Dynamics of adaptation: role of extreme value domains, initial fitness and fitness correlations

A Thesis

Submitted For the Degree of

DOCTOR OF PHILOSOPHY

in the Faculty of Science

by

Sarada Seetharaman

THEORETICAL SCIENCES UNIT
JAWAHARLAL NEHRU CENTRE FOR ADVANCED SCIENTIFIC RESEARCH
Bangalore – 560 064

MARCH 2015

To my family

DECLARATION

I hereby declare that the matter embodied in the thesis entitled " **Dynamics of adaptation: role of extreme value domains, initial fitness and fitness correlations** " 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.

Sarada Seetharaman

CERTIFICATE

I hereby certify that the matter embodied in this thesis entitled " **Dynamics of adaptation: role of extreme value domains, initial fitness and fitness correlations** " has been carried out by Ms. Sarada Seetharaman 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

I express my deep gratitude to my mentor Prof. Kavita Jain for her constant guidance, concern and support. Her keen insights helped me approach problems systematically and valuable suggestions kept me on the right path. Her patience and enthusiasm were always a source of inspiration and her support during my emotional low points helped me tide over them. She is always approachable and understanding and it is a joy to work with her.

I am grateful to Prof. M. R. S. Rao and Prof. K. S. Narayan, the past and present President of JNCASR and Prof. C. N. R. Rao, the founding president of JNCASR for excellent research facilities and creating good scientific environment.

I would like to thank all the faculty members of TSU - Prof. Shobana Narasimhan, Prof. Srikanth Sastry, Prof. Umesh V. Waghmare, Prof. Swapan K. Pati, Prof.Vidhyadhiraja N. S. and Prof. Subir K. Das for giving me the chance to be a part of JNCASR. In these 5 years I have learnt a lot and I am very grateful for the opportunity.

I would also like to express my gratitude to Prof. Swapan K Pati, Prof.

Rama Govindarajan, Prof. Subir Das, Prof. Amitabh Joshi and Prof. Chandan Dasgupta for the classes they have handled for me, the discussion and suggestions regarding my work.

I thank the wardens Prof. Sheeba Vasu and Prof. Subi George for helping me with the hostel accommodation. I thank the administrative officer, Mr. Jayachandra for his help during various points of my stay in JNCASR. I also appreciate the help I have received from administrative staff, complab staff and library staff which allowed me to remain focused on my work. I also thank hostel staff and mess workers for making my life comfortable in JNC.

I acknowledge the help and support of my labmates Gayatri Das, Priyanka, Sona John and Ananthu James. They are responsible for keeping the lab environment friendly and conducive to research.

I acknowledge the enjoyable time I had with my fellow stud reps: Anjali, Deepak, Nikhil and Vybhav.

I would also like to thank my friends in JNCASR- Suman, Sutapa, Shaista, Vinay, Vishwas, Moumita, Shiladitya, Sonia, Jitu, Deb, CD, Meha, Sankalp and Nikhil for helping me with my research and for making my stay in JNC very enjoyable. I especially thank Suman for his help with respect to the thesis writing.

Last but not the least, I express my deep gratitude to my family and my fiancé for the their unwavering support.

Synopsis

Adaptation is an ubiquitous phenomenon occurring in nature. It sometimes helps humans, for example, in the domestication of plants and animals. On the other hand, it can also cause great distress, for example, when drug-resistant microbes emerge [1]. Therefore it is crucial to understand how a population adapts, in order to address the effects caused by it. In asexual populations, adaptation can occur only via beneficial mutations which are changes in the genetic sequence that increase its chances of survival. These are the mutations that we are interested in our work and in order to study their effect, we consider an asexual population of binary sequences that can replicate, mutate and undergo random genetic drift. We consider the adaptation dynamics of this population on rugged fitness landscapes which are endowed with a large number of local fitness peaks [2]. In Fig. 1, we show a schematic example of an asexual population adapting on a rugged fitness landscape after an abrupt change in its environment as a result of which, the population which was previously at a fitness peak drops to a low fitness in the new environment. The population adapts or climbs the fitness landscape

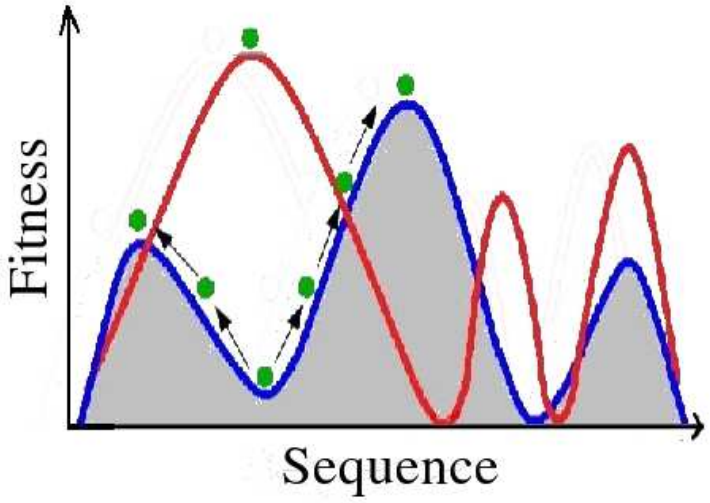


Figure 1: The fitness landscape in the old environment is shown by the broken red curve and in the new environment by the continuous blue one. The small green circles represent the population on the fitness landscape and when the fitness landscape is abruptly changed, it drops from a peak on the old fitness landscape to a lower fitness value in the new one. The arrows show the beneficial mutations because of which the population adapts.

by accumulating beneficial mutations.

The adaptation dynamics of a population depends on the size and frequency of beneficial mutations, or in other words, the distribution of beneficial fitness effects (DBFE). Whether adaptation happens via many mutations conferring small fitness advantage, or a few producing large fitness changes depends on the nature of DBFE. Although initial theoretical works suggested that adaptation occurs mostly by mutations that provide small benefits [3],

recent works suggest that large effect mutations are also possible [4]. The basic idea governing the shape of the DBFE is due to Gillespie [5], who suggested that the mutations conferring higher fitness than the wild type must lie in the right tail of the fitness distribution and so the statistical properties of such extreme fitnesses can be described by an extreme value theory (EVT) which states that the extreme value distribution of independent random variables can be of three types: Weibull which occurs when the fitnesses are right-truncated, Gumbel for distributions decaying faster than a power law and Fréchet for distributions with algebraic tails [6]. because beneficial mutations are rare, accounting for less than 15% of the total mutations and occur at a rate between 10^{-9} to 10^{-8} per cell per generation [7–9], it is a challenging task to measure them experimentally. But, in recent times some success has been achieved and interestingly, all the three EVT distributions have been observed [10–13].

In our work, in order to access all the EVT distributions, we generate the fitness of each sequence from a generalised Pareto distribution defined as

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

In the above equation, the parameter κ can take any real value. The fitness distribution is unbounded for $\kappa \geq 0$ and has an upper bound u at $-1/\kappa$ when $\kappa < 0$. The advantage of using (1) is that all the three extreme value domains

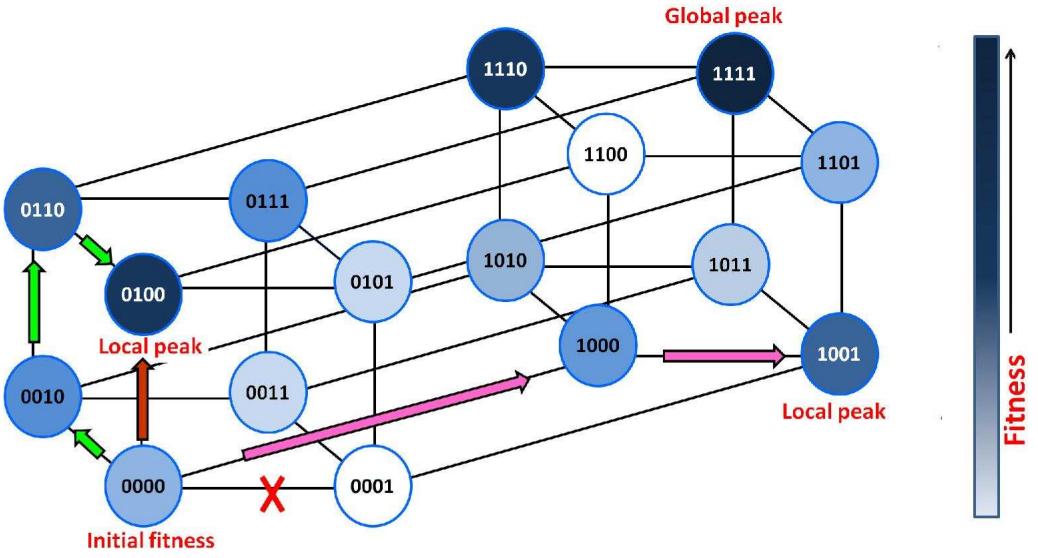


Figure 2: Schematic representation of adaptive walks in a 4-dimensional sequence space, starting from the same initial sequence. The arrows represent the shift of the population from a sequence to a fitter sequence one mutation away.

can be accessed by tuning a single parameter κ. The tails of the fitness distribution when $\kappa < 0$, $\to 0$ and > 0 belong to Weibull, Gumbel and Fréchet distributions respectively. The main aim of our work is to study the adaptation dynamics of the population when the underlying fitness distribution belongs to one of the three EVT domains.

As already mentioned, the population moves on the fitness landscape by means of beneficial mutations and their availability is controlled by the number of mutants appearing in the population. In the weak mutation regime, the number of mutants produced in the population is much less than one per generation. In this case, the population can move from its

initial fitness until a local fitness peak only by means of single mutations that will produce new sequences differing from the current sequence at a single locus. When any mutation that confers a fitness benefit, if not lost due to genetic drift quickly get fixed in the population, then the population is said to be under strong selection. In the strong selection-weak mutation regime (SSWM), the whole population can be represented by a single particle that climbs the fitness landscape by one step mutantations and is termed an *adaptive walk*. The walk ends when the population encounters a local fitness peak. stops. Every mutation that gets fixed is termed as a step in the adaptive walk and the total number of such steps till a local fitness peak is the walk length.

Most of the work in this thesis is in the SSWM regime in which, the dynamics of adaptation depends, in addition to the extreme value domain of the beneficial mutations, on the initial fitness of the population and the fitness correlations between the fitnesses. The two main quantities that we calculated in our work by varying all the parameters mentioned in the previous statement are the average walk length [14, 15] and the average fitness at a step [14, 16]. From the latter, we obtained the fitness difference between successive steps and find that the fitness difference between successive steps decreases in the Weibull domain, increases in the Fréchet domain and is a constant in the Gumbel domain [16]. This is the most important and interesting result of our work as it suggests a simple way to distinguish between the EVT domains.

The thesis is divided into five chapters and now I will briefly describe the contents of each. In *Chapter 1*, we introduce concepts that guided our work and the various terms pertinent to the discussion. We discuss the notion of fitness landscapes and the experiments that suggest that they are partially rugged [2]. Then we explain a theoretical model for a broad class of fitness landscapes known as block model [17] in which fitness correlations can be tuned. We then discuss the experiments that have been carried out to determine the DBFE of a biological population under consideration. In this chapter, we also mention the relevant mutation regimes- the weak mutation regime in which the SSWM model can be used and the strong mutation regime, in which it cannot be. The bulk of our work in this thesis is based in the SSWM regime in which the population climbs the fitness landscape until a local fitness peak is reached by one-step mutations as shown in Fig. 2. In contrast to the weak mutation regime in which monomorphic populations are seen, the population in the strong mutation regime has many better mutants existing in the population at the same time and competing with each other for dominance. In this chapter, we also discuss the results obtained from theoretical works dealing with the strong mutation regime [18, 19]

In *Chapter 2*, we describe the models used in our work here. The bulk of it is concerning the adaptive walks in the SSWM regime in which the number of mutants produced per generation is much smaller than one. During the walk, the probability that the population moves from the current fitness h to one of the fitter one-mutants with fitness $f > h$ at the next step is given

by the transition probability $T(f \leftarrow h)$. We define the quantities of interest and review the known results here. In previous works, the average number of steps in the adaptive walk from its initial fitness to a local fitness peak was determined in the two limiting cases of the adaptive walks namely, greedy walk [20] and random adaptive walk [21]. While in the former, the best of all available mutants is chosen at every step resulting in a constant average walk length independent of the initial fitness, in the latter, any better mutant is chosen with equal probability yielding an average walk length that depends on the logarithm of the rank of the initial fitness. We discuss the results of both these limiting cases of the move rule. In the biologically more realistic model termed the natural selection adaptive walk, the transition probability $T(f \leftarrow h)$ depends on the selection coefficient,$s = \frac{f-h}{h}$ which is the relative fitness difference between h and f. If u is the upper limit of the fitness distribution, the transition probability for an infinitely long sequence can be written as [4, 22]

$$T(f \leftarrow h) = \frac{(1 - e^{-\frac{2(f-h)}{h}})\ p(f)}{\int_h^u dg\ (1 - e^{-\frac{2(g-h)}{h}})\ p(g)}\ ,\ f > h \text{ (full model)} \tag{2}$$

For small selection coefficients, the proportionality factor can be approximated as $T(f \leftarrow h) \propto 2\ s$ and in this case, the transition probability for a sequence of infinite length is given by [4, 23]

$$T(f \leftarrow h) = \frac{(f - h)\ p(f)}{\int_h^u dg\ (g - h)\ p(g)}\ ,\ f > h \text{ (linear model)} \tag{3}$$

For a sequence of fixed length L, the adaptive walk would terminate after a certain number of steps when there are no better mutants available at that step. The probability of the adaptive walk, starting from initial fitness f_0, taking the step $J+1$ and assuming fitness f after already having taken step J with fitness h during the walk can be obtained from the following equation [14]

$$\mathcal{P}_{J+1}(f|f_0) = \int_{f_0}^{f} dh\ T(f \leftarrow h)\ (1 - q^L(h))\ \mathcal{P}_J(h|f_0)\ ,\ J \geq 0 \qquad (4)$$

The above equation is the main equation used in our work for analytical calculations. In (4), $T(f \leftarrow h)$ is given either by (2) or (3), and $q(h)$ is the probability of having a fitness less than h obtained as $q(h) = \int_0^h dg\ p(g)$. The above equation simply means that the probability of the walk taking step $J+1$ and having fitness fitness f is the product of the probability that it has fitness $h < f$ at step J, the probability of the fitness increasing from fitness h to f and the probability that not all L mutants produced at that step have a fitness less than h.

Though (2) is the biologically accurate model, it is analytically difficult to handle and for the sake of calculations, one may resort to (3) [14, 24]. Another reason for using (3) is that the results obtained from it matches the results obtained using (2) in the Weibull domain. We have also tuned the fitness correlations in our model using a block model [17] in which, a sequence of length L is considered to be built of B blocks of length $L_B = L/B$

and the fitness of each block is chosen from (1). The average of the fitness of all the blocks gives the fitness of the sequence. The fitness landscape would be uncorrelated when $B = 1$ and fully correlated when $B = L$ [25].

Although much of the work in this thesis is in the weak mutation regime, we have also studied the adaptation dynamics when the mutation is strong. This part of the work is chiefly numerical. The strong mutation regime is observed when the mutation rate or the population size is high so that many mutants are produced in the population at every time step. In this case, the population can no longer be reduced to a single particle climbing the fitness landscape and we used the Wright -Fisher dynamics to study this parameter regime. In the strong mutation regime, due to the presence of multiple mutants in the population at all times, the population is polymorphic and a step in this case is when for the first time, the probability of reproduction of any mutant exceeds half.

In *Chapter 3*, we consider the full model defined by (2) to study adaptation of infinitely long sequences for which the probability of all the one-step mutants having a lower fitness than the current one can be ignored, and the probability distribution of the population having fitness f at the step J in the full model can be calculated for exponentially and uniformly distributed fitnesses [14]. One of our main results is that the average fitness at step $J + 1$ can be written as

$$\bar{f}_{J+1} = a\ \bar{f}_J + b \quad , J \geq 0 \tag{5}$$

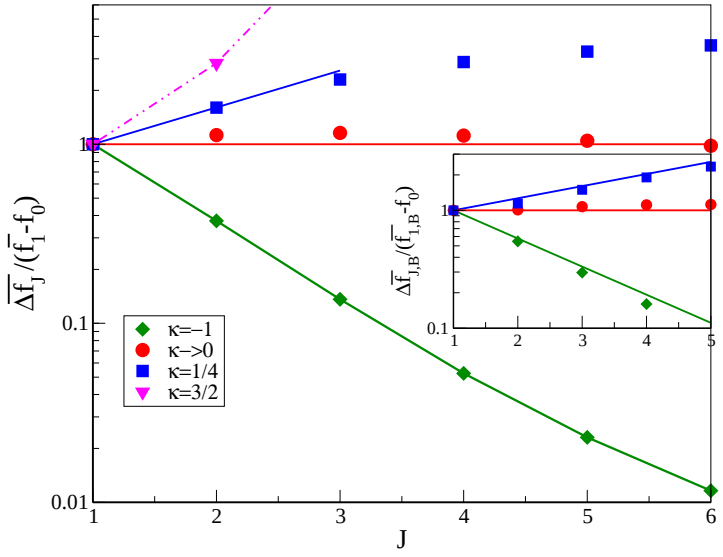


Figure 3: The plot shows (scaled) average fitness difference between successive steps as a function of the number of adaptive substitutions for various κ on uncorrelated (main) and correlated fitness landscapes with $B = 2$ (inset). Taking $f_0 = 0.63, 1, 1.14$ and 2.32 for $\kappa = -1, 0, 1/4$ and $3/2$ respectively, the simulation data are shown as points for $L_B = 1000$ which corresponds to $L = 1000$ and 2000 for independent and correlated fitnesses respectively. The line connecting the data points for $\kappa = 3/2$ is guide to the eye, while the others are obtained from theoretical calculations for uncorrelated and correlated fitnesses.

where the values of a and b depend on κ. For the distributions with finite mean (i.e. $\kappa < 1$), we obtained the results for the fitness difference at any step J of the adaptive walk in the three EVT domains as

$$\overline{\Delta f_J} = \begin{cases} a_-^{J-1}\ ((a_- - 1)f_0 + b_-)\ , & \kappa < 0 \quad (6a) \\ 2\ , & \kappa \to 0 \quad (6b) \\ a_+^{J-1}\ ((a_+ - 1)f_0 + b_+)\ , & 0 < \kappa < 1 \quad (6c) \end{cases}$$

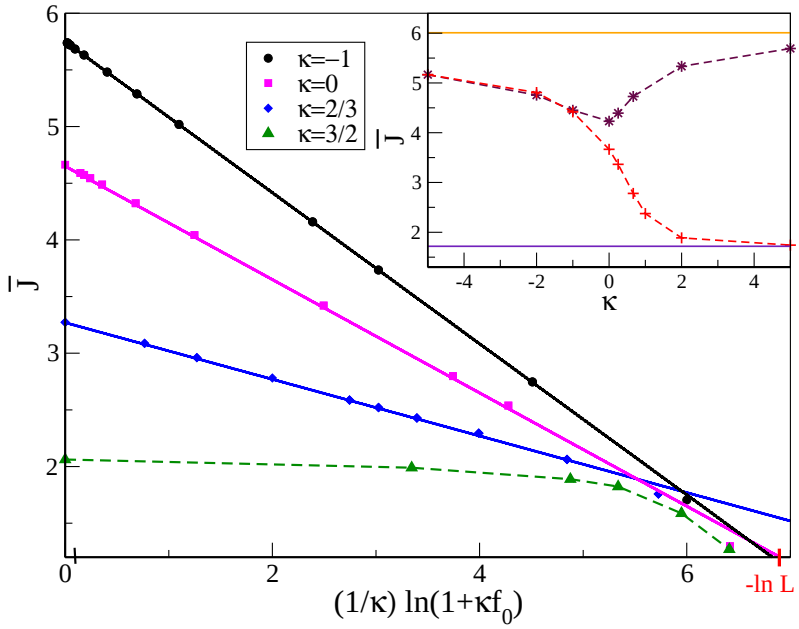


Figure 4: Main: Variation of the average walk length with initial fitness in the linear model on uncorrelated fitness landscapes for various κ. The simulation points are for $L = 1000$ and the lines are obtained from (7) for all $\kappa < 1$, while the one for $\kappa = 3/2$ is a guide to the eye. Inset: Comparison of the average walk length in the full model $(*)$ and the linear model $(+)$ for $(1/\kappa)\ln(1 + \kappa f_0) = 2$. The solid line shows the walk length expressions for the greedy walk and the random adaptive walk respectively.

where $a_- < 1, a_+ > 1$ and f_0 is the initial fitness. The analytical results for the difference in the fitness between successive steps was calculated and shown to have a qualitatively different trend in each EVT domain, as shown in Fig. 3. These results are the most important part of my thesis as the calculations and simulations show that the trends in fitness difference hold not only for the fitness difference between successive steps in the uncorrelated

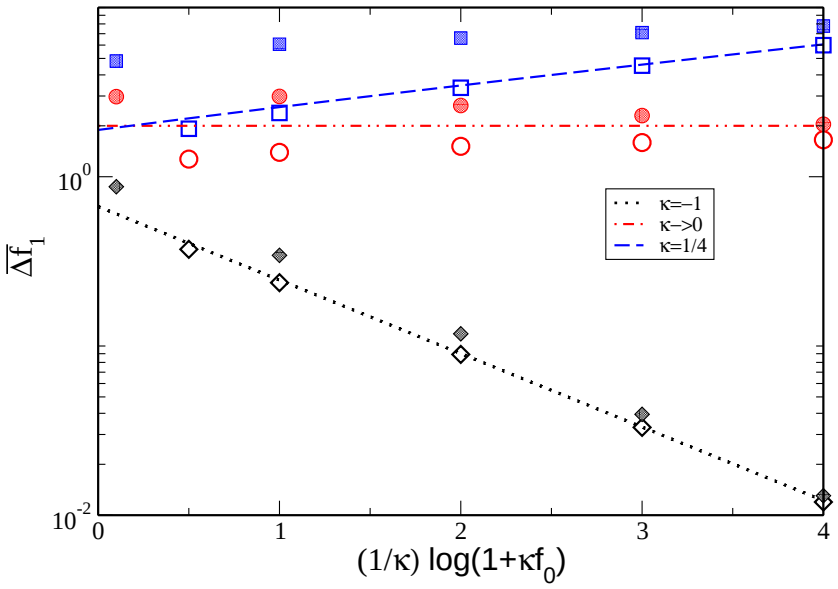


Figure 5: The plot shows the fitness difference at the first step as a function of the initial fitness for different κ and two different $N\mu$. The lines give the theoretical values while the open symbols are the simulation output for $N\mu = 0.02$ and the closed symbols are those for $N\mu = 5$.

fitness landscapes, but also on correlated ones.

