral limit as well. This happens because in the absence of selection, as random genetic drift creates positive and negative linkage disequilibrium (LD) with equal probability, the average LD vanishes [76, 77] and therefore the average number $\bar{j}$ is not affected by recombination. It should however be noted that the higher moments of the number of mutations may depend on both the recombination rate and population size [76].

2.6.2 Effect of drift, selection and recombination

To get an insight into the problem when both selection and population size are finite and recombination is absent, we used a rate matching argument which states that stationarity is achieved when the rate at which the least-loaded fitness class is lost due to deleterious mutations equals the rate at which it is regenerated by beneficial mutations [51]. A similar argument has been used previously by [42], but in a single locus setting, to arrive at the equilibrium fraction of deleterious mutations given in (2.18). In recent years, some analytical understanding of the rate at which an asexual population declines in fitness [67, 34, 66, 75, 78, 68, 69] and adapts

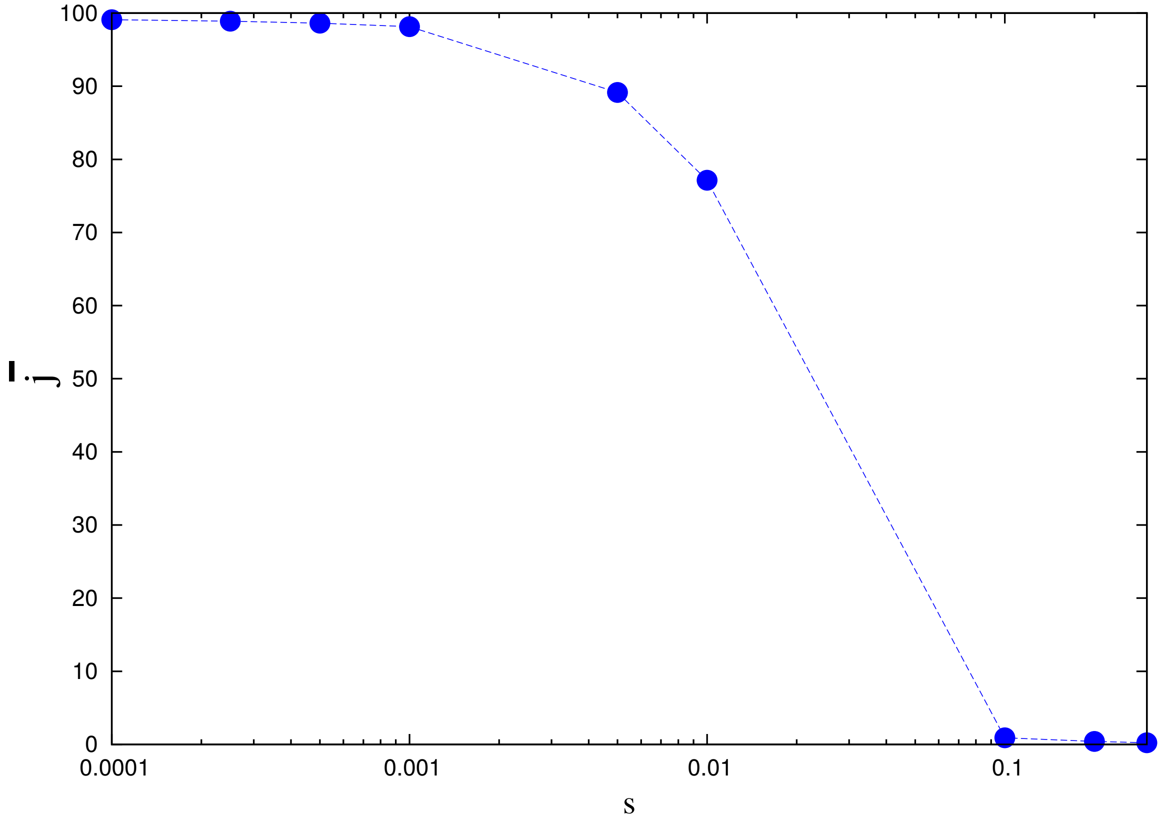


Fig. 2.8 Figure shows the decrease in equilibrium fraction of deleterious mutation with selection coefficient s. Other parameters used are $N = 1000, \mu = 0.001, \nu = 0.00001$ and $L = 100$.

[35, 79, 36, 80, 70] has become available in multilocus models. Using these results and the rate balancing argument described above, we found analytical expressions for the minimum number of deleterious mutations that a finite asexual population under selection carries in the stationary state.

For a nonrecombining population, our main result is that the average fraction q of deleterious mutations decreases from the neutral value (2.7) towards the deterministic fraction (2.5), as population size is increased. If beneficial mutations are rare ($U_b \ll U_d$), as is the case in adapting microbial populations [57], q changes logarithmically with population size. In an adaptation experiment on bacteriophage, it was observed that when the population size is increased by a factor ten, the logarithmic fitness increased mildly [22], which is consistent with the weak N-dependence seen here. Experimental data [55, 56] on Drosophila shows that the mutation rate from preferred to unpreferred codon is twice as much as that for the reverse mutations. In such a case where $U_b \sim U_d$, as the inset of Fig. 2.5 indicates, q decreases faster than the logarithm of population size, but we do not have an analytical form for it. However in the extreme case when $U_b \gg U_d$, we find that the fraction q decreases exponentially fast with the population size. Similar qualitative behaviour, namely, the decrease in $\bar{j}$ with increasing population size is seen when recombination is nonzero, refer Figs. 2.4 and 2.5. When the population size is kept fixed and the selection coefficient is increased, the average fraction of deleterious mutations decreases as one would intuitively expect, see Fig. 2.8. Although the rate balancing argument used here explains the population size dependence of the fraction of deleterious mutations, we have not been able to obtain a complete analytical understanding of its variation with selection coefficient since the s-dependence of the function c in the degeneration rate in (2.12) is not known. We also performed numerical simulations keeping the product Ns constant $(= 10)$, and find that $\bar{j}$ is not a function of Ns unlike the one locus theory prediction (2.17). For $s = 0.005$, we obtained $\bar{j} = 13.7$ which increased to 28.8 on halving s which suggests that it depends more strongly on s than N which is consistent with (2.15).

For a given Ns, we find that the recombination reduces the frequency of the deleterious mutations (also see [81]). As discussed above, in a finite population, due to random genetic drift, both positive and negative LD are created. If LD is positive, the population consists of individuals with extreme fitnesses on which selection can act efficiently and thus removes the LD. On the other hand, when LD is negative, as most of the individuals are likely to have similar fitnesses, selection is ineffective in removing LD. Thus in the presence of selection, the average LD in a nonrecombining population is negative [33, 77]. But once recombination is introduced, it will create individuals with extreme fitnesses thereby helping selection to weed out the deleterious mutations, and thus decreasing $\bar{j}$. The effect is large for intermediate values of Ns since this regime corresponds to both selection and drift having a strong effect. From the results in the neutral and deterministic limit, we expect that the difference in the number of deleterious mutations carried by a nonrecombining and recombining population is nearly zero when $s \ll 1/N$ (weak selection) and $s \gg 1/N$ (strong selection). Thus, as shown in the inset of Fig. 2.4, the maximum advantage of recombination occurs at an intermediate value of selection coefficient as has also been observed in other studies [82].

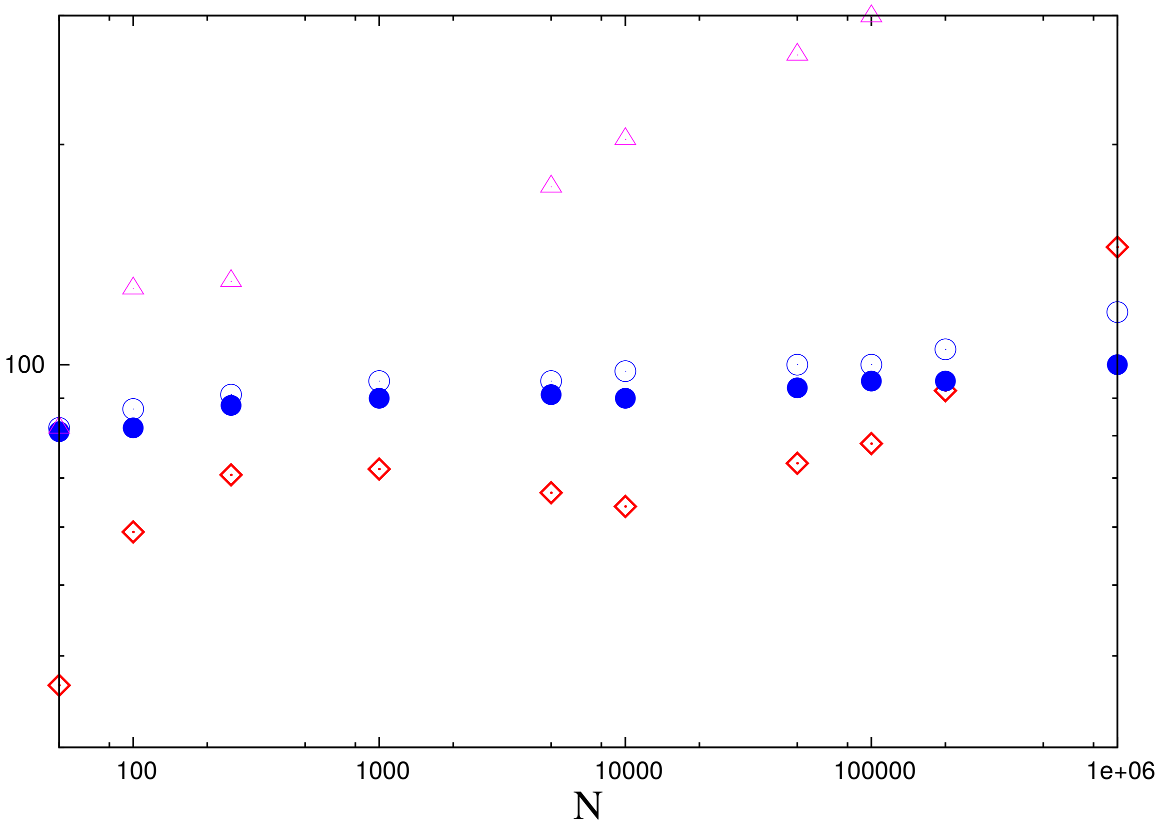


Fig. 2.9 Figure shows the effective population size N_e (◊) given by (2.19) where j_m is shown in Fig. 2.6, and click time of the Muller's ratchet with deleterious mutation rate U_d' and selection coefficient s', when there are finite number of background selection sites with reversible mutations. The click time of the ratchet for (i) a population of size N and without background selection sites (△), (ii) a population of size N with background selection sites (○) and (iii) a population of size N_e without background selection sites (●) are shown. Parameters: $L = 100, s = 0.01, U_d = 0.1, U_b = 0.001, U_d' = 0.015, s' = 0.001$.

Although recombination reduces the number of deleterious mutations, the extent to which it does so depends on how common the beneficial mutations are compared to the deleterious ones. In an adapting asexual population where beneficial mutations occur rarely [57], even slight recombination reduces $\bar{j}$ considerably indicating the advantage of recombination during adaptation [54, 77]. On the other hand, in the codon bias problem where back mutation rates are comparable to the forward ones [55, 56], the fraction of unpreferred codons is given by (2.18) if the loci are assumed to be completely unlinked, but as the inset of Fig. 2.5 shows, linkage increases the unpreferred codon frequency moderately [45–47].

2.7 Applications and Open questions

2.7.1 Effect of background selection

Background selection is a type of Hill-Robertson effect [83] and is known to increase the rate at which the Muller's ratchet clicks [73, 84]. In a finite, nonrecombining population with an infinitely long sequence in which both deleterious and beneficial mutations occur at L *background selection sites*, and deleterious mutations accumulate at rest of the sites [84], we find that the ratchet clicking time is considerably reduced from the situation when there are no background selection sites (see Fig. 2.9). If the background selection sites (BGS) remain at equilibrium in the presence of other linked loci also, they affect the evolutionary dynamics at other sites, and their effect can be quantified by a reduction in the effective population size to the number of individuals carrying the minimum number of deleterious mutations at BGS [83, 73]. Since the minimum number of deleterious mutations in the BGS is j_m, we require the population fraction in the class j_m. For large populations with $Ns \gg 1$ where the deterministic theory is expected to hold, using (2.9) and (2.13), we obtain

$$N_e = Ne^{-\frac{U_d(1-\frac{j_m}{L})}{s}} \tag{2.19}$$

where j_m is a function of population size N. The ratchet time with background selection for a population of size N is found to be well approximated by the ratchet time without it for a population of size N_e as shown in Fig. 2.9. From the results for j_m when $U_b \ll U_d$, we expect N_e in (2.19) to increase linearly with N for small and large populations. But for the intermediate range of population sizes, using (2.15) in (2.19) above, we find the effective population size to be independent of N. These predictions were tested numerically and as shown in Fig. 2.9, the effective population size and the ratchet time remain roughly constant when the actual population size is varied over three orders of magnitude .

2.7.2 Codon usage bias

Our results have implications for the codon usage bias problem in the context of which the model studied here was introduced [41, 42]. To apply the two-allele model to amino acids encoded by more than 2 synonymous codons, the preferred codon is represented by 0 allele, and the other codons collectively by 1 allele. A comparison of two allele model with the 4 allele model [55] showed that these two models give the same allele frequency spectrum when there is no mutational bias, but there is a slight difference when there is a mutational bias (Fig.2 in [55]). Previous numerical results [41, 46] show that the preferred codon frequency changes slower than that predicted by (2.18) which neglects interference effects arising due to linkage. Also, recent studies have observed a reduced level of codon usage bias in genome regions with low frequency of recombination [64]. Our results shown in Fig. 2.5 also support this conclusion and for weak selection, where codon bias problem lies [39], we find that $\bar{j}$ depends weakly on the population size. This can be seen as follows: If we assume that mutations are Poisson distributed, a simple calculation shows that $\bar{j} \sim \ln(N/j_m)$, where j_m is given by (2.15). Thus the interference between linked loci can maintain intermediate codon bias levels for a wide range of population sizes [85, 46].

Here we investigated the effect of linkage using numerical simulations, but an analytical expression for the average $\bar{j}$ as a function of recombination probability is desirable. Also, in the model discussed here, it was assumed that the number of beneficial mutations increases linearly with average number of deleterious mutations. However in an adaptation experiment on bacteriophage [22], a nonlinear relationship between these two quantities has been observed. An extension of these results to the more general cases of compensatory mutations would be interesting.

Chapter 3

Effect of beneficial mutations in an infinitely long genome with constant mutation rates

3.1 Introduction

In the previous Chapter, the role of beneficial mutations in attaining a stationary state for a finite length genome with multiplicative or additive fitness when the mutation rates depend linearly on the sequence fitness was discussed. A complete solution of the genotypic frequency distribution is exactly known [58, 3] for this model, and has been utilized, for example, in modelling codon usage bias [41–43]. In this Chapter, we extend our study to consider a variant of this model in which deleterious and beneficial mutations rates are independent of the sequence fitness. This model has recently appeared in various contexts such as adaptive evolution [86, 36, 70], evolution of sex [51] and evolution of mutation rates [27]. The relationship between these two mutation schemes is discussed in detail in Appendix B.1.

Here, we focus on understanding a model with fitness-independent mutation rates in the deterministic limit ($N \to \infty$), which constitutes an important step towards an understanding of more complex and realistic finite size populations that evolve stochastically. When beneficial mutations are ignored, the exact solution in the stationary state and for the dynamics of the population fraction is known; in particular, at mutation-selection balance, the frequency is Poisson-distributed with a mean given by the ratio of the deleterious mutation rate to selection coefficient [4, 2, 74, 75]. The short time dynamics of the frequency distribution when beneficial mutations are also allowed has been quite well-studied [70] and some approximate results at mutation-selection equilibrium were obtained recently [27]. Here we find an exact expression for the frequency distribution at all times. We also numerically studied the case of finite size population and the preliminary results obtained are discussed in last Appendix.

In the following section, we define the model, discuss some limiting cases and explain the relation of our work to the existing literature. We then proceed to find an exact solution of the stationary state as well as the dynamics using an eigenfunction expansion method in Section 3.3. Besides the exact results, we also provide accurate approximations for the frequency distribution when the selection coefficient is larger or smaller than the mutation rates. Sections 3.4 and 3.5, respectively, deal with these approximations in the stationary state and for the dynamics of the frequency distribution. A discussion of the results follows in the concluding section.

3.2 Model

We consider an infinitely large asexual population of infinitely long diallelic sequences evolving in continuous time. All individuals carrying $j \geq 0$ deleterious mutations relative to the fittest individual are assumed to have the same (Malthusian) fitness $w(j) = -sj, s \geq 0$ and said to belong to the fitness class j. We also assume a single-step mutation scheme in which a deleterious (beneficial) mutation increases (decreases) the fitness class by one and occurs at rate U_d (U_b), but mutations to other classes are ignored. Then the population fraction $X(j,t)$ in the jth fitness class at time t obeys the following differential-difference equations:

$$\dot{X}(0,t) = U_b X(1,t) - U_d X(0,t) + s\mathscr{C}_1(t) X(0,t) \;, \tag{3.1a}$$

$$\dot{X}(j,t) = U_b X(j+1,t) + U_d X(j-1,t) - UX(j,t) - s(j - \mathscr{C}_1(t)) X(j,t) \;,\; j \geq 1 \;. \tag{3.1b}$$

In the above equations, $U = U_d + U_b$ is the total mutation rate and $\mathscr{C}_1(t) = \sum_{j=0}^{\infty} j\, X(j,t)$ is the average number of deleterious mutations in the population at time t. In the stationary state where the LHS of (3.1a) and (3.1b) is zero, we will denote the steady state fraction by $X(j)$.

It is easy to verify that the above set of equations respect the normalisation condition,

$$\sum_{j=0}^{\infty} X(j,t) = 1 \ , \ t \geq 0 \ . \tag{3.2}$$

We also have the boundary condition

$$X(j,t) \xrightarrow{j \to \infty} 0 \ \ , \tag{3.3}$$

which ensures that the total fraction remains finite at all times. To complete the definition of the model, we also need to specify the initial condition $X(j,0)$ for all j. The analysis in Sections 3.3 and 3.4 holds for arbitrary initial conditions but in Section 3.5, we will assume that the population is initially localised in the fitness class j_0.

The time evolution equations above for the frequency $X(j,t)$ are defined for $j \geq 0$ and therefore the maximum fitness is zero. However, (3.1b) has been used without an upper bound on fitness to describe the adaptation dynamics [86, 87, 80, 36, 88, 70]. The latter is a reasonable model at short times for a population initially localised in a fitness class with many deleterious mutations since the frequency in the fitness classes close to the fittest one can then be neglected. For this model, several works [86, 87, 80, 70] have shown that it does not have a traveling wave solution, and either a lower cutoff on the frequency modeling a finite population size [86] or discrete time dynamics [70] are required to obtain it. In [36], although a cutoff for the high-fitness edge is imposed in the deterministic model to account for the finite size of the population, this work also assumes a traveling wave solution for the continuous time model (see their (4) and (5)). However, as in [86, 87, 80, 70], our analysis of short time dynamics described in Section 3.5.1 also does not support a traveling wave behavior.

Although dynamics can be studied on an infinite line, a stationary state does not exist if the fitness is not bounded above. To see this, consider the steady state of (3.1b) by setting the LHS to be zero. Since the frequency in any fitness class must not be negative and the first two terms on the RHS of (3.1b) are positive, their contribution can be balanced if

$$j > \mathscr{C}_1 - U/s = j_* \ , \tag{3.4}$$

thus leading to a maximum fitness corresponding to $-sj_*$. A previous analysis of (3.1b) with unbounded fitnesses finds a negative frequency distribution in the stationary state and claims that “in the deterministic limit there is no true stationary state for arbitrary [beneficial mutation rate]..." (p. 1313, [51]). However as discussed above, the loss of positivity is simply a consequence of the lack of upper bound on the fitness and in Section 3.4, we will show that the model defined by (3.1a) and (3.1b) has a nontrivial steady state. We also mention that if we set the maximum fitness to $-sj_*$ instead of zero, the frequency $X(m) = 0, m < j_*$ and $X(m+j), m \geq j_*$ is given by the solution $X(j)$ of (3.1a) and (3.1b) [2].

Equations (3.1a) and (3.1b) are mathematically nontrivial for two reasons: first, they are nonlinear in the fractions $X(j,t)$ due to the last (selection) term on the RHS and second, they are second order difference equations in j when both mutation rates are nonzero.

(i) When the beneficial mutations are absent ($U_b = 0$), in the stationary state, the boundary equation (3.1a) immediately yields the average number of deleterious mutations, $\mathscr{C}_1 = U_d/s$. This result is very helpful since it renders (3.1b) to be linear in the frequencies and we quickly arrive at the following well known result [4, 2]:

$$X(j) = \frac{e^{-U_d/s}}{j!} \left(\frac{U_d}{s}\right)^j \qquad [U_b = 0] \ . \tag{3.5}$$

The time-dependent frequency has also been obtained using a generating function method and shown to be Poisson-distributed [75].

(ii) When the deleterious mutations are absent ($U_d = 0$), the stationary state is trivial ($X(j) = \delta_{j,0}$). But the short time dynamics can be obtained by extending the method of [75] as described in Section 3.5.1 (also, see [87, 80, 70]).

	$U_b = 0$	$U_d = 0$	$s = 0$	all nonzero
Stationary state	[4, 2]	trivial	none	Equation (3.27)
Dynamics	[75]	[87, 80, 70]	Equation (B.8)	Equation (3.26)

Table 3.1 Summary of the results for the deterministic model defined by (3.1a) and (3.1b) where U_d and U_b, respectively, denote deleterious and beneficial mutation rate and s is the selection coefficient. In all the cases except when deleterious mutations are absent, it is assumed that $U_b < U_d$.

(iii) In the neutral case ($s = 0$), the nonlinear term on the RHS of (3.1a) and (3.1b) vanishes. The stationary state frequency is then easily found to be

$$X(j) = \left(1 - \frac{U_d}{U_b}\right) \left(\frac{U_d}{U_b}\right)^j \qquad [s = 0, U_d < U_b] \ . \tag{3.6}$$

The condition $U_d < U_b$ arises due to the boundary condition (3.3). However, in the parameter regime where $U_b < U_d$, the neutral population does not reach a steady state. In Appendix B.2, we give the exact solution of the neutral dynamics in this parameter regime.

A brief summary of the results in the stationary state and for the dynamics is given in Table 3.1.

3.3 Exact solution of the population frequency by eigenfunction expansion method

We now proceed to find the population fraction $X(j,t)$ when all the three parameters, viz., s, U_b, U_d are nonzero. In the following discussion, we assume that the mutation rate $U_d > U_b$ as in biologically realistic situations [89]. Since the equations (3.1a) and (3.1b) are nonlinear in the fractions $X(j,t)$, we work with the unnormalised variables defined as [90, 91]

$$Z(j,t) = X(j,t)\, e^{-s \int_0^t dt' \, \mathscr{C}_1(t')} \ , \tag{3.7}$$

which obey the following *linear* equations:

$$\dot{Z}(0,t) = U_b Z(1,t) - U_d Z(0,t) \,, \tag{3.8a}$$

$$\dot{Z}(j,t) = U_b Z(j+1,t) + U_d Z(j-1,t) - UZ(j,t) - sjZ(j,t) \,,\ j \geq 1 \,. \tag{3.8b}$$

Summing over j on both sides of (3.7) and using the normalisation condition (3.2), we obtain the following relationship between the average $\mathscr{C}_1(t)$ and the unnormalised frequencies $Z(j,t)$:

$$\sum_{j=0}^{\infty} Z(j,t) = e^{-s\int_0^t dt' \, \mathscr{C}_1(t')} \,. \tag{3.9}$$

Using this in (3.7), we immediately obtain

$$X(j,t) = \frac{Z(j,t)}{\sum_{m=0}^{\infty} Z(m,t)} \,. \tag{3.10}$$

It is convenient to further define (p. 139, [92])

$$Y(j,t) = \left(\frac{U_b}{U_d}\right)^{j/2} e^{(\sqrt{U_d}-\sqrt{U_b})^2 t} \, Z(j,t) \,, \tag{3.11}$$

$$\tau = t\sqrt{U_b U_d} \,, \tag{3.12}$$

$$\gamma = 2 - \sqrt{U_b/U_d} \,, \tag{3.13}$$

$$S = \sqrt{s/U_b} \cdot \sqrt{s/U_d} \,. \tag{3.14}$$

In terms of these variables, we have

$$\frac{\partial Y(0,\tau)}{\partial \tau} = Y(1,\tau) - \gamma Y(0,\tau) \,, \tag{3.15a}$$

$$\frac{\partial Y(j,\tau)}{\partial \tau} = Y(j+1,\tau) + Y(j-1,\tau) - (2+Sj)Y(j,\tau) \,,\ j \geq 1 \,. \tag{3.15b}$$

The above set of equations involving two independent variables, viz., space and time can be solved by the eigenfunction expansion method (Chapter 5 and 6, [92]). Since the differential operator $\partial/\partial\tau$ has eigenfunctions $e^{-\lambda\tau}$ with eigenvalues $-\lambda$, on expanding $Y(j,\tau)$ as a linear combination of these eigenfunctions as

$$Y(j,\tau) = \sum_{\lambda} c_{\lambda} e^{-\lambda\tau} \phi_j(\lambda) \ , \ j \geq 0 \ , \tag{3.16}$$

we obtain difference equations in one independent variable:

$$\phi_1 - (\gamma - \lambda)\phi_0 = 0 \ , \tag{3.17a}$$

$$\phi_{j+1} + \phi_{j-1} - (2 + Sj - \lambda)\phi_j = 0 \ , \ j \geq 1 \ . \tag{3.17b}$$

Equation (3.17b) is an eigenvalue equation for a real symmetric matrix with eigenfunction ϕ_j and eigenvalue $-\lambda$. For such a matrix, it is possible to find a complete set of eigenvectors [93]. Moreover, the eigenvalues are real and the eigenfunctions corresponding to different eigenvalues are orthogonal and can be normalised to unity:

$$\sum_{j=0}^{\infty} \phi_j(\lambda)\phi_j(\lambda') = \delta_{\lambda,\lambda'} \ . \tag{3.18}$$

The eigenvalues are determined by the boundary condition (3.17a) as explained below. The constants c_λ's in (3.16) can be found using the initial condition and are given by

$$c_\lambda = \sum_{m=0}^{\infty} \phi_m(\lambda)\, Y(m,0) \ , \tag{3.19}$$

$$= \sum_{m=0}^{\infty} \phi_m(\lambda) \left(\frac{U_b}{U_d}\right)^{m/2} X(m,0) \ . \tag{3.20}$$

This can be seen by using (3.16) at $t = 0$ and using the orthonormality condition (3.18).

Our remaining task now is to find the eigenfunctions ϕ_j. We remark that if the fitness class j is treated as a continuous variable, (3.17b) reduces to a time-independent Schrödinger equation for a particle in a linear potential for which the eigenfunctions are known to be Airy function (and plane wave when S is zero) [94]. Here we are interested in finding the eigenfunctions in discrete fitness space with (Robin) boundary condition (3.17a). In the neutral case ($S = 0$), exact eigenfunctions and frequency $Y(j,t)$ are obtained in Appendix B.2. When the parameter S is nonzero, the solution of (3.17b) is a linear combination of the Bessel function of first and second kind with order ν and argument z that are denoted by $J_\nu(z)$ and $Y_\nu(z)$, respectively [95] (also, see Appendix B.3):

$$\phi_j(\lambda) = A(\lambda) J_{j+\frac{2-\lambda}{S}}\left(\frac{2}{S}\right) + A'(\lambda) Y_{j+\frac{2-\lambda}{S}}\left(\frac{2}{S}\right) \ , \ j \geq 0 \ . \tag{3.21}$$

It is easy to check that (3.21) satisfies the eigenvalue equation (3.17b) using the recurrence relation for the Bessel function $\mathscr{K}_\nu(z)$ given by (9.1.27, [96])

$$\mathscr{K}_{\nu-1}(z) + \mathscr{K}_{\nu+1}(z) = \frac{2\nu}{z} \mathscr{K}_\nu(z) \ , \tag{3.22}$$

where $\mathscr{K}$ denotes J, Y. To find the constants A, A', we invoke the boundary condition (3.3) that the frequency $X(j,t) \to 0$ as the fitness class $j \to \infty$. From (3.7) and (3.11), it follows that $Z(j,t)$ and $Y(j,t)$ also obey this boundary condition. Using this large j behavior, we find that the coefficient A' is zero since $Y_\nu(z)$ diverges for large ν (9.3.1, [96]). Then using the orthonormality condition (3.18), we find that

$$A^2(\lambda) = \frac{1}{\sum_{j=0}^{\infty} J^2_{j+\frac{2-\lambda}{S}}\left(\frac{2}{S}\right)} \ . \tag{3.23}$$

The eigenvalues are determined using (3.21) in the boundary condition (3.17a) at $j = 0$ and satisfy

$$J_{1+\frac{2-\lambda}{S}}\left(\frac{2}{S}\right) - (\gamma - \lambda) J_{\frac{2-\lambda}{S}}\left(\frac{2}{S}\right) = 0 \ . \tag{3.24}$$

Putting all the pieces together, we finally obtain

$$X(j,t) \propto \left(\frac{U_d}{U_b}\right)^{j/2} \sum_\lambda c_\lambda A(\lambda) J_{j+\frac{2-\lambda}{S}}\left(\frac{2}{S}\right) e^{-\lambda\sqrt{U_b U_d} t} \ , \tag{3.25}$$

where c_λ and $A(\lambda)$ are, respectively, given by (3.20) and (3.23), the eigenvalues by (3.24) and the proportionality constant is determined by the normalisation condition (3.2). If $\lambda_\alpha, \alpha \geq 0$

denotes the $(\alpha+1)$th minimum eigenvalue, the above equation can be rewritten as

$$X(j,t) = \frac{\left(\frac{U_d}{U_b}\right)^{j/2} \sum_{\alpha=0}^{\infty} c_{\lambda_\alpha} A(\lambda_\alpha)\, J_{j+\frac{2-\lambda_\alpha}{S}}\left(\frac{2}{S}\right) e^{-(\lambda_\alpha-\lambda_0)\sqrt{U_b U_d}t}}{\sum_{m=0}^{\infty} \left(\frac{U_d}{U_b}\right)^{m/2} \sum_{\alpha=0}^{\infty} c_{\lambda_\alpha} A(\lambda_\alpha)\, J_{m+\frac{2-\lambda_\alpha}{S}}\left(\frac{2}{S}\right) e^{-(\lambda_\alpha-\lambda_0)\sqrt{U_b U_d}t}} \,. \tag{3.26}$$

This result can be verified by plugging it in (3.1a) and (3.1b) and using the relationship (3.9) between the normalisation constant and the mean.

3.4 Stationary state frequency

To obtain the steady state, we take the limit $t \to \infty$ in (3.26) and find that only the minimum eigenvalue λ_0 contributes to the sum over the eigenvalues and the result is independent of the initial condition. We thus obtain the *exact* stationary state frequency for an infinitely large population evolving under the joint action of deleterious and beneficial mutations and non-epistatic selection to be

$$X(j) = \frac{\left(\frac{U_d}{U_b}\right)^{j/2} J_{j+\frac{2-\lambda_0}{S}}\left(\frac{2}{S}\right)}{\sum_{m=0}^{\infty} \left(\frac{U_d}{U_b}\right)^{m/2} J_{m+\frac{2-\lambda_0}{S}}\left(\frac{2}{S}\right)} \,, \tag{3.27}$$

where λ_0 is the minimum eigenvalue determined from (3.24). Before proceeding further, we note that the Bessel function $J_\nu(z)$ is an oscillatory function in both ν and z. However, since the eigenfunction corresponding to the minimum eigenvalue for a real symmetric matrix with homogeneous boundary condition cannot have zeros (p. 452, [93]), the solution (3.27) satisfying (3.17a) and (3.17b) is guaranteed to be positive for all $j \geq 0$.

Using the above solution in (3.1a), we find that the average number of deleterious mutations in the steady state is given exactly by

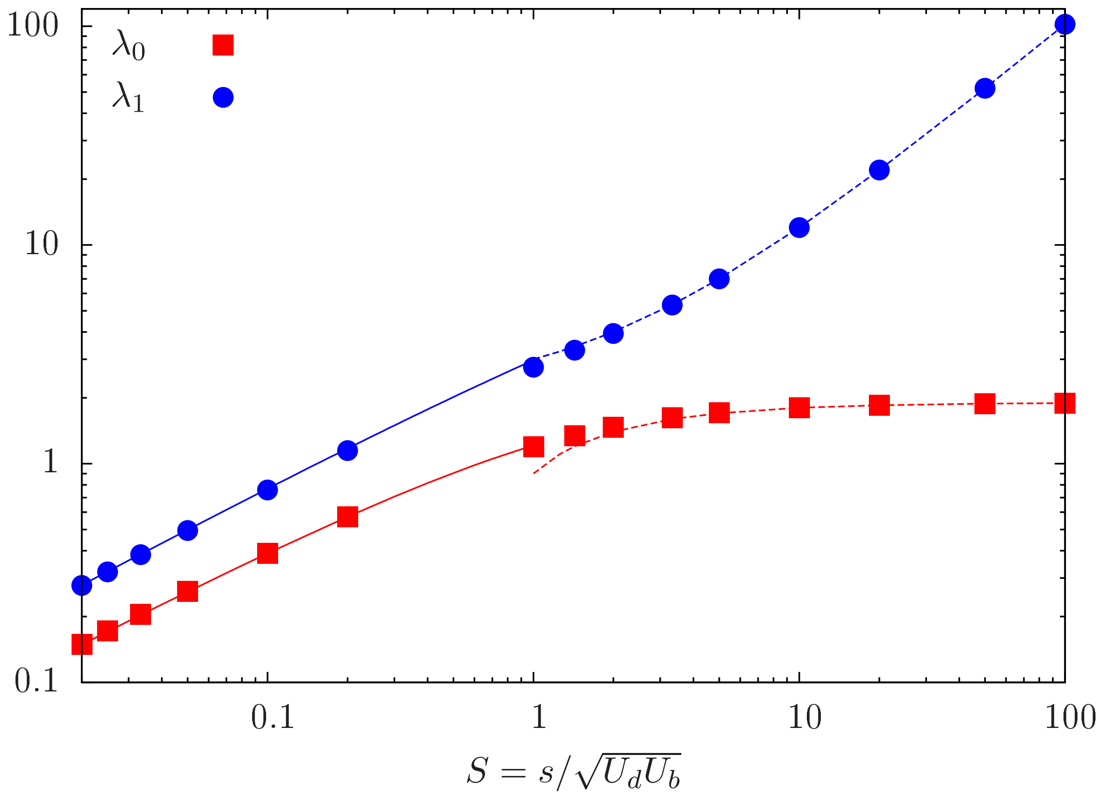


Fig. 3.1 Variation of the minimum and second minimum eigenvalue λ_0 and λ_1 with S. The points are obtained by solving (3.24) numerically and the lines show the approximate expressions (3.37) and (3.31) for λ_0 and (3.48) for λ_1 for $U_b = 0.01\ U_d, s = 0.001$.

$$\mathscr{C}_1 = \frac{U_d}{s} - \frac{\gamma - \lambda_0}{S} \tag{3.28}$$

$$= \frac{U}{s} - \frac{2 - \lambda_0}{S} . \tag{3.29}$$

As Fig. 3.1 shows, the minimum eigenvalue λ_0 initially increases with S and approaches a constant asymptotically. Taking $S \to \infty$ in (3.24) and using that $J_0(0) = 1, J_1(0) = 0$ [96], we find that $\lambda_0 \to \gamma$ (also, see (3.31) below). Then, from (3.28), it follows that beneficial mutations decrease the average number of deleterious mutations as also expected intuitively. Moreover, using the inequalities $\lambda_0 \leq \gamma \leq 2$ in (3.29), it is easily checked that the condition (3.4) for the existence of the stationary state is satisfied.

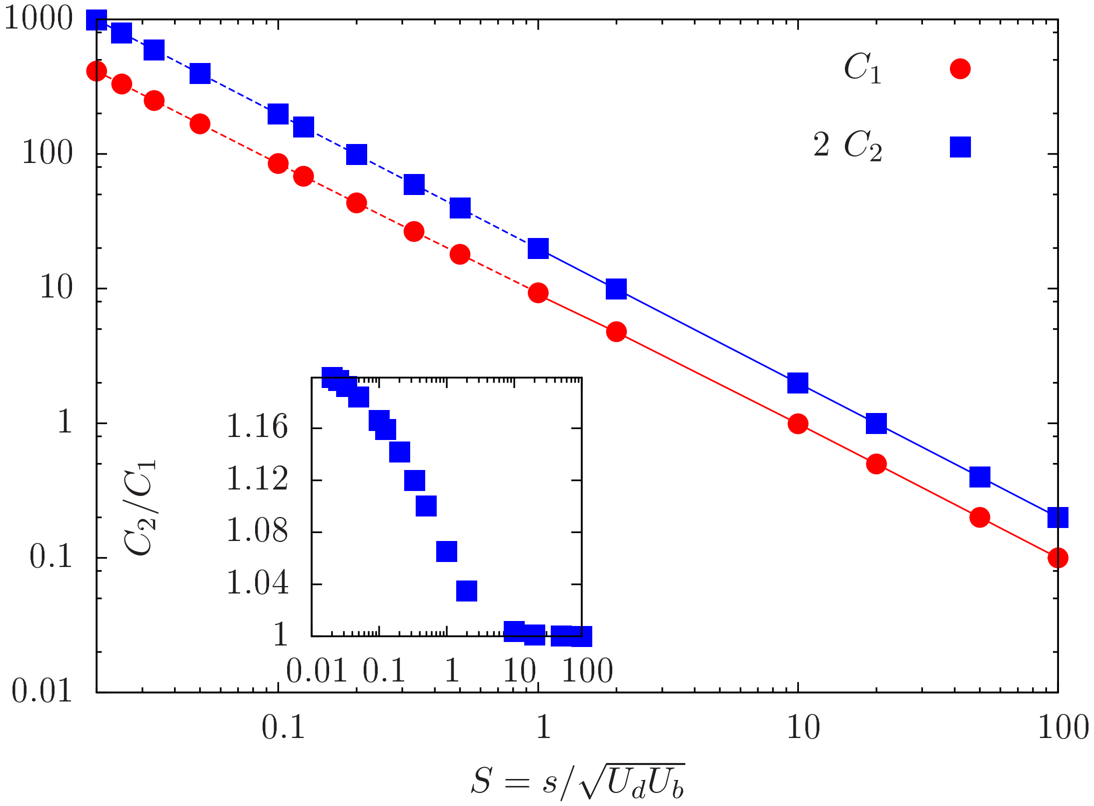


Fig. 3.2 Main figure shows the exact mean $\mathscr{C}_1$ and variance $\mathscr{C}_2$ in the stationary state given, respectively, by (3.28) and (3.30) for $U_b = 0.01\ U_d, s = 0.001$. The lines show the approximate expressions (3.34), (3.38) for mean and (3.35), (3.39) for variance. The variance is scaled by a factor 2 for clarity. The exact ratio $\mathscr{C}_2/\mathscr{C}_1$ shown in the inset supports the non-Poissonian nature of the frequency distribution in the steady state.

The higher order cumulants such as variance and skewness can be found using a cumulant generating function as detailed in Appendix B.4. Alternatively, on multiplying both sides of (3.1b) in the steady state by j and summing over j, we find the stationary state variance $\mathscr{C}_2 = \overline{j^2} - \overline{j}^2$ to be

$$\mathscr{C}_2 = \frac{U_d}{s} - \frac{U_b}{s}(1 - X(0)) \, , \tag{3.30}$$

which shows that beneficial mutations decrease the width of the distribution also. Furthermore, as the inset of Fig. 3.2 shows, the variance to mean ratio is greater than one and therefore the frequency distribution is non-Poissonian when U_b is nonzero.

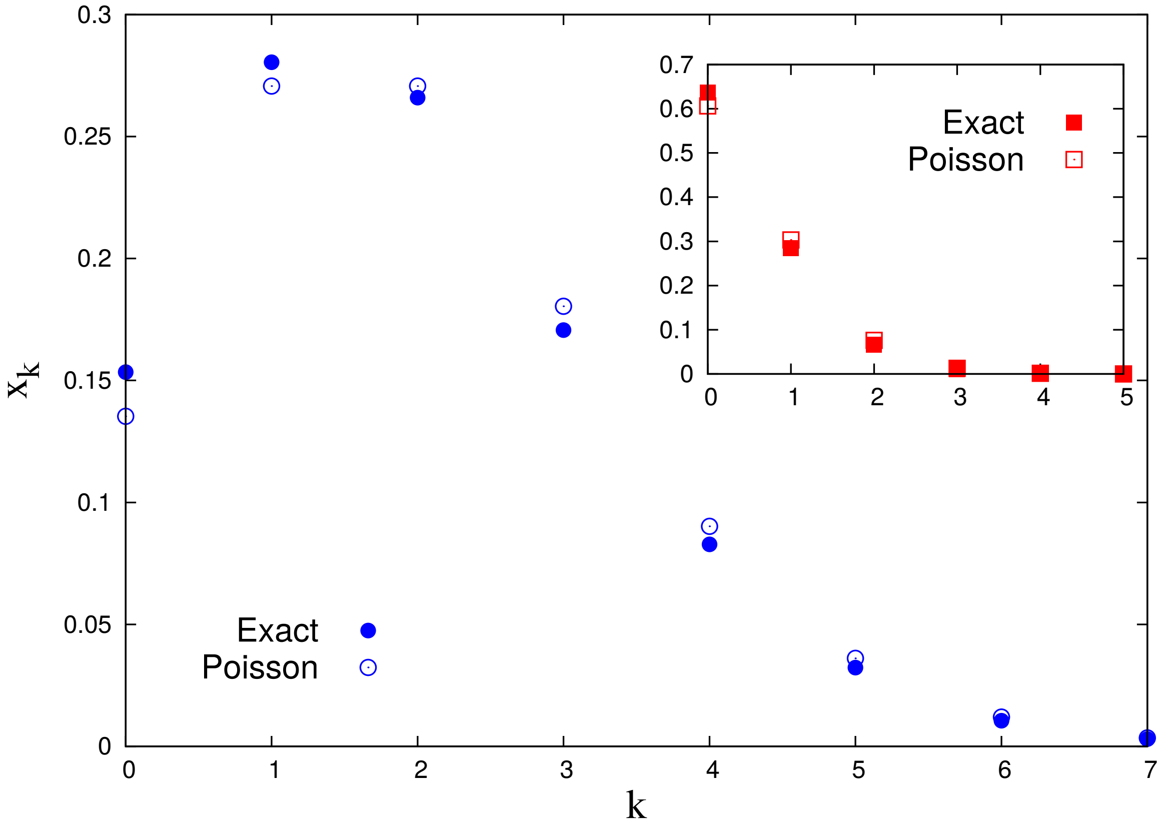


Fig. 3.3 Steady state distribution when $s \gg \sqrt{U_b U_d}$: Comparison of the exact distribution (3.27) and Poisson distribution (3.5) for $U_b = 0.005, s = 0.1, U_d = 0.2$ (main) and $U_b = 0.01, U_d = 0.05, s = 0.1$ (inset).

To obtain some insight into the behavior of the equilibrium frequency given by (3.27) above, we now consider two parameter regimes where the ratio S of the selection coefficient to the mutation rates is large or small relative to one.

3.4.1 When the parameter S is large

The parameter $S \gg 1$ when (i) $U_b < s$ and (ii) either $U_d < s$ or $s < U_d < s^2/U_b$. When $U_b = 0$, as (3.5) shows, the average number of deleterious mutations in the stationary state equals U_d/s. Therefore, when U_b is turned on, we expect that beneficial mutations do not have a significant effect when $U_d < s$ but they can decrease the mean $\mathscr{C}_1$ substantially when $U_d > s$. The analysis given below is in agreement with these expectations.

As described in Appendix B.5 and shown in Fig. 3.1, the minimum eigenvalue λ_0 when S is large is given by

$$\lambda_0 \approx \gamma - S^{-1} \, . \tag{3.31}$$

On plugging (3.31) in (3.27), we obtain

$$X(j) \propto \left(\frac{U_d}{U_b}\right)^{j/2} J_{j+\frac{U_b}{s}(1+\frac{U_d}{s})}\left(\frac{2}{S}\right) \, , \tag{3.32}$$

which, on using the series representation (B.19) of Bessel function, yields

$$X(j) \propto \left(\frac{U_d}{s}\right)^{j} \sum_{m=0}^{\infty} \left(\frac{U_b U_d}{s^2}\right)^{m} \frac{(-1)^m}{m!(m+j+\frac{U_b}{s}(1+\frac{U_d}{s}))!} \, . \tag{3.33}$$

We check that the above solution reduces to (3.5) when $U_b = 0$. To see the effect of beneficial mutations, it is useful to expand (3.33) in a power series in U_b/s as was done recently in [27] and described here briefly in Appendix B.6. This discussion as also Fig. 3.3 show that a nonzero U_b has a significant effect when $U_d > s$.

Furthermore, using (3.31) in the exact expression (3.28) for the average $\mathscr{C}_1$, we find that [27]

$$\mathscr{C}_1 \approx \frac{U_d}{s}\left(1 - \frac{U_b}{s}\right) \, . \tag{3.34}$$

Since U_b/s is small for large S, using (3.5) for the frequency $X(0)$ in (3.30), we find that the variance is well approximated by

$$\mathscr{C}_2 \approx \frac{U_d}{s} - \frac{U_b}{s}(1 - e^{-U_d/s}) \, , \tag{3.35}$$

$$= \begin{cases} \mathscr{C}_1 & , U_d \ll s \, , \\ \frac{U_d - U_b}{s} & , U_d \gg s \, . \end{cases} \tag{3.36}$$

Thus the variance is close to mean (3.34) when $U_d/s \ll 1$ but larger in the opposite parameter regime. The above approximations are tested against the corresponding exact results in Fig. 3.2 and we see a good agreement.

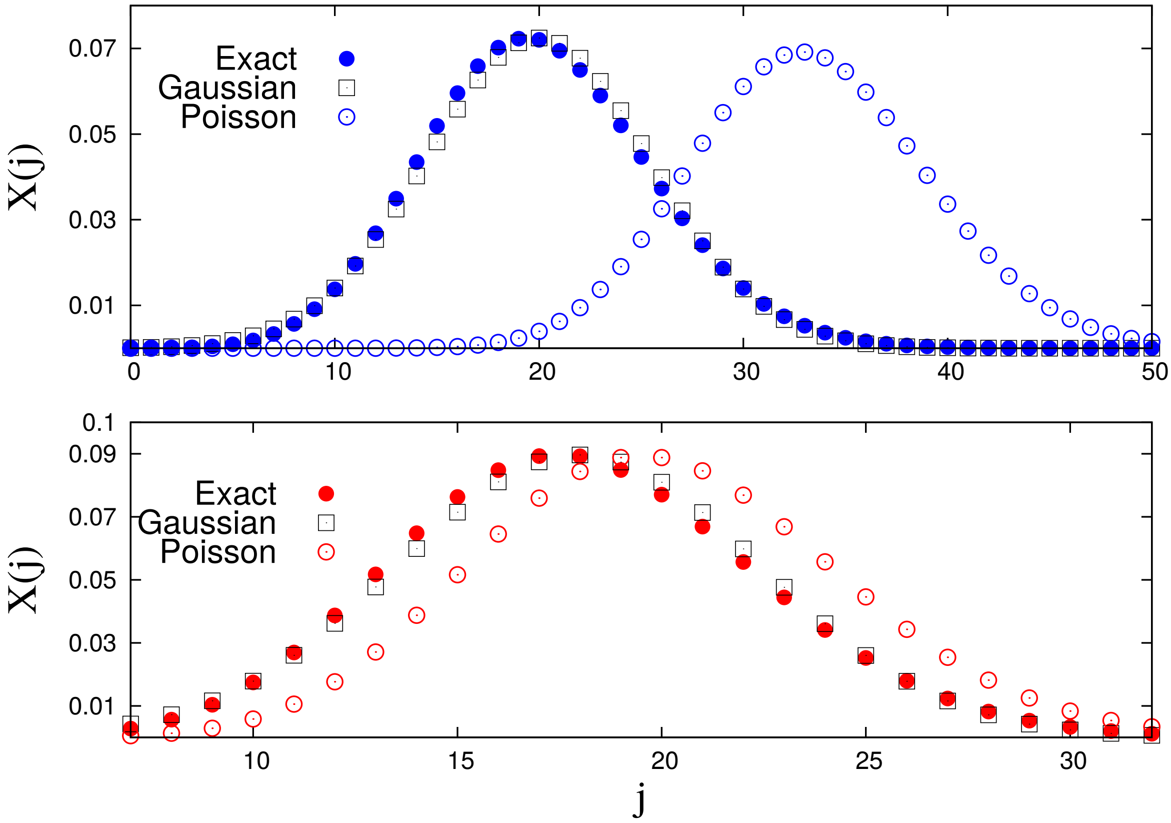


Fig. 3.4 Steady state distribution when $s \ll \sqrt{U_b U_d}$: Comparison of the exact distribution (3.27), Gaussian approximation (3.40) and Poisson distribution (3.5). The parameters in the top and bottom panel are $s = 0.003, U_b = 0.009, U_d = 0.1$ and $U_b = 0.001, s = 0.005, U_d = 0.1$, respectively.

3.4.2 When the parameter S is small

The parameter $S \ll 1$ when (i) $s < U_d$ and (ii) either $s < U_b$ or $U_b < s < \sqrt{U_b U_d}$. A biologically relevant situation where S can be small arises in the case of mutators where mutation rates can be as high as 10^{-2} [97]. Then for selection coefficient in the range $10^{-4} - 10^{-3}$, one obtains $S \sim 0.01 - 0.1$.

For small S, the minimum eigenvalue shown in Fig. 3.1 is calculated in Appendix B.7 and given by

$$\lambda_0 = \left(\frac{9\pi}{8}\right)^{2/3} S^{2/3} - \frac{S}{\gamma - 1} \, . \tag{3.37}$$

Using this in (3.28), we find that

$$\mathscr{C}_1 \approx \frac{U_d - U_b}{s} - \left[\frac{2}{S} - \left(\frac{9\pi}{8}\right)^{2/3} \frac{1}{S^{1/3}} + \frac{1}{\gamma - 1}\right] . \tag{3.38}$$

As our numerical analysis of (3.27) shows that the fraction $X(0) \sim e^{-1/S}$ for small S (also, see Fig. 3.4), the variance (3.30) can be approximated by

$$\mathscr{C}_2 \approx \frac{U_d - U_b}{s} , \tag{3.39}$$

as also seen in (3.35) when $s < U_d$. The above equation also shows that the variance is larger than the mean as illustrated in Fig. 3.2. The above approximations are in good agreement with the exact results, see Fig. 3.2.

An analysis of the frequency distribution (3.27) for small S described in Appendix B.8 suggests a Gaussian approximation for the frequency distribution,

$$X(j) \approx \sqrt{\frac{1}{2\pi\mathscr{C}_2}} \exp\left[-\frac{(j - \mathscr{C}_1)^2}{2\mathscr{C}_2}\right] , \tag{3.40}$$

where the mean and variance are given, respectively, by (3.38) and (3.39). Figure 3.4 compares the above approximation with the exact distribution (3.27) and we see a quite good agreement. However, it should be noted that unlike (3.40), the exact frequency distribution is not symmetric about the mean. The skewness defined as $\mathscr{C}_3/\mathscr{C}_2^{3/2}$, where $\mathscr{C}_3$ is the third cumulant, is a measure of the asymmetry of the distribution. Due to (B.18) in the stationary state, on neglecting the frequency $X(0)$, we find a nonzero $\mathscr{C}_3 = U/s$. Figure 3.4 also shows that both mean and variance are considerably affected by beneficial mutations when $s < U_b < U_d$.