In *Chapter 4*, we discuss the average walk length for a population starting from a fixed initial fitness f_0 using the linear model and full model in all the three EVT domains. When (4) is solved analytically using $T(f \leftarrow h)$ from (3), the average number $\bar{J}(L|f_0)$ of steps for a fixed initial fitness f_0 when (1) has a finite mean is obtained as

$$\bar{J}(L|f_0) = \beta_\kappa \left(\ln L - \frac{1}{\kappa} \ln(1 + \kappa f_0) \right) + c_\kappa \tag{7}$$

where $\beta_\kappa = \frac{1-\kappa}{2-\kappa}$ and c_κ is constant. On the other hand, if the fitness distribution given by (1) has an infinite mean, the average walk length in the linear model becomes independent of the initial fitness. The simulation results obtained compared with the analytical results as shown in Fig. 4. In this chapter, we discuss this transition in dependence of the walk length on the initial fitness. We also calculate the distribution of walk length for exponentially and uniformly distributed fitnesses [14]. For the linear model, we also discuss the dependence of the walk length on the number of blocks in the sequence. We show that while this dependence is linear in the Weibull and Gumbel domains, it is logarithmic in the Fréchet domain. We also compare our results obtained in the linear model with the results obtained numerically from the full model on uncorrelated fitness landscapes as shown in the inset of Fig. 4.

In *Chapter 5*, we consider the strong mutation regime so that many mutants are produced at every time step. The population is no longer monomorphic, but may be spread over many sequences. In our model, we keep the population size and the mutation rate fixed. The fitness difference obtained at the first step for various initial fitnesses is shown in Fig. 5. We can see that even if the numerical values for this quantity varies from the theoretical values obtained in the SSWM regime, its qualitative trends in the three EVT domains still stays [26].

Bibliography

[1] J. J. Bull and S. P. Otto. *Nat. Genet.*, 37:342–343, 2005.

[2] J.A.G.M. de Visser and J. Krug. *Nat Rev Genet*, 15:480490, 2014.

[3] H. A. Orr. *Nature Reviews Genetics*, 6:119–127, 2005.

[4] P. Joyce, D. R. Rokyta, C. J. Beisel, and H. A. Orr. sequences under strong selection and weak mutation. *Genetics*, 180:1627–1643, 2008.

[5] J. H. Gillespie. *Theor. Popul. Biol.*, 23:202–215, 1983.

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

[7] M. Imhof and C. Schlotterer. *Proc. Natl. Acad. Sci. USA*, 98:1113–1117, 2001.

[8] A. Eyre-Walker and P.D. Keightley. *Nat. Rev. Genet.*, 8:610, 2007.

[9] P. D. Sniegowski, P. J. Gerrish, T. Johnson, and A. Shaver. populations. *Phil. Trans. R. Soc. B*, 365 1544:1255–1263, 2010.

[10] R. Sanjuán, A. Moya, and S.F. Elena. *Proc. Natl. Acad. Sci. USA*, 101:8396–8401, 2004.

[11] D.R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H.A. Wichman. *J Mol Evol*, 69:229, 2008.

[12] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. *PLoS Genet*, 8:e1002783, 2012.

[13] 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. *PLoS Genet*, 10(2), 2014.

[14] K. Jain and S. Seetharaman. environments. *Genetics*, 189:1029–1043, 2011.

[15] S Seetharaman and K Jain. landscapes. *Phys. Rev. E*, 90:032703, 2014.

[16] S. Seetharaman and K. Jain. *Evolution*, 68:965975, 2014.

[17] A.S. Perelson and C.A. Macken. *Proc. Natl. Acad. Sci. USA*, 92:9657–9661, 1995.

[18] P. J. Gerrish and R. E. Lenksi. *Genetica*, 102:127–144, 1998.

[19] M.M. Desai and D.S. Fisher. positive selection. *Genetics*, 176:1759–1798, 2007.

[20] H.A. Orr. *J. theor. Biol.*, 220:241–247, 2003.

[21] H. Flyvbjerg and B. Lautrup. *Phys. Rev. A*, 46:6714–6723, 1992.

[22] M. Kimura. *Genetics*, 47:713–719, 1962.

[23] J. B. S. Haldane. *Proc. Camb. Philos. Soc.*, 23:838–844, 1927.

[24] K. Jain. *Europhys. lett*, 96:58006, 2011.

[25] S. Seetharaman and K. Jain. *Phys. Rev. E*, 82:031109, 2010.

[26] S. John and S. Seetharaman. *In preparation.*

Publications

- "Multiple adaptive substitutions during evolution in novel environments", Kavita Jain and **Sarada Seetharaman** , Genetics 189:1029–1043, 2011.
- "Adaptive walks and distribution of beneficial fitness effects ", **Sarada Seetharaman** and Kavita Jain, Evolution 68:965–975, 2014.
- "Length of adaptive walk on uncorrelated and correlated fitness landscapes", **Sarada Seetharaman** and Kavita Jain, Phys. Rev. E 90:032703, 2014.
- "Exploiting the adaptation dynamics to predict the distribution of beneficial fitness effects", **Sarada Seetharaman** and Sona John, arXiv:1503.01984.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Evolution refers to the change in heritable phenotypic changes in a biological population over successive generations. In asexual populations, at the genomic level, there is no mixing between different genetic sequences. These populations evolve by replication which refers to an identical copy of the genetic sequence being produced, mutation or an error in the copying and genetic drift which is the stochastic, random fluctuations in the population fraction size of a genetic sequence. By these means, an asexual population can evolve to undergo adaptation, that is, become more successful to suit the environment, or can suffer extinction which means that their population size has become zero or can experience speciation by means of which two subpopulations from the same ancestor diverge to become different species. In this thesis, we focus on the adaptation of asexual populations and in the next section, we will introduce various terms and concepts relevant to this

work.

1.1 Concepts and definitions

1.1.1 Fitness landscape

Fitness is a quantitative measure of how successful an organism is in a given environment - an organism with high fitness has a better chance of propagation within the population than those with lower fitness. Fitness landscape defined as a map from genetic sequences to (genotypic) fitness is a fundamental concept in the theory of biological evolution [1, 2]. It is a L dimensional hypercube, comprising of all possible sequences of length L in a population along with the fitness associated with each sequence. Empirical measurement of fitness landscapes is very hard since the number of sequences increases exponentially with the sequence length L. In Fig. 1.1, we show a schematic example for the fitness landscapes for sequences of small length. However, as the length of the sequence increases, the fitness landscape can only be given a schematic representation. Even to construct a fitness landscape for a microbe with merely hundred nucleotide sequence, one needs to experimentally measure the fitness of $4^{100} \sim 10^{60}$ sequences which is not possible with the current technology. However some empirical insights have been obtained regarding the qualitative nature of the fitness landscapes in the recent years. Fitness have been measured for various microbes for a small part (up

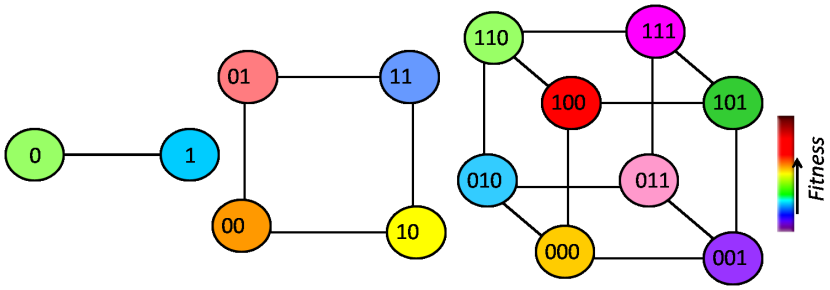


Figure 1.1: The figure shows three examples for the fitness landscapes for sequences of different length.

to ten loci) of the genome which gives information about the local topography of the fitness landscape [3]. Large scale fitness landscapes for about 70,000 HIV sequences have also been constructed [4]. Besides experimentally measuring fitness landscapes directly, the dynamics of adaptation have also been exploited to obtain insights into the structure of the underlying fitness landscape [5–9].

In recent times, it has been possible to track adaptive trajectories for several tens to thousands of generations, especially in microbial populations [9]. Other experiments show that the fitness landscapes can be smooth as evidenced by fast adaptation in some proteins [10] or have multiple peaks as seen in microbial populations that evolve towards different fitness maxima [5,11,12] and enzymes with short uphill paths to the global fitness peak [13]. An example for a smooth end rugged fitness landscape is shown in Fig. 1.2. It has been observed that initially the population evolves quickly and then its fitness increases slowly towards different fitness plateau for the same initial fitness [5,11,12] thus supporting the conclusion that fitness landscapes

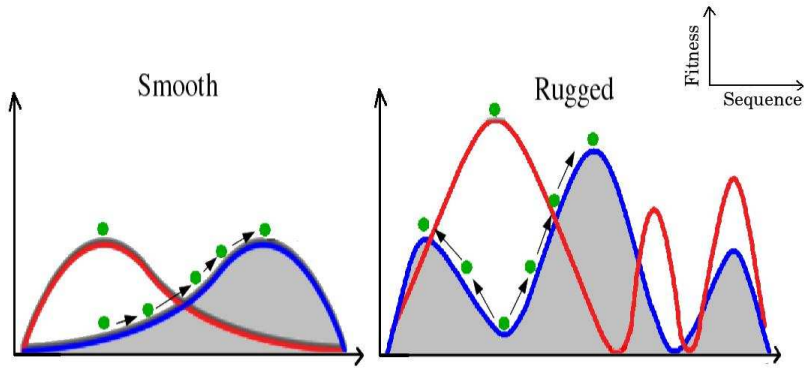


Figure 1.2: The figure shows a schematic example adaptation on fitness landscapes. The red line represents the fitness landscape before the change in environment and the blue one is after the change. The small green circle represent the population that tries to reach a fitness peak in the new environment. The left figure shows an smooth fitness landscape and the right figure shows a rugged one.

are rugged. On such fitness landscapes, while very large populations can reach the global fitness maximum quickly as they produce greater number of mutants, smaller populations stay trapped at a local fitness peak for a long time [8, 14–16]. Detailed studies in which all or a set of mutants from wild type to an optimum are created and their fitness measured [17] have also indicated the smooth [18] and rugged [19,20] nature of the fitness landscapes.

A key result which has emerged from these empirical studies is that the fitness landscapes are quite rugged *i.e.* they are endowed with moderately large number of local fitness peaks which are sequences fitter than their nearest neighbours [21]. A related characteristic of such fitness landscapes is that they are partially correlated [6, 22] which has the effect of reducing the

number of local fitness peaks relative to a fully uncorrelated fitness landscape. The topography of the fitness landscape can be changed by changing the environment. For example, in a *E.coli* population the fitness landscape is expected to be smooth when the carbon source is simple sugar since there is only a single metabolism pathway but becomes rugged with many peaks due to multiple metabolism pathways when the medium is a complex mixture of carbon sources in form of a broth [14].

1.1.2 Mutation

The term refers to the stochastic errors that occur during replication and the process results in the production of a new sequence. The fitness of the mutant sequence may be higher, lower or same as that of the parent sequence, based on which the mutation is termed beneficial, deleterious or neutral respectively. The number of mutations produced in the population, depends on both the population size N and the mutation rate μ and depending on this number, various models have been used to study the population as shall be discussed next [23]. When $N\mu \ll 1$, the population is monomorphic for most times since less than one mutant is produced every generation and we can use the *adaptive walk model* to study the dynamics of the population. On the other hand, if $N\mu > 1$ then the population is polymorphic since a large number of mutants are produced and more than one beneficial mutation may be present in the population. These mutations competing with each other for dominance, is termed *clonal interference* and has been observed

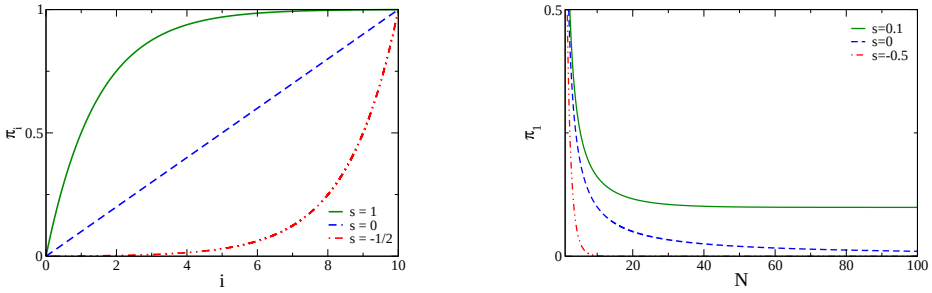
in various experimental populations [24–27]. An extreme case of $N\mu > 1$ is when $N\mu \to \infty$. Here the population becomes a *quasispecies* population which is spread over all the sequences, at all times [23, 28].

1.1.3 Genetic drift

Even beneficial mutations (that produce mutants of high fitness) can get lost from the population due to random sampling termed genetic drift. The effect of this process is strong if the populations size is small. To understand this process better, let us consider the wild type with fitness 1 and the mutant with fitness, $1 + s$. The population size is fixed at N. Here s is the *selection coefficient* giving the relative fitness difference between the mutant and the wild type with respect to the latter such that $s = 0$ corresponds to a neutral mutation while $s > 0$ and $s < 0$ correspond to beneficial and deleterious mutations respectively. Then, the fixation probability π_i that the mutant will sweep through the population at large times starting with an initial number i is given by [29, 30]

$$\pi_i = \frac{1 - (1+s)^{-2i}}{1 - (1+s)^{-2N}} \tag{1.1}$$

From the above equation we can see the stochastic nature of evolution in which, beneficial mutations might get lost if the mutation is rare and deleterious mutation might get fixed if i is large as shown in Fig. 1.3(a). Let us now consider the fate of a rare mutation ($i = 1$) when $N \to \infty$, $s \to 0$ such



(a) (b)

Figure 1.3: (a) The probability of a mutation being fixed in the population when $N = 10$. (b) The probability of a single mutation getting fixed as a function of the population size.

that Ns is finite. The probability of fixation of the mutant can be written as

$$\pi = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \tag{1.2}$$

The probability of fixation of nearly neutral mutations for which $Ns \ll 1$, is $1/N$ as expected. But if $Ns \gg 1$ corresponding to the *strong selection limit*, the probability of fixation of the mutant in the population can be approximated as

$$\pi \approx \begin{cases} 1 - e^{-2s} & \text{if } s > 0 \quad (1.3) \\ 0 & \text{if } s < 0 \quad (1.4) \end{cases}$$

The probability of fixation of a rare mutant as a function of population size N is shown in Fig. 1.3(b).

1.2 Adaptation and distribution of beneficial fitness effects

The increase in the fitness of a population is termed adaptation and, schematically it can be represented as a population climbing the fitness fitness landscape as can be seen in Fig. 1.2. In asexual populations, this process occurs exclusively via beneficial mutations that are not lost due to genetic drift. The problem of adaptive evolution is experimentally challenging because advantageous mutations, which are responsible for adaptation, are rare, accounting for less than 15% of all mutations [31, 32]. However as beneficial mutations contribute substantially to the fate of a population despite their rarity, and play a crucial role in real life scenarios such as the anti-drug resistance developed by microorganisms [33], it is important to know the size and frequency of these mutations. It is important to note that the beneficial mutations occur in the right tail of fitness distributions. The basic idea governing the shape of the distribution of beneficial fitness effects (DBFE) is due to Gillespie [34], who argued that in the event of a small environmental change, as the wild type fitness is expected to remain high, the mutations conferring higher fitness than the wild type will lie in the right tail of the fitness distribution. The statistical properties of such extreme fitnesses are described by an *extreme value theory* (EVT) which states that the extreme value distribution of independent random variables can be of three types: Weibull which

occurs when the fitnesses are right-truncated, Gumbel for distributions decaying faster than a power law and Fréchet for distributions with algebraic tails [35].

Experimentally the distribution of the beneficial mutations (DBFE) has been directly measured, since the size and frequency of the beneficial mutations drive the adaptation of asexual populations. As mentioned above, the distribution of beneficial mutations can belong only to one of the three extreme value distributions and, interestingly, all the three extreme value domains have been observed in recent experiments. Although the exponential distribution for beneficial mutations belonging to the Gumbel domain has been most commonly seen [36–39], the fitness distribution of beneficial mutations belonging to Weibull [40, 41] and Fréchet [42] domains have also been observed. This method directly determines the fitness distribution and, is affected by the rarity of beneficial mutations.

In recent years, the dynamics of adaptation have been studied extensively in experiments [43–45] and various quantities have been tracked. Several quantities such as the fitness rank of the mutant at the first adaptive step [46], the number of adaptive substitutions [40, 47–49] and its dependence on the initial fitness [39,48,49] have been measured, though not with the intention of determining the fitness distribution. However the relation of these properties of adaptation dynamics to the tail of the fitness distribution and hence DBFE is not clear. Clonal interference that is expected in large populations in which $N\mu > 1$ has also been observed in various experimental populations [24–27].

On the theoretical front, the adaptation dynamics have been studied both when $N\mu \ll 1$ and when $N\mu > 1$. In the first case, beneficial mutations are assumed to sweep the population sequentially till the population reaches a local fitness peak and various quantities like the average number of mutations fixed between the starting fitness and a local fitness peak, the average fitness of the mutant fixed have been calculated [37,50–52]. A lot of theoretical work is based on the premise that the relative fitness difference between successive mutations fixed was small [37, 50–57]. However large selection coefficients have been seen in experiments [58, 59] and so far, very few theoretical investigations have taken large effect mutations into account [60, 61]. Also, though fitnesses are known to be correlated, much of the previous work on the subject ignores correlations completely [37, 50–52]. In the second case ($N\mu > 1$), in which more than one mutant is produced every generation, most of the previous studies measured the adaptation rate only when the fitness distribution is exponential [62–66].

1.3 Overview of the thesis

In this thesis, we study the adaptation dynamics of asexual populations on rugged fitness landscapes in all the three EVT domains of DBFE. The motivation of our work is to recognize quantities that show qualitatively different behaviour in each extreme value domain, so that they be used in experiments to determine the DBFE. We use two different models to study this depending

on the number of mutants produced in the population at every generation. When the number of mutants produced in the population is much less than one, we use the adaptive walk model and when it is greater than one, we study the population by using the Wright-Fisher dynamics.

- Adaptive walk model: Here, a small population increases its fitness until a local fitness peak by successively fixing of single beneficial mutations. Every mutation fixed in the population is termed a step in the walk. However, there is a probability that a rare, beneficial mutation might be lost due to genetic drift. So the probability that a mutation gets fixed in the population is proportional to the relative fitness difference between the mutant (f) and the parent (h), as can be discerned from (1.3), as $(1 - e^{-\frac{2(f-h)}{h}})$. Using this probability we calculate the fitness difference between successive mutations in all the three extreme value domains, for both the uncorrelated and correlated fitness landscapes. We find that this fitness difference shows a different trend in each extreme value domain- it increases in the Fréchet domain, is a constant in the Gumbel domain and decrees in the Weibull domain. Moreover, we numerically find that the average length of the adaptive walk depends on the logarithm of the initial fitness rank and the walk is the shortest in the Gumbel domain. We also calculated the average walk length analytically, assuming the relative fitness difference between the mutant and the parent to be small in which case, the dependence of the probability of fixation of a mutation can be considered

to be $\frac{2(f-h)}{h}$ as given in (1.3). In this case, in addition to finding the adaptive walk length dependence on the logarithm of the initial fitness rank, we also find that this is true only when underlying fitness distribution has a finite mean. In other cases, the walk length is a constant independent of the initial fitness. So we see that there is a transition in the behaviour of the walk length depending on the existence of the mean of the fitness distribution. We also calculated the average walk length on correlated fitness landscapes for all the three extreme value distributions.

- Wright Fisher dynamics: When the population size is huge, many mutations are produced at the same time and the associated mutants compete with the other mutants and the wild type for dominance in the population. The subpopulation of every sequence is termed a class. We track the dynamics in this model, for a population of fixed size, using the Wright Fisher dynamics by which each mutant class is allowed to grow stochastically, depending on the population size of the class and its fitness. New mutants are produced in the population at every time step. Using this model we find that irrespective of the number of mutants produced in the population, the fitness difference between successive dominant sequences show a trend identical to what is observed in the adaptive walk model. The fitness difference between successive dominant sequences, not only numerically match the calculated results of the adaptive model, when the number of mutants produced in the

population is much less than one but also, show that even when many mutations are produced every generation, the qualitative trend of the fitness difference increasing in the Fréchet domain, staying a constant in the Gumbel domain and decreasing in the Weibull domain still hold true. We also find that the rate of adaptation shows a strong dependence on the number of mutants produced when the fitness distribution belongs to the Fréchet domain, but is affected weakly otherwise.

1.4 Plan of the thesis

In Chapter 2, we introduce various models that we use in our work. We describe in detail, the block model that is used to model the fitness landscape and to tune the correlations in it. We then move on to the adaptation models. Adaptive walk model is used to address populations in which less than one mutant is produced every generation and any beneficial mutation produced, that manged to escape drift is quickly fixed in the population. So the population is considered to be monomorphic and can be representation by a point particle on the fitness landscape as shown in Fig. 1.2. Every mutation fixed in the population is termed a step in the adaptive walk. The population climbs the fitness landscape by single mutations till a local fitness peak is reached. In this chapter, we have also introduced various quantities that we shall measure and the basic equations that we shall use in later chapters. However, when the number of mutants produced in the population

is high, the equations obtained from the assumptions of the adaptive walk fail and we numerically track the population dynamics using the Wright-Fisher dynamics. In this chapter we explain how every mutant class grows, how we define the leader and the production of mutants.

In Chapter 3, we use the adaptive walk model to measure the fitness difference between successive steps. We obtain analytic expression for this, by assuming the relative fitness difference between successive mutations to be small and arbitrarily large, in all the three extreme value domains. We have presented and described the results on uncorrelated and correlated fitness landscapes. In all these cases, we have backed our calculations using numerical simulations. We find that the fitness difference between successive steps increases, is a constant and decreases, as the walk proceeds in the Fréchet, Gumbel and Weibull domains, respectively. We have also measured the relative fitness difference at every step and we find that unlike fitness difference which shows different trend in each extreme value domain, this quantity decreases as the walk proceeds in all the the three extreme value domains.

In Chapter 4, we present the results for the average walk length of adaptive walks in each extreme value domain. We have carried out extensive calculations for the model that assumes the relative fitness difference to be much less than one. In this case, we find that the average walk length undergoes a transition in its behaviour depending on whether the mean of the

underlying fitness distribution is finite or not. As long as the mean is finite, the walk length has a logarithmic dependence on the rank of the initial fitness. However, as the mean becomes undefined (as can happen in the Fréchet domain) the walk length becomes a constant, independent of the initial fitness. In this chapter, we describe our calculation and results on the correlated fitness landscapes, as well. We have also presented the simulation results for the average walk length from the model that considers the full relative fitness difference and does not make the assumption that it is much less than one. In this case, there is no transition in the behaviour of the average walk length, as it always depends on the initial fitness. We also find that the numerical values obtained here match the analytical calculation results that assume small relative fitness in the Weibull and Gumbel domains, but diverge in the Fréchet domain.

In Chapter 5, we use the Wright-Fisher dynamics to track the dynamics of a finite sized population. Mutations are stochastically produced in the population and their fitness effect can be either beneficial beneficial or deleterious. Though on an average, beneficial mutations are the ones that spread in the population, due to fluctuations deleterious ones may also increase its population fraction. In this chapter we find that the fitness difference between successive dominant sequences, shows similar pattern as what was observed in adaptive walks. Also we find that there is a increase in the rate of adaptation with number of mutants produced in the population, in the Fréchet domain, whereas it is nearly constant in Weibull and Gumbel domains.