3.5 Dynamics of the population frequency

In the last section, we discussed the stationary state and now turn to the dynamics of the frequency distribution. We will focus on the time dependence of the average $\mathscr{C}_1(t)$ starting

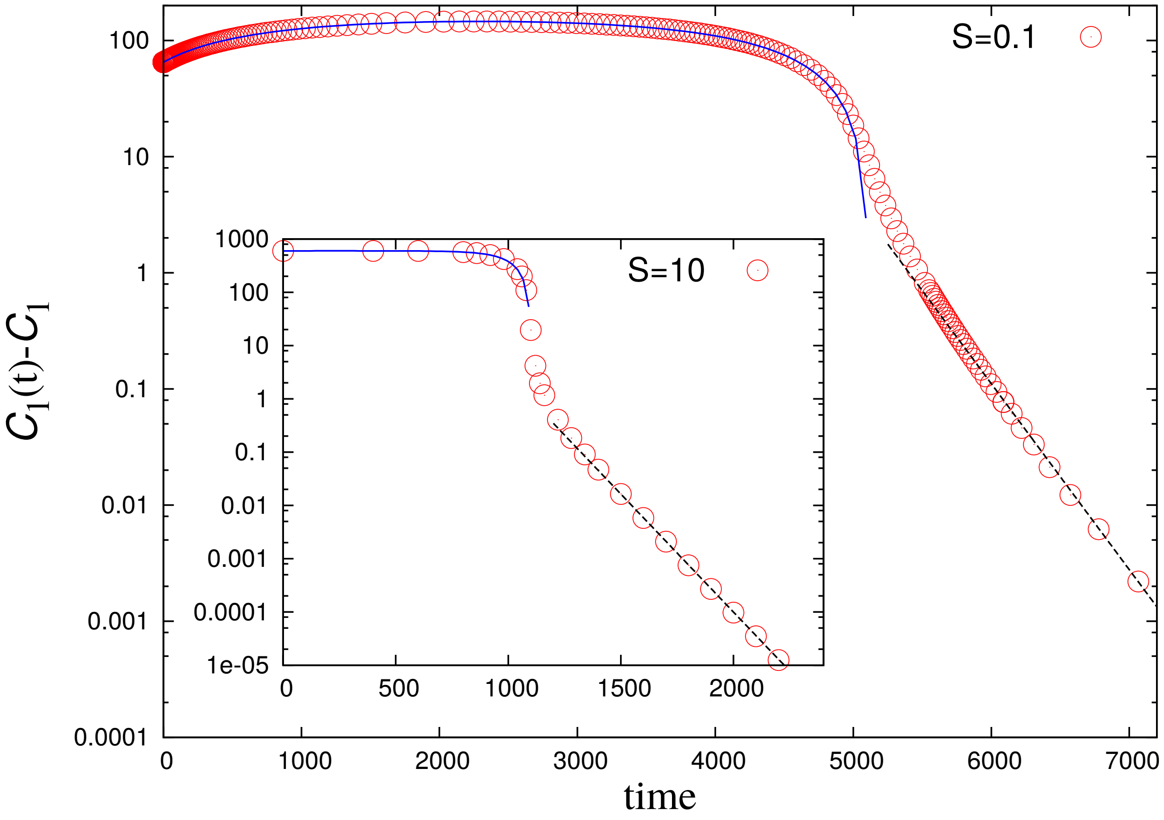


Fig. 3.5 Dynamics of the mean deviation from the stationary state, $\mathscr{C}_1(t) - \mathscr{C}_1$ for two values of $S = s/\sqrt{U_b U_d}$ with initial population located in the fitness class j_0. The exact dynamics obtained by numerically integrating (3.1a) and (3.1b) are shown by points while the solid (blue) lines show the short time dynamics (3.45) and the broken (black) line shows the relaxation dynamics be^{-Rt}, where the relaxation rate R is given by (3.47). The parameters in the main and inset are $U_b = s = 0.001, U_d = 0.1, j_0 = 150, b = 4.515 \times 10^8, \mathscr{C}_1 = 84.8958$ and $U_b = 0.0001, s = U_d = 0.01, j_0 = 600, b = 71561, \mathscr{C}_1 = 0.990147$, respectively.

from a monomorphic initial condition given by

$$X(j,0) = \delta_{j,j_0} \, . \tag{3.41}$$

As the exact expression (3.26) for the time-dependent frequency involves a sum over a large number of eigenvalues, the dynamics are more efficiently studied by solving the differential equations (3.1a) and (3.1b) numerically. The results for the mean thus obtained are shown in Fig. 3.5 for two values of S and, we observe (i) a short time regime where the population is

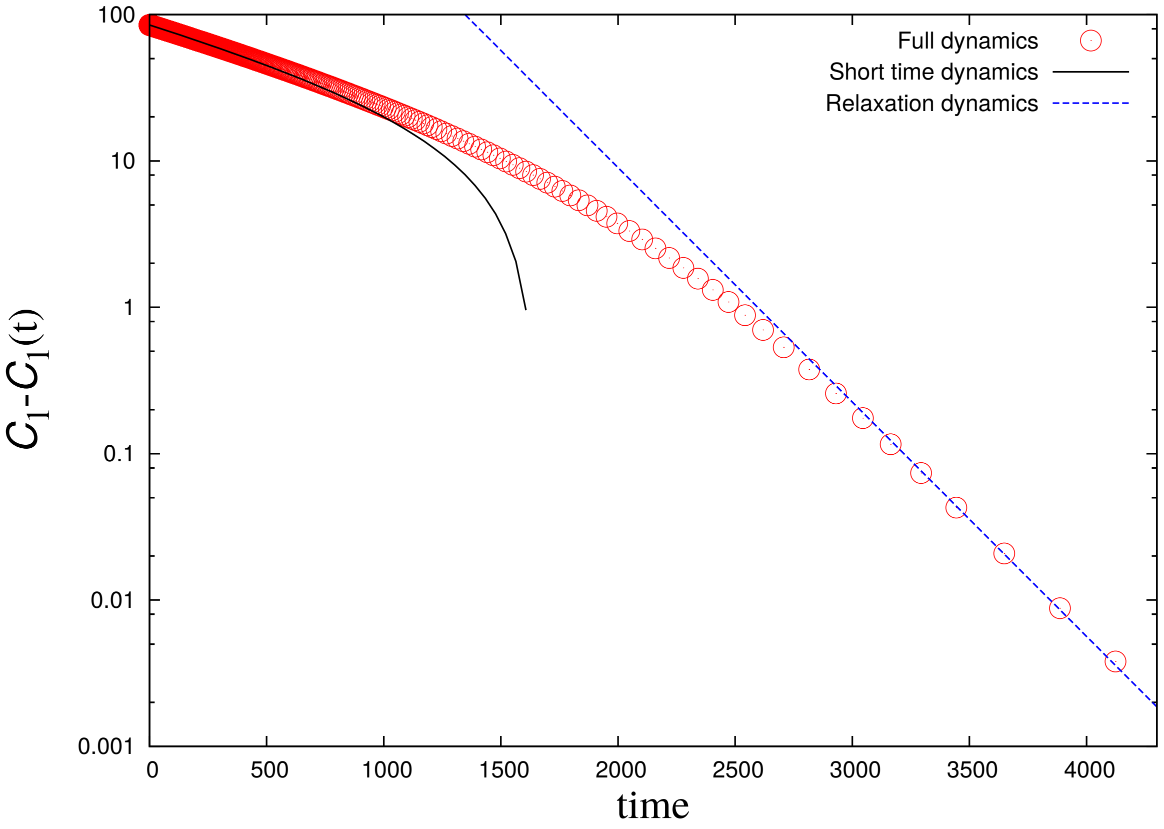


Fig. 3.6 Figure shows the dynamics of mean number of deleterious mutation when the population starts from $j_0 = 0$ at $t = 0$, which is below $\mathscr{C}_1 = 84.8958$. The short time and relaxation dynamics are captured by (3.45) and (3.47) respectively. Other parameters are $U_b = s = 0.001$ and $U_d = 0.1$.

far from the stationary state, (ii) an intermediate time regime where the mean changes quickly and (iii) a long time relaxation regime where the population is close to the steady state.

The initially monomorphic population first spreads over the genotypic space due to mutations followed by an increase in the frequency of high fitness genotypes as a result of selection. At short enough times, one can understand the dynamics away from the stationary state by ignoring the boundary at $j = 0$ (see, Section 3.5.1 below). Note that this holds even if the population is initially located in the zeroth fitness class because the mean initially increases for $U_d > U_b$ as shown in Fig. 3.6. However once the population is close to the stationary state (i.e., $\mathscr{C}_1(t) - \mathscr{C}_1 \lesssim 1$ in Fig. 3.5), the boundary at the zeroth fitness class becomes important. As discussed in Section 3.5.2, at long enough times, it is sufficient to retain the minimum and second minimum eigenvalue in the sum over the eigenvalues in (3.26) to determine the time to relax to the stationary state.

3.5.1 Dynamics far from the stationary state

The dynamics of the nth cumulant $\mathscr{C}_n(t)$ are described in Appendix B.4. For the initial condition (3.41), at short enough times, we can set the frequency in the fittest class to be approximately zero in (B.16) to obtain

$$\dot{\vec{\mathscr{C}}}(t) = -s\hat{D}\vec{\mathscr{C}}(t) + \vec{U} \ , \tag{3.42}$$

where $\vec{\mathscr{C}}$ and $\vec{U}$ are column vectors whose nth element is given by $\mathscr{C}_n$ and $U_d + (-1)^n U_b$ respectively and $\hat{D}$ is an upper shift matrix with matrix element $D_{mn} = \delta_{m+1,n}$. The above equation can be straightforwardly solved for arbitrary initial condition and for (3.41), we obtain

$$\mathscr{C}_n(t) = \frac{U_n}{s}\sinh(st) - \frac{U_{n+1}}{s}(\cosh(st) - 1) + j_0\delta_{n,1} \ . \tag{3.43}$$

Using (3.43) in (B.15), the generating function of the population fraction, $F(\xi,t) = \sum_{j=0}^{\infty} X(j,t)e^{-\xi j}$ can also be obtained and given by

$$\ln F(\xi,t) = -j_0\xi - \frac{U_d}{s}(1-e^{-st})(1-e^{-\xi}) - \frac{U_b}{s}(1-e^{st})(e^{\xi}-1) \ . \tag{3.44}$$

The above result generalises (53) of [87] who obtained it for special values of the parameters.

Due to (3.43), the approximate short time dynamics of mean $\mathscr{C}_1(t)$ and variance $\mathscr{C}_2(t)$ are given by

$$\mathscr{C}_1(t) = \frac{U_d}{s}(1-e^{-st}) + \frac{U_b}{s}(1-e^{st}) + j_0 \ , \tag{3.45}$$

$$\mathscr{C}_2(t) = \frac{U_d}{s}(1-e^{-st}) + \frac{U_b}{s}(e^{st}-1) \ . \tag{3.46}$$

The above equations show that for $t \ll 1/s$, both mean and variance change linearly with time with a slope that depends only on the mutation rates but for longer times, the mean varies exponentially fast at a rate s^{-1}. The short time dynamics of the mean given by (3.45) are valid as long as $|\mathscr{C}_1(t) - \mathscr{C}_1|$ is large and agree with the numerical results shown in Fig. 3.5.

3.5.2 Dynamics close to the stationary state

When beneficial mutations are absent, the frequency $X(j,t)$ is Poisson-distributed with mean $(U_d/s)(1-e^{-st})$ [74, 75] and thus approaches the stationary state at rate s, independent of the deleterious mutation rate. When the beneficial mutation rate U_b is nonzero, the dynamical evolution of the frequency is given by (3.26). At large but finite times, it is a good approximation to retain only the terms containing the minimum and second minimum eigenvalues in the sum over the eigenvalues in (3.26). Thus the frequency $X(j,t)$ relaxes to the steady state exponentially fast at rate

$$R = (\lambda_1 - \lambda_0)\sqrt{U_b U_d} \;, \tag{3.47}$$

where λ_1 is the second minimum eigenvalue. Unlike λ_0, the second minimum eigenvalue λ_1 is an increasing function of S as shown in Fig. 3.1.

The second minimum eigenvalue is calculated in Appendix B.5 and B.7 and given by

$$\lambda_1 = \begin{cases} \left(\frac{21\pi}{8}\right)^{2/3} S^{2/3} - \frac{S}{\gamma - 1} & , \; S \ll 1 \;, \\ S + 2 & , \; S \gg 1 \;. \end{cases} \tag{3.48}$$

This yields the relaxation rate $R = s + U_b$ for large S which shows that the population reaches the stationary distribution faster in the presence of beneficial mutations. For small S, we get $R \sim s^{2/3}(U_b U_d)^{1/6}$ which approaches zero as $s \to 0$ in accordance with the neutral case where the population never reaches a stationary state, see Appendix B.2. The relaxation dynamics of the mean $\mathscr{C}_1(t)$ are in agreement with the above results as shown in Fig. 3.5.

On comparing the results for short and long time dynamics, we find that while the former occurs over a time scale $\sim s^{-1}$ independent of the mutation rates, the relaxation time is determined by both selection and mutation.

3.6 Conclusions and Open questions

In this Chapter, we presented the exact solution (3.26) for the frequency distribution at all times for the model defined by (3.1a) and (3.1b). Our results summarised in Table 3.1 generalise the earlier ones in [4, 2, 75] by including beneficial mutations and extend the treatment in [86, 87, 80, 36, 88, 70] to all times including the stationary state limit. We discussed the biologically realistic situation where the beneficial mutation rate is smaller than its deleterious counterpart [89] but, for completeness, we explore the parameter regime $U_d \ll U_b$ in Appendix B.9.

Here we considered a mutation scheme in which the mutation rate per sequence is same for all sequences, irrespective of their fitness. The evolution of an infinitely large population on additive fitness landscapes when the mutation rates depend linearly on the number of loci carrying deleterious allele has also been studied [58, 3] and the relationship of this mutation scheme with the one studied in this Chapter is elucidated in Appendix B.1. In the fitness-dependent mutation rate model [58, 3], when the number of loci carrying the deleterious allele is small, the beneficial mutation rate vanishes in the limit of infinitely long sequence. This has the immediate consequence that the stationary state properties are not affected by beneficial mutations in this mutation scheme [3].

In contrast, for the model studied here, the general effect of beneficial mutations is to decrease both mean and variance in the stationary state (see, (3.28) and (3.30)) but the extent to which this happens depends on the strength of selection relative to mutations. We find that

(i) when $U_b < U_d < s$, beneficial mutations have a minor effect since the mean number of deleterious mutations in the absence of beneficial mutations is already close to zero,

(ii) when $U_b < s < U_d$, beneficial mutations decrease the mutational load significantly and the frequency is enhanced (diminished) in fitness classes below (above) U_d/s and

(iii) when $s < U_b < U_d$, both mean and variance decrease considerably and the frequency distribution is a nontrivial function.

In this Chapter, we focused on the deterministic evolution, and ignored the effect of random genetic drift. However, we have numerically studied the stochastic evolution of finite size population and as discussed in Appendix B.10, we find that even in the presence of beneficial mutations, the population attains a steady state only when the population size is above a critical value N_c [51]. Future investigations of the finite size population are desirable to obtain an analytical understanding of the critical population size N_c. Since the bulk of the distribution is expected to be described well by deterministic distributions [80, 36, 88, 51], the exact answers obtained for the deterministic system in this Chapter can be used to understand the behavior of the finite population better. A detailed study of the evolution of finite populations exploiting the results presented here will be taken up in future.

Chapter 4

(Dis-)Advantage of Recombination on Rugged Fitness Landscapes

4.1 Introduction

In Chapter 2 [3], we saw that recombination results in higher *steady state* fitness for a finite-sized population evolving on a single peak fitness landscape. However, as discussed in Section 1.2, single peak fitness landscapes are rare and most of the experimentally measured fitness landscapes show intermediate level of ruggedness [19]. In this Chapter, we study the effect of recombination on a class of tunable rugged fitness landscapes to see how our finding of advantage of recombination on a single peak fitness landscape [3] changes when the topology of the fitness landscape changes [7].

A key distinction between evolution on smooth vs. rugged fitness landscapes is that finite populations evolving on the latter may not attain a steady state, even over fairly long time scales, because of the tendency to get trapped at local peaks. Here, we monitor the fitness effect over multiple time scales – at both short times (when fitness is changing rapidly) and intermediate to long times (when it changes very slowly due to trapping in a local fitness peak). We use extensive simulations with the Wright-Fisher model in the presence of other evolutionary forces such as mutation, selection, and drift. We find that the effect of recombination has a complex dependence on these forces and it changes from advantageous to disadvantageous as the parameter regime is varied. We attempt to give an intuitive explanation for the existence of different parameter regimes of advantage and disadvantage of recombination. We also explore how the initial state of the population – its initial 'position' on the fitness landscape, as well as the extent of genetic variation it harbors – crucially determines whether recombination can increase population fitness or not. This sort of strong dependence on the initial conditions has remained largely unexplored in previous works [9, 8].

The Chapter is organized as follows: we first describe the methods used in this study and then discuss our main results for the effect of recombination on population fitness and its dependence on various parameters. Finally, we discuss the results obtained for different initial conditions.

4.2 Materials and Methods

We consider the Rough Mount Fuji (RMF) fitness landscapes with fitness function given as [16, 98, 99, 20, 18, 9]

$$W(\{\sigma\}) = cd_{\sigma,\sigma_*} + B\eta(\sigma). \tag{4.1}$$

Here, d_{σ,σ_*} is the Hamming distance between σ and the reference sequence $\sigma_* = (1,1,1,1...1)$, which is equal to the number of zeros in the sequence and η is a random variable drawn independently for each sequence from a normal distribution with mean zero and variance one. Thus, the RMF fitness landscape is parametrized by the ratio $\theta = B/c$ (also see Section 1.2.3). In the following simulations, c is set to 1 and the ruggedness is tuned by varying B .

Once the fitness landscape is created, we use the usual Wright-Fisher simulation method described in Chapter 1 to study the population dynamics. As a first step, we define the genome as a binary string of length L. Fitness values of all possible 2^L sequences are assigned from the fitness function given in(4.1). A population of size N is introduced to this fitness landscape and allowed to evolve under the action of evolutionary forces such as mutations, selection, random genetic drift, and recombination. The two main initial conditions we consider are: (i) a random initial condition where each individual is assigned a random sequence (this corresponds to a situation with maximum initial variation) (ii) a monomorphic initial condition where all individuals are assigned an identical genome and thus have the same fitness value (this corresponds to a situation with zero initial variation). This genome can either be chosen randomly or may represent a special point on the fitness landscape such as the global peak, a local peak, or a valley. During reproduction, each individual is allowed to go through three steps, which include a recombination step with a rate r, a mutation step in which each locus has a probability μ to mutate to the other allelic state, and a selection step (also, see Section 1.4.2.1). The specific question we want to address in this study is: What is the effect of recombination in different regimes of this high-dimensional parameter space (N, L, μ, r, θ)?

The population fitness

$$\overline{w}(t) = \frac{1}{N} \sum_{individuals,i} w_i(t), \tag{4.2}$$

is measured at each generation. We average $\overline{w}(t)$ over several histories and monitor it over multiple time scales. The advantage of recombination is measured as the difference between $\overline{w_r}(t) - \overline{w_0}(t)$, where $\overline{w_r}(t)$ and $\overline{w_0}(t)$ are average fitness with and without recombination, respectively. The relative fitness advantage at time t is given as

$$\Delta w = \frac{w_r - w_0}{w_0}, \tag{4.3}$$

where, w_r and w_0 are the fitnesses of recombining and nonrecombining genotype at a fixed time point t.

The two extreme limits of the smooth ($B = 0$) and the maximally rugged fitness landscapes ($c = 0$) are relatively well studied.

1. $B = 0$: Our results discussed in Chapter 2 [3], show that recombination has a clear advantage in reducing the mutation load at equilibrium state.

2. $c = 0$: In this limit, recombination shows a short time advantage that vanishes with time because of the lower probability of escaping from local maxima [100].

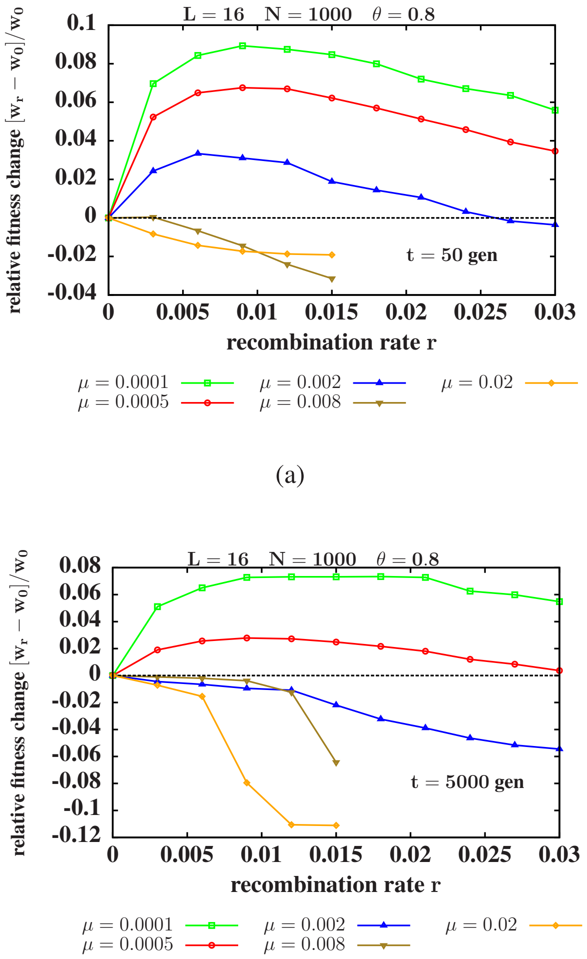
The escape probability is measured by starting with all the population at a local peak in h (∼1000) numerical experiments and counting the cases where the population fitness changes (n) within tmax (∼ 100000) generations. The probability of escaping the local peak is defined as n/h. Although realistic fitness landscapes are characterized by an intermediate level of ruggedness [18, 16, 19], the effect of recombination on moderately rugged fitness landscapes is much less explored. Previous study by Nowak et al.,[9] which looked into this reported that recombination has a short term advantage and the effect reverses due to the tendency of recombining population to get trapped at local peaks. In our study, we analyse how these results depend on the initial condition that adds to the completeness of these already existing results.

4.3 Results

We first discuss the effect of different evolutionary forces on the (dis-)advantage of recombination. Since the population does not attain a steady state, the effects are expected to be dynamic. Thus, we analyze the short time as well as long time effects and a comparison with these two is also discussed. All results presented in this section assume a random initial condition, where each individual is assigned a sequence randomly from the 2^L possible sequences (maximum initial variation) [8].

4.3.1 Recombination rate

The advantage of recombination changes non-monotonically with recombination rate for small mutation rate which is shown in Fig. 4.1. This observation is consistent with the zero mutation rate results obtained in a previous study [8]. This effect is easy to understand, because when mutation rate is low recombination is preferred as it is the only force that can create variation in the population. However, once recombination crosses an optimum value, it becomes disadvantageous because it can destroy the favorable configuration (local peak genotype) by chance before selection could act on it and increase its frequency.



(a)

(b)

Fig. 4.1 Figure shows the advantage of recombination with recombination rate r for different mutation rates. Panel (a) short time effect with Δw measured at 50th generation and (b) long time effect with Δw measured at 5000th generation. Other parameters are $L = 16$, $N = 1000$ and $\theta = 0.8$.

4.3.2 Mutation rate

Figure 4.1 also shows a transition in the effect of recombination from an advantageous phase to a disadvantageous phase as mutation rate increases. A schematic diagram of the effect of recombination in the $\mu - r$ plane is shown in Fig. 4.2. For a fixed recombination rate, the transition from one phase to another occurs at a critical mutation rate μ^*.

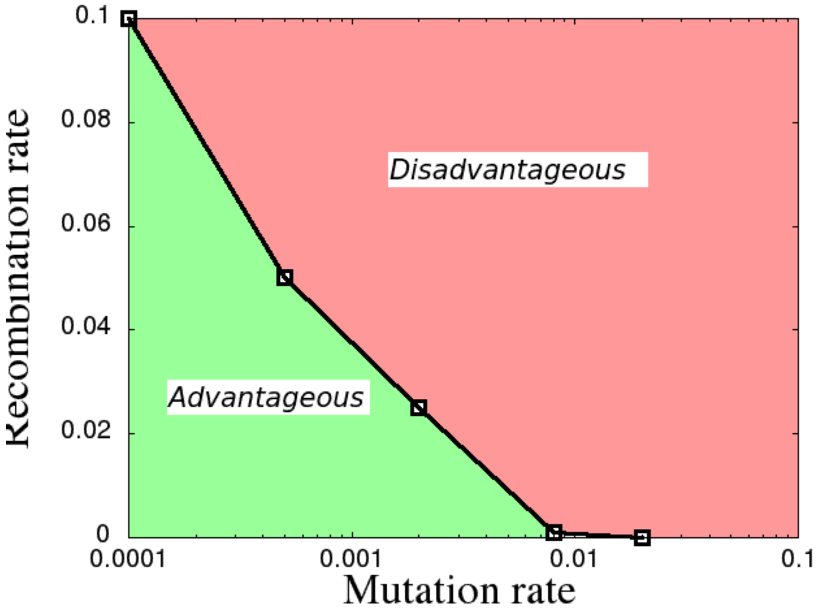
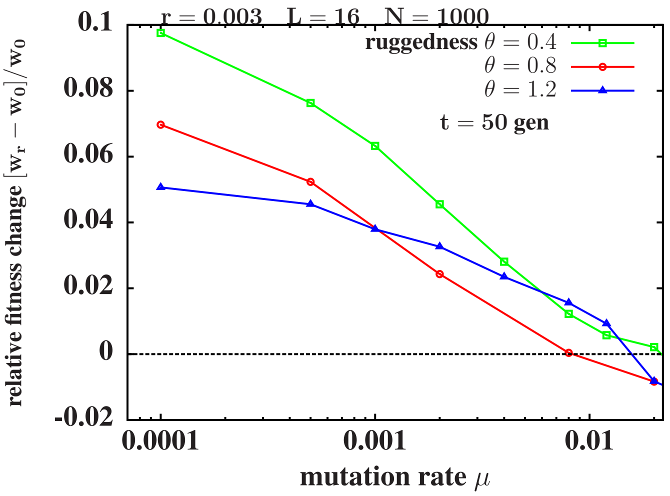


Fig. 4.2 Figure shows the transition of effect of recombination in $\mu - r$ plane. Effect changes from an advantageous phase to a disadvantageous phase as mutation rate increases beyond a critical value μ^*.

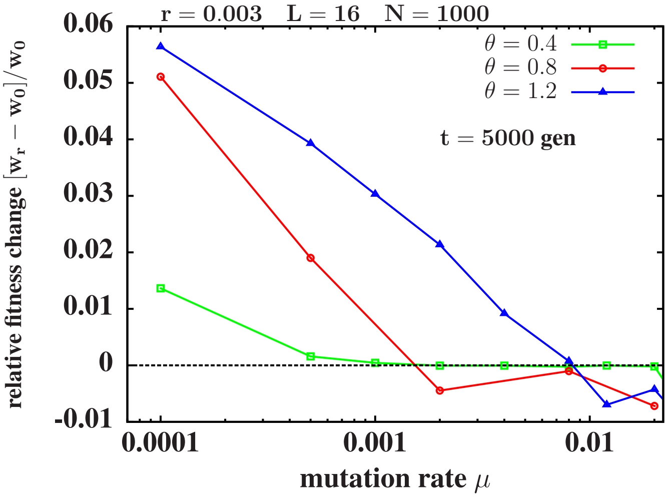
Recombination increases existing genetic variation, in particular, by generating fit combinations (local peak sequences) out of existing variation, which in the short run would not be accessible by mutation alone. This is particularly true when mutation rates are low. However, at higher mutation rates, many of these local peaks (as well as the global peak) become more accessible by mutation itself. In fact, in populations where adaptation via mutation is relatively efficient, recombination can actually have a detrimental effect by breaking up favorable combinations much faster than they are generated. This explains why even the short term advantage of recombination declines with mutation rate (see Fig. 4.1(a)). Therefore, if mutational force is sufficiently high to create enough variation, recombination becomes disadvantageous.

4.3.3 Ruggedness

Figure 4.3 shows the effect of ruggedness on the advantage of recombination. In short time, fitness landscape with a low level of ruggedness shows a higher advantage of recombination which is shown in Fig. 4.3(a). However, this advantage vanishes fast and at long time, advantage persists only in a fitness landscape with a high level of ruggedness as shown in Fig. 4.3(b).



(a)



(b)

Fig. 4.3 Figure shows the effect of recombination with mutation rate for different ruggedness parameter $\theta = 0.4, 0.8$ and 1.4. Other parameters are $r = 0.003, L = 16$ and $N = 1000$.

4.3.4 Significance of μ^*

All these observations suggest that for a particular fitness landscape, there is a mutation rate μ^*, such that any increase in genetic variation (either by a further increase in mutation rate or by introduction of recombination) leads to a decrease in mean fitness. The transition point

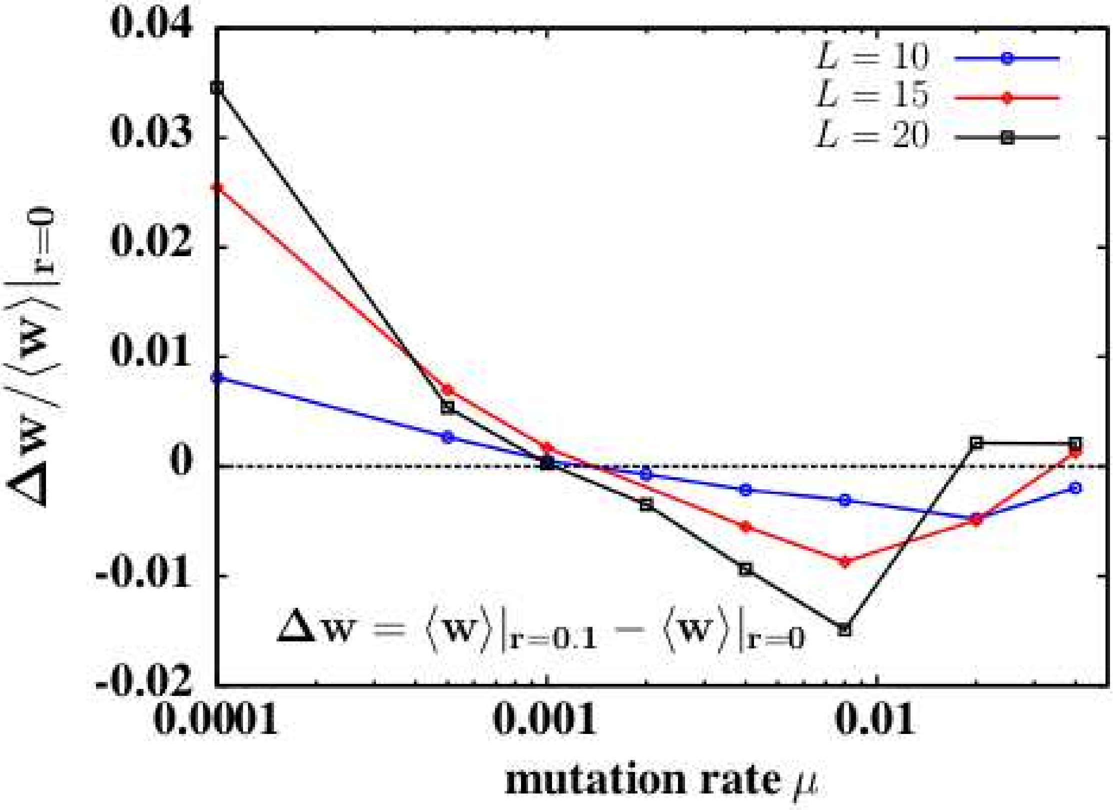


Fig. 4.4 Long term (dis-)advantage $\Delta w / \langle w \rangle|_{r=0}$ of recombination, $r = 0.1$ as a function of mutation rate μ for different L for ruggedness parameter $\theta = 0.5$.

μ^* is dynamic and its rate of decay is expected to depend on the ruggedness of the fitness landscape. As time increases, μ^* shifts towards the lower bound resulting in a long time disadvantage of recombination even for small mutation rates. The rate of decay of μ^* to zero depends on the ruggedness of the fitness landscape: it is expected to be slower in a maximally rugged fitness landscape compared to the one with a low level of ruggedness. More careful study of the dynamics of μ^* and the effect of ruggedness will be taken up in future.

4.3.5 Effect of varying genome length L

Figure 4.4 shows the fractional change $\Delta w / \langle w \rangle|_{r=0}$ in fitness due to recombination as a function of μ for different values of L. At small μ, where recombination is advantageous, the advantage becomes higher with increasing L. At intermediate μ, where recombination is disadvantageous, the disadvantage becomes greater with increasing L. At very large μ, recombination again becomes advantageous, with the advantage being more apparent at larger L. Thus, the characteristic $U-$ shaped curve obtained on plotting $\Delta w/\langle w\rangle|_{r=0}$ vs. μ shows much steeper variation as L is increased. This data is obtained with the assumption of one break point, but having more than one break point does not change the qualitative behaviour.

4.4 Initial condition dependence

The trends observed here, with the combined effect of all these five parameters, showed a strong dependence on initial condition as well. All the results discussed above are based on the assumption of an initial condition with maximum variation. Now, we study the effect of other initial conditions in which population is clustered about- global peak, local peak, in a valley, or any random point in the fitness landscape.

4.4.1 Global peak

Recombination shows no effect in low mutation rates and it becomes disadvantageous at high mutation rate. The disadvantage is higher for high recombination rate which is shown in Fig. 4.5.

4.4.2 Local peak

Even though the recombining genotype shows high variation, it has lower probability to escape from the local peak shown in Fig. 4.6(b). This effect of getting stuck in local peak results in a long time disadvantage of recombination which is consistent with the results obtained in the previous studies [9].

4.4.3 Local valley

In small mutation rates, recombination shows an initial short term advantage in attaining a local peak. But, it gets stuck in local peak very fast and the recombination becomes disadvantageous as in the previous case.

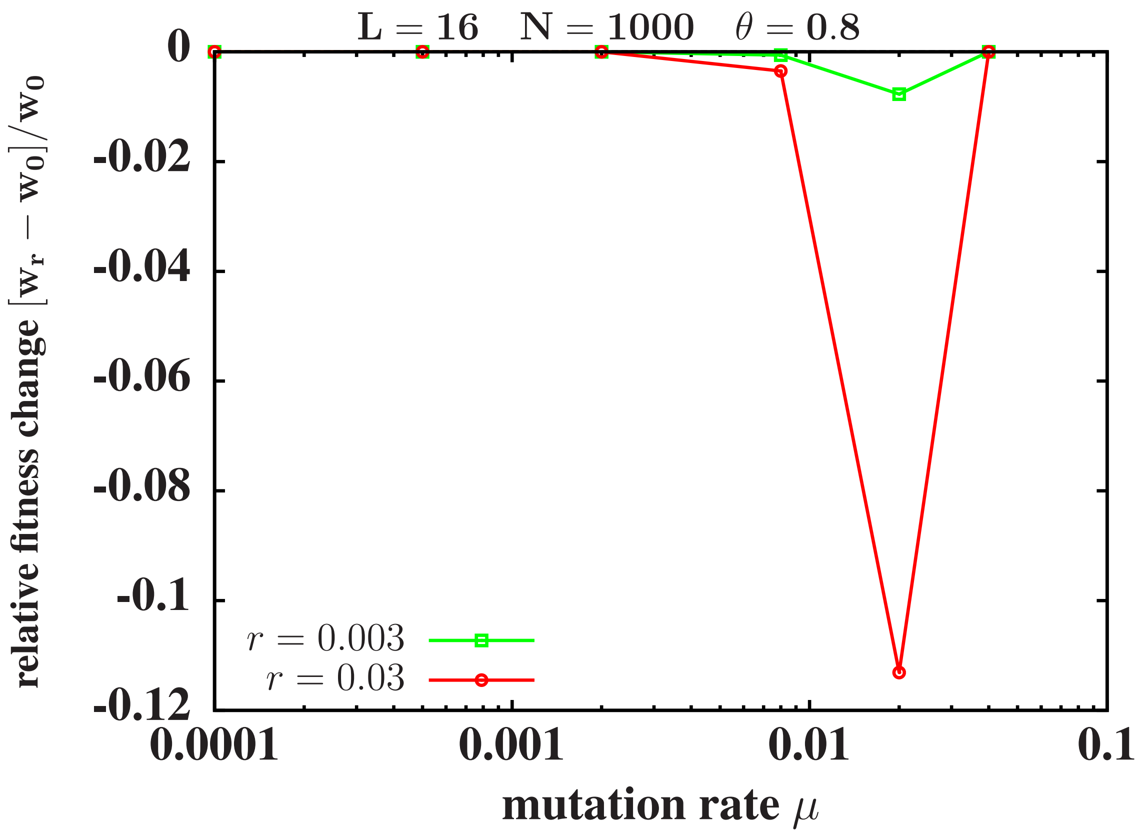
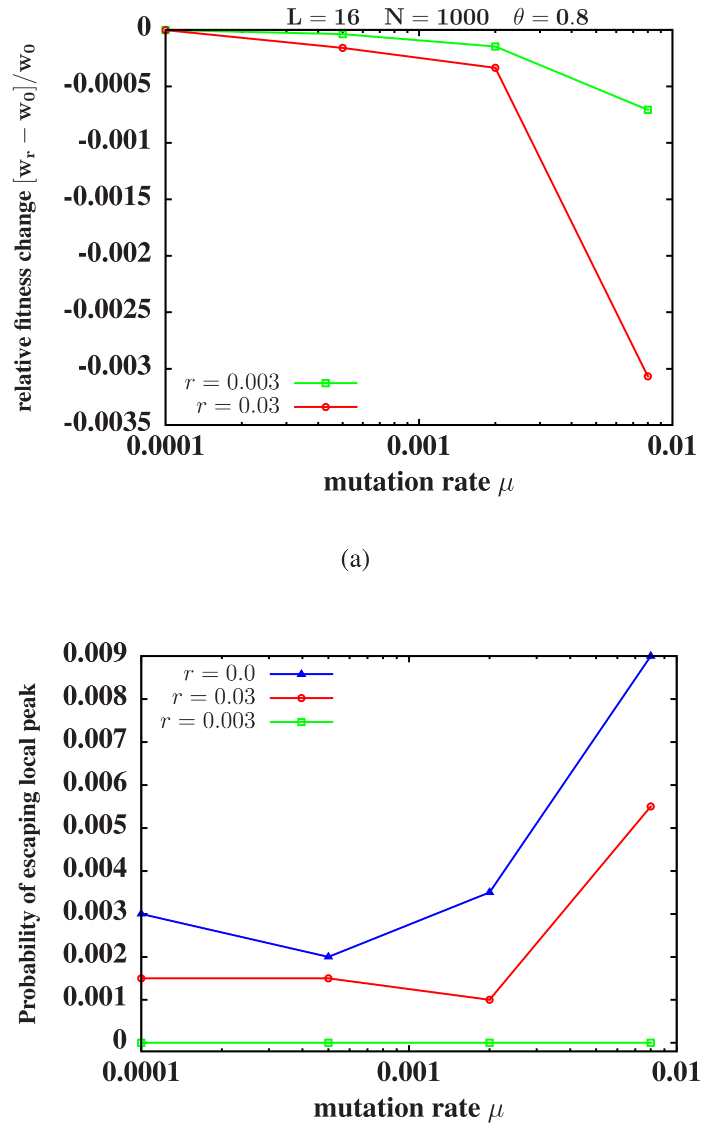


Fig. 4.5 Figure shows the disadvantage of recombination when the population is at global peak, for two different recombination rates. Other parameters are $L = 16, N = 1000$ and $\theta = 0.8$.

4.4.4 Populations with variable initial spread on the fitness landscape

We consider populations with different levels of initial variation, ranging from a delta function initial condition to sequences distributed normally with width d around a reference sequence, where d represents the pairwise difference between the chosen sequence and the reference sequence.

Recombination has a short time advantage that vanishes with time as shown in Fig. 4.7. The advantage is maximum for high initial variation and it persists for a longer time. The trend observed is consistent with both high and low mutation rate, but in high mutation rate advantage vanishes faster as expected.



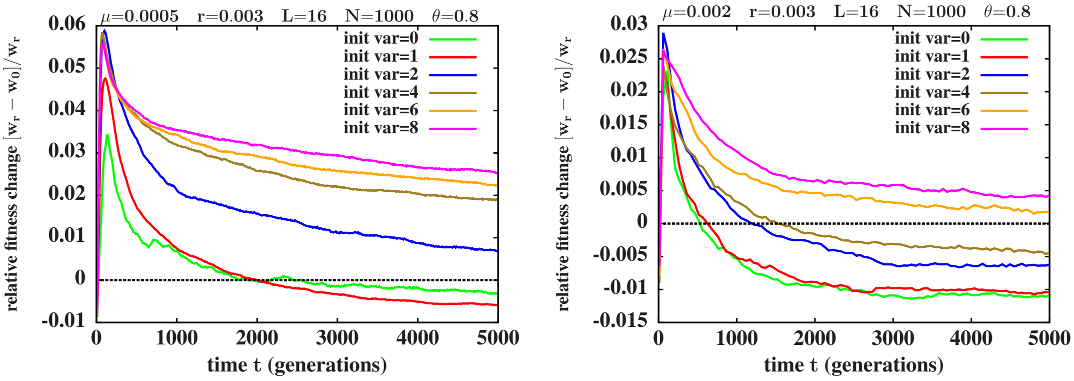
(a)

(b)

Fig. 4.6 Figure shows effect of recombination when the population is at local peak.Other parameters are $L = 16$, $N = 1000$ and $\theta = 0.8$.

4.5 Conclusions

Whether or not recombination is advantageous depends on where the population is in the (N, L, μ, r, θ) parameter space. We explored the variation with μ, r, and θ and found that: (i) Fitness changes non-monotonically with r for a wide range of μ and θ values: recombination can be



(a) (b)

Fig. 4.7 Figure shows the effect of recombination with time for different level of initial variation. We consider two cases with mutation rate $\mu = 0.0005$ and $\mu = 0.002$, shown in (a) and (b) respectively. Other parameters $r = 0.003, L = 16, N = 1000$ and, $\theta = 0.8$ are kept fixed.

advantageous at small rates and disadvantageous at large rates. (ii) For small mutation rates μ, recombination at a small rate is always beneficial, but this benefit decreases with μ. Beyond a critical value μ^*, recombination at howsoever small a rate is detrimental. (iii) The range of μ and r over which recombination confers advantage increases with ruggedness parameter B/c. (iv) The change in fitness (both positive and negative) due to recombination becomes higher in magnitude with increasing L.

(Dis-)advantage of recombination depends strongly on the strength of mutation that creates variation. The short time and long time dynamics of average fitness increase (adaptation) show different dependence on recombination in the presence of mutations. The short time dynamics is determined by the potential to generate fit combinations from the existing initial variation and quickly access the nearest local peaks. In general, recombination has an advantage in this regime. The long time dynamics is determined mainly by the ability to escape from a local peak and reach the global peak. Recombination shows a disadvantage in this regime as it reduces the probability to escape from the local peak as shown in Fig. 4.6(b).

Another important effect we observe is the strong dependence on the initial condition, which suggests that the advantage or disadvantage of recombination depends on the ability to reach the local/global peak. To attain a local/global peak, it is necessary to create variation in the population. Once the population attains a peak, recombination seems to have a negative effect on the probability to escape from the peak and therefore shows disadvantages. In short, recombination shows a short term advantage when mutation rate is weak and this advantage dies once the population attains a local/global peak. This short term advantage is the highest when the initial variation is high and the ruggedness is low. However, the advantage persists longer in very rugged fitness landscape with low mutation rates.

Chapter 5

Exploiting the adaptation dynamics to predict the distribution of beneficial fitness effects

5.1 Introduction

In this Chapter, we study the second question of interest mentioned in Chapter 1, namely, the dynamics of adaptation process and its relationship to the distribution of beneficial fitness effects (DBFEs)[6].

Microbial populations have to constantly adapt in order to survive in a changing environment. For example, a bacterial population exposed to a new antibiotic must evolve in order to exist [101]. In asexual populations, this process of adaptation is driven only by rare beneficial mutations [12] which provide fitness advantage. So in order to survive in new environment, enough beneficial mutations should be available and the beneficial mutations should confer sufficient fitness advantage. While the first factor depends on the mutation rate and population size, the second factor is determined by the underlying fitness distributions. Even though we have some understanding about the mutation rate of different microbial populations, the full fitness distribution is more complex and relatively little is known about it. But for moderately adapted populations (i.e. fitness of the wildtype is high enough), rare beneficial mutations

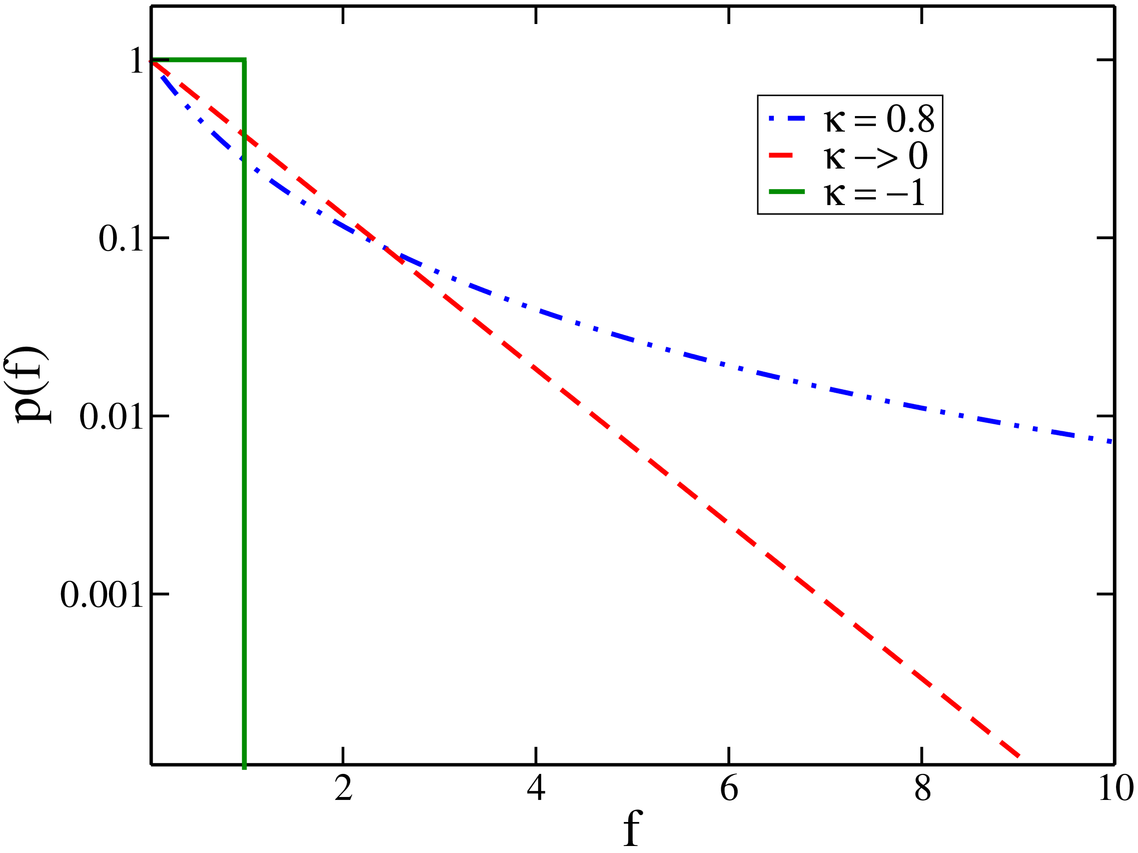


Fig. 5.1 The figure shows the distribution of beneficial fitness effects $p(f)$ with fitness f for the three EVT domains, given by (5.1) for various κ. Here κ is the tuning parameter with $\kappa > 0$, $\kappa \to 0$ and $\kappa < 0$ corresponding to Fréchet, Gumbel and Weibull domains respectively.

which occur in the tail of the fitness distribution can be described by the extreme value theory (EVT) as proposed first by Gillespie [102]. The EVT states that the extreme tail of all distributions of uncorrelated random variables (fitness, in this case) can be of three types only. Depending on whether the tail of underlying fitness distribution is truncated or decaying faster than a power law or as a power law, the EVT distribution would belong to Weibull or Gumbel or Fréchet domain, respectively [103].

All the three EVT domains can be obtained from the generalized Pareto distribution given as

$$p(f) = (1+\kappa f)^{-\frac{1+\kappa}{\kappa}}, \tag{5.1}$$

where κ is the tuning parameter.

Quantities	DBFE domains: Low mutation regime			DBFE domains: High mutation regime		
	Weibull	Gumbel	Fréchet	Weibull	Gumbel	Fréchet
$\overline{\Delta f_{step}}$	[10]	[10]	[10]	[6]	[6]	[6]
$\bar{\mathscr{F}}(t)$	[6]	[13]	[13]	[6]	[13]	[6]

Table 5.1 Here, $\overline{\Delta f_{step}}$ is the average fitness difference between the present leader and the new beneficial mutation that gets established and $\bar{\mathscr{F}}(t)$ is the rate of change of fitness.

One example from each of the three EVT domains is shown in Fig. 5.1, which shows the distribution of beneficial effects $p(f)$ with fitness f. The three types of EVT domains are classified according to the value of κ. Here negative κ belongs to the Weibull domain, while $\kappa = 0$ corresponds to Gumbel domain and positive κ to Fréchet domain. Interestingly, all the three DBFEs have been observed in experiments on microbial populations [11, 104–112]. While the exponential distribution belonging to the Gumbel domain has been most commonly seen [11, 104, 105, 107], in recent times, the distribution of beneficial mutations belonging to Weibull [108, 112] and Fréchet [109] domains have also been observed.

Recent theoretical studies have shown analytically and numerically that qualitatively different patterns occur in the adaptation dynamics of populations in different EVT domains of DBFEs in low mutation regime [113–117, 10, 118, 119]. Specifically, it has been shown that the fitness gain in a fixation event follows the pattern of diminishing returns in Weibull domain, constant returns in Gumbel domain and accelerating returns in Fréchet domain, and thus indicates that this quantity can be used to predict the DBFE. But these observations are restricted to strong selection-weak mutation (SSWM) regime in which the genetic variation in the population is minimal, that is, only one beneficial mutation is present in the population in the time interval between its appearance and fixation [104]. It is then natural to ask whether the relationship between the adaptation dynamics and the DBFE mentioned above holds for large populations as well, where there might be more than one beneficial mutation competing for dominance in the population. The main aim of our study is to address this question and to see if the fitness gain in a fixation event can be used for predicting the DBFE in a more general scenario.

Here we are mainly concerned with the populations in which a large number of mutants are produced at every generation. Hence, more than one beneficial mutation is expected to be present at the same time [1, 35, 120, 80, 28]. In this case, the beneficial mutations will compete with each other as has been observed in different experimental populations [121–124]. In this high mutation regime, as a result of the competition among the beneficial mutations, the rate of adaptation slows down. The fitness advantage due to the mutations that get fixed is much higher, since the availability of more mutations results in allowing only the best (fittest) mutation to get fixed [70]. A clear comparison of the population fraction of new mutants appearing in the population for two mutation regimes is given in Fig. 5.2. In Fig. 5.2(a) we see that the population in the SSWM regime is more or less monomorphic with only one mutant present at a time in all the three EVT domains. However, in high mutation regime, population is polymorphic with more than one mutant produced in it at every generation as shown in Fig. 5.2(b). In fact, a large amount of genetic variation is observed in the case of bounded distributions corresponding to $\kappa < 0$ in (5.1) resulting in strong competition between the beneficial mutants.

Here, we have used Wright-Fisher dynamics described in Chapter 1 to study the adaptation dynamics in high and low mutation regimes for the three EVT domains of DBFE. The main motivation of this study is to look for quantities which can be used to distinguish between the DBFEs using the properties of adaptation dynamics as opposed to the direct measurements of DBFEs. Our most important and interesting result is concerned with fitness difference between mutations that spread in the population which shows qualitatively different trends in three EVT domains and thus helps in distinguishing the DBFEs.

We have also studied another quantity which is the rate of change of fitness with time, and observed that this shows quantitatively different behaviour for different EVT domains of the DBFEs. Though some results for the rate of change of fitness are already known in the literature [13], we measured it for all the three cases (Weibull, Gumbel and Fréchet) and identified that this can be used to distinguish the DBFEs in both SSWM and high mutation regimes. To obtain a complete picture, a comparison of our study with the existing literature is given in Table 5.1 below.