Bibliography

[1] S. Gavrilets. *Fitness Landscapes and the Origin of Species*. Princeton University Press, New Jersey, 2004.

[2] J.A.G.M. de Visser and J. Krug. *Nat Rev Genet*, 15:480490, 2014.

[3] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. M. de Visser. *J. Stat. Mech.*, -:P01005, 2013.

[4] T. Hinkley, J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. Whitcomb, C. Petropoulos, and S. Bonhoeffer. *Nat Genet*, 43:487–489, 2011.

[5] C. L. Burch and L. Chao. *Nature*, 406:625–628, 2000.

[6] C. R. Miller, P. Joyce, and H.A. Wichman. *Genetics*, 187:185–202, 2011.

[7] C. L. Burch and L. Chao. *Genetics*, 151:921–927, 1999.

[8] K. Jain, J. Krug, and S.-C. Park. *Evolution*, 65:1945, 2011.

[9] J. E. Barrick and R. E. Lenski. *Nat. Rev. Genet*, 14:827, 2013.

[10] P.A. Romero and F.H. Arnold. *Nat. Rev. Mol. Cell Biol.*, 10:866–876, 2009.

[11] R. Korona, C. H. Nakatsu, L. J. Forney, and R. E. Lenski. *PNAS*, 91:9037–9041, 1994.

[12] G. Fernandez, B. Clotet, and M.A. Martinez. *J. Virol.*, 81:2485–2496, 2007.

[13] A. S. Perelson and C. A. Macken. *PNAS*, 92:9657–9661, 1995.

[14] D. E. Rozen, M.G.J.L. Habets, A. Handel, and J. A. G. M. de Visser. *PLoS ONE*, 3 (3):e1715, 2008.

[15] Y. Iwasa, F. Michor, and M. A. Nowak. *Genetics*, 166:1571–1579, 2004.

[16] D. M. Weinreich, R. A. Watson, and L. Chao. *Evolution*, 59:1165–1174, 2005.

[17] F.J. Poelwijk, D.J. Kivet, D.M. Weinreich, and S.J. Tans. *Nature*, 445:383, 2007.

[18] M. Lunzer, S. P. Miller, R. Felsheim, and A. M. Dean. *Science*, 310:499 – 501, 2005.

[19] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl. *Science*, 312:111–114, 2006.

[20] J.A.G.M. de Visser, S.-C. Park, and J. Krug. *Am. Nat.*, 174:S15–S30, 2009.

[21] J. Neidhart, G. S. Szendro, and J. Krug. *Genetics*, 198:699–721, 2014.

[22] C. Carneiro and D.L. Hartl. *PNAS*, 107:1747–1751, 2010.

[23] K. Jain and S. Seetharaman. *J. Nonlin. Math. Phys.*, 18:321–338, 2011.

[24] J.A.G.M. de Visser and D. E. Rozen. *Genetics*, 172:2093–2100, 2006.

[25] J.A.G.M. de Visser, C.W. Zeyl, P.J. Gerrish, J.L. Blanchard, and R.E. Lenski. *Science*, 283:404–406, 1999.

[26] R. Miralles, P. J. Gerrish, A. Moya, and S.F. Elena. *Science*, 285:813–815, 1999.

[27] D.E. Rozen, J.A.G.M. de Visser, and P. J. Gerrish. *Curr Biol.*, 12:1040–1045, 2002.

[28] S. Seetharaman and K. Jain. *Phys. Rev. E*, 82:031109, 2010.

[29] R. A. Fisher. *The genetical theory of natural selection*. Oxford: Clarendon Press, 1930.

[30] M. Kimura. *Genetics*, 47:713–719, 1962.

[31] A. Eyre-Walker and P.D. Keightley. *Nat. Rev. Genet.*, 8:610, 2007.

[32] P. D. Sniegowski and P. J. Gerrish. *Phil. Trans. R. Soc. B*, 365:1255–1263, 2010.

[33] J. J. Bull and S. P. Otto. *Nat. Genet.*, 37:342–343, 2005.

[34] J. H. Gillespie. *Theor. Popul. Biol.*, 23:202–215, 1983.

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

[36] R. Sanjuán, A. Moya, and S.F. Elena. *PNAS*, 101:8396–8401, 2004.

[37] D. R. Rokyta, C. J. Beisel, and P. Joyce. *J Theor Biol.*, 243:114–120, 2006.

[38] R. Kassen and T. Bataillon. *Nat. Genet.*, 38:484–488, 2006.

[39] R. C. MacLean, G. G. Perron, and A. Gardner. *Genetics*, 186:1345–1354, 2010.

[40] D. R. Rokyta, Z. Abdo, and H. A. Wichman. *J Mol Evol*, 69:229, 2009.

[41] T. Bataillon, T. Zhang, and R. Kassen. *Genetics*, 189:939–949, 2011.

[42] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. *PLoS Genet.*, 8:e1002783, 2012.

[43] S. F. Elena and R. E. Lenski. *Nat. Rev. Genet.*, 4:457–469, 2003.

[44] 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. *PLoS Genet*, 10(2), 2014.

[45] C. Bank, T. H. Ryan, D. J. Jeffrey, and N.A.B. Daniel. *Mol. Biol. Evol.*, 2014.

[46] D. R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. *Nat. Genet.*, 37:441–444, 2005.

[47] S.E. Schoustra, T. Bataillon, D.R. Gifford, and R. Kassen. *PLoS Biol*, 7 (11):e1000250, 2009.

[48] D. R. Gifford, S. E. Schoustra, and R. Kassen. *Evolution*, 65:3070–3078, 2011.

[49] A. Sousa, S. Magalhães, and I. Gordo. *Mol. Biol. Evol.*, 29:1417–1428, 2012.

[50] P. Joyce, D. R. Rokyta, C. J. Beisel, and H. A. Orr. *Genetics*, 180:1627–1643, 2008.

[51] J. Neidhart and J. Krug. *Phys. Rev. Lett.*, 107:178102, 2011.

[52] K. Jain. *EPL*, 96:58006, 2011.

[53] H. A. Orr. *Evolution*, 56:1317–1330, 2002.

[54] H. A. Orr. *Evolution*, 60:1113, 2006.

[55] S. Kryazhimskiy, G. Tkačik, and J. B. Plotkin. *PNAS*, 106:18638–18643, 2009.

[56] K. Jain and S. Seetharaman. *Genetics*, 189:1029–1043, 2011.

[57] J. A. L. Filho, F. G. B. Moreira, P. R. A. Campos, and V. M. Oliveira. *J. Stat. Mech.*, -:P02014, 2012.

[58] J. J. Bull, M. R. Badgett, and H. A. Wichman. *Mol. Biol. Evol.*, 17:942–950, 2000.

[59] R. D. H. Barrett, R. Craig MacLean, and G. Bell. *Biol. Lett.*, 2:236–238, 2006.

[60] J. M. Heffernan and L. M. Wahl. *Theo. Pop. Biol.*, 62:349356, 2002.

[61] R.D.H. Barrett, L.K. M'Gonigle, and S.P. Otto. *Genetics*, 174:2071–2079, 2006.

[62] P. J. Gerrish and R. E. Lenski. *Genetica*, 102:127–144, 1998.

[63] S.-C. Park and J. Krug. *PNAS*, 104:18135–18140, 2007.

[64] M.M. Desai and D.S. Fisher. *Genetics*, 176:1759–1798, 2007.

[65] S.-C. Park, D. Simon, and J. Krug. *J. Stat. Phys.*, 138:381–410, 2010.

[66] P.R.A. Campos and L. M. Wahl. *Evolution*, 64(7):1973–1983, 2010.

Chapter 2

Models

2.1 Introduction

Adaptation is associated with a population climbing the fitness landscape as schematically shown in Fig. 1.2. In asexual populations, this process happens solely by means of beneficial mutations. The problem of adaptive evolution is challenging because advantageous mutations, which are responsible for adaptation, are rare [1]. It has been observed that initially the population evolves quickly and then its fitness increases slowly towards different fitness plateau for the same initial fitness [2–4] thus supporting the conclusion that fitness landscapes are rugged. Depending on the number of mutants produced in the population at every time step, we use two different models to study its adaptation dynamics. In this chapter, we shall introduce the concept of fitness landscape and explain how its ruggedness can be tuned. We also

explain the two adaptation models we use in this thesis.

2.2 Block model

All possible sequences of length L in a population, along with their associated fitness comprise the $L+1$ dimensional fitness landscape, whose L dimensions represent the sequence space along with one more dimension to indicate the fitness of each sequence. The fitness landscape can be defined on a hypercube to represent the sequence space, whose each vertex corresponds to a binary sequence. Examples for this is shown in Fig. 1.1 for $L \leq 3$ and in Fig. 2.1 for $L = 4$. We study adaptation on rugged fitness landscapes that are characterized by many local fitness maxima using a *block model* [5] in which a sequence of length L is split into B blocks of equal length $L_B = L/B$. The partitioning of a sequence is motivated by the domain structure of proteins [6] and paired-unpaired regions in RNA secondary structure [7]. In proteins, the domains that perform essential enzymatic functions are more likely to be stable and in RNA secondary structure, the paired regions may have a lower free energy than the unpaired ones.

The 2^{L_B} fitnesses of each of the B blocks is chosen from the fitness distribution, $p(f)$ and if interactions between the blocks are neglected [6], the fitness of the whole sequence can be written as the average of the block fitnesses [5]. Fitness correlations arise because of common blocks between two sequences and can be changed by tuning the number of blocks in the

sequence. The two limits namely $B = 1$ and $B = L$ produce fully uncorrelated and fully correlated fitness landscapes respectively. A measure of the ruggedness of a fitness landscape is the number of local fitness peaks (defined as sequences fitter than all of their one mutant neighbours) which decreases as the fitness correlations increase [5]. On correlated fitness landscapes, a local fitness peak is reached when the fitness of each block is a local fitness maximum. As the probability of one of the 2^{L_B} sequences being a local fitness peak is $2^{L_B}/(L_B+1)$, the average number of local peaks on a correlated fitness landscape is given by $\left(\frac{2^{L_B}}{L_B+1}\right)^B$ [5]. Thus on a fully correlated fitness landscape, there is only one local (same as global) fitness peak, whereas on fully uncorrelated fitness landscapes, there are on an average $2^L/(L+1)$ local fitness maxima. In general, different blocks in a sequence may have different fitness and a random variable chosen from a fitness distribution $p(f)$ may be assigned to each block.

A result from extreme value theory that we will need for subsequent discussion, states that the typical value f of the mth best fitness amongst L independent fitnesses can be determined by equating the rank m to the average number of fitnesses higher than f. This is given by [8]

$$L \int_f^u dg\ p(g) = m \tag{2.1}$$

where u is the upper limit of the fitness distribution. Setting $m = 1$ in the above equation, we get the largest value of L random variables, or in other

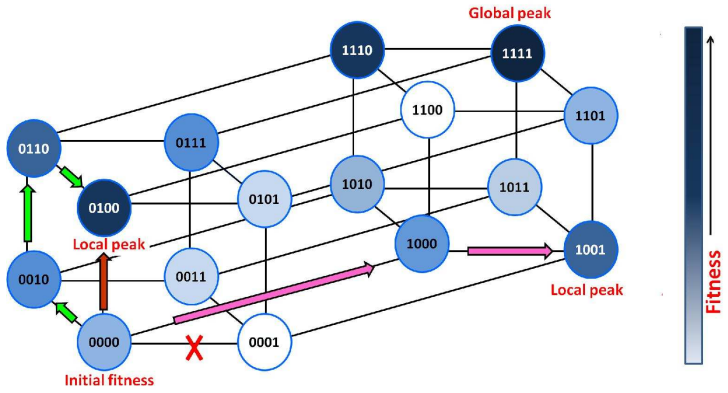


Figure 2.1: Schematic representation of adaptive walks in a 4-dimensional sequence space, starting from the same initial sequence. The arrows represent the shift of the population from a sequence to a fitter sequence one mutation away (refer text for details).

words, the typical fitness $\tilde{f}_B$ of a local fitness peak. The typical local peak fitness $\tilde{f}_B$ of a sequence of length L is the average of B random variables, each of which is the best of L_B random variables. We obtain the expression for this from the equation

$$L_B \int_{\tilde{f}_B}^{u} df \; p(f) = 1 \tag{2.2}$$

On uncorrelated fitness landscapes where $B = 1$, we drop the subscript B and refer to the average fitness of a local fitness peak as just $\tilde{f}$.

Fitness distributions: Each sequence is assigned a fitness which is an independent and identically distributed (i.i.d.) random variable chosen from a probability distribution. Experiments indicate that deleterious and neutral mutations account for most of the weight in the fitness distribution, but a

significant fraction comes from the beneficial mutations as well [1]. Since the adaptation process is governed by these rare beneficial mutations, we need to consider the upper tail of the fitness distribution [9] which immediately suggests the use of the extreme value theory and the related peak-over-thresholds formulation described below [8, 10].

Consider the conditional cumulative distribution $P_{f_T}(f)$ for the fitness f chosen from the distribution $\hat{p}(f)$ above a large threshold f_T which here refers to the wild type fitness. Formally, we have

$$\begin{aligned} P_{f_T}(f) &= \text{Prob}(F - f_T < f | F > f_T) && (2.3) \\ &= 1 - \frac{\hat{q}(f + f_T)}{\hat{q}(f_T)} && (2.4) \end{aligned}$$

where $\hat{q}(f) = \int_f dg\ \hat{p}(g)$. For large enough thresholds, the above cumulative distribution approaches the Generalised Pareto Distribution (GPD) [8]:

$$P_{f_T}(f) \overset{\text{large } f_T}{\longrightarrow} P(f, \tau) = 1 - \left[1 + \frac{\kappa f}{\tau}\right]^{-1/\kappa}, \quad -\infty < \kappa < \infty \qquad (2.5)$$

where τ is a scale factor and the shape parameter κ can take any real value. The limiting distribution with positive κ corresponds to a power law distribution, and is obtained when $\hat{p}(f)$ itself decays algebraically. When $\kappa < 0$, the fitness distribution (2.5) makes sense when $f < -\tau/\kappa$ and therefore such a distribution is bounded above. This class of distributions appears when $\hat{p}(f)$

is truncated. Finally, the limit $\kappa \to 0$ gives an exponentially decaying function which is obtained from unbounded distributions decaying faster than a power law. For example, for the fitness distribution $\hat{p}(f) = cf^{c-1}\, e^{-f^c}, c > 0$, the conditional distribution works out to be

$$P_{f_T}(f) = 1 - \frac{e^{-(f+f_T)^c}}{e^{-f_T^c}} \tag{2.6}$$

$$\approx 1 - e^{-cf_T^{c-1}f}\ ,\ f_T \gg 1 \tag{2.7}$$

Thus the tail of the conditional distribution is an exponential, and the threshold fitness f_T and the exponent c characterizing the tail of the full distribution $\hat{p}(f)$ appear in the scale factor τ. In summary, the distribution $p(f,\tau) = dP(f,\tau)/df$ of *beneficial* mutations for i.i.d. fitnesses is a GPD, or in the language of the extreme value theory (EVT), the distribution of the beneficial fitness effects (DBFE) $p(f,\tau)$ can be of only three types viz., Weibull ($\kappa < 0$), Gumbel ($\kappa \to 0$) and Fréchet ($\kappa > 0$) [8]. Experimentally, DBFE belonging to all the EVT domains have been observed [11–20]. We will set $\tau = 1$ in the rest of this thesis and denote the fitness distribution by

$$p(f) = (1 + \kappa f)^{-\frac{1+\kappa}{\kappa}} \tag{2.8}$$

As we are interested in adaptive changes, an uncorrelated fitness landscape is generated by choosing fitnesses independently from $p(f)$. Fig. 2.2 shows the fitness distribution for various κ. The average fitness of a local peak as

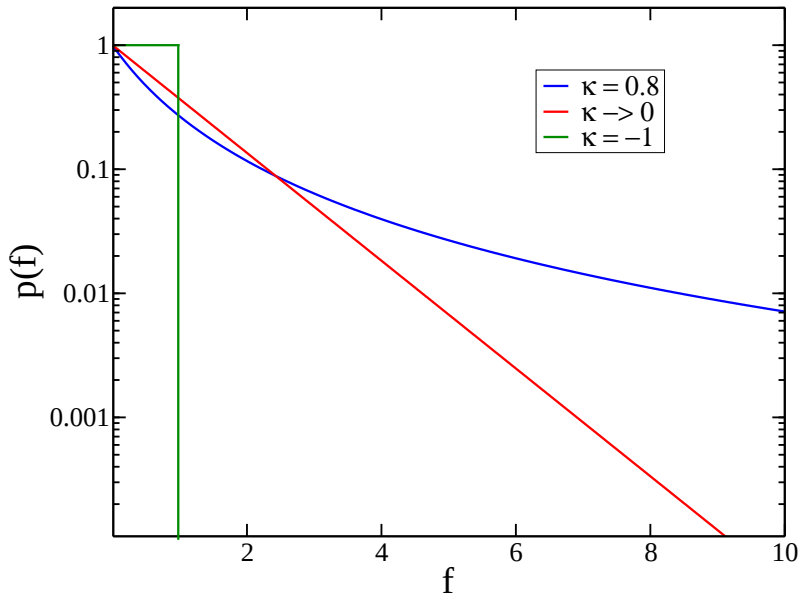


Figure 2.2: The figure shows the fitness distribution for various κ.

obtained from (2.2) then immediately yields

$$\tilde{f} = \frac{L^{\kappa} - 1}{\kappa} \tag{2.9}$$

Similarly, on correlated fitness landscapes the average fitness of a local peak is given by

$$\tilde{f}_B = \frac{L_B^{\kappa} - 1}{\kappa} \tag{2.10}$$

2.3 Adaptive walk model

This model is used in the strong selection-weak mutation (SSWM) regime in which the number of mutants produced per generation is much less than

one. In this case, for a population of size N with a mutation rate mu, although a single-mutant with selection coefficient s arises on an average every $(N\mu)^{-1}$ generations, the waiting time to its fixation is $(N\mu\pi(s))^{-1}$ generations, where the $\pi(s)$ is the fixation probability introduced in Chapter 1. If the initial fitness is small, several beneficial alleles each with a different selective effect are possible, and Gillespie showed the probability that the one with selection coefficient s will sweep through the population is proportional to $\pi(s)$ [9]. In the adaptive walk model, the mutation rates are small enough so that only those sequences that are one mutation away from the currently occupied sequence can be accessed. Once a beneficial mutant is fixed in the population, the new wild type produces a *novel* neighborhood of mutants that are single mutation away from it. Again one of the beneficial mutants sweeps through the population and replaces the current wild type. This substitution process goes on until the population encounters a local fitness peak, as double and higher order mutants are ignored.

Here we consider an asexual population initially localized at a sequence with fitness f_0 in which only the beneficial mutations spread through the population while the neutral and deleterious ones are quickly lost [9]. As illustrated in Fig. 2.1 for sequence space of dimension 4, starting from the sequence $\{0000\}$, at the first step in the walk, the population has three fitter neighbors viz. $\{0010\}, \{0100\}$ and $\{1000\}$, and it chooses one of them according to a stochastic rule described below. After the first step is taken,

the population again scans its nearest neighbors and walks to a fitter neighbor. This process repeats until a local fitness peak is reached whereupon the adaptive walk terminates since the next beneficial mutation is at least two mutations away which is not accessible in the weak mutation regime. The number of steps taken from the initial sequence to a local fitness peak is termed as the *walk length*. In Fig. 2.1, two walks to the local fitness peak $\{0100\}$ with length one and three are shown. Of course, an adaptive walk to a different local fitness peak (say, sequence $\{1001\}$) is also possible.

We now discuss the stochastic rules by which a nearest fitter sequence may be chosen [21]. Perhaps the simplest algorithm is the *greedy adaptive walk* (GAW) in which the fittest mutant is chosen at any step in the walk. In contrast, in the *random adaptive walk* (RAW), any fitter one-mutant is equally likely to be chosen. Here we are interested in the biologically relevant situation where, as one would intuitively expect, a mutant which is much fitter than the wild type has a higher chance of sweeping through the population than a mutant which is mildly fitter. From the population genetics theory [22], it is known that in a large adapting asexual population, if h is the fitness of the wild type and $f > h$ is the fitness of the mutant, the probability that the mutant will take over the population is given by (1.3) as

$$\pi(f, h) = 1 - \exp\left[-\frac{2(f-h)}{h}\right], \tag{2.11}$$

since the selection coefficient is given by $s = \frac{f-h}{h}$. Thus, as in Fig. 2.1, when

several of the L nearest mutants are beneficial, the population moves to one of them with a probability proportional to π. The normalised transition probability is then given by [23–25]

$$T(f \leftarrow h) = \frac{1 - e^{-\frac{2(f-h)}{h}}}{\sum_{g>h} 1 - e^{-\frac{2(g-h)}{h}}} \tag{2.12}$$

$$\simeq \frac{p(f)\ (1 - e^{-\frac{2(f-h)}{h}})}{\int_h^u dg\ p(g)\ (1 - e^{-\frac{2(f-h)}{h}})}\ \ (full\ model) \tag{2.13}$$

For all κ, (2.13) is applicable when L is large and we can see that the equation is clearly nonlinear in the fitnesses. However our previous work [25] shows that when $\kappa \leq 0$, the relative fitness difference $s = (f - h)/h$ between the mutations encountered is small, and we may therefore write $\pi(f, h) \approx 2s$ [9, 23, 24, 26] which gives us

$$T(f \leftarrow h) = \frac{f - h}{\sum_{g>h} g - h}\ \ (linear\ model) \tag{2.14}$$

For the above equation, we can use an integral approximation similar to (2.13) only when $\kappa < 1$, since the mean of $p(f)$ diverges beyond this range (see Section 3.2.1). In this thesis, we shall refer to the model that uses (2.13) as the *full model* and the one that uses (2.14) as the linear model.

Since in many experiments the population is founded using a single ancestor thus keeping the initial fitness fixed [27, 28], we consider the adaptation process starting from a fitness f_0. On correlated fitness landscapes, if a sequence is divided into B blocks and the initial fitness of the bth block is $f_0^{(b)}$,

the initial fitness of the whole sequence is given by

$$f_0 = \frac{1}{B}\sum_{b=1}^{B} f_0^{(b)} \tag{2.15}$$

On uncorrelated fitness landscapes, the probability distribution $\mathcal{P}_J(f|f_0)$ that the population has fitness f at the Jth step of the adaptive walk given that it started with fitness f_0 obeys the following recursion equation [24]

$$\mathcal{P}_{J+1}(f|f_0) = \int_{f_0}^{f} dh\ T(f \leftarrow h)\ (1 - q^L(h))\ \mathcal{P}_J(h|f_0)\ ,\ J \geq 0 \tag{2.16}$$

where $q(f) = \int_0^f dg\ p(g) = 1 - (1 + \kappa f)^{-1/\kappa}$ gives the probability of having a fitness less than f and $T(f \leftarrow h)$ is given by (2.13) or by (2.14). Equation (2.16) simply means that the population moves from fitness h to a higher fitness f at the next step with probability (2.13) or (2.14) provided at least one fitter mutant is available, the probability of whose is given by $1 - q^L(h)$. Equation (2.16) can be used to write a second order differential equation in f for the distribution $P_J(f|f_0)$ defined through $\mathcal{P}_J(f|f_0) = p(f)P_J(f|f_0)$ which is given by [24]

$$P''_{J+1}(f|f_0) = \frac{p(f)(1 - q^L(f))}{\int_f^u dg\ (g - f)\ p(g)} P_J(f|f_0)\ ,\ J \geq 1 \tag{2.17}$$

where the prime refers to a derivative with respect to (w.r.t.) f. For monomorphic initial condition with fixed fitness f_0, we have the boundary

conditions

$$\mathcal{P}_J(f|f_0) = \delta(f - f_0)\delta_{J,0} \tag{2.18}$$

$$\mathcal{P}'_1(f_0|f_0) = \frac{p(f_0)(1 - q^L(f_0))}{\int_{f_0}^{u} dg\ (g - f_0)\ p(g)} \tag{2.19}$$

Equation (2.18) is self explanatory and (2.19) is obtained by applying (2.18) on the first derivative of (2.16) w.r.t. f [24].

The main quantities that we are interested in is the average fitness at each step of the adaptive walk, for sequence that are infinitely long, which can be obtained from the distribution $\mathcal{P}_J(f|f_0)$ as

$$\bar{f}_J(f_0) = \int_{f_0}^{u} df\ f\ \mathcal{P}_J(f|f_0) \tag{2.20}$$

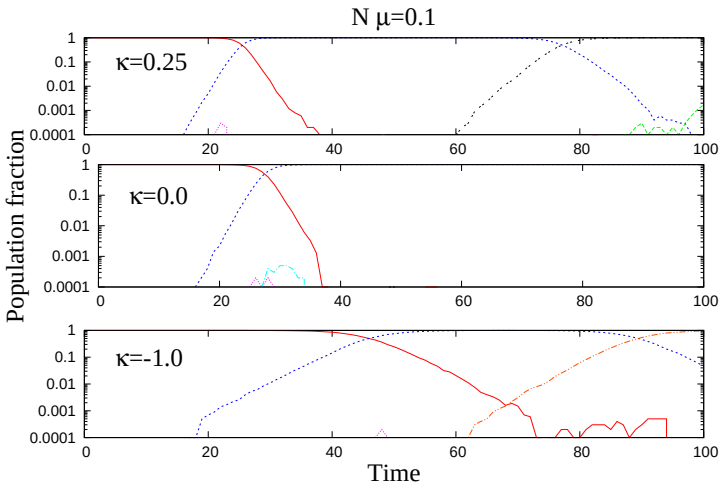
A related quantity is the selection coefficient which at step J is given by

$$s_J = \frac{f_J - f_{J-1}}{f_{J-1}}\ ,\ J > 0 \tag{2.21}$$

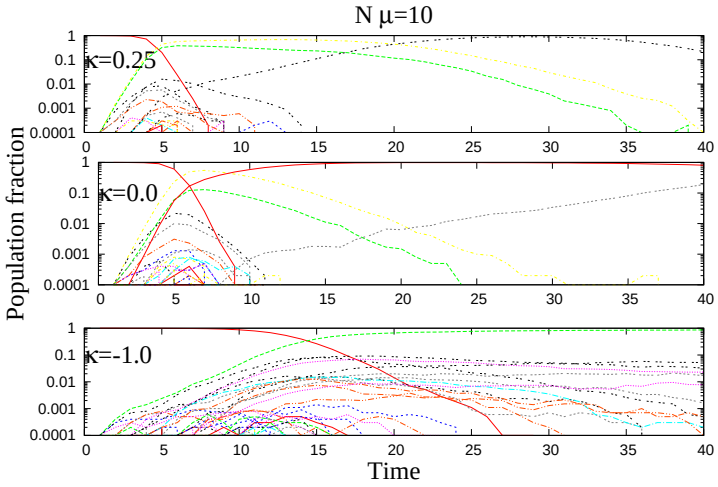
In this thesis, we use the above formalism to obtain an expression for the fitness of a step in the adaptive walk for both the linear model and the full model as explained in Chapter 3. However, we could obtain the analytical expression for the walk length only for the linear model and have only simulation results for the full model as presented in Chapter 4. The linear model is interesting to study, not only because it is amenable to analysis, but also because the results obtained here appear in other systems [29] viz.

models of deterministically evolving populations [30–32] and the Jepsen gas that describes a system of particles with random velocities undergoing elastic collisions [33, 34].

In computer simulations of the dynamics of the adaptive walk, we started with a sequence of length L and initial fitness f_0, and considered uncorrelated ($B = 1$) and weakly correlated fitnesses with $B > 1$. In the former case, the initial fitness of the sequence is fixed and in the latter case, B random variables are generated independently from (2.8) and they are accepted as block fitnesses if their sum is $Bf_0 \pm \delta$ where $\delta \sim 0.01 f_0$. At each step of the adaptive walk, we generate L new fitnesses that are chosen from (2.8) and one of them is chosen to be fixed according to the transition probability (2.13) or 2.14. While in the case of uncorrelated fitnesses, the fitness of the whole sequence changes at each step, when $B > 1$ the fitness of only one of the blocks is changed. The fitnesses sampled during the walk are not stored as for large L, the number of one mutant neighbors probed in previous steps can be ignored in comparison to L [23, 36, 37]. In our simulations, the fitness and selection coefficient of each step are averaged over only those walks that proceed until that step. In all the simulations on uncorrelated fitness landscapes, the data were averaged over 10^6 independent realizations of the fitness landscape and 10^5 for the correlated ones.



(a)



(b)

Figure 2.3: Population fraction of different classes in SSWM ($N\mu = 0.1$) and clonal interference ($N\mu = 10$) regimes for all three DBFE domains.

2.4 Wright Fisher model

For populations in which a large number of mutants are produced at every generation, the genetic variation of the population is also high and more

than one beneficial mutation is expected to be present at the same time [38–41]. In this case, the beneficial mutations will compete with each other as has been observed in different experimental populations [42–45]. In this clonal interference regime, because of the competition among the beneficial mutations, the rate of adaptation slows down and also 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 [46]. A clear comparison between the two mutation regimes for different DBFE is shown in Fig. 2.3. In Fig. 2.3(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, whereas the population is polymorphic when more than one mutant is produced in it at every generation as shown in Fig. 2.3(b). Moreover, we notice that the maximum amount of genetic variation is observed in the case of bounded distributions corresponding to $\kappa = -1$ resulting in strong clonal interference effects.

We track the dynamics of a population of self-replicating, infinitely long binary sequences of fixed size using the standard Wright-Fisher process [40, 46]. In our work, the population size is held constant at $N = 10^4$, unless specified otherwise and the mutation probability per sequence is given by μ. Every occupied sequence, is counted as a *class* and labeled when it arises in the population. Initially, the whole population is in class 1, which is the initial *leader* and its fitness is fixed and specified. At every time step, out of N sequences, m_t are chosen from a binomial distribution with mean $N\mu$

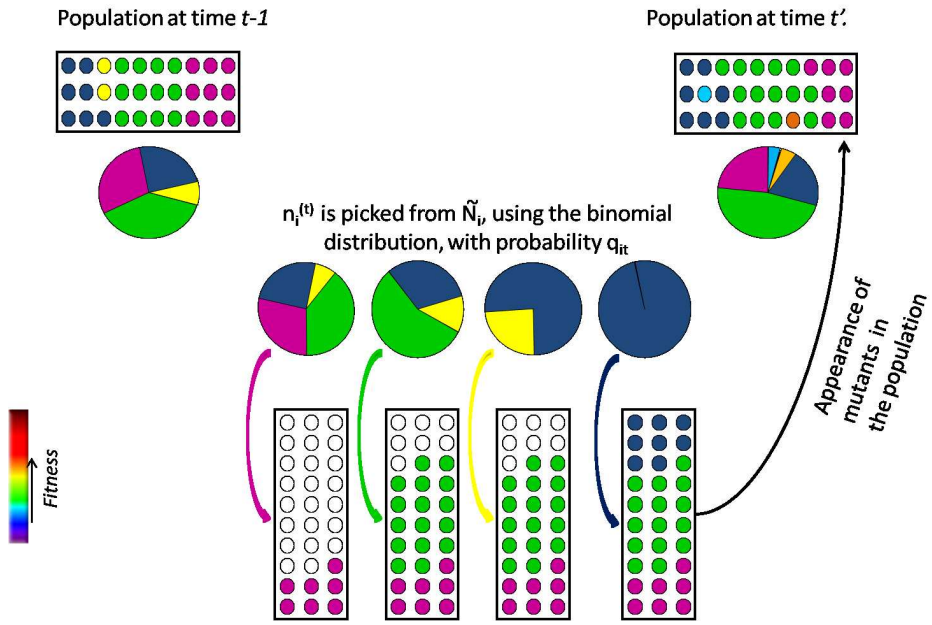


Figure 2.4: Schematic representation of the algorithm used in the strong mutation regime.

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. A schematic representation of the algorithm we have used is shown in Fig. 2.4.

At any time t, the number of classes present in the population is given by $\mathcal{N}_c^{(t)}$, and the population size and fitness of each class, i, where $1 \leq i \leq \mathcal{N}_c$, is denoted by $n_i^{(t)}$ and f_i, respectively. The normalized probability of each class at every time step, $p_i^{(t)}$ 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

$$p_i^{(t)} = \frac{n_i^{(t)}\ f_i}{\Sigma_{j=1}^{\mathcal{N}_c^{(t)}} n_j^{(t)}\ f_j} \tag{2.22}$$

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 (2.22). The different classes populate the next time step based on the multinomial distribution

$$P(n_i^{(t')}, n_2^{(t')}..n_{\mathcal{N}_c}^{(t')}) = N! \prod_{j=1}^{\mathcal{N}_c^{(t')}} \frac{\left(p_j^{(t)}\right)^{n_j^{(t')}}}{n_j^{(t')}!} \tag{2.23}$$

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 the above algorithm along with the constraint by converting (2.23) to a binomial distribution for every class, $1 \leq i < \mathcal{N}_c^{(t)}$ as $n_i^{(t')} = \binom{\tilde{N}_i}{n_i^{(t')}} q_{it}^{n_i^{(t')}} (1-q_{it})^{\tilde{N}_i - n_i^{(t')}}$ and by setting the population of the last class as $n_{\mathcal{N}_c^{(t)}}^{(t')} = N - \sum_{i=1}^{\mathcal{N}_c^{(t)}-1}\ n_i^{(t')}$. In the previous equation, $q_{it} = \frac{p_i^{(t)}}{\Sigma_{j=i}^{\mathcal{N}_c} p_j^{(t)}}$ 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 vary μ to access both the SSWM (low mutation) and the clonal interference (high mutation) regime. In our simulations unless

specified otherwise, in the low and high mutation regimes, $N\mu = 0.01$ and $N\mu = 50$, respectively.

A new class is assigned to each mutant and its fitness is chosen from a generalized Pareto distribution (2.8) [8], in order to access all the three extreme value domains. In this work, the fitness of the mutants is independently chosen from (2.8) thus making the fitness of the mutant, f_m an uncorrelated variable, which may be greater or smaller than the parent fitness, f_p, and we analyze 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 [40, 46] 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 $f_m = (1 + s)f_p$. 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 decrease thus producing different patterns for the fitness increment in each EVT domain.

Whenever a class goes extinct, the classes below it are moved up, and the number of classes in the population is reduced by one. The normalized probability, (2.22) of any classes exceeding half corresponds to a leader change and the new leader determined now, belongs to the class whose normalized probability exceeded half. Every change of leader is counted as a *step*. While

in the clonal interference 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, 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. 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.

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 adaptation are averaged over all walks in that simulation run. We shall refer to the number of classes present in the population at any time as $\mathcal{N}_c$.

In the adaptive walk model, since only mutations one mutation away are successively fixed in the population, the model is applicable only till a local fitness peak whose average fitness is given by (2.9). However in the Wright-Fisher model the population can produce multiple mutations in a step and the maximum fitness that can be encountered in this model after t_{max} time steps is obtained from [8]

$$t_{max} N\mu \int_{f_{max}}^{u} p(f) = 1 \tag{2.24}$$

from which we get

$$f_{max} = \frac{(t_{max} N\mu)^{\kappa} - 1}{\kappa} \tag{2.25}$$

Here, we set the total number of iterations is 10^5 in every simulation run and the dynamics is tracked for $t_{max} = 10^4$ time steps.

Bibliography

[1] A. Eyre-Walker and P.D. Keightley. *Nat. Rev. Genet.*, 8:610, 2007.

[2] R. Korona, C. H. Nakatsu, L. J. Forney, and R. E. Lenski. *PNAS*, 91:9037–9041, 1994.

[3] C. L. Burch and L. Chao. *Nature*, 406:625–628, 2000.

[4] G. Fernandez, B. Clotet, and M.A. Martinez. *J. Virol.*, 81:2485–2496, 2007.

[5] A. S. Perelson and C. A. Macken. *PNAS*, 92:9657–9661, 1995.

[6] C. P. Ponting and R. R. Russell. *Annu. Rev. Biophys. Biomol. Struct.*, 31:45–71, 2002.

[7] R. T. Batey, R. P. Rambo, and J. A. Doudna. *Angew. Chem. Int. Ed.*, 38:2326, 1999.

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

[9] J. H. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, Oxford, 1991.

[10] P. Joyce, D. R. Rokyta, C. J. Beisel, and H. A. Orr. *Genetics*, 180:1627–1643, 2008.

[11] R. Sanjuán, A. Moya, and S.F. Elena. *PNAS*, 101:8396–8401, 2004.

[12] D. R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. *Nat. Genet.*, 37:441–444, 2005.

[13] R. Kassen and T. Bataillon. *Nat. Genet.*, 38:484–488, 2006.

[14] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman. *J Mol Evol*, 69:229, 2008.

[15] R. C. MacLean and A. Buckling. *PLoS Genetics*, 5:e1000406, 2009.

[16] T. Bataillon, T. Zhang, and R. Kassen. *Genetics*, 189:939–949, 2011.

[17] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. *PLoS Genet.*, 8:e1002783, 2012.

[18] 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. *PLoS Genet*, 10(2), 2014.

[19] C. Bank, T. H. Ryan, D. J. Jeffrey, and N.A.B. Daniel. *Mol. Biol. Evol.*, 2014.

[20] D. R. Rokyta, Z. Abdo, and H. A. Wichman. *J Mol Evol*, 69:229, 2009.

[21] H. A. Orr. *J. theor. Biol.*, 220:241–247, 2003.

[22] B. Charlesworth and D. Charlesworth. *Elements of evolutionary genetics*. Roberts and Company Publishers, 2010.

[23] H. A. Orr. *Evolution*, 56:1317–1330, 2002.

[24] K. Jain and S. Seetharaman. *Genetics*, 189:1029–1043, 2011.

[25] S. Seetharaman and K. Jain. *Evolution*, 68:965–975, 2014.

[26] J. H. Gillespie. *Theor. Popul. Biol.*, 23:202–215, 1983.

[27] S.E. Schoustra, T. Bataillon, D.R. Gifford, and R. Kassen. *PLoS Biol*, 7 (11):e1000250, 2009.

[28] D. R. Gifford, S. E. Schoustra, and R. Kassen. *Evolution*, 65:3070–3078, 2011.

[29] J. Neidhart and J. Krug. *Phys. Rev. Lett.*, 107:178102, 2011.

[30] K. Jain and J. Krug. *J. Stat. Mech.: Theor. Exp.*, page P04008, 2005.

[31] C. Sire, S. Majumdar, and D. S. Dean. *J. Stat. Mech.: Theor. Exp.*, page L07001, 2006.

[32] K. Jain and S. Seetharaman. *J. Nonlin. Math. Phys.*, 18:321–338, 2011.

[33] I. Bena and S.N. Majumdar. *Phys. Rev. E*, 75:051103, 2007.

[34] S. Sabhapandit, I. Bena, and S.N. Majumdar. *J. Stat. Mech.*, page P05012, 2008.

[35] D. R. Rokyta, C. J. Beisel, and P. Joyce. *J Theor Biol.*, 243:114–120, 2006.

[36] H. Flyvbjerg and B. Lautrup. *Phys. Rev. A*, 46:6714–6723, 1992.

[37] S. Seetharaman. M.S. thesis, JNCASR, Bangalore, 2011.

[38] H. J. Muller. *Mutation Res.*, 1:2–9, 1964.

[39] P. J. Gerrish and R. E. Lenski. *Genetica*, 102:127–144, 1998.

[40] S.-C. Park and J. Krug. *PNAS*, 104:18135–18140, 2007.

[41] M.M. Desai and D.S. Fisher. *Genetics*, 176:1759–1798, 2007.

[42] J.A.G.M. de Visser and D. E. Rozen. *Genetics*, 172:2093–2100, 2006.

[43] J.A.G.M. de Visser, C.W. Zeyl, P.J. Gerrish, J.L. Blanchard, and R.E. Lenski. *Science*, 283:404–406, 1999.

[44] R. Miralles, P. J. Gerrish, A. Moya, and S.F. Elena. *Science*, 285:813–815, 1999.

[45] D.E. Rozen, J.A.G.M. de Visser, and P. J. Gerrish. *Curr Biol.*, 12:1040–1045, 2002.

[46] S.-C. Park, D. Simon, and J. Krug. *J. Stat. Phys.*, 138:381–410, 2010.

Chapter 3

Fitness evolution during adaptive walk

3.1 Introduction

A fundamental question in the study of adaptive dynamics is whether adaptation happens via many mutations conferring small fitness advantage, or a few producing large fitness changes. Although initial theoretical works suggested that adaptation occurs mostly by mutations that provide small benefits [1, 2], it has been recently realised that large effect mutations are also possible [3]. The basic idea governing the shape of the distribution of beneficial fitness effects (DBFE) is due to Gillespie [4], who argued that in the event of a small environmental change, as the wild type fitness is expected to remain high, the mutations conferring higher fitness than the wild type will

lie in the right tail of the fitness distribution. As discussed in Chapter 2, the extreme value distribution of such independent extreme fitnesses can be only of three types: Weibull which occurs when the fitnesses are right-truncated, Gumbel for distributions decaying faster than a power law and Fréchet for distributions with algebraic tails [5].

In Chapter 2, we introduced the strong selection-weak mutation (SSWM) regime in which an asexual population performs an adaptive walk on the fitness landscape. In the adaptive walk model, if the population is fixed at a sequence with fitness h, a mutant with fitness $f > h$ substitutes it with a probability proportional to the fixation probability $\pi(s)$ where $s = (f-h)/h$. For long sequences, when the selection coefficient is assumed to be small, we use the linear model for which the normalised transition probability is given in (2.14) and we use the full model whose normalised transition probability is given in (2.13) for all values of the selection coefficient [6]. In our work, we use the block model discussed in Chapter 2 to model the fitness landscape and generate the fitness for each sequence from (2.8). In this chapter, we will study how the fitness of a maladapted asexual population evolves with the fixation of beneficial mutations using both the linear model and the full model.

3.2 Evolution of fitness fixed in the linear model

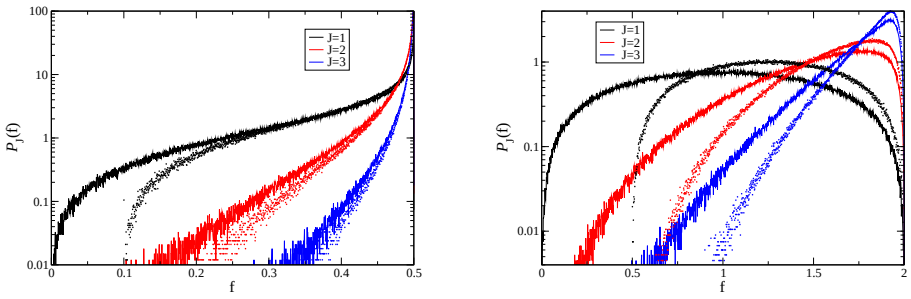
3.2.1 On uncorrelated fitness landscapes

Transition in the behavior of fitness fixed:

Given that fitness is h at a step in the adaptive walk, the probability, $\mathcal{P}_J(f)$ that the adaptive walk fixes a fitness f at step J can be obtained from the transition probability (2.14) which gives, $T(f \leftarrow h) \propto (f-h)p(f)$ favoring large fitness differences. But as the fitness distribution $p(f)$ is a decreasing (increasing) function of f for $\kappa > -1$ ($\kappa < -1$), the transition probability is nonmonotonic in f with the most probable fitness $f^* = 2 + \kappa h$ for $\kappa > -1$ but monotonically increasing for $\kappa \leq -1$. This property is reflected in the distribution $\mathcal{P}_J(f)$ (shown in Fig. 3.1 for $\kappa = -1/2$ and -2) which peaks at higher fitnesses as f_0 increases for $\kappa > -1$ while for $\kappa < -1$, irrespective of f_0, the most probable fitness occurs at the upper limit u of the fitness distribution.

Due to (2.20), the average fitness fixed at the next step given by

$$\int_h^u df\ f\ T(f \leftarrow h) = \frac{\int_h^u df\ f\ (f-h)p(f)}{\int_h^u dg\ (g-h)p(g)} \tag{3.1}$$



(a) (b)

Figure 3.1: Plot of $\mathcal{P}_J(f)$ for (a) $\kappa = -2$ and (b) $\kappa = -1/2$ to show the behavior of the most probable fitness when $L = 1000$. The lines show the simulation data for initial fitness $f_0 = 0$ and points for $f_0 > 0$ in both cases.

for $\kappa < 1$ however displays a different dependence on the GPD exponent. The numerator on the RHS of the above equation contains the second moment of the fitness distribution which is infinite for $\kappa \geq 1/2$. We show that the fitness of a step far from a local fitness peak has a dependence on the length of the walk when the second moment of $p(f)$ does not exist (when $\kappa \geq 1/2$), whereas there is no such dependence otherwise (when $\kappa < 1/2$). We shall discuss this transition in the rest of this section. Moreover for $\kappa < 1/2$, the fitness difference

$$\int_h^u df\ f\ T(f \leftarrow h) - h = \frac{\int_h^u df\ (f-h)^2 p(f)}{\int_h^u dg\ (g-h)p(g)} \tag{3.2}$$

has the important property that it *increases* with fitness h for positive κ. For an infinite sequence in which fitter mutants are always available, these two results taken together suggest that the fitness jumps keep increasing during

the adaptive process for fat-tailed fitness distributions.

Below we discuss how the average fitness increases with the number of adaptive substitutions and with initial fitness on uncorrelated fitness landscapes.