We also measured quantities like the genetic variation and the number of mutations in the most populated sequence. All of these quantities are discussed in Section 5.3 and 5.4. We suggest that the distinct trends shown by the above mentioned quantities can be used to predict DBFEs from experimental studies on adaptation. The relevance of our work to experiments is also explored in Section 5.7.

5.2 Materials and Methods

We track the dynamics of a population of self-replicating (asexual), infinitely long binary sequences of fixed size using the standard Wright-Fisher process [120, 70], which is described in Chapter 1. In our work, the population size is held constant at $N = 10^4$, unless specified otherwise and the total mutation probability (beneficial and deleterious) per sequence is given by μ. Every occupied sequence is counted as a *class* and labelled when it arises in the population. Initially, the whole population is in class 1 whose fitness is fixed and specified in every simulation run. We have used the term leader to refer to the class whose normalised probability of reproduction (product of population fraction and fitness) is greater than half. In that case, clearly, class 1 is the initial leader since the whole population is localized there. At every time step, out of N sequences, m_t are chosen from a binomial distribution with mean $N\mu$ as mutants. Every mutant produced increases the number of classes in the population by one, and with time, the mutants may produce their own set of further mutants. The population fraction of each class may grow or go extinct, as can be observed in Fig. 5.2. At any time t, the number of classes present in the population is given by $\mathscr{N}_c(t)$, and the population size and fitness of each class, i, where $1 \leq i \leq \mathscr{N}_c$, is denoted by $n(i,t)$ and $f(i)$, respectively. The normalized probability of each class at every time step, $\tilde{p}(i,t)$ contributing offspring to the population at the next time step, depends on the population size of the class at the present time step and the fitness of the class as

$$\tilde{p}(i,t) = \frac{n(i,t)\, f(i)}{\Sigma_{j=1}^{\mathscr{N}_c(t)} n(j,t) f(j)}. \tag{5.2}$$

Note that though the fitness of the class is the same as long as it persists in the population, its size may vary at every time step, thus changing its probability of reproduction as given by (5.2). The different classes are populated in the next time step based on the multinomial distribution

$$P(n(1,t'),n(2,t')..n(\mathcal{N}_c,t')) = N! \prod_{j=1}^{\mathcal{N}_c(t)} \frac{[\tilde{p}(j,t)]^{n(j,t)}}{n(j,t)!} \tag{5.3}$$

where $t' = t+1$. The above equation is subject to the constraint $\Sigma_{j=1}^{\mathcal{N}_c(t)} n(j,t') = N$. In our simulations, we implement (5.3) along with the above constraint by converting (5.3) to a binomial distribution for every class, $1 \leq i < \mathcal{N}_c(t)$ as

$$n(i,t') = \binom{\tilde{N}(i)}{n(i,t)} q(i,t)^{n(i,t)} (1-q(i,t))^{\tilde{N}(i)-n(i,t)} \tag{5.4}$$

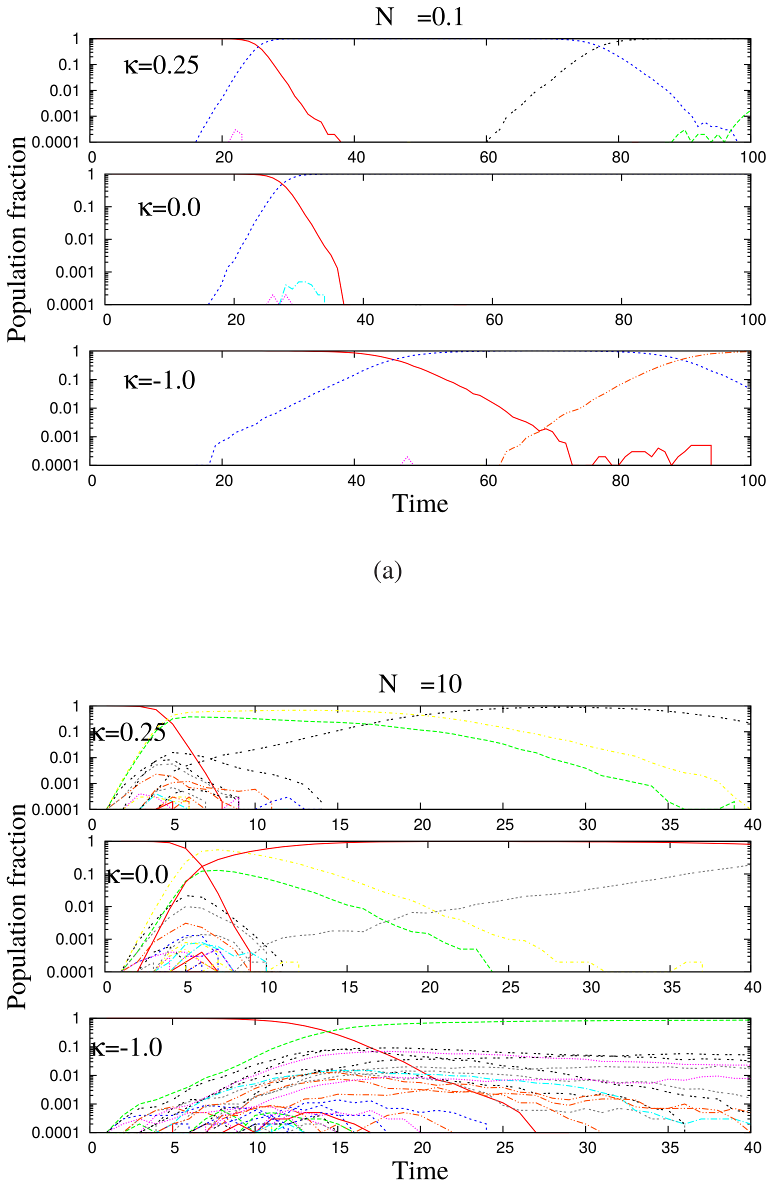
We set the population size of the last class as $n(\mathcal{N}_c(t),t') = N - \Sigma_{i=1}^{\mathcal{N}_c(t)-1}\ n(i,t')$. In (5.4),

$$q(i,t) = \frac{\tilde{p}(i,t)}{\Sigma_{j=i}^{\mathcal{N}_c(t)} \tilde{p}(j,t)} \tag{5.5}$$

and $\tilde{N}(i) = N - \Sigma_{j=1}^{i-1} n(j,t)$.

At every time step, once the classes are populated based on the algorithm described above, m_t sequences are chosen as mutants based on the binomial distribution with mean $N\mu$. Every new mutant class that appears in the population reduces the population size of the class in which it arose by one. In our work, we have varied μ to access both the SSWM (low mutation) and the high mutation regime. In our simulations unless specified otherwise, $N\mu = 0.01$ in low (SSWM) and $N\mu = 50$ in high mutation regimes.

A new class is assigned to each mutant and its fitness is chosen from a generalized Pareto distribution [103] given in (5.1). The advantage of using (5.1) is that we can access all three EVT domains of DBFE by changing κ. The distributions whose $\kappa < 0$ belong to the Weibull domain, while $\kappa = 0$ belong to the Gumbel domain, and $\kappa > 0$ belong to the Fréchet domain, respectively. The frequency distribution of beneficial effects $p(f)$ for various values of κ is shown in Fig. 5.1. The upper bound u for the distributions chosen from (5.1) is infinity when $\kappa \geq 0$ and equals $-1/\kappa$ for $\kappa < 0$. In this work, the fitness of the mutants is independently chosen from (5.1) thus making the fitness of the mutant, F_m an uncorrelated variable, which may be greater or smaller than the parent fitness, F_p. This model known as the House-of-



(a)

(b)

Fig. 5.2 Population fraction of different mutant classes are shown as different coloured lines. Where, (a) shows the SSWM ($N\mu = 0.1$, low mutation rate) regime and (b) shows the high mutation ($N\mu = 10$) regime for all three EVT domains of DBFE.

Cards model [125, 13] is based on the assumption that genetic organization of an individual is highly delicate and can be completely disturbed by any small change in the configuration. We have analyzed the results to see how they vary between the three EVT domains and different mutation rates.

In the allocation of the fitness to any mutant, our work differs from the other works on clonal interference [120, 70] wherein the fitness of the mutant is hiked above the parent fitness by the selection coefficients (s) which may be held constant or chosen from a distribution as $F_m = (1+s)F_p$. Unlike the model we have used in this work (as explained above), in this case there is a strong correlation between the mutant fitness F_m and the parent fitness F_p. In those cases, the mutant fitness is always greater than the parent fitness and on an average, a double or higher mutant is fitter than a single mutant. This is in contrast with our work since in ours, as the fitness of the parent increases, the number of better mutants available decreases thus producing different patterns for the fitness increment in each EVT domain.

In our model, whenever a mutant class goes extinct, the classes below it is moved up, and the number of classes in the population is reduced by one. The normalized probability of reproduction given in (5.2) of a class exceeding half corresponds to a leader change. The new leader determined now belongs to the class whose normalized probability exceeded half. We have also explored other criteria for defining the leader as the most populated class and find that our main results are robust with respect to the change in criteria.

Every change of leader is counted as a *step*. In the high mutation regime the population is spread over many sequences and a sequence can produce two or more mutants each of which may become leaders at different time steps. However, in the SSWM regime, the whole population is localized at a single sequence with a fixed fitness and can only move to a different sequence with higher fitness one mutation away. Thus every new leader arises from the previous leader, as can be observed in Fig. 5.2(a). When a better sequence appearing in the population does not get lost due to genetic drift, it quickly gets fixed. Further mutations that may lead to future leaders appear in this genetic background. The change in the fitness of the population is the same as the change in fitness of the leader. In this case, every move of the population (leader) from one sequence to another is termed a step in the adaptive walk [126, 114, 127, 128], whereas in high mutation regime, the population is polymorphic as can be seen from Fig. 5.2(b) and the leader change is not obvious from the figure.

Various quantities like the fitness difference between successive leaders and the average number of mutations in the leader are averaged only over the walks that take the step. Other quantities like the number of classes present at any time point and the rate of change of fitness are averaged over all time steps in that simulation run.

In this paper, the total number of iterations is 10^5 in every simulation run and the dynamics is tracked for finite time limit of 10^4 generations which we shall refer to as t_{max}. In this time span, if we assume that there is only one mutant which has fitness greater than f_{max} then the maximum fitness value, f_{max} that arises in the population can be calculated as

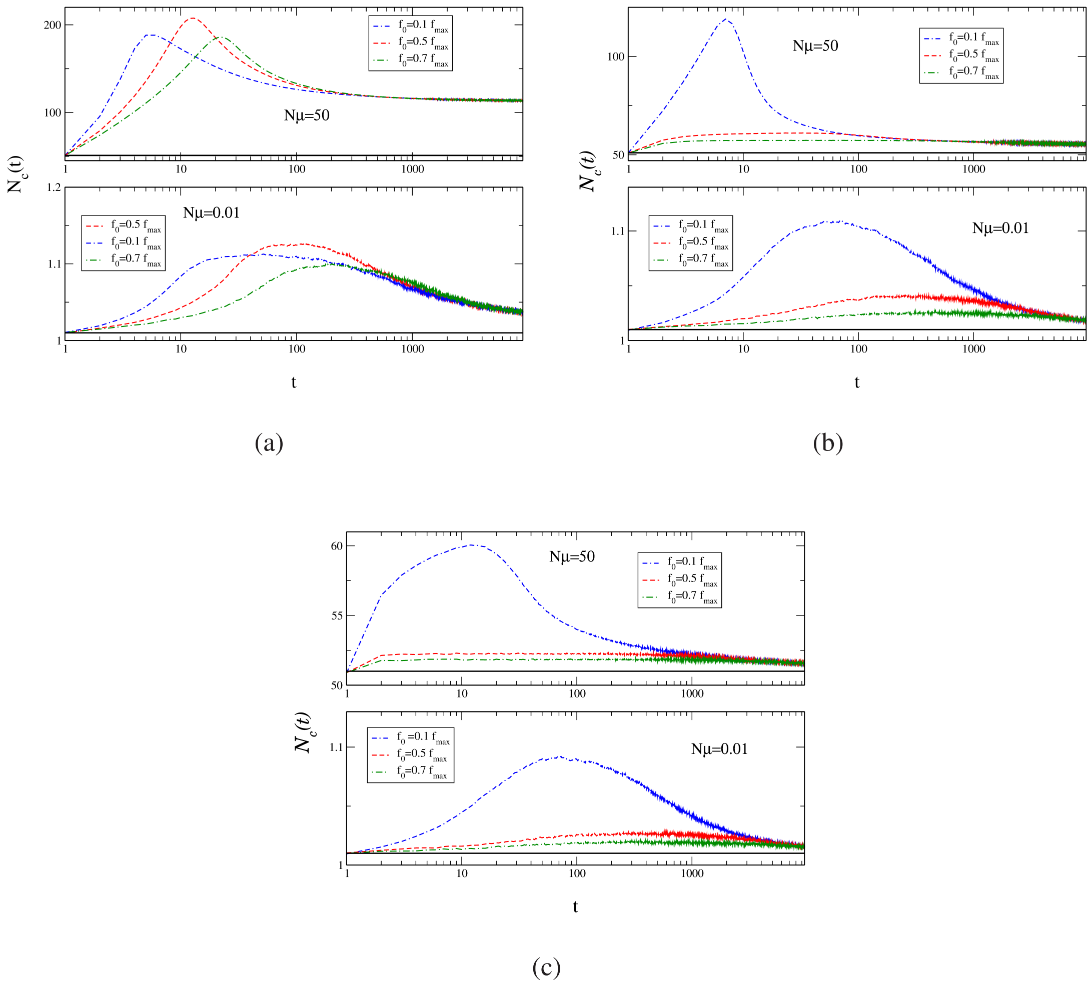
$$t_{max} N \mu \int_{f_{max}}^{u} p(f) df = 1 \tag{5.6}$$

where u is the upper limit of the fitness distribution equalling (-1/κ) for bounded distributions and infinity for unbounded ones [103]. From the above integral, we get

$$f_{max} = \frac{(t_{max} N \mu)^{\kappa} - 1}{\kappa}. \tag{5.7}$$

5.3 The number of classes in the population

For a population of fixed size, the number of classes in the population is expected to increase with the mutation rate. The average genetic variation defined here as the average number of classes ($\mathscr{N}_c$) present in the population is shown in Fig. 5.3 for all the three domains of DBFE. The top and bottom panels of the figure show the data corresponding to the high and low mutation regimes respectively. In both the mutation regimes, we see that the average number of classes increases during the initial time steps and decreases at later times when the classes with lower fitness are eliminated by the fitter ones. The maximum number of classes existing in the population for the first case, as shown in Fig. 5.3(a), does not belong to the lowest initial fitness, but to a slightly higher initial fitness. This could be because when the initial fitness is low, its class is quickly replaced by a fitter mutant and all further mutants arise on this new background must compete with this fitter class.



(a)

(b)

(c)

Fig. 5.3 The plot shows the average number of classes in the population as function of time for various initial fitnesses. The fitnesses are chosen from (5.1) with (a) $\kappa = -1$ (b) $\kappa \to 0$ and (c) $\kappa = 1/4$. For each κ value, the plot shows $\mathscr{N}_c(t)$ in both the high mutation (top panels) and low mutation (bottom panels) regimes. The straight line in all plots shows $N\mu + 1$.

In the low mutation regime, the population for the most time is localized at a single sequence and produces $N\mu$ mutants at every time step. So in this case, the average number of classes approach a constant $N\mu + 1$ at large times as can be seen in the bottom panels of Fig. 5.3. These panels also indicate that the value of this constant increases with decreasing κ. This is because in the case of bounded distributions with $\kappa < 0$, the fitness of beneficial mutant produced is expected to be closer to the parent fitness. In other words, mutations are nearly neutral and thus it takes longer time to take over the population as shown in Fig. 5.2(a).

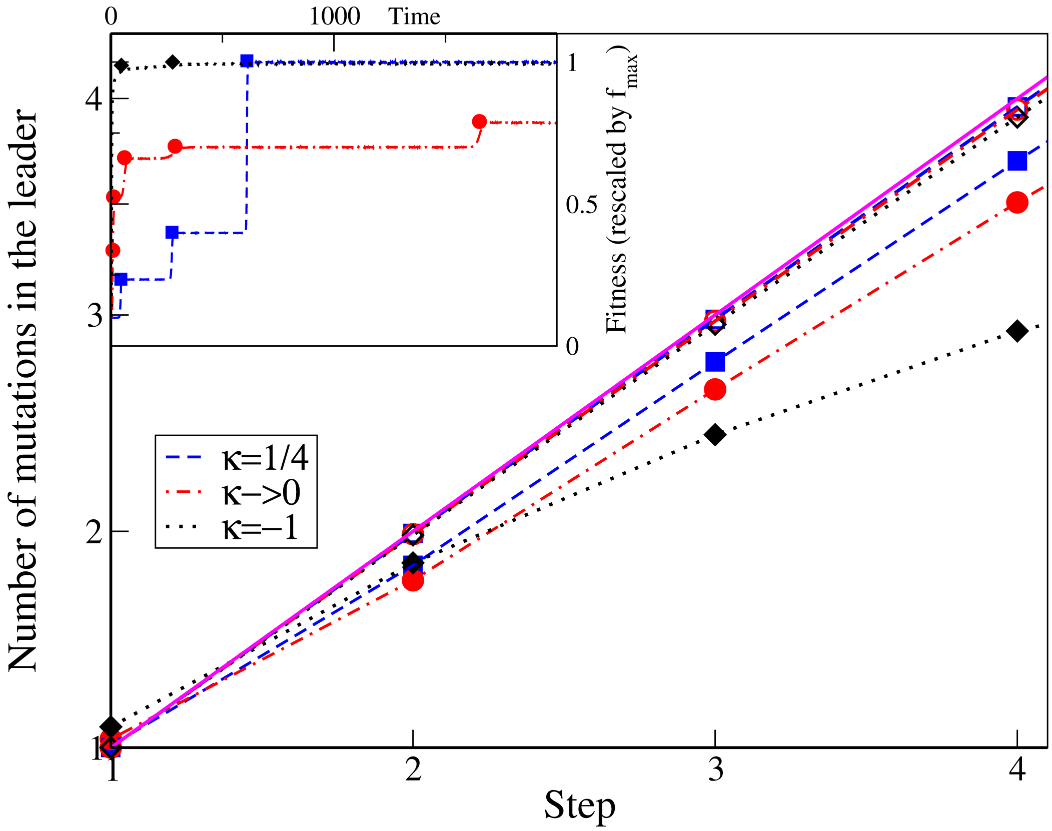


Fig. 5.4 The main plot shows the number of mutations in the leader at any step for various κ and mutation rates. The simulation data are represented by points while the broken lines connect the data points. The solid line shows $y = x$. In the inset, from a single simulation run, the fitness of the whole population as a function of time is shown by broken lines and the fitness of the leader whenever the leader changes is shown by symbols.

This results in a larger number of mutants in Weibull domain which can be observed in the bottom panel of Fig. 5.3(a). We can clearly see from the top panels of Fig. 5.3 that number of classes increases with decreasing κ even in high mutation regime. Also, the average number of classes present at a time is much higher in this regime. This makes sense because the fitness of the classes belonging to $\kappa = -1$ cannot be very different from each other (can only vary between 0 and 1) which makes it possible for many of them to exist in the population. The maximum fitness of the classes belonging to $\kappa = 1/4$ distribution will, on an average be much higher than all others (since the distribution is unbounded with a fat tail), thus out-competing the others in the population.

5.4 Number of mutations in the leader

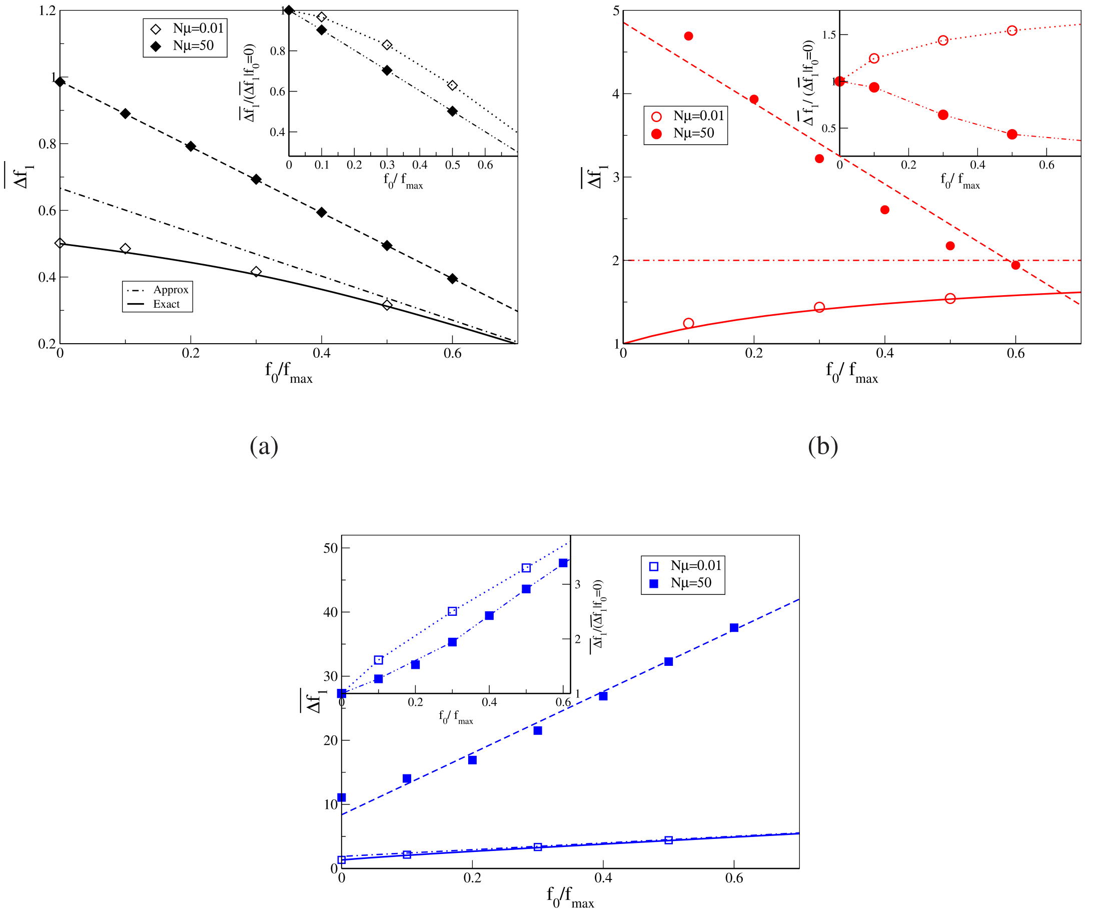
In the low mutation regime, the average number of mutations in the leader is expected to be very close to the step number since the genetic variation in the population is low and any mutation that escapes drift quickly takes over the population [102]. We verify this point via simulations as depicted in Fig. 5.4. We find that the mutation number equals the step in all the three EVT domains of the DBFE in the low mutation regime for the initial steps.

However in the high mutation regime, the number of mutations in the leader of any step differs between the three DBFE domains. When the mutation rate is increased, the genetic variation of the population and the significance of clonal interference also increases. In the high mutation regime, the number of mutations in the leader is found to be less than the step number in all the three DBFE domains. This is because there is a chance that different mutants originating from the same parent class can become the leader of the population at different times. This decrease from the step number is the minimum for the fat-tailed distributions and maximum for the truncated ones, as shown in Fig. 5.4. This result is consistent with the number of classes present in the population as discussed in the previous section. In the Fréchet domain, since the clonal interference is minimal, mostly a mutant originating from the present leader will become the next one. In the Weibull domain, due to the large number of classes present in the population, mutants originating from the same class can become the leaders at different time points.

5.5 Fitness and fitness difference

From our simulations, we find that the average fitness of the first mutant fixed in the population, $\bar{f}_1$ increases linearly with the initial fitness, f_0 for all κ in the low mutation regime and for $\kappa \neq 0$ in the high mutation regime. So we can write

$$\bar{f}_1 = a_{\kappa}^{(N\mu)} f_0 + b_{\kappa}^{(N\mu)} \tag{5.8}$$



(a)

(b)

(c)

Fig. 5.5 The main plot shows the fitness difference at the first step as a function of the initial fitness for various $N\mu$. The fitnesses are chosen from (5.1) with (a) $\kappa = -1$ (b) $\kappa \to 0$ and (c) $\kappa = 1/4$. The solid lines in the main plot are obtained by numerically evaluating the integral given by (5.9), while the dotted lines are the approximate results that can be obtained for the results when the initial fitness is high, in the low mutation regime. The broken lines for $\kappa \neq 0$ are lines of best fit as mentioned in the text. The broken line for $\kappa \to 0$ is used for connecting the data points. The inset shows the fitness difference at the first step as a comparative measure of the fitness difference obtained at the first step when $f_0 = 0$. Here, the lines are used for connecting the data points.

where the coefficients $a_\kappa^{(N\mu)}$ and $b_\kappa^{(N\mu)}$ are constants. In the low mutation regime, where the population for most times is monomorphic, the adaptive walk model has been used to

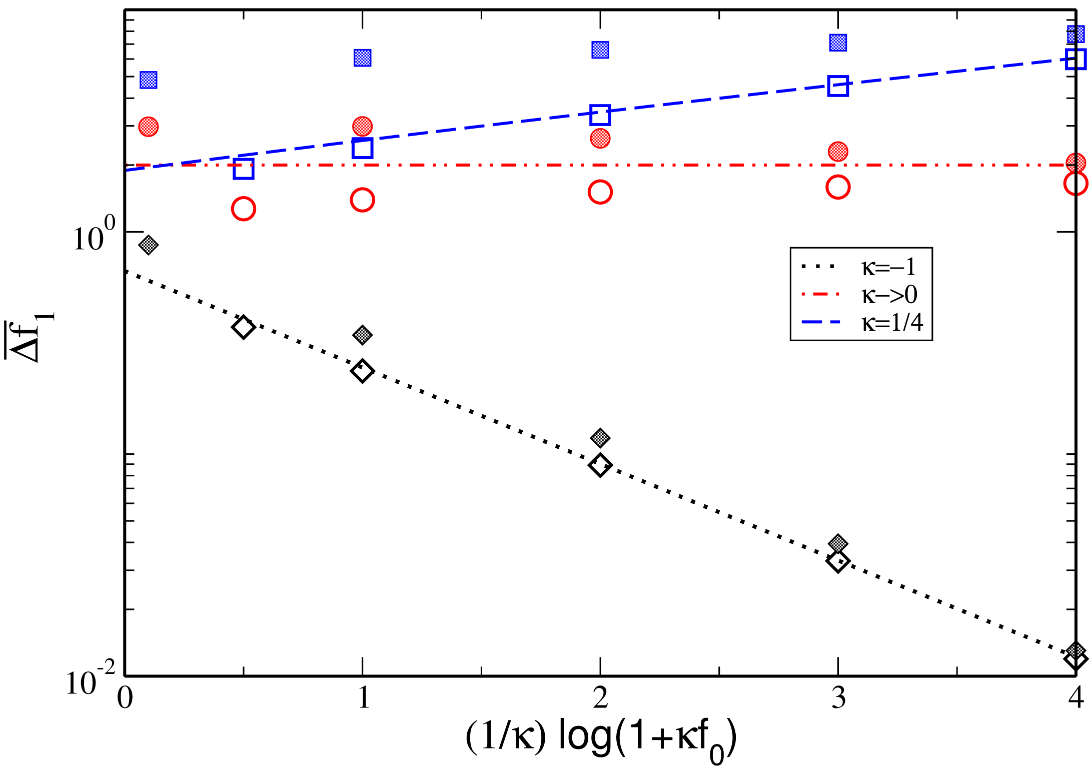


Fig. 5.6 The plot shows the fitness difference at the first step as a function of the initial fitness for different κ and two different $N\mu$. The lines give the theoretical values while the open symbols are the simulation output for $N\mu = 0.02$ and the closed symbols are those for $N\mu = 5$.

analytically obtain the fitness at the first step, $\bar{f}_1$ as [117, 10]

$$\bar{f}_1 = \int_{f_0}^{u} df\, T(f \leftarrow f_0) f \tag{5.9}$$

where the transition probability is given by

$$T(f \leftarrow f_0) = \frac{(1 - e^{-\frac{2(f-f_0)}{f_0}})p(f)}{\int_{f_0}^{u} dg \left(1 - e^{-\frac{2(g-f_0)}{f_0}}\right) p(g)}. \tag{5.10}$$

In this model, from (5.9), the coefficient $a_\kappa^{(N\mu\ll1)}$ was obtained as $0.33, 1.0$ and 1.6 for $\kappa = -1, 0,$ and $1/4$ respectively. The corresponding $b_\kappa^{(N\mu\ll1)}$ for the aforementioned κ were 0.66, 2.0 and 1.89 [10]. In the high mutation regime where the adaptive walk model is not applicable, we obtained the values for the coefficients in (5.8) numerically. We find that for large f_0, $a_\kappa^{(50)}$ equals 0.004 and 1.5 and $b_\kappa^{(50)}$ equals 0.99 and 9.1 for $\kappa = -1$ and $1/4$ respectively.