Average fitness for $\kappa < 1/2$:

If the population reaches a local fitness peak after $\bar{J}$ steps and when the number of adaptive substitutions $J \ll \bar{J}$ or the initial fitness is far from the local fitness peak, the average fitness fixed at the Jth step for a sequence of length L is well approximated by the corresponding quantity for an infinitely long sequence [6]. Taking the infinite sequence limit in (2.17) and using the definition (2.20) for average fitness, we have

$$\bar{f}_{J+1}(f_0) \;=\; \int_{f_0}^{u} dh \int_{h}^{u} df\, f\; T(f \leftarrow h)\; \Phi_J(h|f_0) \tag{3.3}$$

where $\Phi_J(f|f_0) \equiv \mathrm{Lim}_{L\to\infty}\mathcal{P}_J(f|f_0)$ and we have interchanged the order of integration to arrive at the last equation. In the linear model where the transition probability is given by (2.14), as the numerator in the integrand contains the second moment of the distribution $p(f)$ which is undefined for $\kappa \geq 1/2$, the above equation is valid for $\kappa < 1/2$ only. On performing the

integrals in (3.3), we have

$$\bar{f}_{J+1}(f_0) = \int_{f_0}^{u} dh \, \frac{2+h}{1-2\kappa} \, \Phi_J(h|f_0) \tag{3.4}$$

$$= \frac{2+\bar{f}_J(f_0)}{1-2\kappa} \tag{3.5}$$

where we have used the fact that for $\kappa < 1$, the adaptive walk goes on indefinitely for an infinitely long sequence [6]. The solution of the above equation is given by

$$\bar{f}_J(f_0) = \left(\frac{1+\kappa f_0}{\kappa}\right)(1-2\kappa)^{-J} - \frac{1}{\kappa} \tag{3.6}$$

The above result for zero initial fitness matches Eq. 33 of [3] for high initial rank. Equation (3.6) predicts that the final fitness u is approached exponentially for bounded distributions but the fitness increases with the number of substitutions linearly for exponentially distributed fitnesses and exponentially for unbounded distributions with $0 < \kappa < 1/2$.

It is useful to consider the fitness improvement $\overline{\Delta f_J}$ during the successive steps defined as

$$\overline{\Delta f_J} = \overline{f_J - f_{J-1}} \tag{3.7}$$

where the overbar represents averaging over only those walks that reach the Jth step. For infinitely long sequences, as the Jth step is definitely taken [6],

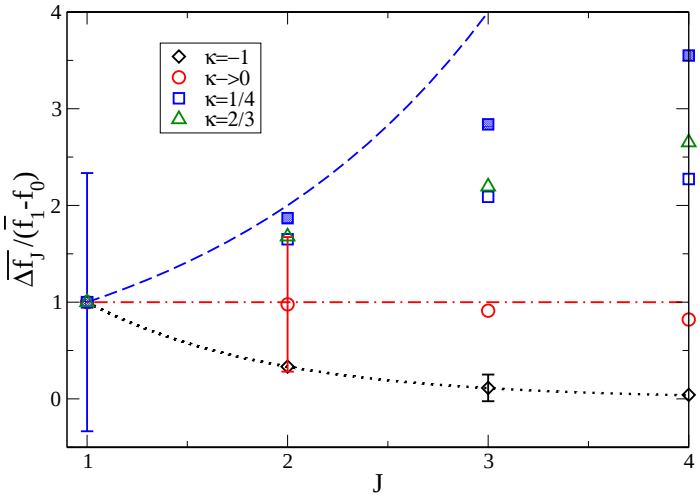


Figure 3.2: The plot shows (scaled) average fitness difference between successive steps as a function of the number of adaptive substitutions for $L = 1000$ (open symbols) and $f_0 = 0$ on uncorrelated fitness landscapes. The theoretical prediction (3.8) is shown by lines and the simulation data by points. The filled boxes are the simulation data for $L = 10000$ and $\kappa = 1/4$ to show that the agreement with theoretical prediction (3.8) improves with increasing L. The standard deviation about the mean fitness difference is shown by error bars for a few representative points.

we have

$$\overline{\Delta f_J} = \bar{f}_J - \bar{f}_{J-1} = 2(1 + \kappa f_0)(1 - 2\kappa)^{-J} \tag{3.8}$$

A similar expression for fitness effects has been obtained by Joyce *et. al.*, [3] but its consequences were not discussed. The above result has also been obtained for the special case of exponentially distributed fitnesses and zero

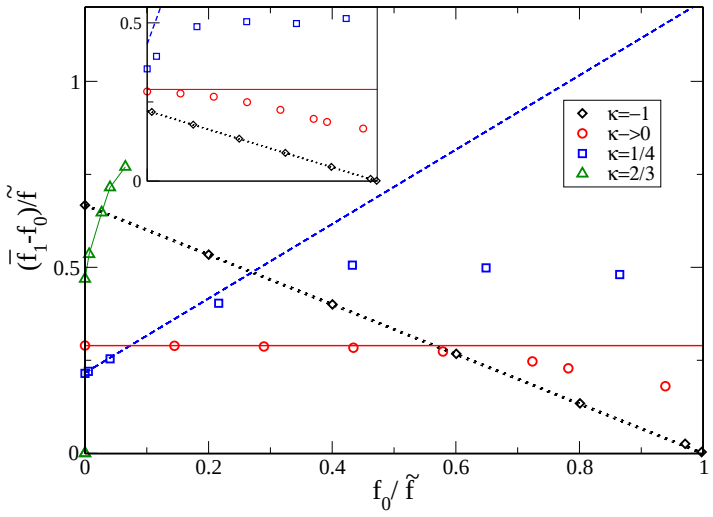


Figure 3.3: The plot shows (scaled) average fitness difference between successive steps as a function of initial fitness on uncorrelated fitness landscape for various κ and $L = 1000$ when $J = 1$ (main plot) and 2 (inset). The points give the simulation data and the lines are the theoretical prediction (3.8) for $\kappa < 1/2$. For $\kappa = 2/3$, the data has been scaled down by a factor two for clarity and the lines are guide to the eye.

initial fitness by [7]. For fixed initial fitness, (3.8) shows that for $\kappa < 0$, the fitness benefit decreases exponentially as the walk proceeds (*diminishing returns*) while for $\kappa \to 0$, the average fitness increases linearly during successive steps conferring constant benefit (*constant returns*) and for $0 < \kappa < 1/2$, the fitness difference increases exponentially fast with each step conferring higher benefit than the previous one (*accelerating returns*). Similar qualitative trends are seen when the initial fitness is varied since the fitness gain changes linearly with f_0 and the sign of the slope changes as κ crosses zero. In

Figs. 3.2 and 3.3, the simulation results and the above theoretical prediction (3.8) for infinitely long sequence are compared and we see a good agreement when the local fitness maximum is far away. We have also measured the standard deviation about the mean fitness fixed and our simulations indicate that as the walk proceeds, the fluctuations increase for $\kappa > 0$ but remain almost a constant for $\kappa \to 0$ and decrease for $\kappa < 0$.

Average fitness for $1/2 \leq \kappa < 1$:

When the second moment of the fitness distribution becomes infinite, we work with a sequence of finite length to find how the average fitness diverges with L. Since the adaptation process is over when the fitness fixed is of the order of the fitness of the local fitness optimum, we truncate the fitness distribution (2.8) at the local fitness maximum $\tilde{f}$ given by (2.9) [5].

Proceeding in a manner similar to that used to obtain (3.3) above, we find that for large but finite L,

$$\bar{f}_{J+1}(f_0) \approx -\frac{2+\bar{f}_J(f_0)}{2\kappa-1} + \frac{(1-\kappa)\ L^{2\kappa-1}}{(2\kappa-1)\kappa^2} \int_{f_0}^{\tilde{f}} dh\ \Phi_J(h)\ (1+\kappa h)^{\frac{1-\kappa}{\kappa}} \qquad (3.9)$$

The second term on the RHS can be neglected when $\kappa < 1/2$ and we recover (3.5). At the first step in the walk, using (2.18), we immediately get

$$\bar{f}_1 = -\frac{2+f_0}{2\kappa-1} + \frac{(1-\kappa)(1+\kappa f_0)^{\frac{1-\kappa}{\kappa}}}{\kappa^2(2\kappa-1)} L^{2\kappa-1} \qquad (3.10)$$

which shows that for large L, the average fitness $\bar{f}_1$ scales as $L^{2\kappa-1}$. Beyond

the first step, an exact expression for $\Phi_J(h)$ is not available but if we assume that this distribution decays faster than $h^{-1/\kappa}$, one can replace the upper limit in the integral on the RHS of (3.9) by infinity to get

$$\bar{f}_{J+1}(f_0) \approx -\frac{2+\bar{f}_J(f_0)}{2\kappa-1} + \mathcal{A}_J(\kappa, f_0) L^{2\kappa-1} \tag{3.11}$$

where $\mathcal{A}_J(\kappa, f_0)$ represents the resulting integral and the other prefactors. The above equation shows how the fitness fixed differs in its character for κ below and above half.

For $\kappa < 1/2$, as the right hand side (RHS) of (3.11) becomes independent of the sequence length (and hence $\tilde{f}$) for large L, the fitness fixed during the initial steps in the walk depends on the initial fitness but not on the local peak fitness. In contrast, for $1/2 \leq \kappa < 1$, the fitness fixed depends on both the initial fitness and the local peak fitness. From (2.9) and (3.11), we see that the fitness fixed increases as $\tilde{f}^{\frac{2\kappa-1}{\kappa}}$ which scales sublinearly with $\tilde{f}$ for $\kappa < 1$. We also find that for $\kappa > 1$, the population immediately jumps to a fitness close to the local optimum irrespective of the initial fitness and therefore we expect the fitness fixed in this parameter regime to depend only on the fitness of the local optimum. These inferences are indeed consistent with the results of the numerical simulations shown in Fig. 3.4. Equation (3.11) also suggests that the fitness at the Jth step increases with the sequence length as $L^{2\kappa-1}$ which is supported by numerical simulations shown in Fig. 3.4 for $\kappa = 2/3$. Also the fitness difference between successive steps increases with

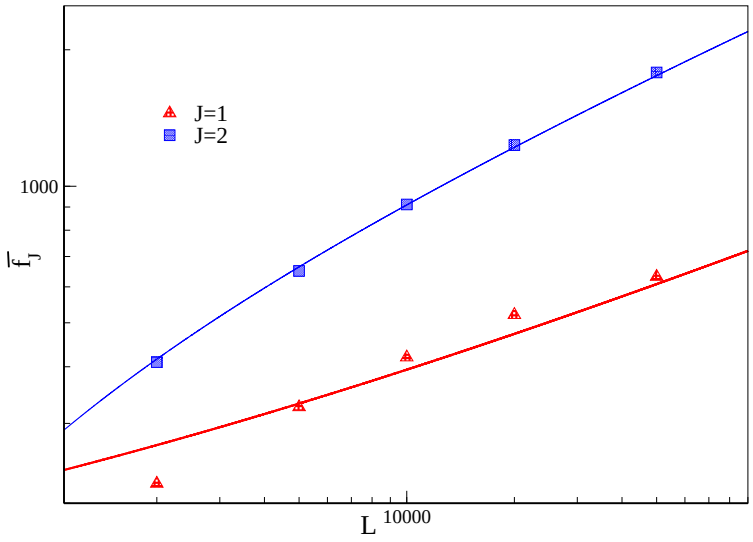


Figure 3.4: Numerical data for the average fitness fixed during the walk as a function of L when $\kappa = 2/3$, $B = 1$ and $f_0 = 1$ to support the expectation that it scales as $L^{2\kappa-1}$ (refer (3.11)). The lines are best fit to the curve of the form $\bar{f}_J = A_1(J) + A_2(J)L^{2\kappa-1}$.

both J and f_0 as shown in Figs. 3.2 and 3.3.

The fitness evolution on correlated fitness landscapes is discussed in 3.2.2 within a simple approximation. Although the results thus obtained do not match the simulation results (displayed in Fig. 3.5) quantitatively, the correct qualitative behavior is captured.

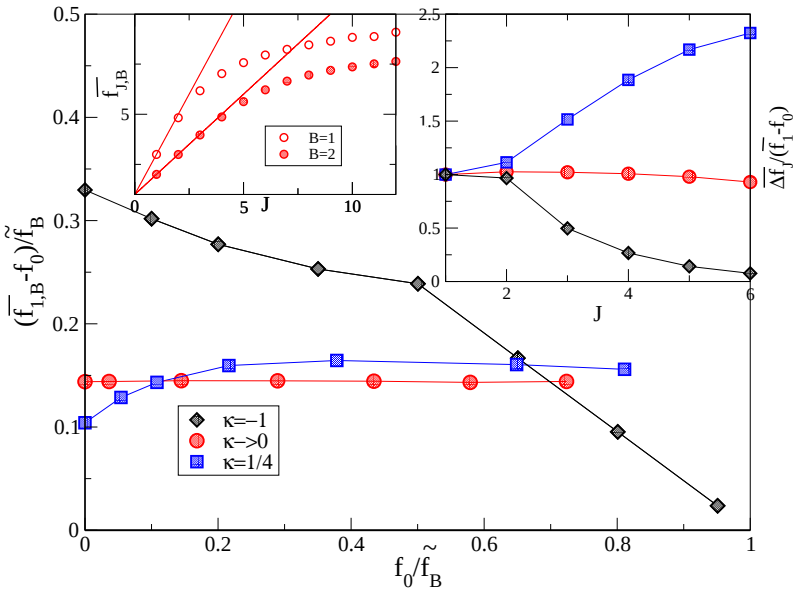


Figure 3.5: Main: (Scaled) average fitness difference at the first step as a function of initial fitness for various κ. Top left inset: Fitness evolution during the course of the adaptive walk for exponentially distributed fitnesses with $f_0 = 1$. The lines are the theoretical prediction (3.16) and the points give the simulation data. Top right inset: (Scaled) average fitness difference between successive steps as a function of the number of adaptive substitutions when $f_0 = 0$. In all the plots, the block number $B = 2$ and sequence length $L = 2000$. Unless mentioned otherwise, the points show the simulation data while the lines are guide to the eye.

3.2.2 Effect of correlations on fitness evolution

For a given set $\{f_0^{(b)}\}$ of initial block fitnesses, the average fitness fixed when the sequence is partitioned into B blocks is given by

$$\bar{f}_{J,B} = \frac{1}{B}\sum_{i=1}^{B} \frac{\overline{f_J^{(i)} + \sum_{k\neq i} f_{J-1}^{(k)}}}{B} \tag{3.12}$$

where the overbar represents averaging with respect to (w.r.t.) the distribution $D_J(i) \equiv \mathcal{P}_J(f_{J-1}^{(1)}, ..., f_J^{(i)}, ..., f_{J-1}^{(B)} | \{f_0^{(b)}\})$ that at step J, an adaptive substitution occurs in the ith block. Assuming that this distribution can be factorized over the blocks, we find that

$$\bar{f}_{J,B} = \sum_{i=1}^{B} \frac{\bar{f}_J^{(i)}}{B^2} + \sum_{i=1}^{B} \frac{\sum_{k \neq i} \bar{f}_{J-1}^{(k)}}{B^2} \tag{3.13}$$

$$= \frac{\bar{f}_J - \bar{f}_{J-1}}{B} + \bar{f}_{J-1} \tag{3.14}$$

where the last equation is obtained on averaging over the initial block fitnesses under the constraint (2.15) and $\bar{f}_J$ is the average fitness fixed at the Jth step on uncorrelated fitness landscape. At the first step in the walk, using (3.6) in (3.14), we immediately get

$$\bar{f}_{1,B} = f_0 + \frac{2}{B} \left(\frac{1 + \kappa f_0}{1 - 2\kappa} \right) \tag{3.15}$$

which states that the fitness difference at the first step depends linearly on the initial sequence fitness for nonzero κ. However our simulation data shown in Fig. 3.5 does not agree with the above expectation except for exponentially distributed fitnesses which suggests that the factorization property for the distribution $D_J(i)$ does not hold in general. For $\kappa \to 0$, due to (3.14) and (3.6), we have

$$\bar{f}_{J,B} = f_0 + \frac{2J}{B} \tag{3.16}$$

At first few steps, the above expression is consistent with the simulation results as can be seen in Fig. 3.5. For $\kappa \neq 0$, although the fitness difference does not obey (3.14), the trend with κ shown in Fig. 3.5 is similar to that in the uncorrelated case where the fitness difference increases and decreases for $\kappa > 0$ and $\kappa < 0$ respectively.

3.2.3 Mean selection coefficient during the walk

The average selection coefficient fixed at step J also exhibits a similar behavior as the fitness. To see this, consider the distribution $\mathcal{S}_J(s|f_0)$ of selection coefficient s at the Jth step in the walk which can be determined using (2.16) for an infinitely long sequence as

$$\mathcal{S}_J(s|f_0) = \int_{f_0}^{u} df \int_{f_0}^{u} dh\ \delta\left(s - \frac{f-h}{h}\right)\ T(f \leftarrow h)\ \Phi_{J-1}(h|f_0) \quad (3.17)$$

$$= \int_{f_0}^{\frac{u}{s+1}} dh\ h\ T(h(s+1) \leftarrow h)\ \Phi_{J-1}(h|f_0)\ ,\ J \geq 1 \quad (3.18)$$

In the last equation, the upper limit of the integral is obtained using the fact that the fitness f at the Jth step can not exceed u. Then the average selection coefficient can be written as

$$\bar{s}_J(f_0) = \int_{f_0}^{u} dh\ h\ \Phi_{J-1}(h|f_0) \int_{0}^{\frac{u}{h}-1} ds\ s\ T(h(s+1) \leftarrow h) \quad (3.19)$$

The average selection coefficient in the linear model displayed in Fig. 3.6 for various κ shows that it decreases with increasing J or f_0. For $\kappa \leq 0$, due

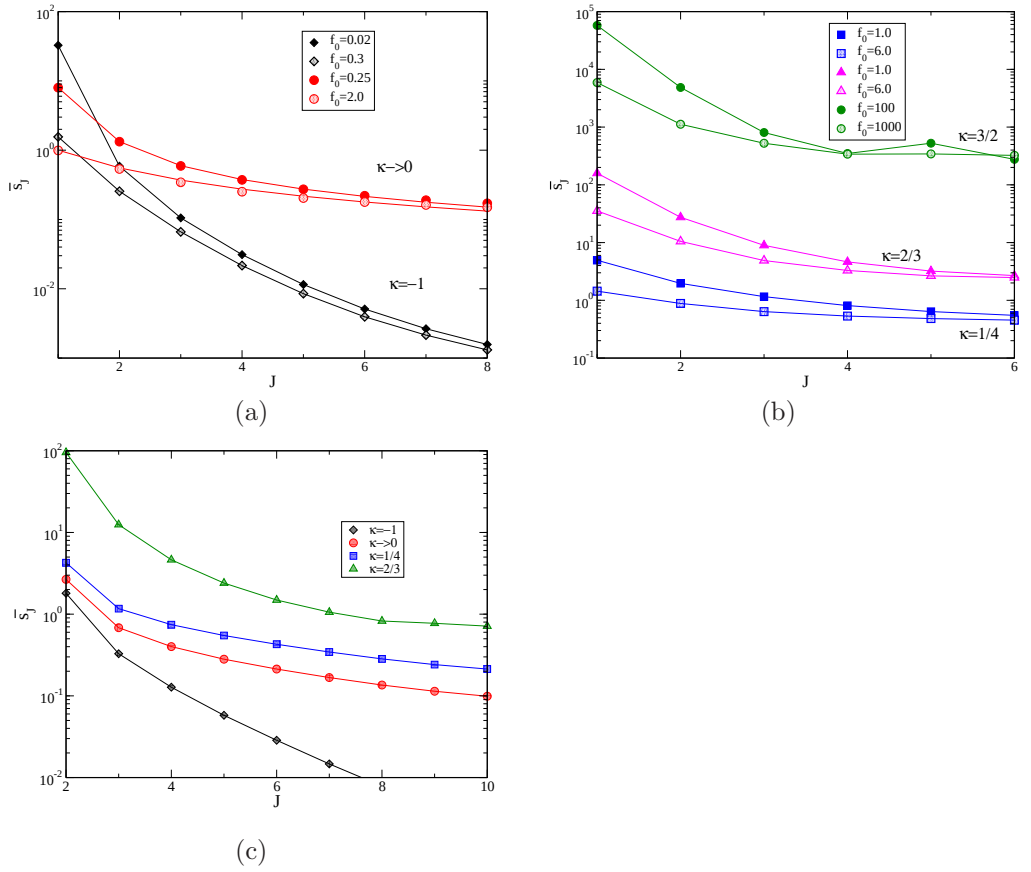


Figure 3.6: Average selection coefficient for various κ and initial fitness f_0 with $L = 1000$ on (a), (b) uncorrelated and (c) correlated fitness landscapes ($B = 2, f_0 = 0$). The theoretical prediction (3.24) is shown for exponentially distributed uncorrelated fitness while in all the other cases, the lines are guide to the eye.

to the upper bound of $p(f)$, the fitness difference is a nonincreasing function of J and f_0. So one may expect the selection coefficient (2.21) to decrease but for $\kappa > 0$, although the fitness benefit increases, the selection coefficient still decreases. Moreover like average fitness $\bar{f}_J$, the average selection coefficient $\bar{s}_J$ is also undefined for $\kappa \geq 1/2$. Using (3.18) for the distribution of selection

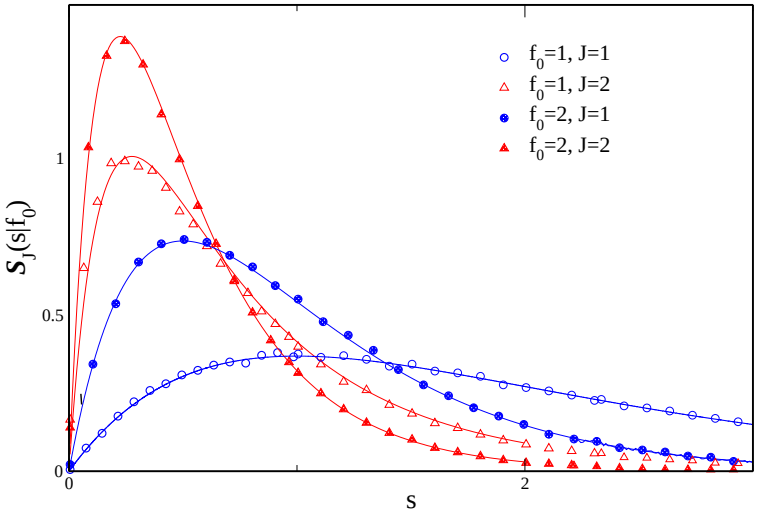


Figure 3.7: Distribution of the selection coefficient for exponentially distributed uncorrelated fitnesses obtained numerically (points) for $L = 1000$ compared against the theoretical result (3.23) for infinitely long sequence.

coefficient in the limit $L \to \infty$, we obtain

$$\bar{s}_J(f_0) = \int_{f_0}^{u} dh \, \frac{(1-\kappa)h^2}{(1+\kappa h)^{\frac{\kappa-1}{\kappa}}} \Phi_{J-1}(h|f_0) \int_0^{\frac{u}{h}-1} ds \; s^2 \; p\,[h(s+1)] \quad (3.20)$$

As the inner integral over s in the last equation is undefined for $\kappa \geq 1/2$, we get

$$\bar{s}_J(f_0) = \frac{2\kappa}{1-2\kappa} + \frac{2}{1-2\kappa} \int_{f_0}^{u} \frac{dh}{h} \Phi_{J-1}(h|f_0) \;, \;\; \kappa < 1/2 \quad (3.21)$$

Note that while the expectation value of fitness h is involved in the expression (3.4) for average fitness, the average of $1/h$ appears in the above equation.