The interesting result from our work is that, irrespective of the number of mutants produced in the population, the difference $\overline{\Delta f_{step}} = \bar{f}_1 - f_0$ between the fitness of the first step and the initial fitness displays different qualitative trends: increases for positive κ, approaches a constant when $\kappa = 0$ and decreases for negative κ, as shown in Figs. 5.5 and 5.6.

We can better understand these increasing and decreasing trends by the following heuristic argument. In both the low and high mutation regimes, for large f_0, the fitness at the first step f_1 increases linearly with the initial fitness as given in (5.8) and so, we can write the selection coefficient defined as the relative fitness difference, at the first step as

$$s = \frac{\bar{f}_1 - f_0}{f_0} = \frac{(a_\kappa^{(N\mu)} - 1) f_0}{f_0} + \frac{b_\kappa^{(N\mu)}}{f_0}, \quad \text{for all} \quad \kappa\,, N\mu \tag{5.11}$$

In an adapting population, since the fitness of the first step is greater than the initial fitness, the selection coefficient is always positive. As the fitness distributions belonging to the Fréchet domain are unbounded with fat tails, high f_0 values can be considered in which case, the second term on the right hand side (RHS) of (5.11) can be ignored and we can write $s \approx (a_\kappa^{(N\mu)} - 1) > 0$. Thus for $\kappa > 0$, $a_\kappa^{(N\mu)} > 1$ and therefore it follows that the fitness difference at the first step increases with f_0. On the other hand, since the distribution belonging to the Weibull domain are truncated, we can invoke the following inequality to explain the decrease in fitness difference with increasing f_0:

$$\bar{f}_1 - f_0 < u - f_0, \tag{5.12}$$

where u is the upper limit of the fitness distribution. With increasing f_0, the RHS of the above equation decreases which shows that as the initial fitness increases, $\bar{f}_1 - f_0$ has to necessarily decrease. Thus the qualitative trends discussed above appear to be determined by the behaviour of the tail (bounded/unbounded), and not by the details of the model.

Also, it is interesting to note that while the data points for the exponentially decaying distribution ($\kappa = 0$) increase and seem to be approaching a constant in the low mutation regime, the data in the high mutation regime seems to be reducing to approach the same constant. Our simulation results shown in Fig. 5.5 not only match the predicted theoretical values and validate the claim of different qualitative trends in each EVT domain in the SSWM regime but also show that the trends hold irrespective of the number of mutants produced in the population. This result suggests that the qualitatively different trends of the fitness difference (increasing, constant and decreasing with initial fitness in the Fréchet, Gumbel and Weibull domain, respectively) can be used to distinguish between the EVT domains in a more general scenario.

Though the fitness difference at the first step is greater in the high mutation regime, when compared with the results in the low mutation regime, when we look at the fitness difference at the first step scaled by the fitness difference obtained when the initial fitness is zero (insets of Fig. 5.5, we see that this increase is slower in the high mutation regime compared to the results obtained in the low mutation regime. This indicates that as the mutation rate increases, though the number of mutants accessed is higher, the difference in fitness compared to a lower initial fitness is not proportionally higher and is in fact lower for all the fitness distributions.

5.6 Rate of change of fitness with time

Besides the fitness increment at a fixed event of leader change, we also measured the fitness as a function of time as shown in Fig. 5.7. We observed that even though the fitness increases with time in all the three EVT domains, the rate at which the fitness increases depends strongly on the DBFE. This rate has an initial fast transient phase, after which it slows down.

The initial transient phase is strongly dependent on the initial condition as well as the mutation rate as shown in Fig. 5.8. The increase in fitness is fastest for the lowest initial condition, but it approaches the same fitness value as in the case of higher initial fitness in

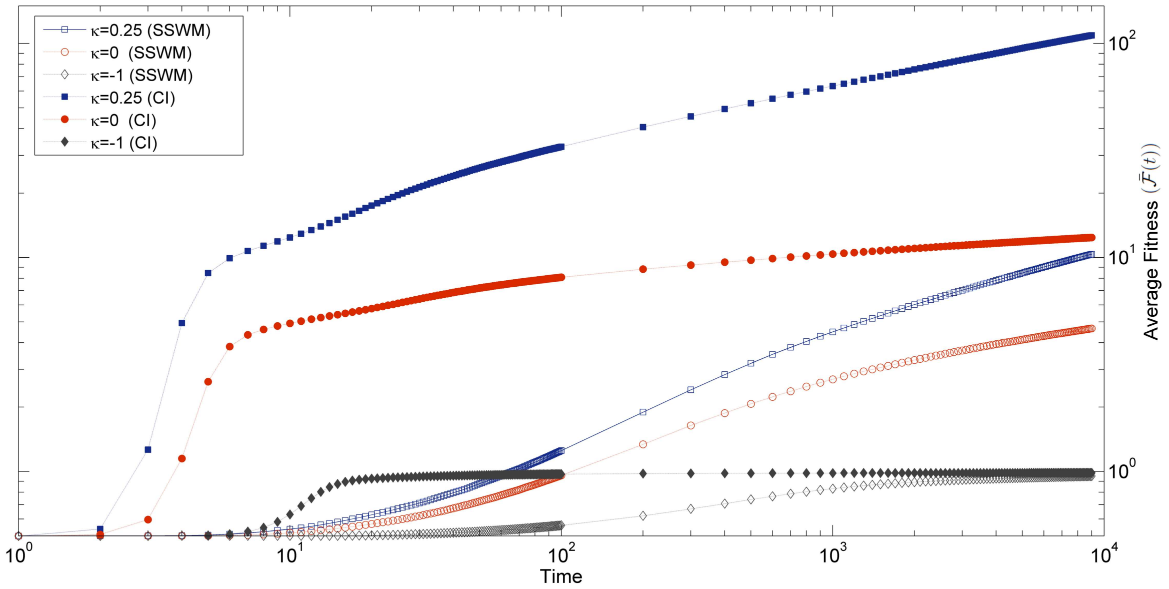


Fig. 5.7 Figure shows the average fitness increase with time for three different values of κ in the SSWM regime ($N\mu = 0.01$), and in high mutation regime($N\mu = 50$). In all the cases, the population starts with the same initial fitness $f_0 = 0.5$.

few generations. The time taken for populations of different initial fitness to reach the same fitness value depends on the mutation rate: for $N\mu \gg 1$, it takes about 20 generations, whereas for $N\mu \ll 1$, it is approximately 200 generations. Even after this transient phase, the rate of increase in average fitness ($\bar{\mathcal{F}}(t)$) with time depends on the mutation rate as shown in Fig. 5.7. This is because of the fact that, when a large number of mutations are available at the same time, a highly fit mutant can invade the population and give a large fitness increment. So the fitness of a highly fit mutant sequence would be greater in the high mutation regime compared to the one in low mutation regime. The maximum fitness value reached in 9000 generations, in the case of Fréchet distribution is about 10 times more for high mutation regime, which is consistent with the expectation from (5.7). Even beyond this point we noticed that the fitness is still increasing. In the same way, Gumbel distribution also shows a significant increase in maximum fitness reached in high mutation regime compared to the SSWM regime (about 4 times). Here also we found that the fitness is still increasing beyond the time point till which we tracked the dynamics. The bounded distribution (Weibull) reaches near the upper bound in

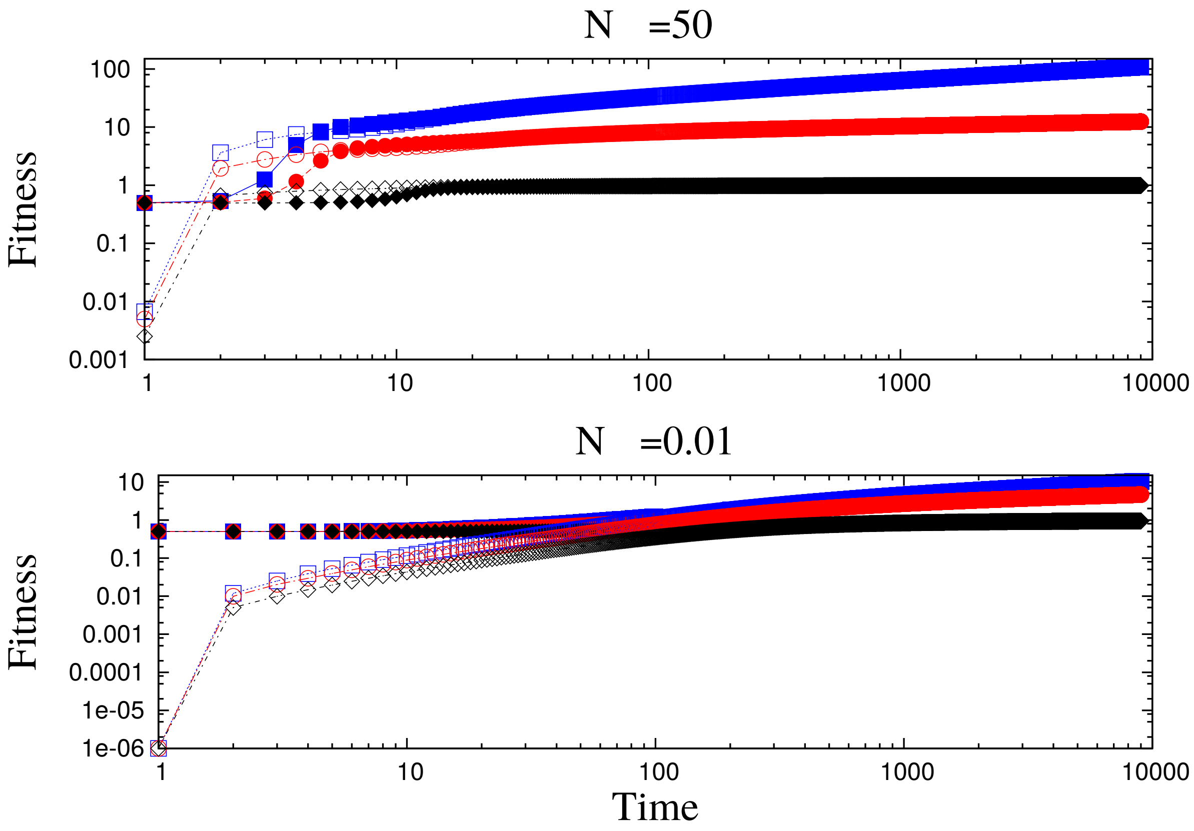


Fig. 5.8 The figure shows the average fitness of the population for various κ in both the low and high mutation regimes. Two different initial conditions $f_0 = 0$ (open symbols) and $f_0 = 0.5$ (closed symbols) are considered.

SSWM and evolves slowly. But the fitness reaches a fitness plateau in high mutation regime and rate of adaptation becomes zero as can be seen in Fig. 5.7.

From this we observe that the rate of change of fitness depends strongly on the properties of the underlying DBFE, which suggests that looking at this quantity can help us in distinguishing the DBFEs. So we measured the fitness increment defined as

$$\Delta\bar{\mathscr{F}}(t) = \langle \bar{\mathscr{F}}(t+1) - \bar{\mathscr{F}}(t) \rangle \tag{5.13}$$

at each step. The $\Delta\bar{\mathscr{F}}(t)$ initially increases, then slowly decreases and settles down to a zero as shown in Fig. 5.9. If we denote this function as

$$\Delta\bar{\mathscr{F}}(t) = \frac{A}{t^{\alpha}} \tag{5.14}$$

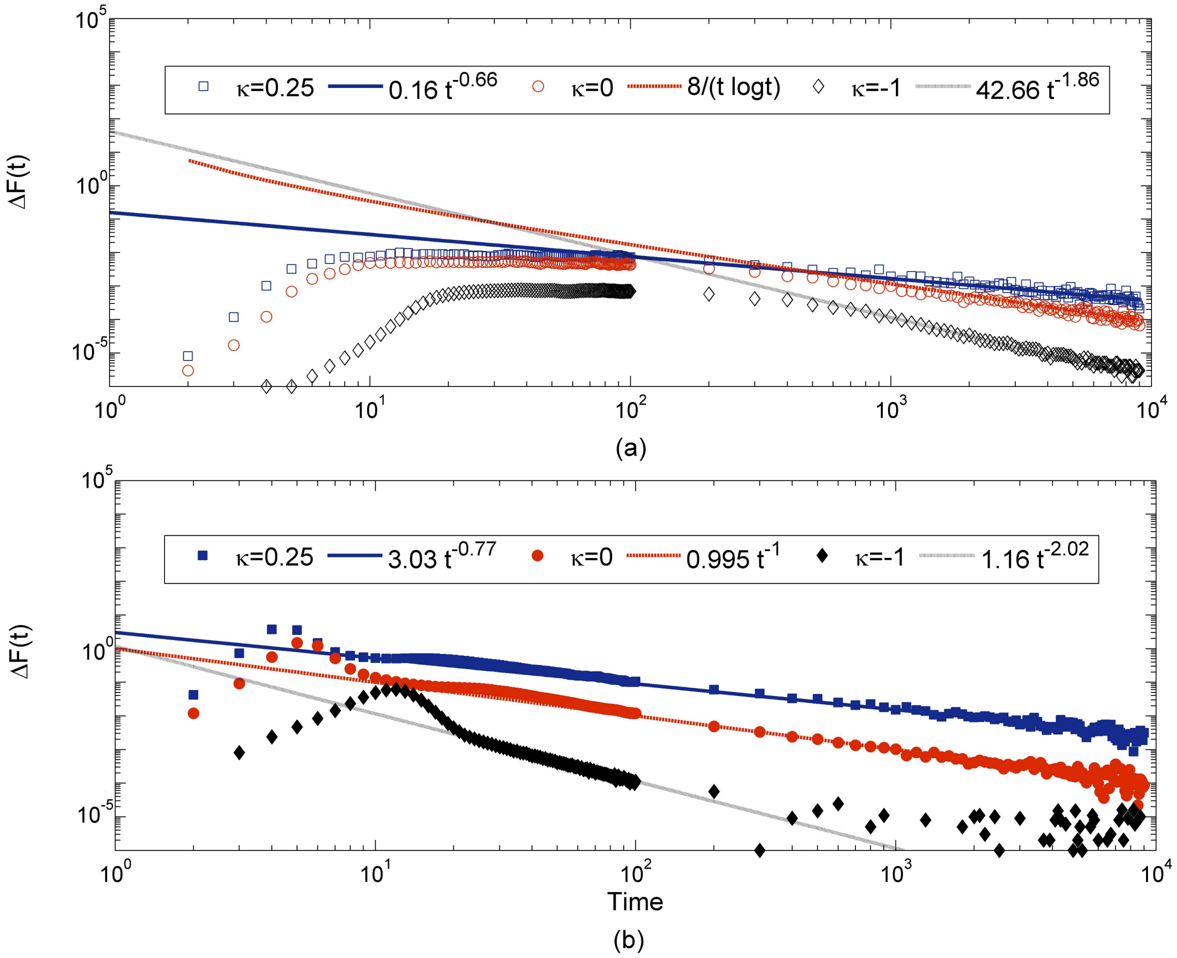


Fig. 5.9 Figure shows the fitness increment in each time step for three different values of κ in two mutation regimes (SSWM and high mutation). In each case the data is fitted with the theoretically expected function given in (5.14), except for exponential distribution for which we used the theoretical prediction by Park and Krug [13]. In all the cases, the population starts with the same initial fitness $f_0 = 0.5$.

where A is a constant and the exponent α can be used to distinguish the DBFE, since, as explained below, exponent α is found to be greater (smaller) than one in Weibull (Fréchet) domain, but is close to one in Gumbel domain.

In the SSWM regime, from Fig. 5.9(a), we can see that each type of DBFE considered shows a different rate of decay. Weibull domain has a faster decay with $\alpha = 1.86$, Gumbel domain has $\alpha \approx 1$ [13] and Fréchet domain $\alpha = 0.66$ [13]. We observed that the same trend holds in high mutation rate regime as well, where α values are slightly larger in all cases. In this regime also $\alpha = 2.02$, 1 and 0.76 for Weibull, Gumbel and Fréchet domains respectively as shown in Fig. 5.9(b). In the high mutation regime, in the case of Weibull distributions fitness reaches a plateau in few generations, after which its rate of change goes to zero, as can be observed in Fig. 5.9(b). The theoretical prediction for the fitness at every time step for the unbounded distributions belonging to the Gumbel and Frèchet domains was obtained by Park and Krug [13] in the low mutation regime. The comparison of our simulation data with these predictions shows a very good agreement in Gumbel domain and in Fréchet domain (up to a constant). In this work, we have considered the bounded distribution also and observed that its rate of decrease is faster with an exponent greater than one, which was not considered in the previous studies. We observed that even in high mutation regime, the exponent α shows the same behaviour. In this regime the rate of change of fitness has been calculated only for exponential distribution belonging to the Gumbel domain [13] and their prediction matches with our data. In this work, we have obtained a complete picture by studying the rate of change of fitness numerically for the other two EVT domains as well.

Thus, the second main finding from our study is that in all DBFEs, the fitness difference at each time step decreases with time as given by (5.14) and we can distinguish between the three EVT domains of DBFEs by looking at the exponent α. A comparison of our results with the existing literature is given in Table 5.1.

5.7 Discussion

The main purpose of our work is to determine the quantities which can be used to distinguish the different extreme value domains of the DBFE. Previous studies [10, 119] have found that in an adapting population, the fitness gain at each fixation event shows qualitatively different trends in the three DBFE domain, when the number of mutants produced in the population is much less than one at every generation ($N\mu \ll 1$). The focus of this work is to explore the parameter regime in which the number of mutants produced is much above one ($N\mu \gg 1$). When the mutation rate is high, the population becomes polymorphic and the better mutants existing in the population compete with each other. From our study we have observed that

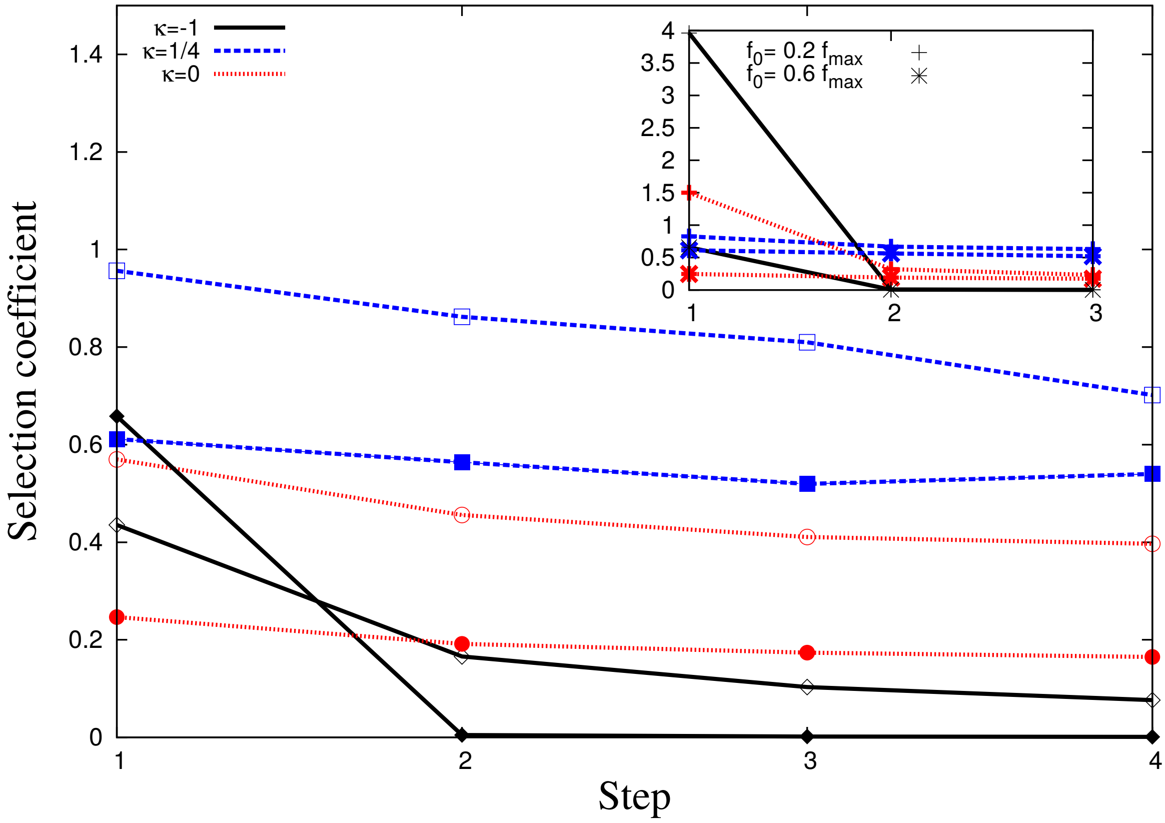


Fig. 5.10 The main figure shows the selection coefficient as a function of step for all three κ values. We considered two different $N\mu$ where open symbols and closed symbols are for $N\mu = 0.01$ and $N\mu = 50$, respectively. The inset shows the selection coefficient of various steps for two different the initial fitnesses $f_0 = 0.2 f_{max}$ and $f_0 = 0.6 f_{max}$, where f_{max} is calculated using (5.7) in the high mutation regime.

the qualitative trends found for fitness difference when a new mutation establishes in the low mutation regime hold irrespective of the number of mutants produced. Thus this study suggests that fitness difference between the successive mutations that spreads in the population is a very important and robust quantity that can be used to predict the DBFEs in a more general scenario.

From our simulations, we see that as the initial fitness is increased the fitness difference at the first step given by $\overline{\Delta f_{step}}$ reduces, approaches a constant or increases with initial fitness in the Weibull, Gumbel and Fréchet domains, respectively. We can understand these trends by a heuristic reasoning as discussed in detail in the Section 5.5. This argument explains the increase in $\overline{\Delta f_{step}}$ with f_0 for unbounded power law distribution and shows that the trends are determined by the behaviour of the tail (bounded/unbounded), and not by the details of the model.

Another important measure in understanding the dynamics of adaptation is the rate at which it occurs. Most of the previous studies which measured the adaptation rate have only considered exponentially distributed fitness distributions [35, 120, 80, 70, 129]. A previous study by Park and Krug [13] also considered DBFEs belonging to Fréchet domain but only in the SSWM regime (see Table 5.1). In this work, we have extended the previous studies by numerically measuring the rate of change of fitness for bounded distributions also. We have measured the rate of change of fitness in all the three EVT domains of the DBFE in both low and high mutation regimes. We observed that in all the cases, the rate of change of fitness decreases with time as $\sim t^{-\alpha}$, where $\alpha > 1$ for Weibull, $\alpha \approx 1$ for Gumbel [13] and $\alpha < 1$ for Fréchet domains [13].

Experimentally, the distribution of beneficial fitness effects can be inferred by two methods. In the first method, mutations are introduced in the wildtype sequence and those that confer a fitness advantage are separated and their distribution of fitness effects are determined. By this method, DBFE belonging to all the EVT domains have been observed [11, 104–112]. In contrast, here we focus on learning about DBFE via adaptation dynamics. Though many works have tracked the dynamics of the population during adaptation [104, 130–133], in most of them only the selection coefficient of the mutant fixed was measured. In our study, we have observed that the selection coefficient as given by (5.11) always decreases, with the increasing initial fitness or increasing steps as shown in Fig. 5.10. Hence this quantity is not useful to distinguish between the EVT domains. However, from our study we observe that the fitness difference between steps shows different patterns depending on the EVT domain of the DBFEs in both the high and low mutation regimes and can be used to distinguish between the EVT domains.

In this work, we have numerically shown that the fitness returns in each EVT domain is very robust and holds even when the number of mutations produced is large ($N\mu \gg 1$). Fitness difference can be measured in experiments, for example as in [107]. We suggest that experiments can predict the EVT domain of DBFE by measuring the fitness difference between successive mutations fixed in the population, or even from the fitness of the first mutation, when the initial fitness is varied. However currently experimental studies that measure both fitness and DBFE in the same study are not available but it is highly desirable to have such studies to test our predictions.

The criteria used for choosing the leader, namely, the class which has a population fraction greater than $1/2$, is not good in high mutation regime, because more than one mutant class with fraction close to $1/2$ can be present in the population at a time, and in that case it is difficult to identify the leader in an experimental population with this criteria. It will be interesting to check the robustness of our results for different criteria, which can be used in experiments with some genetic markers.

Chapter 6

Conclusions

The main focus of this study is to understand the combined effect of beneficial and deleterious mutations in the presence of other evolutionary forces, in particular, we study the effect of beneficial mutations in two biological questions, namely, the evolution of sex and recombination and the dynamics of adaptation process.

In Chapters 2 and 3, we studied the effect of beneficial mutation on the irreversible accumulation of deleterious mutation (Muller's ratchet)[1] for two different mutation schemes. The main findings are:

(i) Whether or not beneficial mutations can stop Muller's ratchet depends on the mutational scheme used. Beneficial mutations always stop the Muller's ratchet if the mutation rates are fitness-dependent (as discussed in Chapter 2) [3]. But, when the mutation rates are fitness independent (as in Chapter 3), for a fixed beneficial mutation rate, Muller's ratchet stops only when the population size is above a critical value N_c.

(ii) The equilibrium fraction of deleterious mutations (mutational load) reduces with recombination. This reduction is appreciable when beneficial mutations are rare, as in the case of adapting microbial populations, whereas it has a mild or moderate effect on codon usage bias where the mutation rates between the preferred and unpreferred codons are comparable. These results are discussed in Chapter 2.

(iii) On the technical front, exact solution of the frequency distribution in deterministic models were found in Chapters 2 and 3.

In Chapter 4, we asked the question how the advantage of recombination on a single peak fitness landscape changes when the topology of fitness landscape changes. We found that on rugged fitness landscapes,

(i) Small rate of recombination is advantageous when mutational forces are weak.

(ii) Recombination has a short time advantage, which is determined by the potential to generate fit genotypes from the existing initial variation, but it becomes disadvantageous at long times, as it reduces the probability to escape from the local fitness peak [9].

(ii) Effect of recombination shows strong dependence on the initial condition, as the short term advantage is highest when the initial variation is high and the ruggedness is low.

In Chapter 5, we studied the adaptation dynamics of asexual populations to address the second question of interest, namely, the relationship between the adaptation dynamics and the distribution of beneficial fitness effects (DBFEs). We found that:

(i) Distinguishable trends shown by average fitness difference between successive steps, in three extreme value domains of DBFEs in the low mutation regime [119] holds good for high mutation regime as well.

(ii) Rate of adaptation also shows distinct trends for different DBFEs in both high and low mutation regimes.

References

[1] H. J. Muller. The relation of recombination to mutational advance. *Mut. Res.*, 1:2–9, 1964.

[2] J. Haigh. The accumulation of deleterious genes in a population - Muller's ratchet. *Theor. Pop. Biol.*, 14:251–267, 1978.

[3] S. John and K. Jain. Effect of drift, selection and recombination on the equilibrium frequency of deleterious mutations. *J. Theor. Biol.*, 365:238–246, 2015.

[4] M. Kimura and T. Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54:1337–1351, 1966.

[5] K. Jain and S. John. Deterministic evolution of an asexual population under the action of beneficial and deleterious mutations on additive fitness landscapes. *arXiv:1604.03676*, 2016.

[6] S. John and S. Seetharaman. Exploiting the adaptation dynamics to predict the distribution of beneficial fitness effects. *PLoS One*, 11(3):e0151795, 2016.

[7] H. Sachdeva and S. John. (Dis-)Advantage of recombination on rugged fitness landscapes, (in preparation).

[8] R. A. Neher and B. I. Shraiman. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *Proc. Natl. Acad. Sci. U.S.A.*, 106(16):6866–6871, 2009.

[9] S. Nowak, J. Neidhart, I. G. Szendro, and J. Krug. Multidimensional epistasis and the transitory advantage of sex. *PLoS Comp. Biol.*, 10(9):e1003836, 2014.

[10] S. Seetharaman and K. Jain. Adaptive walks and distribution of beneficial fitness effects. *Evolution*, 68:965–975, 2014.

[11] R. Sanjuán, A. Moya, and S.F. Elena. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. USA*, 101:8396–8401, 2004.

[12] A. Eyre-Walker and P.D. Keightley. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.*, 8:610, 2007.

[13] S.-C. Park and J. Krug. Evolution in random fitness landscapes: the infinite sites model. *J. Stat. Mech.: Theor. and Exp.*, 4:P04014, 2008.

[14] J. Wakeley. The limits of theoretical population genetics. *Genetics*, 169:1–7, 2005.

[15] S. Kryazhimskiy, D. P. Rice, and M. M. Desai. Population subdivision and adaptation in asexual populations of *Saccharomyces cerevisiae*. *Evolution*, 66(6):1931–1941, 2012.

[16] F.J. Poelwijk, D.J. Kivet, D.M. Weinreich, and S.J. Tans. Empirical fitness landscapes reveal accessible paths. *Nature*, 445:383, 2007.

[17] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, and A. K. Chakraborty. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[18] J.A.G.M. de Visser and J. Krug. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Gen.*, 15:480–490, 2014.

[19] J. V. Cleve and D. B. Weissman. Measuring ruggedness in fitness landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 112(24):7345–7346, 2015.

[20] J. Neidhart, I. G. Szendro, and J. Krug. Adaptation in tunably rugged fitness landscapes: The Rough Mount Fuji model. *Genetics*, 198(2):699–721, 2014.

[21] N. H. Barton, D. E. G. Briggs, J. A. Eisen, D. B. Goldstein, and N. H. Patel. *Evolution*. Cold Spring Harb., 2007.

[22] O.K. Silander, O. Tenaillon, and L. Chao. Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS Biology*, 5(4):e94, 2007.

[23] J.H. Miller. Spontaneous mutators in bacteria: Insights into pathways of mutagenesis and repair. *Annu. Rev. Microbiol.*, 50:625–643, 1996.

[24] A. Oliver, R. Cantón, P. Campo, F. Baquero, and J. Blázquez. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science*, 288:1251–1253, 2000.

[25] W. Tröbner and R. Piechocki. Selection against hypermutability in *Escherichia coli* during long-term evolution. *Mol. Gen. Genet.*, 198:177–178, 1984.

[26] M. Lynch. The lower bound to the evolution of mutation rates. *Genome Evol. Biol.*, 3:1107–1118, 2011.

[27] A. James and K. Jain. Fixation probability of rare nonmutator and evolution of mutation rates. *Ecol. and Evol.*, 6:755–764, 2016.

[28] K. Jain, J. Krug, and S.-C. Park. Evolutionary advantage of small populations on complex fitness landscapes. *Evolution*, 65:19–45, 2011.

[29] A.S. Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336:435 – 440, 1988.

[30] S.P. Otto and T. Lenormand. Evolution of sex: Resolving the paradox of sex and recombination. *Nat. Rev. Genet.*, 3:252–261, 2002.

[31] W. R. Rice. Evolution of sex: Experimental tests of the adaptive significance of sexual recombination. *Nat. Rev. Genet.*, 3:241–251, 2002.

[32] J.A.G.M. de Visser and S.F. Elena. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat. Rev. Genet.*, 8:139–149, 2007.

[33] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78:737–756, 1974.

[34] I. Gordo and B. Charlesworth. The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics*, 154:1379–1387, 2000.

[35] P. J. Gerrish and R. E. Lenski. The fate of competing beneficial mutations in an asexual populations. *Genetica*, 102:127–144, 1998.

[36] I.M. Rouzine, E. Brunet, and C.O. Wilke. The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theor. Pop. Biol.*, 73:24–46, 2008.

[37] W.J. Ewens. *Mathematical Population Genetics*. Springer, Berlin, 2004.

[38] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[39] R. Hershberg and D. A. Petrov. Selection on codon bias. *Annu. Rev. Genet.*, 42:287–299, 2008.

[40] J. B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, 12:32–42, 2011.

[41] W.-H. Li. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.*, 24:337–345, 1987.

[42] M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 149:897–907, 1991.

[43] G. A. T. McVean and B. Charlesworth. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res. Camb.*, 74:145–158, 1999.

[44] W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genet. Res., Camb.*, 8:269–294, 1966.

[45] J. M. Comeron, M. Kreitman, and M. Aguadé. Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. *Genetics*, 151:239–249, 1999.

[46] G. A. T. McVean and B. Charlesworth. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics*, 155:929–944, 2000.

[47] B. Charlesworth, A. J. Betancourt, and V. B. Kaiser. Genetic recombination and molecular evolution. *Cold Spring Harb. Symp. Quant. Biol.*, 74:177–186, 2009.

[48] V. B. Kaiser and B. Charlesworth. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends in Genetics*, 25:339–348, 2009.

[49] R. Lande. Risk of population extinction from fixation of deleterious and reverse mutations. *Genetica*, 102-103:21–27, 1998.

[50] M. C. Whitlock. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution*, 54:1855–1861, 2000.

[51] S. Goyal, D. J. Balick, E. R. Jerison, R. A. Neher, B. I. Shraiman, and M. M. Desai. Dynamic mutation selection balance as an evolutionary attractor. *Genetics*, 191:1309–1319, 2012.

[52] D. K. Howe and D. R. Denver. Muller's ratchet and compensatory mutation in *Caenorhabditis briggsae* mitochondrial genome evolution. *BMC Evolutionary Biology*, 8:62, 2008.

[53] S. Estes and M. Lynch. Rapid fitness recovery in mutationally degraded lines of *Caenorhabditis elegans*. *Evolution*, 57:1022–1030, 2003.

[54] N.H. Barton and B. Charlesworth. Why sex and recombination? *Science*, 281:1986–1990, 1998.

[55] K. Zeng. A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. *Mol. Biol. Evol.*, 27 (6):1327–1337, 2010.

[56] D. R. Schrider, D. Houle, M. Lynch, and M. W. Hahn. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, 194:937–954, 2013.

[57] P. D. Sniegowski and P. J. Gerrish. Beneficial mutations and the dynamics of adaptation in asexual populations. *Phil. Trans. R. Soc. B*, 365:1255–1263, 2010.

[58] G. Woodcock and P. G. Higgs. Population evolution on a multiplicative single-peak fitness landscape. *J. Theor. Biol.*, 179:61–73, 1996.

[59] T. Wiehe. Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. *Genet. Res. Camb.*, 69:127–136, 1997.

[60] W. Feller. *An introduction to probability theory and its applications, Vol. I.* John Wiley and Sons, 2000.

[61] P. Pfaffelhuber, P.R. Staab, and A. Wakolbinger. Muller's ratchet with compensatory mutations. *Ann. Appl. Prob.*, 22:2108 – 2132, 2012.

[62] I. Eshel and M.W. Feldman. On the evolutionary effect of recombination. *Theor. Pop. Biol.*, 1:88–100, 1970.

[63] R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, New York, 2008.

[64] J. L. Campos, K. Zeng, D. J. Parker, B. Charlesworth, and P. R. Handdrill. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in Drosophila melanogaster. *Mol. Bio. Evol.*, 30:811–823, 2012.

[65] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. Rates of spontaneous mutation. *Genetics*, 148:1667–1686, 1998.

[66] K. Jain. Loss of least-loaded class in asexual populations due to drift and epistasis. *Genetics*, 179:2125–2134, 2008.

[67] W. Stephan, L. Chao, and J. G. Smale. The advance of Muller's ratchet in a haploid asexual population - approximate solutions based on diffusion theory. *Genet. Res. Camb.*, 61:225–232, 1993.

[68] R. Neher and B. Shraiman. Fluctuations of fitness distributions and the rate of Muller's ratchet. *Genetics*, 191:1283 – 1293, 2012.

[69] J. J. Metzger and S. Eule. Distribution of the fittest individuals and the rate of Muller's ratchet in a model with overlapping generations. *PLoS Comp. Biol.*, 9:e1003303, 2013.

[70] S.-C. Park, D. Simon, and J. Krug. The speed of evolution in large asexual populations. *J. Stat. Phys.*, 138:381–410, 2010.

[71] M. Kimura, T. Maruyama, and J. F. Crow. The mutation load in small populations. *Genetics*, 48:1303–1312, 1963.

[72] A. Kondrashov. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J. Theor. Biol.*, 175:583–594, 1995.

[73] I. Gordo and B. Charlesworth. The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes. *Genet. Res.*, 78:149–161, 2001.

[74] L. P. Maia, D. F. Botelho, and J. F. Fontanari. Analytical solution of the evolution dynamics on a multiplicative-fitness landscape. *J. Math. Biol.*, 47:453–456, 2003.

[75] A. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. How often does the ratchet click? facts, heuristics, asymptotics. In J. Blath, P. Mörters, and M. Scheutzow, editors, *Trends*

in Stochastic Analysis, London Mathematical Society Lecture Note Series 353, pages 365–390. Cambridge University Press, 2009.

[76] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, 38:226–231, 1968.

[77] L. Hadany and J. M. Comeron. Why are sex and recombination so common? *Ann. N. Y. Acad. Sci*, 1133:26–43, 2008.

[78] D. Waxman and L. Loewe. A stochastic model for a single click of Muller's ratchet. *J. Theor. Biol.*, 264:1120–1132, 2010.

[79] C. O. Wilke. The speed of adaptation in large asexual populations. *Genetics*, 167:2045–2053, 2004.

[80] M.M. Desai and D.S. Fisher. Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics*, 176:1759–1798, 2007.

[81] N. H. Barton. Genetic linkage and natural selection. *Philos. Trans. R. Soc. B*, 365:2559–2569, 2010.

[82] I. Gordo and P. R. A. Campos. Sex and deleterious mutations. *Genetics*, 179:621–626, 2008.

[83] B. Charlesworth. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190:5–22, 2012.

[84] V. B. Kaiser and B. Charlesworth. Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. *Genetics*, 185:339–348, 2010.

[85] J. R. Powell and E. N. Moriyama. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.*, 94:7784–7790, 1997.

[86] L. S. Tsimring, H. Levine, and D. A. Kessler. RNA virus evolution via a fitness-space model. *Phys. Rev. Lett.*, 76:4440–4443, 1996.

[87] D. A. Kessler, H. Levine, D. Ridgway, and L. Tsimring. Evolution on a smooth landscape. *J. Stat. Phys.*, 87:519–544, 1997.

[88] E Brunet, I. M. Rouzine, and C. O. Wilke. The stochastic edge in adaptive evolution. *Genetics*, 179:603–620, 2008.

[89] L. Perfeito, L. Fernandes, C. Mota, and I. Gordo. Adaptive mutations in bacteria: high rate and small effects. *Science*, 317:813–815, 2007.

[90] C.J. Thompson and J.L. McBride. On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. *Math. Biosci.*, 21:127, 1974.

[91] K. Jain and S. Seetharaman. Nonlinear deterministic equations in biological evolution. *J. Nonlin. Math. Phys.*, 18:321–338, 2011.

[92] N. G. van Kampen. *Stochatic processes in physics and chemistry*. North Holland Personal Library, 1997.

[93] R. Courant and D. Hilbert. *Methods of Mathematical Physics, Vol.1*. Dover, 1953.

[94] S. Flügge. *Practical quantum mechanics*. Springer-Verlag, 1974.

[95] M. Ehrhardt and R. E. Mickens. Solutions to the discrete Airy equation: application to parabolic equation calculations. *J. Comp. Appl. Math.*, 172:183–206, 2004.

[96] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1964.

[97] S. Wielgoss, J. E. Barrick, 0. Tenaillon, M. J. Wiser, W. J. Dittmar, S. Cruveiller, B. Chane-Woon-Ming, R. E. Lenski, and D. Schneider. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl. Acad. Sci. U.S.A.*, 110:222–227, 2013.

[98] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug. Evolutionary accessibility of mutational pathways. *PLoS Comp. Biol.*, 7:e1002134, 2011.

[99] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. M. de Visser. Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech.*, -:P01005, 2013.

[100] J.A.G.M. de Visser, S.-C. Park, and J. Krug. Exploring the effect of sex on an empirical fitness landscape. *Am. Nat.*, 174:S15–S30, 2009.

[101] J. J. Bull and S. P. Otto. The first steps in adaptive evolution. *Nat. Genet.*, 37:342–343, 2005.

[102] J. H. Gillespie. A simple stochastic gene substitution process. *Theor. Popul. Biol.*, 23:202–215, 1983.

[103] D. Sornette. *Critical Phenomena in Natural Sciences*. Springer, Berlin, 2000.

[104] D.R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Genet.*, 37:441–444, 2005.

[105] R. Kassen and T. Bataillon. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nat. Genet.*, 38:484–488, 2006.

[106] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman. Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.*, 69:229, 2008.

[107] R. C. MacLean and A. Buckling. The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*. *PLoS Genet.*, 5(3):e1000406, 2009.

[108] T. Bataillon, T. Zhang, and R. Kassen. Cost of adaptation and fitness effects of beneficial mutations in *Pseudomonas fluorescens*. *Genetics*, 189:939–949, 2011.

[109] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. Quantifying the adaptive potential of an antibiotic resistance enzyme. *PLoS Genet.*, 8(6):e1002783, 2012.

[110] M. Foll, Y. P. Poh, N. Renzette, A. Ferrer-Admetlla, C. Bank, S. Hyunjin, M. Anna-Sapfo, E. Gregory, L. Ping, W. Daniel, R. C. Daniel, B. Z. Konstantin, N. B. Daniel, P. W. Jennifer, F. K. Timothy, A. S. Celia, W. F. Robert, and D. J. Jeffrey. Influenza virus drug resistance: A time-sampled population genetics perspective. *PLoS Genet.*, 10(2):e1004185, 2014.

[111] C. Bank, T. H. Ryan, D. J. Jeffrey, and N.A.B. Daniel. A systematic survey of an intragenic epistatic landscape. *Mol. Biol. Evol.*, 32(1):229–238, 2014.

[112] D. R. Rokyta, Z. Abdo, and H. A. Wichman. The genetics of adaptation for eight microvirid bacteriophages. *J. Mol. Evol.*, 69:229, 2009.

[113] H. A. Orr. The population genetics of adaptation: The adaptation of DNA sequences. *Evolution*, 56:1317–1330, 2002.

[114] H. A. Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163:1519–1526, 2003.

[115] H. A. Orr. A minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.*, 220:241–247, 2003.

[116] H. A. Orr. The population genetics of adaptation on correlated fitness landscapes: The block model. *Evolution*, 60:1113, 2006.

[117] K. Jain and S. Seetharaman. Multiple adaptive substitutions during evolution in novel environments. *Genetics*, 189:1029–1043, 2011.

[118] S. Seetharaman and K. Jain. Length of adaptive walk on uncorrelated and correlated fitness landscapes. *Phys. Rev. E*, 90:32703, 2014.

[119] S. Seetharaman. *Adaptation on rugged fitness landscapes*. M.S. thesis, JNCASR, Bangalore, 2011.

[120] S.-C. Park and J. Krug. Clonal interference in large populations. *Proc. Natl. Acad. Sci. U.S.A.*, 104:18135–18140, 2007.

[121] J.A.G.M. de Visser and D. E. Rozen. Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics*, 172:2093–2100, 2006.

[122] J.A.G.M. de Visser, C.W. Zeyl, P.J. Gerrish, J.L. Blanchard, and R.E. Lenski. Diminishing returns from mutation supply rate in asexual populations. *Science*, 283:404–406, 1999.

[123] R. Miralles, P. J. Gerrish, A. Moya, and S.F. Elena. Clonal interference and the evolution of RNA viruses. *Science*, 285:813–815, 1999.

[124] D.E. Rozen, J.A.G.M. de Visser, and P. J. Gerrish. Fitness effects of fixed beneficial mutations in microbial populations. *Curr. Biol.*, 12:1040–1045, 2002.

[125] J. F. C. Kingman. A simple model for the balance between selection and mutation. *J. Appl. Prob.*, 15:1, 1978.

[126] C.O. Wilke and T. Martinetz. Adaptive walks on time-dependent fitness landscapes. *Phys. Rev. E*, 60:2154–2159, 1999.

[127] N.A. Rosenberg. A sharp minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.*, 237:17–22, 2005.

[128] S. Kryazhimskiy, G. Tkačik, and J. B. Plotkin. The dynamics of adaptation on correlated fitness landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 106:18638–18643, 2009.

[129] P.R.A. Campos and L. M. Wahl. The adaptation rate of asexuals: deleterious mutations, clonal interference and population bottlemecks. *Evolution*, 64(7):1973–1983, 2010.

[130] S.E. Schoustra, T. Bataillon, D.R. Gifford, and R. Kassen. The properties of adaptive walks in evolving populations of fungus. *PLoS Biol.*, 7(11):e1000250, 2009.

[131] R. C. MacLean, G. G. Perron, and A. Gardner. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in *Pseudomonas aeruginosa*. *Genetics*, 186:1345–1354, 2010.

[132] D. R. Gifford, S. E. Schoustra, and R. Kassen. The length of adaptive walks is insensitive to starting fitness in *Aspergillus nidulans*. *Evolution*, 65:3070–3078, 2011.

[133] A. Sousa, S. Magalhães, and I. Gordo. Cost of antibiotic resistance and the geometry of adaptation. *Mol. Biol. Evol.*, 29:1417–1428, 2012.

[134] K. Jain and J. Krug. Adaptation in simple and complex fitness landscapes. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural Approaches to Sequence Evolution: Molecules, Networks and Populations*, pages 299–340. Springer, Berlin, 2007.

Appendix A

A.1 Deterministic dynamics and stationary state

Equation (2.1) is nonlinear in the population fraction due to the first term on the RHS. This nonlinearity can be eliminated by a change of variables from $X(j,t)$ to an unnormalised population variable $Z(j,t)$ which is defined as [134, 91]

$$Z(j,t) = X(j,t)\, e^{\int_0^t dt' \, \overline{w}(t')} \tag{A.1}$$

Then the unnormalised population fraction obeys the following *linear* differential equation:

$$\begin{aligned} \frac{\partial Z(j)}{\partial t} &= -sj\, Z(j,t) - [(L-j)\mu + j\nu]Z(j,t) \\ &+ (L-j+1)\mu Z(j-1,t) + (j+1)\nu Z(j+1,t) \end{aligned} \tag{A.2}$$

with boundary conditions

$$Z(-1,t) = Z(L+1,t) = 0 \tag{A.3}$$

at all times. The RHS of (A.2) is a three-term recursion relation (in j) with variable coefficients, which is usually not easy to solve.

Inspired by the results of [58], we assume that the population fraction $Z(j,t)$ is of the following form

$$Z(j,t) = \binom{L}{j} x_1^j(t)\, x_2^{L-j}(t) \tag{A.4}$$

where x_1, x_2 are calculated below. The normalised fraction $X(j,t)$ is then given by [134, 91]

$$
\begin{aligned}
X(j,t) &= \frac{Z(j,t)}{\sum_{j'=0}^{L} Z(j',t)} && \text{(A.5)} \\
&= \binom{L}{j} x^j(t)\,(1-x(t))^{L-j} && \text{(A.6)}
\end{aligned}
$$

where $x = x_1/(x_1+x_2)$ lies between zero and one. It should be noted that the above form for the population fraction of a fitness class implies that each locus in the sequence contributes *independently* to the population fraction of a sequence.

Using the ansatz (A.4) in the boundary conditions (A.3), we find that x_1, x_2 obey linear, coupled differential equations which can be expressed as

$$
\frac{\partial}{\partial t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\nu - s & \mu \\ \nu & -\mu \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{A.7}
$$

On using the ansatz (A.4) in the bulk equation (A.2) for which $0 < j < L$, we get

$$
\frac{j}{x_1}\left[\frac{\partial x_1}{\partial t} + (\nu + s)x_1 - \mu x_2\right] + \frac{L-j}{x_2}\left[\frac{\partial x_2}{\partial t} - \nu x_1 + \mu x_2\right] = 0 \tag{A.8}
$$

However due to (A.7), the coefficient of j and $L-j$ equals zero for any $0 < j < L$. Thus the ansatz (A.4) is consistent with the bulk equations, and the problem reduces to solving the matrix equation (A.7). By going to the diagonal basis, we obtain

$$
\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} \frac{2\mu}{\nu-\mu+s+\sqrt{(\mu-s-\nu)^2+4\mu\nu}} & \frac{2\mu}{\nu-\mu+s-\sqrt{(\mu-s-\nu)^2+4\mu\nu}} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} e^{\lambda_+ t}\,\tilde{x}_1(0) \\ e^{\lambda_- t}\,\tilde{x}_2(0) \end{pmatrix} \tag{A.9}
$$

where the column vectors in the matrix above are the eigenvectors of the matrix on the RHS of (A.7) corresponding to the eigenvalues $\lambda_\pm$, which are given by

$$
\lambda_\pm = \frac{-\nu - \mu - s \pm \sqrt{(\mu - s - \nu)^2 + 4\mu\nu}}{2} \tag{A.10}
$$

and $\tilde{x}_1(0), \tilde{x}_2(0)$ can be found using the initial condition $X(j,0)$.

In the steady state, the population fraction is obtained by taking the limit $t \to \infty$ in the expressions of $x_1(t), x_2(t)$ obtained above. Using the fact that the eigenvalue λ_- in (A.10) is negative, we find that the steady state fraction x is given by (2.5).