For exponentially distributed fitnesses, on solving the differential equation (2.17) in the infinite sequence length limit [6], we find that the distribution $\Phi_J(f|f_0)$ is given by

$$\Phi_J(f|f_0) = \frac{e^{-(f-f_0)}(f-f_0)^{2J-1}}{(2J-1)!} \tag{3.22}$$

thus leading to the distribution of selection coefficient in (3.18) for an infinitely long sequence as

$$S_J(s|f_0) \stackrel{L\to\infty}{\longrightarrow} \frac{se^{-sf_0}}{(1+s)^{2J}}\left[((1+s)f_0+2J-2)^2+2J-2\right] \tag{3.23}$$

The above result for an infinitely long sequence compares well with the simulation results for a finite sequence shown in Fig. 3.7. We find that the distribution is nonmonotonic in s and decays faster with increasing J or f_0. From (3.23), we also obtain

$$\bar{s}_J = \frac{2}{(2J-3)!}\int_0^\infty dy\ \frac{y^{2J-3}}{y+f_0}e^{-y} = 2e^{f_0}E_{2J-2}(f_0)\ ,\ J>1 \tag{3.24}$$

where $E_n(x)$ is the exponential integral function. Using the asymptotic expansion of exponential integral [8], we find that $\bar{s}_J \sim 2/f_0$ for large f_0 and therefore the selection coefficient decreases with initial fitness. For large J, we have $\bar{s}_J \sim 1/J$ on using the representation of $E_n(x)$ for large n [8]. Thus the mean selection coefficient decreases with increasing initial fitness and during the course of the walk.

We have obtained an analytical expression for the distribution $\Phi_J(f|f_0)$ which is available only for special cases such as the first step in the walk, which we shall discuss now.

At the first step in the walk, using (2.18) in (3.18), we immediately have

$$S_1(s|f_0) \stackrel{L\to\infty}{\longrightarrow} \frac{(1-\kappa)f_0^2 s}{(1+\kappa f_0)^{\frac{\kappa-1}{\kappa}}} p((s+1)f_0)\ ,\ 1+\kappa(s+1)f_0 > 0 \tag{3.25}$$

which is a nonmonotonic function of s for $\kappa > -1$ but increases monotonically for $\kappa \leq -1$. The above equation also gives

$$\bar{s}_1(f_0) = \frac{2\kappa + 2f_0^{-1}}{1-2\kappa}\ ,\ \kappa < 1/2 \tag{3.26}$$

which decreases as the initial fitness increases (see Fig. 3.6). If the above distribution (3.25) is averaged over the initial fitness also, the mean selection coefficient diverges for all $\kappa < 1$. On integrating (3.25) over the fitness distribution with $\kappa \geq 0$, we obtain

$$S_1(s) \stackrel{L\to\infty}{\longrightarrow} s(1-\kappa)\int_0^\infty df_0 \frac{f_0^2}{(1+\kappa f_0)^2}\ (1+\kappa(s+1)f_0)^{-\frac{1+\kappa}{\kappa}} \tag{3.27}$$

It is useful to consider the integral

$$I(s) = s\int_0^\infty dx\ \frac{x^2}{(1+x)^2}\ (1+xs)^{-a} \tag{3.28}$$

$$\approx s^{-1/\kappa}\int_0^\infty dx\ \frac{x^{2-a}}{(1+x)^2} \tag{3.29}$$

where we have assumed $s \gg 1$ and $a = (1+\kappa)/\kappa$. We also have

$$I(s) = s^{-2} \int_0^\infty dx \, \frac{x^2}{(1+xs^{-1})^2} \, (1+x)^{-a} \tag{3.30}$$

$$\approx s^{-2} \int_0^\infty dx \, x^2 \, (1+x)^{-a} \tag{3.31}$$

where the last integral is finite provided $\kappa < 1/2$. Splitting the integral $I(x)$ as follows:

$$I(s) = s \int_0^{1/s} + \int_{1/s}^1 + \int_1^\infty \tag{3.32}$$

$$\approx s \int_1^\infty dx (1+xs^{-1})^{-2} + \int_0^{1/s} dx \frac{x^2}{(1+x)^2} \tag{3.33}$$

$$= \mathcal{O}(s^{-1/\kappa}) + \mathcal{O}(s^{-2}) \tag{3.34}$$

Thus for $\kappa < 1/2$, the distribution of the selection coefficient decays as $\sim s^{-2}$ while for $1/2 < \kappa < 1$, it decays as $s^{-1/\kappa}$ giving a diverging mean selection coefficient for all $\kappa < 1$.

3.3 Evolution of the fitness fixed in the full model

In the biologically relevant full model of the adaptive walk, we find that while the fitness fixed during adaptation increases in all the three extreme value domains [6], the average difference $\overline{\Delta f_J}$ between fitnesses fixed at step

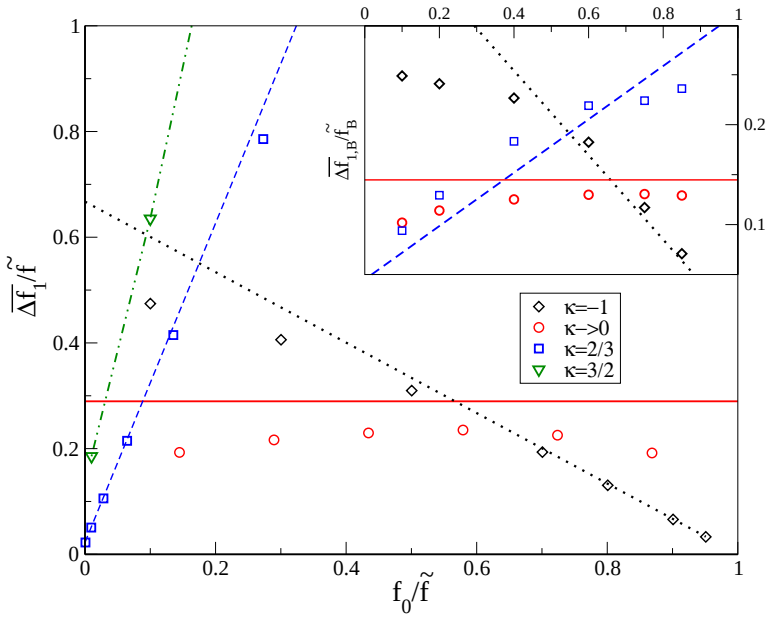


Figure 3.8: The plot shows (scaled) average fitness difference at the first step as a function of initial fitness for various κ on uncorrelated (main) and correlated fitness landscapes with $B = 2$ (inset). In both the plots, $L_B = 1000$ which corresponds to $L = 1000$ and 2000 for uncorrelated and correlated fitnesses respectively. The points give the simulation data and the line connecting the data points are obtained from (3.45) and (3.46) for uncorrelated and correlated fitnesses respectively for $\kappa < 1$. The data points for $\kappa = 3/2$ are scaled down by 10^2 for clarity and the line connecting the data is guide to the eye.

$J - 1$ and J exhibits interesting trends that can be exploited to distinguish between them (refer Figs. 3.8 and 3.9).

3.3.1 On uncorrelated fitness landscapes

Transition in the behaviour of fitness fixed:

To find the average fitness and selection coefficient, we consider the probability distribution $\mathcal{P}_J(f|f_0)$ of the population having fitness f at the Jth step of the adaptive walk, given that it started with fitness f_0. On uncorrelated fitness landscapes, it obeys the recursion equation (2.16) and here we use the transition probability given by (2.13) [6, 9].

The average fitness fixed at the Jth step is given by $\bar{f}_J(f_0) = \int_{f_0}^{u} df\ f\ \mathcal{P}_J(f|f_0)$. Far from a local fitness peak, the average fitness fixed for a sequence of length L is well approximated by the corresponding quantity for an infinitely long sequence [6] as given in (3.3) for the transition probability (2.13). In (3.3), for a given h, we use the the transition probability (2.13) which reduces to (2.14) varying as $(f-h)\ p(f)$ for $f \ll 3h/2$ (small selection coefficient) which varies as or is just $p(f)$ when selective effects are large. As the dominant contribution to the inner integral in (3.3) comes from the large-f behavior of the integrand, the integral over f is seen to be proportional to the mean of the fitness distribution $p(f)$ which, we recall, is undefined for $\kappa \geq 1$. This result means that the fitness fixed is independent of the sequence length L for $\kappa < 1$, but increases with L otherwise.

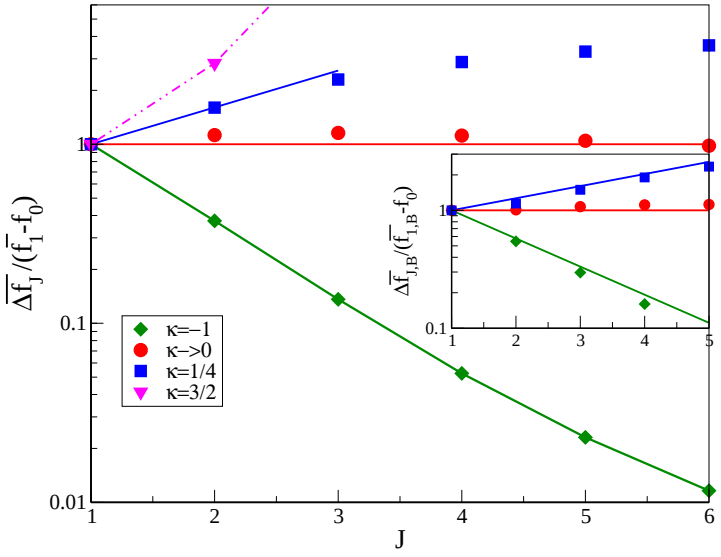


Figure 3.9: The plot shows (scaled) average fitness difference between successive steps as a function of the number of adaptive substitutions for various κ on uncorrelated (main) and correlated fitness landscapes with $B = 2$ (inset). Taking $f_0 = 0.63, 1, 1.14$ and 2.32 for $\kappa = -1, 0, 1/4$ and $3/2$ respectively, the simulation data are shown as points for $L_B = 1000$ which corresponds to $L = 1000$ and 2000 for independent and correlated fitnesses respectively. The line connecting the data points for $\kappa = 3/2$ is guide to the eye, while the others are obtained from (3.45) and (3.46) for uncorrelated and correlated fitnesses respectively.

Fitness increment between steps:

The fitness improvement $\overline{\Delta f_J}$ during the successive steps is given by (3.7). For infinitely long sequences, as the Jth step is definitely taken, we have $\overline{\Delta f_J} = \bar{f}_J - \bar{f}_{J-1}$. Thus it is sufficient to study the behaviour of the fitness fixed at each step which we discuss next.

We find that $\overline{\Delta f_J}$ decreases during the walk in the Weibull domain and

increases in the Fréchet domain. A similar behavior is seen at a fixed step in the walk when initial fitness is varied. A heuristic understanding of the latter result can be obtained for uncorrelated fitnesses using a simple back-of-the-envelope calculation of the average fitness $\bar{f}_1$ at the first step, which is given by $\int_{f_0}^{u} df\ f\ T(f \leftarrow f_0)$. We first note that if the fitness distribution decays slowly, fitnesses much larger than initial fitness can occur with appreciable frequency and thus the selection coefficients can be large. On the other hand, for bounded distributions, the selection coefficient can be at most $u/f_0 - 1$ which is below unity for $f_0 > u/2$. Indeed as Fig. 3.10 shows, the selection coefficients fixed are large (small) for positive (negative) κ. As a result, the fixation probability $\pi(s)$ can be approximated by unity in the Fréchet domain, while $\pi(s) \approx 2s$ in the Weibull domain. A quick calculation gives $\bar{f}_1 \sim f_0/(1-2\kappa)$, $\kappa < 0$ which is linear in f_0 with a slope below unity. On the other hand, in the Fréchet domain, a transition occurs in the behavior of the fitness fixed at $\kappa = 1$ where the mean of the distribution $p(f)$ becomes infinite. We find that the average fitness is infinite for $\kappa \geq 1$ but for $0 < \kappa < 1$, the fitness $\bar{f}_1 \sim f_0/(1-\kappa)$ which also increases with f_0 but with a slope above unity. The key point that emerges from these simple calculations (and detailed ones later in this section) is that the average fitness at the first step is of the form $af_0 + b$ where the slope a is above (below) one for positive (negative) κ. The result for the fitness difference claimed above then immediately follows.

To understand the behavior at higher steps in the adaptive walk, more

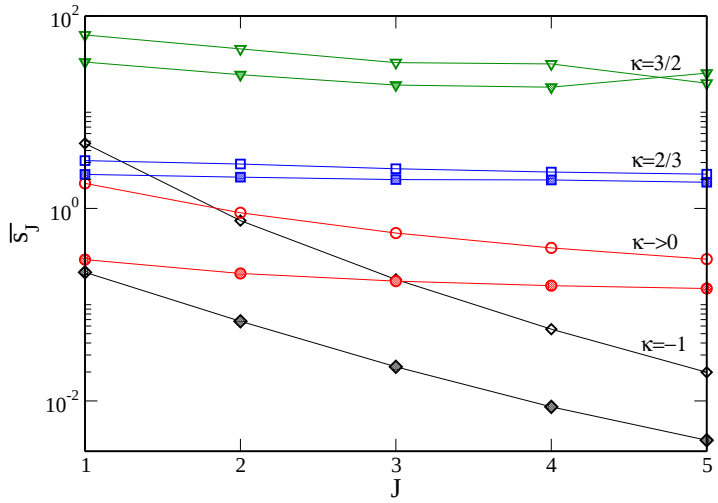


Figure 3.10: The plot shows the average selection coefficient fixed during the course of the walk on uncorrelated fitness landscapes for various κ and $L = 1000$. The open and shaded symbols are respectively for $f_0 = 0.1\tilde{f}$ and $0.75\tilde{f}$ where $\tilde{f}$ is the average fitness of a local fitness peak given by (2.9). The points are the simulation data, while the lines are guide to the eye. The data for $\kappa = 3/2$ is scaled down by a factor 10 for clarity.

work is required as described below.

Gumbel domain: We can obtain the expression for the average fitness at step J by performing the inner integral in (3.3) for $\kappa \to 0$, to get

$$\bar{f}_{J+1}(f_0) = \int_{f_0}^{\infty} dh\ \Phi_J(h|f_0)\ \frac{h^2 + 4h + 2}{h + 2} \tag{3.35}$$

The above equation does not close in the average fitness fixed i.e. the RHS contains the average of quantities which can not be written in terms of $\bar{f}_J$.

However for large initial fitness f_0, as $h \gg 1$, we can write

$$\bar{f}_{J+1}(f_0) = \int_{f_0}^{\infty} dh\ \Phi_J(h|f_0)\ \left(h + 2 + \mathcal{O}(h^{-1})\right) \tag{3.36}$$

$$= \bar{f}_J + 2 + \mathcal{O}\left(\overline{f_J^{-1}}\right) \tag{3.37}$$

where we have used that the adaptive walk goes on indefinitely for an infinitely long sequence [6]. As the average fitness increases during adaptation, one may expect the average of inverse fitness to decrease. Neglecting the last term on the RHS of (3.37), we immediately find the solution of the resulting recursion equation to be

$$\bar{f}_J = 2J + f_0 \tag{3.38}$$

Weibull domain: The inner integral in (3.3) can be done exactly, but the resulting expression is too complicated and we omit the general expression here. For the special case of $\kappa = -1$, we get

$$\bar{f}_{J+1}(f_0) = \int_{f_0}^{1} dh\ \Phi_J(h|f_0)\ \frac{2e^{2/h}(1-h)^2 + e^2 h(2+h)(\Gamma(2, 2-2/h) - 1)}{e^{2/h}(6h-4) - 2e^2 h} \tag{3.39}$$

where $\Gamma(a, x)$ is the incomplete gamma function [8]. The above equation demonstrates that as in the Gumbel domain, the recursion relation for $\bar{f}_J$ does not close here also. Since $p(f)$ has an upper bound when $\kappa < 0$, the selection coefficient which is the relative fitness difference, is well below one

when the initial fitness f_0 is close to u as can be seen in Fig. 3.10. In the inner integral on the RHS of (3.3), the selection coefficient is smaller than half if the fitness $f \ll 3h/2$ which is ensured if $f_0 > 2u/3$. These observations suggest that in the Weibull domain, the small selection coefficient can be assumed to be small. On using (2.13) in (3.3), we find that the recursion equation closes in the average fitness and given by $\bar{f}_{J+1} = a_- \bar{f}_J + b_-$ where

$$a_- = (1-2\kappa)^{-1} \tag{3.40}$$

$$b_- = 2(1-2\kappa)^{-1} \tag{3.41}$$

On iterating the recursion equation, we find the average fitness to be

$$\bar{f}_J = a_-^J f_0 + \frac{b_-}{1-a_-}\,(1-a_-^J) \tag{3.42}$$

It is evident that for negative κ, the coefficient $a_- < 1$. It is easily verified that (3.38) is obtained from the above equation when $\kappa \to 0$.

Fréchet domain: For $\kappa < 1$ and large f_0, proceeding in manner similar to that in the Gumbel domain, we find that the average fitness at step J is of the form $\bar{f}_{J+1} \approx a_+ \bar{f}_J + b_+$ where

$$a_+ = \frac{\kappa - e^2(1-\kappa)E_{\frac{1}{\kappa}}(2)}{2e^2\kappa(1-\kappa)E_{\frac{1}{\kappa}}(2)} \tag{3.43}$$

$$b_+ = \frac{\kappa - e^2(1+\kappa)E_{\frac{1}{\kappa}}(2) - 2e^4\kappa(1-\kappa)E^2_{\frac{1}{\kappa}}(2)}{2e^4\kappa^2(1-\kappa)E^2_{\frac{1}{\kappa}}(2)} \tag{3.44}$$

and $E_n(x)$ is the exponential integral [8]. For $\kappa \to 0$, using the large n representation of $E_n(x)$ [8] in the above expressions for a_+ and b_+, it can be checked that the result (3.38) in the Gumbel domain is obtained.

For infinitely long sequences, on using the results previously obtained in the section, we have

$$\overline{\Delta f_J} = \begin{cases} a_-^{J-1}\ ((a_- - 1)f_0 + b_-)\ ,\ \kappa < 0 & \text{(3.45a)} \\ 2\ ,\ \kappa \to 0 & \text{(3.45b)} \\ a_+^{J-1}\ ((a_+ - 1)f_0 + b_+)\ ,\ 0 < \kappa < 1 & \text{(3.45c)} \end{cases}$$

where $a_- < 1, a_+ > 1$. For fixed initial fitness, the above equation shows that for $\kappa < 0$, the fitness benefit decreases exponentially as the walk proceeds (*diminishing returns*) while for $\kappa \to 0$, the fitness gain is same (*constant returns*) and for $0 < \kappa < 1$, it increases exponentially fast with each step conferring higher benefit than the previous one (*accelerating returns*). Similar qualitative trends are seen when the initial fitness is varied: the fitness increment decreases (increases) linearly with f_0 for negative (positive) κ. In Figs. 3.8 and 3.9, the simulation results and the above theoretical prediction (3.45) for infinitely long sequence are compared and we see a good agreement when the initial fitness is sufficiently large but local fitness maximum is far away. The latter condition is satisfied when the number of adaptive substitutions and the initial fitness are smaller than the average length of the walk and the average fitness of a local maximum respectively. The results of our numerical simulations in Figs. 3.8 and 3.9 also show that the fitness

difference between successive steps increases with both f_0 and J for $\kappa \geq 1$.

In both the linear model and the full model, from (3.45) and (3.8), we first note that in each of the three extreme value domains, fitness difference displays the same qualitative trend, irrespective of whether the correct asymptotic behavior of transition probability is taken into account. However the result (3.8) matches with (3.45) in the Weibull and Gumbel domains but not in the Fréchet domain. This is because the selection coefficient, shown in Fig. 3.10 for two initial fitnesses, decreases with f_0 for $\kappa \leq 0$ and at sufficiently large f_0, selective effects can be assumed to be small. But for $\kappa > 0$, the selection coefficient remains high even for large initial fitnesses and therefore we do not expect the small selection coefficient approximation to work here. The behavior of the selection coefficient can be immediately obtained at the first step in the walk using (3.45a)-(3.45c) since $\bar{s}_1 = \overline{\Delta f_1}/f_0$ and we find that $\bar{s}_1$ decays to zero with increasing f_0 for $\kappa \leq 0$, but to a finite constant $a_+ - 1$ for $0 < \kappa < 1$. On comparing (3.45c) and (3.8), we find that the value of the exponent κ at which a transition occurs in the behavior of the fitness fixed is different. Moreover the growth rate a_+, which takes values in the range $1.1 - 27.5$ as κ is increased from 0.05 to 0.95, is smaller than the corresponding rate $(1 - 2\kappa)^{-1}$ in (3.8) because the transition probability (2.13) decays faster than (2.14) for large fitnesses.

In the full model, since the inner integral over s in (3.19) is undefined for $\kappa \geq 1$ for the same reasons as described earlier in this section for average fitness, we find that the average selection coefficient also undergoes a transition

at $\kappa = 1$.

3.3.2 On correlated fitness landscapes

When a sequence is partitioned into B blocks, the fitness of the sequence at any step is determined by the *joint* distribution of the fitness of the block that acquired one beneficial mutation and the fitnesses of the rest of the blocks at the preceding step. But as it is difficult to work with this distribution, here we use the approximation that the joint distribution can be factorized over the blocks. In other words, we assume that the blocks evolve independently which is a reasonable approximation for weakly correlated fitnesses. Since J substitutions in a sequence partitioned in B blocks can be obtained if each block acquires J/B mutations, it immediately follows that

$$\overline{\Delta f_{J,B}} \approx \bar{f}_{J/B} - \bar{f}_{(J/B)-1} \ , \ J > 0 \tag{3.46}$$

where $\bar{f}_J$ is the average sequence fitness at the Jth step on uncorrelated fitness landscapes. Using the results of the Section 3.3.1 in the above equation, we find that the trend of fitness difference for correlated fitnesses is the same as that in the uncorrelated case. This result is consistent with the simulation data shown in the inset of Figs. 3.8 and 3.9 where the fitness difference increases and decreases for $\kappa > 0$ and < 0 respectively. The selection coefficient decreases with increasing correlations in all extreme value domains. Our numerical data, for a parameter set in which the same initial

and final fitness is chosen for uncorrelated and correlated fitnesses, shows that the selection coefficient at the first step reduced from 1.4 to 0.8 for $\kappa = -1$, 0.7 to 0.4 for $\kappa \to 0$ and 2.8 to 0.4 for $\kappa = 2/3$, as the block number increased from one to two.

3.4 Discussion

In this chapter, we investigated how the statistical properties of the adaptive walk relate to the tail behavior of the fitness distribution. The sign of the exponent κ in the fitness distribution (2.8) determines the nature of the DBFE which can be of three types namely Weibull, Gumbel and Fréchet. It is important to note that this classification of the extreme value domains is applicable only if the fitnesses are completely uncorrelated or at most, weakly correlated [10]. For strongly correlated fitnesses, a classification of extreme value domains on the basis of the behavior of the tail of the fitness distribution is not available and it is not clear if it even exists. For these reasons, here we studied the adaptive walk properties on weakly correlated fitness landscapes [11].

The exponent κ in (2.8) has been measured in experiments and interestingly, all the three extreme value domains for uncorrelated fitnesses have been seen. Although many early studies supported the Gumbel domain [12–15],

recently Weibull [16, 17] and Fréchet domain [18–20] have also been documented. It has been suggested that as the beneficial mutations are expensive due to pleiotropic constraints, fat-tailed fitness distributions whose extreme value statistics lies in the Fréchet domain can occur if such constraints are limited, and bounded ones that lie in the Weibull domain for severe constraints [18]. Experiments suggest that the fitness landscapes are correlated [21–23] but we know little about the fitness correlations quantitatively. Sign epistasis which is a characteristic feature of rugged fitness landscapes with many local fitness maxima has also been documented in several recent experiments [24–27].

3.4.1 Comparison to previous works

Our theoretical analysis here differs in an important way from the previous studies on adaptive walks [3, 6, 7, 28–33] that assume selective effects to be small and therefore work with (2.14) in which the transition probability is linear in the selection coefficient. Here we not only work with (2.14) but also, have presented results for (2.13) which is a nonlinear function of the selection coefficient. Our numerical data in Fig. 3.10 on selection coefficient fixed shows that large selection coefficients can arise in any extreme value domain when the initial fitness is small, as is the case in adaptation experiments in stressful environment [34, 35], or in a moderately fit population if the underlying fitness distribution is slowly decaying as is the case in the Fréchet domain. Here we focus on the latter situation and therefore our formulae

hold for moderately high initial fitnesses that are far from a local fitness optimum.

Our main conclusion is that small selection coefficient approximation can be safely employed in the Weibull domain, but the analysis in the Gumbel and especially Fréchet domain requires that large effect mutations are taken into account. We find that regardless of whether we assume mutations to have small or large effect, a transition in the behavior of the fitness fixed occurs at a certain value of exponent κ below which the fitness fixed during the initial steps in the walk does not depend on the average fitness of a local peak, and above which it does. The transition point is given by $\kappa = 1$ where mean of the fitness distribution becomes infinite if transition probability (2.13) is used but by $\kappa = 1/2$ which corresponds to an infinite variance, if selection coefficient is assumed to be small.

3.4.2 Evolution of fitness and selection coefficient

Although the fitness fixed during the adaptive walk increases with the number of substitutions and with initial fitness in all extreme value domains, the fitness difference (3.45) between successive steps depends on how fast the fitness distribution (2.8) decays (refer Figs. 3.8 and 3.9). In the Weibull domain, the fitness benefits decrease as the walk proceeds or the starting fitness is increased. In contrast, in the Fréchet domain, increasing adaptive substitutions or initial fitness leads to increasing fitness gain. This behavior of fitness increments is robust with respect to fitness correlations as attested

by the insets of Figs. 3.8 and 3.9, and holds irrespective of whether the correct asymptotic behavior of the fixation probability is taken into account. To get some insight into the behavior of average fitness difference with initial fitness, we recall that the selection coefficient is bounded above for truncated distributions but not for the unbounded ones. As a result, in the Weibull domain, we have $f_1 - f_0 \leq u - f_0$ which suggests that $\overline{\Delta f_1}$ decreases with f_0. This is in contrast to the behavior in the Fréchet domain where the selection coefficient is large (and positive) and hence $f_1 - f_0 \propto f_0$ which increases linearly with f_0. Treating the fitness at the Jth step as the initial fitness for the next step, the patterns during the course of the walk can also be understood.

We believe that experimental measurements of fitness difference as a function of the initial fitness in a population evolving under strong selection-weak mutation conditions can give an insight into the domain of the DBFE. Although negative correlation between initial fitness and fitness gain has been observed [15,34] and increasing fitness gain in successive steps has been seen in small populations [36], how these results correlate with the tail behavior of the beneficial mutations in these studies is not known. On the other hand, the adaptive walk properties have not been studied in experiments that measure the exponent κ [12–18]. In [37], since the local fitness optimum to which the population approaches is fixed, a truncated fitness distribution is expected. It would be interesting to check if our prediction that the fitness difference decreases with increasing initial fitness for bounded distributions

is supported by the fitness data in their experiment.

We also studied the behavior of average selection coefficient and obtained a general result that in all the three extreme value domains, it decreases during the course of the walk and with initial fitness and fitness correlations, see Fig. 3.10 (also refer [7]). Since the fitnesses are closely related on correlated fitness landscapes, the fitness difference and hence selection coefficient is expected to decrease with increasing correlations. The behavior with initial fitness can also be rationalized using the fact that the fitness difference grows at most linearly with initial fitness and therefore for large f_0, selection coefficient decreases. These results are consistent with those obtained in the experimental studies on *Aspergillus nidulans* [38,39] in which the mean selection coefficient is observed to decrease as the walk proceeds, and *Escherichia coli* [37] in which the mean selective effect is found to be larger for poorer initial condition.

3.4.3 Beyond the SSWM regime

In the experiments discussed above, the population size is $> 10^4$ [17, 37–40], the smallest selection coefficient detected is $\sim 10^{-3}$ [37,39] and the mutation rate per base pair for the microbes used in the experiments namely, bacteriophage $\phi X174$ [13], *Escherichia coli* [37], *Aspergillus nidulans* [38] is of the order $10^{-7} - 10^{-11}$ [41]. Thus these experiments are in the strong selection-weak mutation regime where adaptive walk model studied here is defined. However when the population size is large enough that the weak mutation

condition fails, clonal interference occurs in which two or more independent beneficial mutations arise in the population and compete with each other for dominance [42–45]. In Chapter 5, we shall discuss whether the trends in the fitness difference discussed here are also exhibited by populations with competing beneficial mutations especially during the early adaptation stage.

Bibliography

[1] R. A. Fisher. *The genetical theory of natural selection*. Oxford: Clarendon Press, 1930.

[2] H. A. Orr. *Genetics*, 163:15191526, 2003.

[3] P. Joyce, D. R. Rokyta, C. J. Beisel, and H. A. Orr. *Genetics*, 180:1627–1643, 2008.

[4] J. H. Gillespie. *Theor. Popul. Biol.*, 23:202–215, 1983.

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

[6] K. Jain and S. Seetharaman. *Genetics*, 189:1029–1043, 2011.

[7] S. Kryazhimskiy, G. Tkačik, and J. B. Plotkin. *PNAS*, 106:18638–18643, 2009.

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

[9] H. Flyvbjerg and B. Lautrup. *Phys. Rev. A*, 46:6714–6723, 1992.

[10] M. Clusel and E. Bertin. *Int. J. Mod. Phys. B*, 22:3311–3368, 2008.

[11] A. S. Perelson and C. A. Macken. *PNAS*, 92:9657–9661, 1995.

[12] R. Sanjuán, A. Moya, and S.F. Elena. *PNAS*, 101:8396–8401, 2004.

[13] D. R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. *Nat. Genet.*, 37:441–444, 2005.

[14] R. Kassen and T. Bataillon. *Nat. Genet.*, 38:484–488, 2006.

[15] R. C. MacLean, G. G. Perron, and A. Gardner. *Genetics*, 186:1345–1354, 2010.

[16] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman. *J Mol Evol*, 69:229, 2008.

[17] T. Bataillon, T. Zhang, and R. Kassen. *Genetics*, 189:939–949, 2011.

[18] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. *PLoS Genet.*, 8:e1002783, 2012.

[19] 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. *PLoS Genet*, 10(2), 2014.

[20] C. Bank, T. H. Ryan, D. J. Jeffrey, and N.A.B. Daniel. *Mol. Biol. Evol.*, 2014.

[21] C. Carneiro and D.L. Hartl. *PNAS*, 107:1747–1751, 2010.

[22] C. R. Miller, P. Joyce, and H.A. Wichman. *Genetics*, 187:185–202, 2011.

[23] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. M. de Visser. *J. Stat. Mech.*, -:P01005, 2013.

[24] F.J. Poelwijk, D.J. Kivet, D.M. Weinreich, and S.J. Tans. *Nature*, 445:383, 2007.

[25] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug. *PLoS Comp. Biol.*, 7:e1002134, 2011.

[26] D. J. Kvitek and G. Sherlock. *PLoS Genetics*, 7:e1002056, 2011.

[27] J. Lalić and S. F. Elena. *Heredity*, 109:71–77, 2012.

[28] H. A. Orr. *Evolution*, 56:1317–1330, 2002.

[29] H. A. Orr. *Evolution*, 60:1113, 2006.

[30] D. R. Rokyta, C. J. Beisel, and P. Joyce. *J Theor Biol.*, 243:114–120, 2006.

[31] J. Neidhart and J. Krug. *Phys. Rev. Lett.*, 107:178102, 2011.

[32] K. Jain. *EPL*, 96:58006, 2011.

[33] J. A. L. Filho, F. G. B. Moreira, P. R. A. Campos, and V. M. Oliveira. *J. Stat. Mech.*, -:P02014, 2012.

[34] J. J. Bull, M. R. Badgett, and H. A. Wichman. *Mol. Biol. Evol.*, 17:942–950, 2000.

[35] R. D. H. Barrett, R. Craig MacLean, and G. Bell. *Biol. Lett.*, 2:236–238, 2006.

[36] C. L. Burch and L. Chao. *Genetics*, 151:921–927, 1999.

[37] A. Sousa, S. Magalhães, and I. Gordo. *Mol. Biol. Evol.*, 29:1417–1428, 2012.

[38] S.E. Schoustra, T. Bataillon, D.R. Gifford, and R. Kassen. *PLoS Biol*, 7 (11):e1000250, 2009.

[39] D. R. Gifford, S. E. Schoustra, and R. Kassen. *Evolution*, 65:3070–3078, 2011.

[40] D. R. Rokyta, Z. Abdo, and H. A. Wichman. *J Mol Evol*, 69:229, 2009.

[41] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. *Genetics*, 148:1667–1686, 1998.

[42] P. J. Gerrish and R. E. Lenski. *Genetica*, 102:127–144, 1998.

[43] M.M. Desai and D.S. Fisher. *Genetics*, 176:1759–1798, 2007.

[44] J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim. *Nature*, 461:1243–1247, 2009.

[45] I. Gordo and P. R. A. Campos. *Biol Lett*, 9:20120239, 2012.

Chapter 4

Length of adaptive walk

4.1 Introduction

In this chapter, we address how the number of adaptive steps that a population takes before it gets trapped at a local fitness peak depends on the properties of the underlying fitness landscape. Similar to Chapter 3, here as well, we consider an asexual population in the *strong selection-weak mutation* regime where, the entire population can be represented by a single point in the fitness landscape, and performs an uphill *adaptive walk*. However, unlike in Chapter 3 where infinitely long sequences were considered, here we consider sequences of finite length for which the adaptive walk terminates once the population reaches a local fitness peak since a better fitness is at least two mutations away [1, 2]. The properties of the adaptive walk depend on the distribution of beneficial mutations which can be found by appealing to

the extreme value theory (EVT) [2] since beneficial mutations are rare and therefore lie in the tail of the full fitness distribution [3, 4]. For independent and identically distributed (i.i.d.) random variables, the EVT states that the distribution of the tails can belong to one of the three domains namely Weibull, Gumbel and Fréchet [5]. Here we study how the walk length depends on these three extreme value domains. Although fitnesses are known to be correlated, much of the previous work on the subject ignores correlations completely [6–9]. Here we also investigate how correlations affect the number of adaptive substitutions. Motivated by recent experiments on adaptive walks in which a maladapted population starts at different fitness [10–12], we also analyze the dependence of the walk length on the initial fitness.

The average walk length calculated using the two extreme cases of stochastic rules, the greedy adaptive walk (GAW) and the random adaptive walk (RAW) described in Chapter 2, show no dependence on the EVT domain. The average walk length $\bar{J}$ of the GAW has been calculated by appealing to the theory of records, and for infinitely long sequences, it turns out that [13]

$$\bar{J}_{GAW} = e - 1 \approx 1.718 \tag{4.1}$$

for any fitness distribution. On the other hand, in the case of RAW the average length of the walk diverges with the sequence length [14]. More

precisely, the average walk length for zero initial fitness is given by [14]

$$\bar{J}_{RAW} \approx \ln L + 1.099 \tag{4.2}$$

and is independent of the choice of the fitness distribution. But neither GAW or RAW is biologically realistic and according to the population genetics theory [15], the probability that a beneficial mutation will spread through the population increases with the relative fitness difference between the mutant and the parent exponentially fast towards unity. Two variants of the linear model have been studied: while [1] and [8] considered adaptation in a single fixed neighborhood where the mutants are produced only at the first step and the same are retained all through the walk, the model studied in [6, 7, 9, 16, 22] assumes that a new set of L fitnesses (corresponding to the fitness of the one mutant neighbours) are generated at each step of the walk. Though we shall use the latter model here, it is interesting to note that most results for the walk length are robust with respect to this assumption. Using (2.13), we numerically find that the adaptive walk is shortest in the Gumbel domain [16]. However when the relative fitness difference is assumed to be small, this probability is proportional to the relative fitness difference, and in this case, we find that the adaptive walks are shortest in the Fréchet domain and longest in the Weibull domain. Although the assumption of small fitness differences is biologically incorrect, especially in the Fréchet domain, it is still interesting to consider this model as it connects to other systems [8], such as

deterministically evolving populations [17, 18] and a gas of particles undergoing elastic collisions [19, 20], and lends itself to analytical calculations. We calculate the average walk length in all the three EVT domains for uncorrelated fitnesses, and show that it depends logarithmically on the initial rank of the population. Using results from the large deviation theory [21], we also obtain analytical expressions for the walk length for correlated fitnesses, and find that the walk lasts longer on correlated fitness landscapes as they have fewer local fitness peaks.

4.2 Walk length in the linear model

4.2.1 On uncorrelated fitness landscapes

For zero initial fitness, it has been shown that if the mean $\bar{f}$ of the fitness distribution $p(f)$ is finite, the walk length increases with the length of the sequence but remains constant otherwise [9, 22]. To understand this transition at $\kappa = 1$ above which $\bar{f}$ is infinite, here we present a simple argument and refer the reader to [9] for details. For $\kappa < 1$, as the transition probability (2.14) is nonzero for finite fitness differences, the adaptive walk goes on indefinitely for infinitely long sequence or in other words, the adaptive walk length diverges with the sequence length L. A calculation for zero initial fitness and large L shows that the walk length cumulants increase logarithmically with the sequence length [9]. In particular, the mean walk length $\bar{J}$

increases as [8, 9, 22]

$$\bar{J}(L|f_0 = 0) \approx \beta_\kappa \ln L \tag{4.3}$$

where

$$\beta_\kappa = \frac{1-\kappa}{2-\kappa} \ , \ \kappa < 1 \tag{4.4}$$

which shows that the walks are shorter for slowly decaying fitness distributions. For $\kappa > 1$, as the mean of the fitness distribution is infinite, the normalization sum in the denominator on the right hand side (RHS) of (2.14) is dominated by the largest value $\tilde{f}$ amongst L i.i.d. random variables (refer (2.9)). This implies that the transition occurs to one of the highly fit sequences with fitness of order $\tilde{f}$. Since the number of such sequences is of order unity, the walk terminates in a few steps resulting in a constant walk length.

As shown in Fig. 4.1, a similar transition is seen at $\kappa = 1$ when the sequence length is kept fixed and the initial fitness is varied. We now generalist the calculation in [9] for zero initial fitness to find how the average walk length changes with the initial fitness when $\kappa < 1$. Since the mean of the fitness distribution is finite when κ is below unity, for long sequences, we can calculate the walk length using (2.14) as detailed below.

An adaptive walk will stop at step J if all the L neighboring sequences have a fitness lower than that of the currently occupied sequence. Thus if $Q_J(L|f_0)$ is the probability that the adaptive walk of a sequence of length L

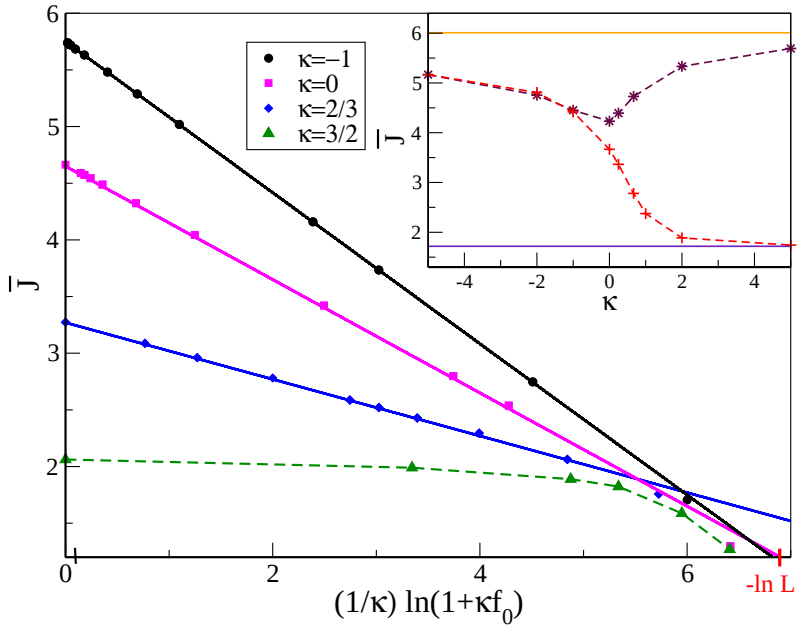


Figure 4.1: Main: Variation of the average walk length with initial fitness in the linear model on uncorrelated fitness landscapes for various κ. The simulation points are for $L = 1000$ and the lines are obtained from (4.18) for all $\kappa < 1$, while the one for $\kappa = 3/2$ is a guide to the eye. Inset: Comparison of the average walk length in the full model ($*$) and the linear model ($+$) for $(1/\kappa)\ln(1+\kappa f_0) = 2$. The solid line shows the walk length expressions (4.1) (bottom) and (4.2) (top) for the greedy walk and the random adaptive walk respectively.

lasts exactly J steps, we can write [22]

$$Q_J(L|f_0) = \int_{f_0}^{u} df\ q^L(f)\ \mathcal{P}_J(f|f_0) \tag{4.5}$$

where $\mathcal{P}_J(f|f_0)$ is the probability distribution of the fitness f at the Jth step, given the initial fitness f_0 which satisfies (2.16), and $q(f)$ is the cumulative

probability of having a fitness lower than f. For the transition probability (2.14), the integral equation (2.16) for the distribution $\mathcal{P}_J(f|f_0)$ appearing in (4.5) can be recast as a second order differential equation for the distribution $P_J(f|f_0)$ defined through $\mathcal{P}_J(f|f_0) = p(f)P_J(f|f_0)$ as given in (2.17) [22]. Although we are unable to analyse (2.17) when L is finite, as explained below, it is possible to extract useful information from it when the sequence is infinitely long and using the fact that for a finite sequence, there is a characteristic fitness scale $\tilde{f}$ given by (2.9).

We first introduce the generating function $G(x, f) = \sum_{J=1}^{\infty} P_J(f)\ x^J, x < 1$ which, due to (2.17), obeys the following differential equation:

$$G''(x, f) = \frac{x(1-\kappa)(1-q^L(f))}{(1+\kappa f)^2} G(x, f) \tag{4.6}$$

and is subject to the initial conditions (2.18) and (2.19). In the above equation, the cumulative probability $q^L(f)$ of the maximum value distribution is a smoothly varying function that increases from zero to one, as the fitness f increases and belongs to one of the three EVT domains. For the cumulative fitness distribution (2.5), we find that for large L [5]

$$q^L(f) \approx e^{-(\frac{1+\kappa f}{1+\kappa \tilde{f}})^{-\frac{1}{\kappa}}} = \begin{cases} e^{-z^{-\frac{1}{\kappa}}} & ,\ \kappa < 0 \quad \text{(Weibull)} \qquad (4.7a) \\ e^{-e^{-z}} & ,\ \kappa \to 0 \quad \text{(Gumbel)} \qquad (4.7b) \\ e^{-z^{-\frac{1}{\kappa}}} & ,\ \kappa > 0 \quad \text{(Fréchet)} \qquad (4.7c) \end{cases}$$

where

$$z(f) = \begin{cases} f - \tilde{f} & , \kappa \to 0 \quad (4.8a) \\ (1+\kappa f)(1+\kappa \tilde{f})^{-1} & , \kappa \neq 0 \quad (4.8b) \end{cases}$$

It is useful to consider (4.6) as a function of z defined above. If $\tilde{z} \equiv z(\tilde{f})$, the general solution of the differential equation (4.6) may be written as

$$G(x,z) = \begin{cases} a_1 g_1(x,z) + a_2 g_2(x,z) & , z < \tilde{z} \quad (4.9a) \\ b_1 h_1(x,z) + b_2 h_2(x,z) & , z > \tilde{z} \quad (4.9b) \end{cases}$$

where g_i, h_i satisfy (4.6), and the constants a_1, a_2 are determined in A using the initial conditions at $z_0 \equiv z(f_0) < \tilde{z}$. The other constants of integration b_1, b_2 can be found by matching the solution $G(x,z)$ and its first derivative (w.r.t. z) at $z = \tilde{z}$. Noting that $\tilde{z}$ is constant in L and f_0 but z_0 depends on them, we find that the constants b_1, b_2 are of the form

$$b_i = b_{i1}(x) a_1(z_0) + b_{i2}(x) a_2(z_0) \ , \ i = 1,2 \quad (4.10)$$

To find the properties of the walk length, we next define a generating function H for the walk length distribution (4.5) as

$$H(x,L) = \sum_{J=1}^{\infty} Q_J(L|f_0)\ x^J \quad (4.11)$$

$$= \int_{z(f_0)}^{z(u)} dz\ p(z)\ \frac{dz}{df}\ q^L(z)\ G(x,z) \quad (4.12)$$

On approximating $q^L(z)$ for $z < \tilde{z}$ by zero, we get

$$H(x,L) \approx \int_{\tilde{z}}^{z(u)} dz\ p(z)\ \frac{dz}{df}\ q_>^L(z)\ G_>(x,z) \tag{4.13}$$

where the subscript $>$ is used to denote the quantities when $z > \tilde{z}$. Using (4.8) and (4.10), we can extract the z_0-dependence of the generating function and find that

$$H(x,L) = \begin{cases} a_1(z_0)R_1(x) + a_2(z_0)R_2(x) & ,\ \kappa \to 0 \quad (4.14a) \\ \dfrac{\kappa}{1+\kappa\tilde{f}}(a_1(z_0)R_1(x) + a_2(z_0)R_2(x)) & ,\ \kappa \neq 0 \quad (4.14b) \end{cases}$$

where

$$R_i(x) = \int_{\tilde{z}}^{z(u)} dz\ p(z)\ q_>^L(z)\ \sum_{j=1}^{2} b_{ji}\ h_j(x,z) \tag{4.15}$$

is independent of L and f_0. Furthermore, from the explicit expressions for a_1 and a_2 discussed in Appendix A, we see that a_2 decays more rapidly with L than a_1, and therefore we may neglect the second term on the RHS of (4.14a) and (4.14b) for large L. Since the nth cumulant μ_n of the walk length is given by [5]

$$\mu_n(L) = \left.\frac{d^n \ln H}{dX^n}\right|_{X=0} \tag{4.16}$$

where $X = \ln x$, to leading order in L, we finally obtain

$$\mu_n(L) \approx \begin{cases} (\ln L - f_0)\left.\dfrac{d^n}{dX^n} e^{X/2}\right|_{X=0} & ,\ \kappa \to 0 \quad (4.17a) \\ \dfrac{1}{2\kappa}\ln\left(\dfrac{L^\kappa}{1+\kappa f_0}\right)\left.\dfrac{d^n}{dX^n}\sqrt{\kappa^2 + 4e^X(1-\kappa)}\right|_{X=0} & ,\ \kappa \neq 0 \quad (4.17b) \end{cases}$$

Setting $n = 1$ in our final result (4.17), we find the average walk length to be

$$\bar{J}(L|f_0) = \beta_\kappa \left(\ln L - \frac{1}{\kappa} \ln(1 + \kappa f_0) \right) + c_\kappa \tag{4.18}$$

where β_κ is given by (4.4) and the constant c_κ in which the subleading corrections in L are subsumed is determined numerically. We check that the results of [22] and [9] for $f_0 = 0$ are reproduced from the above equation. We also note that since the *typical* rank m of a fitness (with the fittest ranked one) is given by [5]

$$m_0 = \frac{L}{(1 + \kappa f_0)^{\frac{1}{\kappa}}} = \left(\frac{1 + \kappa \tilde{f}}{1 + \kappa f_0} \right)^{\frac{1}{\kappa}} \tag{4.19}$$

our result (4.18) gives $\bar{J} = \beta_\kappa \ln m_0 + c_\kappa$. Thus the effect of nonzero initial fitness is to replace the sequence length L in (4.3) for zero initial fitness (where all the mutants are fitter) by the average number of mutants present at the beginning of the walk. The logarithmic dependence of the walk length on the initial rank has been obtained in [1, 8] using a model in which both the initial rank m and the mutational neighborhood are *fixed*. Here instead the initial fitness is fixed, but the initial rank is a random variable and a new suite of mutants is generated at every step in the walk. The fact that the same basic result is obtained in the deterministic and stochastic model shows that the stochastic effects are rather unimportant on an average as noted in previous works as well [6, 8].