A.2 Moran model for neutral, nonrecombining population

For the Moran process defined in the main text, the probability distribution $P(n(i),t)$ of the number of individuals in the fitness class i evolves according to the following equation:

$$\begin{aligned}
&\frac{\partial P(n(i),t)}{\partial t} \\
&= \sum_{j \neq i} \Bigg[\sum_{n(j)=1}^{N-n(i)} P(n(i)+1, n(j)-1, t)\; R(n(i)+1 \to n(i), n(j)-1 \to n(j)) \\
&- \sum_{n(j)=0}^{N-n(i)} P(n(i), n(j), t)\; R(n(i) \to n(i)-1, n(j) \to n(j)+1) \\
&+ \sum_{n(j)=1}^{N-n(i)} P(n(i)-1, n(j)+1, t)\; R(n(i)-1 \to n(i), n(j)+1 \to n(j)) \\
&- \sum_{n(j)=0}^{N-n(i)} P(n(i), n(j), t)\; R(n(i) \to n(i)+1, n(j) \to n(j)-1) \Bigg]
\end{aligned} \tag{A.11}$$

where R is the rate at which a birth-and-death event occurs, and $P(n(i),n(j),t)$ is the joint distribution of the number of individuals in the ith and jth fitness class. Using the above equation, it can be seen that the average number of individuals in the fitness class i given by $\bar{n}(i,t) = \sum_{n(i)=1}^{N} n(i) P(n(i),t)$ changes as

$$\begin{aligned}
\frac{\partial \bar{n}(i,t)}{\partial t} = \sum_{j \neq i} \sum_{n(i),n(j)} & [R(n(i) \to n(i)+1, n(j) \to n(j)-1) P(n(i),n(j),t) \\
& - R(n(i) \to n(i)-1, n(j) \to n(j)+1) P(n(i),n(j),t)]
\end{aligned} \tag{A.12}$$

We next find the rates at which the birth-and-death process occurs. For class $i, j = 0, ..., L$ and $i \neq j$, we have

$$\begin{aligned} & R(n(i) \rightarrow n(i)+1, n(j) \rightarrow n(j)-1) \\ = \; & (1-\mu_i-\nu_i)\frac{n(i)}{N}\frac{n(j)}{N} + \mu_{i-1}\frac{n(i-1)}{N}\frac{n(j)}{N} + \nu_{i+1}\frac{n(i+1)}{N}\frac{n(j)}{N} \end{aligned} \tag{A.13}$$

with $n(-1) = n(L+1) = 0$. In the above equation, the first term on the RHS gives the probability of the event that a birth occurs in the ith class, the offspring does not mutate and a death occurs in the jth class, while the second and third term give the probability that a birth occurs in a class neighboring the ith class, the offspring acquires a mutation and a death occurs in the jth class. On using the above equation in (A.12), after some simple algebra, we get

$$\frac{\partial \bar{n}(i,t)}{\partial t} = \mu_{i-1}\bar{n}(i-1,t) + \nu_{i+1}\bar{n}(i+1,t) - (\mu_i+\nu_i)\bar{n}(i,t) \;,\; 0 \leq i \leq L \tag{A.14}$$

which can be easily solved in the stationary state to give (2.10).

A.3 Deterministic solution in the absence of beneficial mutations

Consider an infinitely large, nonrecombining population when only deleterious mutations are allowed. Let J be the least-loaded fitness class so that the frequency $X_J^{(0)}(k,t) = 0 \;,\; k < J$ at all times. Then the evolution (2.1) reduces to

$$\begin{aligned} \frac{\partial X_J^{(0)}(j,t)}{\partial t} &= -(sj + \bar{w}_J(t))X_J^{(0)}(j,t) - (L-j)\mu X_J^{(0)}(j,t) \\ &+ (L-j+1)\mu X_J^{(0)}(j-1,t) \;,\; J \leq j \leq L \end{aligned} \tag{A.15}$$

where the average fitness $\bar{w}_J(t) = -s\sum_{k=J}^{L} kX_J^{(0)}(k,t)$. In the stationary state, the equation for $j = J$ gives

$$\bar{w}_J = -(L-J)\mu - sJ \tag{A.16}$$

On iterating the two-term recursion relation for $X_J^{(0)}(j)$, we obtain

$$X_J^{(0)}(j) = \binom{L-J}{j-J} \left(\frac{\mu}{s}\right)^{j-J} \left(1-\frac{\mu}{s}\right)^{L-j} \; , \; \mu < s \tag{A.17}$$

For $J = 0$, (2.8) is recovered.

Appendix B

B.1 Comparison of mutation schemes

As already described in Section 2.2, in an infinitely large population of finite diallelic sequences of length L in which the wild type allele mutates with rate μ and the back mutation occurs with rate ν, the frequency $X(j,t)$ of a sequence with j deleterious mutations and fitness $w(j) = -sj$ evolves in continuous time as [58, 3]

$$\dot{X}(j,t) = (j+1)\nu X(j+1,t) + (L-j+1)\mu X(j-1,t) - [(L-k)\mu + j\nu]X(j,t) - s(k-\bar{j})X(j,t)\,, \tag{B.1}$$

where $x_{-1} = x_{L+1} = 0$. In the limit $\mu, \nu \to 0$ and $L \to \infty$, one can define the deleterious and beneficial mutation rate per sequence as $U_d = L\mu$ and $U_b = L\nu$, and rewrite the above equation as

$$\dot{X}(j,t) = \varepsilon_{j+1}U_bX(j+1,t) + (1-\varepsilon_{j-1})U_dX(j-1,t) - [(1-\varepsilon_j)U_d + \varepsilon_jU_b]X(j,t) - s(j-\bar{j})X(j,t)\,, \tag{B.2}$$

where $\varepsilon_j = j/L$. In a well adapted population in which the number of loci carrying the deleterious allele is small, the back mutations to the wild type allele can be ignored. More precisely, when the number of deleterious mutations scales sub-linearly with L, the fraction $\varepsilon_j \to 0$ for an infinitely long sequence and we obtain the model defined by (3.1a) and (3.1b) with $U_b = 0$. Similarly, in a maladapted population in which $\varepsilon_j \to 1$, the model (B.2) reduces to the one studied in Chapter 3 with $U_d = 0$. The model defined by (3.1a) and (3.1b) thus interpolates between the two limits of the model (B.2) described above.

B.2 Neutral dynamics using eigenfunction expansion method

The treatment below for the unnormalised frequency $Y(j,t)$ essentially follows Chapter 7 of [92] and here we briefly describe our results. For $S = 0$, the solution of the eigenvalue equation (3.17b) is given by

$$\phi_j = C_+ a_+^j + C_- a_-^j \ , \ j \geq 0 \ , \tag{B.3}$$

where $a_\pm$ are solutions of the quadratic equation $a^2 + (\lambda - 2)a + 1 = 0$ and the coefficients C_+ and C_- are related due to the boundary equation (3.17a). Since $a_+ a_- = 1$, it is convenient to write $a_\pm = e^{\pm iq}$ where q is real. The latter condition is required to ensure that the eigenfunction ϕ_j does not diverge at large j (see (3.3)). Using $a_+ + a_- = 2 - \lambda$, we find that the eigenvalues form a continuous spectrum and are given by

$$\lambda = 4 \sin^2 \left(\frac{q}{2} \right) \ , \ 0 \leq q \leq \pi \ . \tag{B.4}$$

Since the ratio C_+/C_- determined using (3.24) has unit modulus, we can write

$$\frac{C_+}{C_-} = -\frac{\gamma - 2 + e^{iq}}{\gamma - 2 + e^{-iq}} = e^{i2\eta(q)} \ , \tag{B.5}$$

and finally arrive at

$$\phi_j(q) = \sqrt{\frac{2}{\pi}} \cos(qj + \eta(q)) \ , \tag{B.6}$$

where we have used the orthonormality condition (3.18) to determine the proportionality constant. For the initial condition $X(j,0) = \delta_{j,0}$, on replacing the sums in (3.16) and (3.20) by integrals (as the eigenvalues are continuous), we obtain

$$\begin{aligned} Y(j,t) &= \int_0^{\pi} \phi_j(q)\phi_0(q) e^{-2(1-\cos q)\sqrt{U_b U_d} t} dq \;, && \text{(B.7)} \\ &= \frac{2}{\pi}\int_0^{\pi} dq \sin q \frac{(\gamma-2)\sin(qj)+\sin(qj+q)}{(\gamma-2)^2+2(\gamma-2)\cos q+1}\; e^{-2(1-\cos q)\sqrt{U_b U_d} t} \;, && \text{(B.8)} \end{aligned}$$

where the last expression follows on using (B.5) in (B.6). The above integral does not appear to be exactly solvable, but in the scaling limit $q \to 0, j,t \to \infty$ with $q^2 t$ and qj finite, the above equation simplifies to give

$$\begin{aligned} Y(j,t) &\approx \frac{2}{\pi}\int_0^{\infty} dq \frac{(\gamma-1)q\sin(qj)+q^2\cos(qj)}{(\gamma-1)^2} e^{-q^2\sqrt{U_b U_d} t} \;, && \text{(B.9)} \\ &= \frac{1}{2\sqrt{\pi}} \frac{j}{\gamma-1} \frac{e^{-\frac{j^2}{4t\sqrt{U_b U_d}}}}{(t\sqrt{U_b U_d})^{3/2}} \;. && \text{(B.10)} \end{aligned}$$

B.3 Some properties of the Bessel functions

If $\mathscr{K}$ denotes J, Y, the Bessel function $\mathscr{K}_\nu(z)$ is defined as the solution of the following differential equation (9.1.1,[96])

$$z^2 \frac{d^2\mathscr{K}_\nu(z)}{dz^2} + z\frac{d\mathscr{K}_\nu(z)}{dz} + (z^2-\nu^2)\mathscr{K}_\nu(z) = 0 \;. \qquad \text{(B.11)}$$

The Bessel function of the second kind $Y_\nu(z)$ is related to the Bessel function of the first kind $J_\nu(z)$ by (9.1.2,[96])

$$Y_\nu(z) = \frac{\cos(\nu\pi)J_\nu(z) - J_{-\nu}(z)}{\sin(\nu\pi)} \;. \qquad \text{(B.12)}$$

For the series representation of $J_\nu(z)$, see (B.19) below; the asymptotic expansions of $J_\nu(z)$ for $z > \nu$ and $z < \nu$ are given in (B.25) and (B.31), respectively.

B.4 Cumulants of the number of deleterious mutations

Following [75], we first define the generating function of the frequency as

$$F(\xi,t)=\sum_{k=0}^{\infty}X(j,t)e^{-\xi j}\,. \tag{B.13}$$

Multiplying (3.1a) and (3.1b) by $e^{-\xi j}$ and summing over j, we obtain

$$\frac{d\ln F(\xi,t)}{dt}=s\mathscr{C}_1(t)+s\frac{d\ln F(\xi,t)}{d\xi}-U_d(1-e^{-\xi})+U_b(e^{\xi}-1)\left(1-\frac{X(0)}{F(\xi,t)}\right)\,. \tag{B.14}$$

The nth cumulant $\mathscr{C}_n(t), n=1,2,...$ of the number of deleterious mutations is related to $F(\xi,t)$ through

$$\ln F(\xi,t)=\sum_{n=1}^{\infty}\mathscr{C}_n(t)\frac{(-\xi)^n}{n!}\,. \tag{B.15}$$

Using the above equation in (B.14), we get

$$\begin{aligned}\sum_{n=1}^{\infty}\dot{\mathscr{C}}_n(t)\frac{(-\xi)^n}{n!} &= -s\sum_{n=1}^{\infty}\mathscr{C}_{n+1}(t)\frac{(-\xi)^n}{n!}+U_d\sum_{n=1}^{\infty}\frac{(-\xi)^n}{n!}\\ &+ U_b\left(\sum_{n=1}^{\infty}\frac{\xi^n}{n!}\right)\left(1-\frac{X(0,t)}{F(\xi,t)}\right)\,.\end{aligned} \tag{B.16}$$

Matching the coefficient of ξ^n for $n=1,2$ on both sides, we find that

$$\dot{\mathscr{C}}_1(t) = -s\mathscr{C}_2(t)+U_d-U_b(1-X(0,t))\,, \tag{B.17}$$

$$\dot{\mathscr{C}}_2(t) = -s\mathscr{C}_3(t)+U-U_bX(0,t)(1+2\mathscr{C}_1(t))\,. \tag{B.18}$$

B.5 Approximate expressions for the eigenvalues when S is large

The Bessel function of the first kind has the following series representation (9.1.10, [96]):

$$J_\nu(z) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!(\nu+m)!} \left(\frac{z}{2}\right)^{2m+\nu} . \tag{B.19}$$

For large S, keeping the first two terms in the above series and using it in the eigenvalue equation (3.24), we get

$$(\lambda-\gamma)(\lambda-S-2) \approx 1 . \tag{B.20}$$

Solving the above quadratic equation, we obtain the first two eigenvalues λ_0 and λ_1 given in (3.31) and (3.48) respectively. Figure 3.1 shows a comparison between the exact eigenvalues obtained numerically using (3.24) and the above approximations. Our numerical analysis of (3.24) also suggests that the αth eigenvalue is given by

$$\lambda_\alpha \approx \alpha S + 2 \ , \ \alpha = 1, 2, ... \tag{B.21}$$

B.6 Approximate expression for the stationary distribution when S is large

Noting that the terms corresponding to $m = 0, 1$ in the summand on the RHS of (3.33) contribute to the leading order term in U_b/s, we obtain

$$X(j) \propto \frac{\mu^j}{j!} \left[1 + \frac{U_b}{s}\left((\gamma_{EM} - H_j)(1+\mu) - \frac{\mu}{j+1}\right)\right] , \tag{B.22}$$

where $\mu = U_d/s$. In the above equation, we have used that $j!/(j+\varepsilon)! \approx 1 + \varepsilon(\gamma_{EM} - H_j)$ for small ε where $\gamma_{EM} \approx 0.577...$ and $H_j = \sum_{i=1}^{j} i^{-1}$ is the Harmonic number. On fixing the proportionality constant using (3.2), we recover (A7) of [27]:

$$X(j)=\frac{e^{-\mu}\mu^{j}}{j!}\left[1+\frac{U_b}{s}\left(\sum_{m=0}^{\infty}\frac{e^{-\mu}\mu^{m}}{m!}(H_m(1+\mu)+\frac{\mu}{m+1})-H_j(1+\mu)-\frac{\mu}{j+1}\right)\right]. \tag{B.23}$$

For $U_d < s$, the above expression shows that the distribution is close to the Poisson distribution (3.5). However, for $U_d > s$, we obtain [27]

$$X(j)\approx\frac{e^{-U_d/s}}{j!}\left(\frac{U_d}{s}\right)^{j}\left[1+\frac{U_bU_d}{s^2}\ln\left(\frac{U_d}{sj}\right)\right], \tag{B.24}$$

on using $\sum_{m=0}^{\infty}z^m H_m/m! \approx e^z \ln z$ for large z and $H_j \approx \ln j + \gamma_{EM}$ for large j in (B.23). This result shows that the beneficial mutations enhance the frequency in fitness classes $j < U_d/s$ and diminish it in higher ones (also, see Fig. 3.3).

B.7 Approximate expressions for the eigenvalues when S is small

For the Bessel function $J_\nu(z), z > \nu$, the asymptotic expansion for large orders is given by (9.3.3, [96])

$$J_\nu(\nu\sec\beta)\sim\frac{\cos\left(\nu(\tan\beta-\beta)-(\pi/4)\right)}{\sqrt{(\nu/2)\pi\tan\beta}}, \; 0<\beta<\pi/2\,. \tag{B.25}$$

As shown in Fig. 3.1, the eigenvalues λ_0, λ_1 are an increasing function of S and approach zero as $S \to 0$ (also, see (B.4) for the neutral case). Then using (B.25) in (3.24) and carrying out a small λ expansion, we obtain

$$\frac{\cos(\frac{2\lambda^{3/2}}{3S}-\frac{\pi}{4}-\sqrt{\lambda})}{\cos(\frac{2\lambda^{3/2}}{3S}-\frac{\pi}{4})}=\gamma-\lambda\,. \tag{B.26}$$

After some algebra, the above simplifies to

$$\tan\left(\frac{2}{3}\frac{\lambda^{3/2}}{S}\right)=-\frac{\gamma-1+\sqrt{\lambda}-\lambda}{\gamma-1-\sqrt{\lambda}-\lambda}\,. \tag{B.27}$$

The above equation immediately suggests that the eigenvalue $\lambda \sim S^{2/3}$ so that the RHS can be nonzero and finite. Guided by this observation and a numerical analysis of (3.24), we expect that the αth eigenvalue is of the following form:

$$\lambda_\alpha = \lambda_\alpha^{(0)} S^{2/3} + \lambda_\alpha^{(1)} S \; . \tag{B.28}$$

Substituting this in (B.27) and expanding both sides of the equation for small S, we find that

$$\lambda_\alpha^{(0)} = \left(\frac{3\pi(4\alpha+3)}{8} \right)^{2/3} , \tag{B.29}$$

$$\lambda_\alpha^{(1)} = -\frac{1}{\gamma-1} \; . \tag{B.30}$$

As we have assumed λ to be small to arrive at (B.26), the above results for the eigenvalues are valid for small α. For larger α, our numerical analysis of (3.24) suggests that the eigenvalues increase linearly with S.

B.8 Approximate expression for the stationary distribution when S is small

The asymptotic expansion of the Bessel function $J_\nu(z), z < \nu$ for large orders is given by (9.3.1, [96])

$$J_\nu(z) \sim \frac{1}{\sqrt{2\pi\nu}} \left(\frac{ze}{2\nu} \right)^\nu \; . \tag{B.31}$$

Using this in (3.27), we obtain the steady state frequency to be

$$X(j) \propto \left(\sqrt{\frac{U_d}{U_b}} \right)^{j+\delta} \frac{1}{\sqrt{j+\delta}} \left(\frac{e/S}{j+\delta} \right)^{j+\delta} , \tag{B.32}$$

where $\delta = (2-\lambda_0)/S$ and λ_0 is given by (3.37). A Gaussian approximation for the above expression can be obtained by writing $X(j) \propto e^{I(j)}$ and expanding $I(j)$ about its turning point $\tilde{j} = (U_d/s) - \delta$ up to quadratic orders in the deviation $j - \tilde{j}$. On fixing the normalisation, we

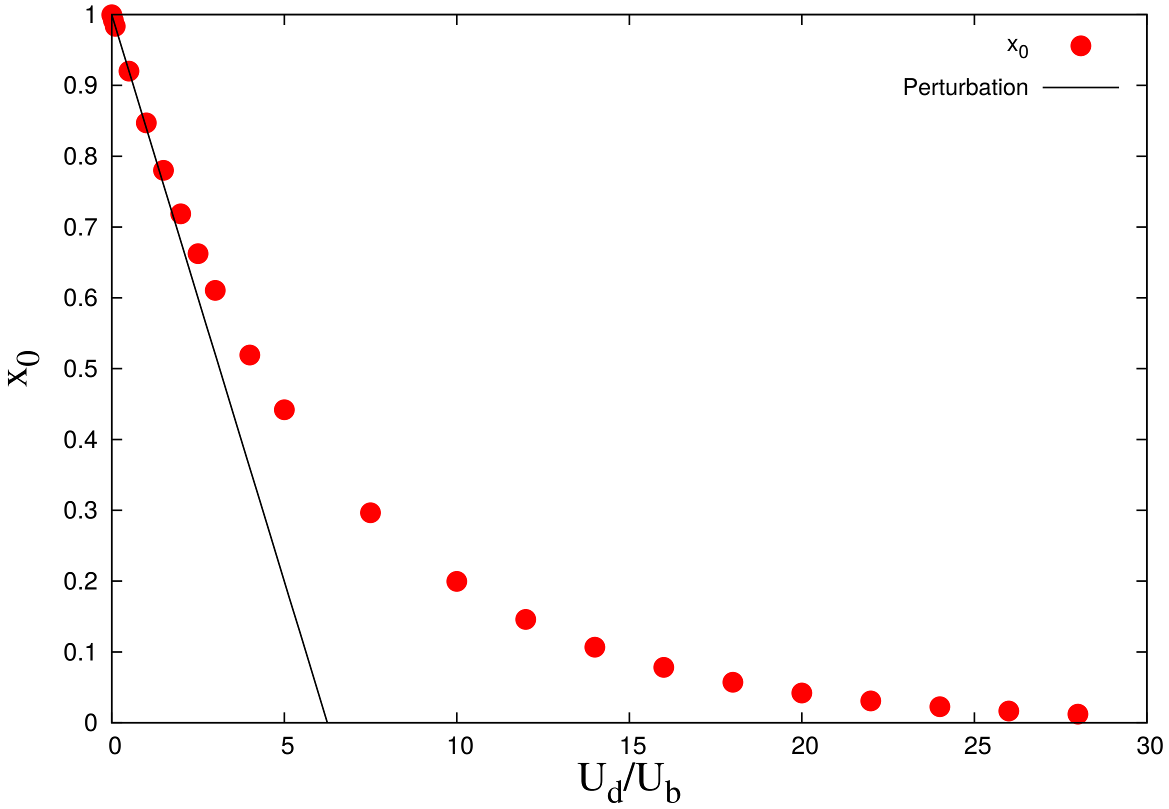


Fig. B.1 Figure shows the decrease in the population fraction in the zeroth fitness class with deleterious mutation rate (U_d) for a fixed value of $s = 0.05$ and $U_b = 0.01$. The solid line shows the result of the perturbation theory given in (B.38) for $U_d < U_b$.

obtain

$$X(j) \approx \sqrt{\frac{s}{2\pi U_d}} \exp\left[-\frac{s}{2U_d}\left(j - \frac{U_d}{s} + \frac{2-\lambda_0}{S}\right)^2\right] . \tag{B.33}$$

The mean and variance of the above distribution differs from (3.38) and (3.39) by a factor U_b/s and is therefore a good approximation when $U_b \ll U_d$.

B.9 Stationary state distribution when the deleterious mutation rate is smaller than the beneficial one

For completeness, here we consider the parameter regime in which $U_d < U_b$ within a perturbation theory in U_d. We begin by expanding the steady state fraction in a power series in U_d

as

$$X(j) = \sum_{n=0}^{\infty} U_d^n \frac{X^{(n)}(j)}{n!} \,, \tag{B.34}$$

where $X^{(n)}(j)$ is the nth derivative of $X(j)$ with respect to U_d evaluated at $U_d = 0$. When the deleterious mutation rate is zero, as the entire population is in the zeroth fitness class, we have $X^{(0)}(j) = \delta_{j,0}$. As a result, the mean $\mathscr{C}_1^{(0)} = \sum_{j=0}^{\infty} jX^{(0)}(j) = 0$. Using (B.34) in (3.1a) and (3.1b) in the steady state and retaining terms to leading order in U_d, we obtain

$$s\mathscr{C}_1^{(1)} = 1 - U_b X^{(1)}(1) \tag{B.35}$$

$$U_b X^{(1)}(2) = (U_b + s)X^{(1)}(1) - 1 \tag{B.36}$$

$$U_b X^{(1)}(j) = (U_b + s(j-1))X^{(1)}(j-1) \,,\; j \geq 3 \,. \tag{B.37}$$

The last equation implies that the steady state fraction is a monotonically increasing function of j; however, since each frequency is bounded above by unity and the total fraction must also add up to one, to obtain a sensible result to linear order in U_d, the fraction $X^{(1)}(j)$ must be zero for all $j \geq 2$. This immediately yields

$$X(0) = 1 - \frac{U_d}{U_b + s} + \mathscr{O}(U_d^2), \tag{B.38}$$

$$X(1) = \frac{U_d}{U_b + s} + \mathscr{O}(U_d^2), \tag{B.39}$$

$$X(j) = \mathscr{O}(U_d^2) \,,\; j \geq 2 \,. \tag{B.40}$$

Thus the fraction in the zeroth fitness class decreases linearly with U_d when the deleterious mutation rate is smaller than the beneficial one; in contrast, the fraction $X(0)$ decays exponentially or faster with U_d when $U_d > U_b$ (see Fig. B.1 and Sections 3.4.1 and 3.4.2).

B.10 Finite Size population

As real populations are finite and evolve stochastically, here we present our preliminary results taking random genetic drift into account. The dynamics of the population is modeled in the

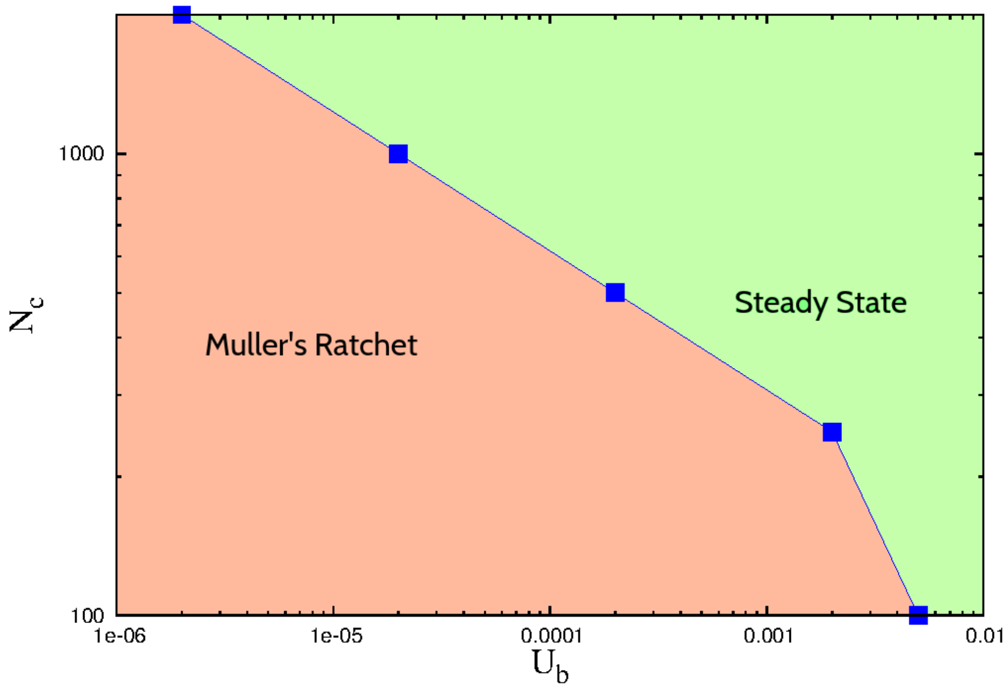


Fig. B.2 Figure shows the transition from a ratcheting state to a steady state as the population size crosses the critical value N_c. Other parameters are $U_d = 0.1$ and $s = 0.05$.

same way as described in Section 2.2, except that the mutation probabilities for beneficial and deleterious mutations are drawn from a Poisson distribution with mean U_d and U_b. We find that unlike for the mutation scheme studied in Chapter 2 where a finite population always attains a steady state in the presence of beneficial mutations, here Muller's ratchet stops only when the population size is above a critical value N_c [51]. Our numerical results for the dependence of N_c on beneficial mutation rate is shown in Fig. B.2 and we find that with increasing beneficial mutation rate, the minimum population size required to halt the Muller's ratchet decreases. A complete analytical understanding of these results is however not currently available.