Our numerical results for the average walk length on uncorrelated fitness landscapes are compared with (4.18) in Fig. 4.1 where the numerical fits for constants c_κ for $\kappa = -1$, 0 and 2/3 are 1.15, 1.21 and 1.55 respectively. We see a good match between the simulation data and (4.18) except when the initial fitnesses are close to the local fitness optimum where the simulation data lies below the theoretical results. This discrepancy may be due to the fact that the approximation $q(f) = 0$ is good for fitnesses far below the local fitness peak, while we have used it for all $f < \tilde{f}$ to arrive at (4.13).

4.2.2 On correlated fitness landscapes

In the above discussion, we have assumed that the sequence fitnesses are uncorrelated. We now discuss how the walk length changes when correlated fitnesses generated using a block model (described in Chapter 2) are considered. The initial fitness of the whole sequence built of B blocks each of fitness $f_0^{(b)}$, $1 \leq b \leq B$, is given in 2.15. Since the block fitnesses evolve independently, the average walk length is the sum of the mutations accumulated by each block [22, 24]. Thus the average walk length $\bar{J}_B$ for a sequence composed of B blocks is given by

$$\bar{J}_B(L|f_0) = \sum_{b=1}^{B} \bar{J}(L_B|f_0^{(b)}) \tag{4.20}$$

where $\bar{J}(L_B|f_0^{(b)})$ is the average walk length for a sequence of length L_B with initial fitness $f_0^{(b)}$ on uncorrelated fitness landscapes. In the simplest

situation where the initial fitness $f_0^{(b)}$ of each block is same, we immediately have [22, 24]

$$\bar{J}_B(L|f_0) = B\bar{J}(L_B|f_0) \tag{4.21}$$

However if the block fitnesses are random variables that satisfy (2.15), an average over the joint distribution $P_B(\{f_0^{(b)}\})$ of block fitnesses is also required. We thus have

$$\bar{J}_B(L|f_0) = \int_0^u df_0^{(1)} ... \int_0^u df_0^{(B)} P_B(\{f_0^{(b)}\}) \sum_{b=1}^{B} \bar{J}(L_B|f_0^{(b)}) \tag{4.22}$$

Since the block fitnesses are i.i.d. random variables subject to the constraint (2.15), the distribution of block fitnesses can be written as

$$P_B(\{f_0^{(b)}\}) = \frac{\prod_{b=1}^{B} p(f_0^{(b)})}{\mathcal{N}_B(Bf_0)} \; \delta(Bf_0 - \sum_{i=1}^{B} f_0^{(i)}) \tag{4.23}$$

where the normalization constant $\mathcal{N}_B(X)$ is the distribution of the sum of B random variables given by

$$\begin{aligned} \mathcal{N}_B(X) &= \int_0^u df_0^{(1)} ... \int_0^u df_0^{(B)} \prod_{b=1}^{B} p(f_0^{(b)}) \; \delta\left(X - \sum_{i=1}^{B} f_0^{(i)}\right) && (4.24) \\ &= \int_0^X df \; p(f) \; \mathcal{N}_{B-1}(X - f) && (4.25) \end{aligned}$$

with $\mathcal{N}_0(f) = \delta(f)$. Thus we can express the average walk length as

$$\bar{J}_B(L|f_0) = B(\beta_\kappa \ln L_B + c_\kappa) - \frac{\beta_\kappa B}{\kappa} \frac{\int_{l_1}^{l_2} df\ p(f) \ln(1+\kappa f)\ \mathcal{N}_{B-1}(Bf_0 - f)}{\mathcal{N}_B(Bf_0)} \tag{4.26}$$

where the integration limits are $l_1 = 0, l_2 = Bf_0$ in the Gumbel and Fréchet domains. In the Weibull domain, three cases arise: (i) if $Bf_0 < u$, the limits are $l_1 = 0, l_2 = Bf_0$, (ii) if $u < Bf_0 < (B-1)u$, we have $l_1 = 0, l_2 = u$ and (iii) if $(B-1)u < Bf_0 < Bu$, the limits are $l_1 = Bf_0 - (B-1)u, l_2 = u$.

Exactly solvable case

For exponentially distributed fitnesses, the distribution $\mathcal{N}_B(X)$ in (4.24) is known exactly to be [25]

$$\mathcal{N}_B(X) = e^{-X} \frac{X^{B-1}}{(B-1)!} \tag{4.27}$$

Taking the limit $\kappa \to 0$ in (4.26), we find the average walk length as

$$\begin{aligned} \bar{J}_B(L|f_0) &= B(\beta_0 \ln L_B + c_0) - B\beta_0 \frac{\int_0^{Bf_0} df\ e^{-f}\ f\ \mathcal{N}_{B-1}(Bf_0 - f)}{\mathcal{N}_B(Bf_0)} && (4.28) \\ &= B(\beta_0 \ln L_B + c_0) - B\beta_0 \frac{e^{-Bf_0}(Bf_0)^B}{B!\ \mathcal{N}_B(Bf_0)} && (4.29) \\ &= B\bar{J}(L_B|f_0) && (4.30) \end{aligned}$$

which is the same as that in the case where each block fitness is f_0 (refer (4.21)).

Weakly correlated fitnesses

For $\kappa \neq 0$, it appears difficult to obtain exact expressions for the walk length for correlated fitnesses. The case of two independent blocks ($B = 2$) presents the simplest model for correlated fitnesses, and we discuss this here. The distribution $\mathcal{N}_2(X)$ of two random variables is given by

$$\mathcal{N}_2(X) = \begin{cases} \int_0^X dg\ p(g)\ p(X-g)\ ,\ X < u & \text{(4.31a)} \\ \int_{X-u}^u dg\ p(g)\ p(X-g)\ ,\ X > u & \text{(4.31b)} \end{cases}$$

For $2f_0 < u$, using (4.31a) in the expression (4.26), we get

$$\begin{aligned} \frac{\bar{J}_2(L|f_0)}{B} &= \beta_\kappa \ln L_B + c_\kappa - \frac{\beta_\kappa}{\kappa} \frac{\int_0^{2f_0} df\ p(f) \ln(1+\kappa f)\ p(2f_0 - f)}{\int_0^{2f_0} df\ p(f)\ p(2f_0 - f)} & \text{(4.32)} \\ &= \beta_\kappa \ln L_B + c_\kappa - \frac{\beta_\kappa}{\kappa} \ln(1+\kappa f_0) + \frac{\beta_\kappa}{2\kappa} \mathcal{I}_\kappa(w_0) & \text{(4.33)} \end{aligned}$$

where the integral

$$\mathcal{I}_\kappa(w_0) = \frac{\int_1^{w_0} dz\ \ln z\ z^{\frac{1-\kappa}{\kappa}} (1-z^{-1})^{-1/2}}{\int_1^{w_0} dz\ z^{\frac{1-\kappa}{\kappa}} (1-z^{-1})^{-1/2}} \qquad \text{(4.34)}$$

with $w_0 = (1+\kappa f_0)^2/(1+2\kappa f_0)$. Note that for large initial fitnesses $f_0 \sim u/2$, the function $w_0 \gg 1$.

Fréchet class: For positive κ and large f_0, an approximate expression for the

integral $\mathcal{I}_\kappa(w_0)$ can be obtained after an integration by parts, and we get

$$\frac{\bar{J}_2(L|f_0)}{B} \approx \bar{J}(L_B|f_0) + \frac{\beta_\kappa}{2\kappa}(\ln w_0 - \kappa) \tag{4.35}$$

$$\approx \beta_\kappa \ln L_B + c_\kappa - \frac{\beta_\kappa}{2\kappa}(\ln(\kappa f_0) + \ln 2 + \kappa) \tag{4.36}$$

Weibull class: The integral $\mathcal{I}_\kappa(w_0)$ can be calculated exactly for uniformly distributed fitnesses and is given by

$$\mathcal{I}_{-1}(w_0) = 2 + \ln w_0 - 2\sqrt{\frac{w_0}{w_0 - 1}}\sinh^{-1}(\sqrt{w_0 - 1}) \tag{4.37}$$

$$\approx 2(1 - \ln 2) - \frac{\ln w_0}{2w_0}\ ,\ w_0 \gg 1 \tag{4.38}$$

For arbitrary negative κ, we note that the integral $\mathcal{I}_\kappa(w_0)$ is finite when $w_0 \to \infty$ and can be written in terms of the harmonic number $H_n = n\sum_{i=1}^{\infty}(i(n+i))^{-1}$ [26]. An integration by parts then yields

$$\mathcal{I}_\kappa(w_0) \approx H_{-\frac{1}{\kappa}-\frac{1}{2}} - H_{-\frac{1}{\kappa}-1} + \frac{\kappa\Gamma(\frac{1}{2}-\frac{1}{\kappa})}{\sqrt{\pi}\Gamma(-\frac{1}{\kappa})}\ \frac{\ln w_0}{w_0^{1/|\kappa|}} \tag{4.39}$$

which matches the result for $\kappa = -1$ as $H_{1/2} = 2 - \ln 4$. Ignoring the last term on the RHS of the above equation which decays with f_0, we find that the average walk length can be written as

$$\frac{\bar{J}_2(L|f_0)}{B} = \bar{J}(L_B|f_0) + \frac{\beta_\kappa}{2\kappa}(H_{-\frac{1}{\kappa}-\frac{1}{2}} - H_{-\frac{1}{\kappa}-1})\ ,\ f_0 \lesssim u/2 \tag{4.40}$$

For $f_0 > u/2$ where $\mathcal{N}_2(X)$ is given by (4.31b), the integrals can be done

exactly and we have

$$\frac{\bar{J}_2(L|f_0)}{B} = \beta_\kappa \ln L_B + c_\kappa - \frac{\beta_\kappa}{\kappa} \frac{\int_{-1}^{1} dh\ (1-h^2)^{-\frac{1+\kappa}{\kappa}}\ (\ln(1+\kappa f_0) + \ln(1+h))}{\int_{-1}^{1} dh\ (1-h^2)^{-\frac{1+\kappa}{\kappa}}} \tag{4.41}$$

$$= \beta_\kappa \ln L_B - \frac{\beta_\kappa}{\kappa} \ln(1+\kappa f_0) + c_\kappa + \frac{\beta_\kappa}{2\kappa}(H_{-\frac{1}{\kappa}-\frac{1}{2}} - H_{-\frac{1}{\kappa}-1}) \tag{4.42}$$

$$= \bar{J}(L_B|f_0) + \frac{\beta_\kappa}{2\kappa}(H_{-\frac{1}{\kappa}-\frac{1}{2}} - H_{-\frac{1}{\kappa}-1})\ ,\ f_0 > u/2 \tag{4.43}$$

For bounded distributions, although the walk length is continuous at initial fitness equal to $u/2$, it is interesting to note that it is not differentiable. For uniformly distributed fitnesses where exact expressions for the walk length can be calculated, the average walk length obtained from (4.33) and (4.43) is found to be the same at $f_0 = 1/2$. The first derivative of the walk length (with respect to f_0) is given by

$$\frac{d\bar{J}_2}{df_0} = \begin{cases} \frac{2\beta_{-1}}{f_0^2}\left[f_0 + \frac{1}{2}\ln(1-2f_0)\right] & ,\ f_0 < 1/2 \quad \text{(4.44a)} \\ -\frac{2\beta_{-1}}{1-f_0} & ,\ f_0 > 1/2 \quad \text{(4.44b)} \end{cases}$$

From the above equation, we see that while the derivative at $f_0 = 1/2$ obtained from (4.44a) is undefined, the expression (4.44b) yields a finite constant. For general $\kappa < 0$, the derivative of the walk length calculated using (4.43) is seen to be finite, while it diverges when (4.33) is used.

On strongly correlated fitness landscapes

We now turn to the situation when the block number $B \gg 1$. To calculate the integral in (4.26), let us first consider the integrand

$$\mathcal{F}(f) = p(f) \ln(1 + \kappa f)^{1/\kappa} \, \mathcal{N}_{B-1}(Bf_0 - f) \tag{4.45}$$

The first two factors on the RHS are obviously independent of B and f_0. However for all $\kappa < 1$ where the fitness distribution has a finite mean $\bar{f}$, the last factor peaks about the mean $B(f_0 - \bar{f})$ of the sum distribution which increases with both B and f_0. Then for large enough B and f_0, the integrand $\mathcal{F}(f)$ gets a contribution from the *lower tail* of the sum distribution instead of the region around its mean. The behavior of the tail of the sum distribution can be obtained by applying a large deviation principle if the fitness distribution possesses all finite moments as is the case for $\kappa \leq 0$. However for power law distributions with $\kappa > 0$, the $(1/\kappa)$-th and higher moments diverge and the large deviation principle is not applicable, and in this case, we use the result that the sum distribution decays as the fitness distribution itself [5, 21]. The fact that the central limit theorem for the sum distribution does not capture the correct behavior of the integral under question is illustrated in Appendix B for exponentially distributed fitnesses.

To calculate the walk length using the large deviation theory, we first consider a normalised distribution with support on the interval $[0, u]$ defined

as

$$g(t) = \kappa(\alpha - 1)(1 + \kappa t)^{-\alpha} \tag{4.46}$$

where $\alpha < 1, u = -1/\kappa$ for $\kappa < 0$ and $\alpha > 1, u = \infty$ when $\kappa \geq 0$. Then the distribution of the sum of B i.i.d. random variables chosen from $g(t)$ is given by

$$I_B(X;\alpha) = \int_0^u dt^{(1)} ... \int_0^u dt^{(B)} \prod_{j=1}^{B} g(t^{(j)})\ \delta\left(X - \sum_{i=1}^{B} t^{(i)}\right) \tag{4.47}$$

Differentiating on both sides w.r.t. α, we get

$$\frac{\partial I_B(X;\alpha)}{\partial \alpha} = \frac{B I_B(X;\alpha)}{\alpha - 1} - B \int_0^{\min(X,u)} dt\ g(t) \ln(1 + \kappa t)\ I_{B-1}(X - t;\alpha) \tag{4.48}$$

The upper limit in the above integral is X for unbounded distributions. But for bounded distributions, when correlations are strong (large B), the limits in case (ii) described below (4.26) apply. On dividing the above equation by $I_B(X;\alpha)$, it follows that the average walk length (4.26) can be written as

$$\bar{J}_B(L|f_0) = B(\beta_\kappa \ln L_B + c_\kappa) - \frac{\beta_\kappa B}{\kappa}\left(\frac{1}{\alpha - 1} - \frac{\partial}{\partial \alpha}\frac{\ln I_B(Bf_0;\alpha)}{B}\right)\Bigg|_{\alpha = 1 + \frac{1}{\kappa}} \tag{4.49}$$

Our task is now reduced to finding the sum distribution $I_B(X)$ for the various EVT domains which we describe below.

Weibull class: According to the large deviation principle, for large B, the

distribution $I_B(X)$ is of the form [21],

$$I_B(X \simeq Bx) \sim e^{Br(x)} \tag{4.50}$$

where the rate function $r(x)$ can be determined as described below. On using the integral representation of the Dirac delta function in (4.47), we get

$$\begin{aligned} I_B(X) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\ e^{ikX}\left(\int_{-\infty}^{\infty} dy\ e^{-iky}g(y)\right)^B && (4.51)\\ &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\omega\ e^{B(\omega x+\ln\tilde{g}(\omega))} && (4.52)\end{aligned}$$

where $\tilde{g}(\omega) = \int_0^\infty dt\ g(t)\ e^{-\omega t}$ is the Laplace transform of the distribution function $g(t)$. Evaluating the RHS of (4.52) using the saddle point method for large B [23], we get

$$\frac{\ln I_B(X)}{B} = r(x) = \omega_* x + \ln\tilde{g}(\omega_*) \tag{4.53}$$

where the saddle point ω_* is real and given by

$$\left.\frac{d\ln\tilde{g}}{d\omega}\right|_{\omega=\omega_*} = -x \tag{4.54}$$

The Laplace transform of the distribution $g(t)$ in (4.46) is given by

$$\tilde{g}(\omega) = e^{\eta}\left[(\alpha-1)E_\alpha(\eta) + \eta^{\alpha-1}\Gamma(2-\alpha)\right] \tag{4.55}$$

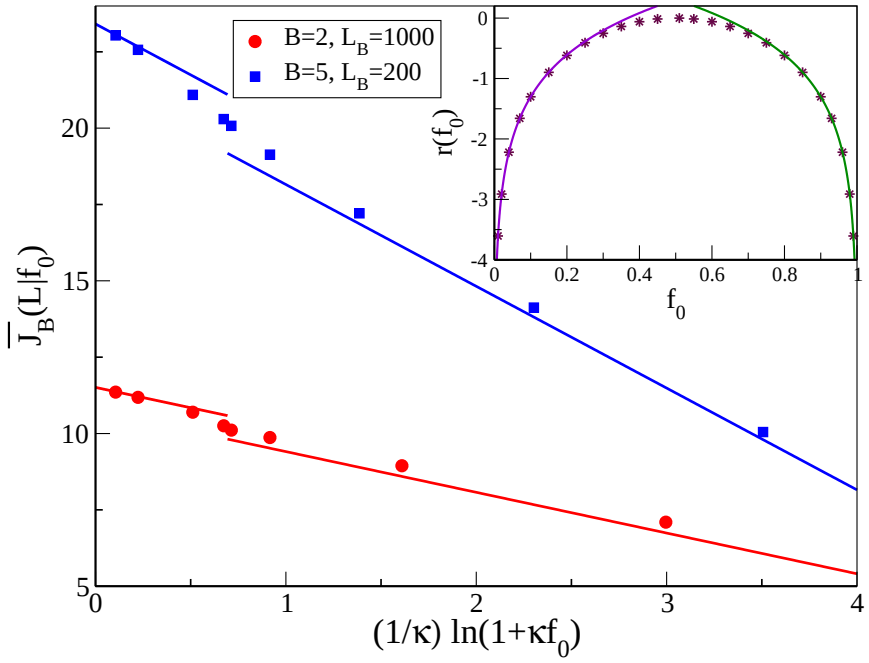


Figure 4.2: Main: Plot shows the variation of the average walk length with initial fitness for the linear model on correlated fitness landscapes for various B when $\kappa = -1$. The theoretical predictions (4.61) and (4.63) (lines) are compared against the simulation data (points). Inset: Plot shows the rate function for $\kappa = -1$ obtained using (4.53) and (4.56) (points) and the analytical formulae (4.58) and (4.59).

and the function $\omega_*(f_0)$ is a solution of the equation

$$\mathcal{T}(\omega_*) = \frac{(\alpha-1)\omega^2(E_{\alpha-1}(\eta) - E_\alpha(\eta)) - \eta^\alpha\kappa^2(\eta+\alpha-1)\Gamma(2-\alpha)}{\omega\kappa((\alpha-1)\omega E_\alpha(\eta) + \eta^\alpha\kappa\Gamma(2-\alpha))}\Bigg|_{\omega=\omega_*} = f_0 \tag{4.56}$$

where $\eta = \omega/\kappa$, $E_\alpha(\eta) = \int_1^\infty dx\ e^{-\eta x}x^{-\alpha}$ is the exponential integral and $\Gamma(n+1) = n!$ is the gamma function. The function $\mathcal{T}(\omega_*)$ in the above equation decreases from its maximum value $-1/\kappa$ to zero as ω_* is increased

from $-\infty$ to ∞. Using the asymptotic expansion of the exponential integral [27], we find that

$$\mathcal{T}(\omega_*) = \begin{cases} -\kappa^{-1} + (1-\alpha)\omega_*^{-1} \ , \ \omega_* \to -\infty & (4.57a) \\ \omega_*^{-1} \ , \ \omega_* \to \infty & (4.57b) \end{cases}$$

When the initial fitness is large (small), f_0 equals the left hand side (LHS) of (4.56) when ω_* is negative (positive). Then using (4.57a) and (4.57b) in (4.53), we find the rate function to be

$$r(f_0) \approx \begin{cases} 1 + \ln((\alpha-1)\kappa f_0) + \ln(1-\alpha\kappa f_0) & (4.58) \\ 1 - \alpha - (1-\alpha)\ln\left(\dfrac{1-\alpha}{1+\kappa f_0}\right) + \ln(\Gamma(2-\alpha)) & (4.59) \end{cases}$$

(4.58) and (4.59) are for $f_0 \ll \mathcal{T}(0)$ and $f_0 \gg \mathcal{T}(0)$ respectively where $\mathcal{T}(0) = (1-\kappa)^{-1}$. The above expression for the rate function is compared against the results from numerical simulations for uniformly distributed fitnesses in the inset of Fig. 4.2, and we see a good agreement for $f_0 < 0.3$ and > 0.7. For small f_0, using (4.49), we obtain

$$\begin{aligned} \frac{\bar{J}_B(L|f_0)}{B} &= \beta_\kappa \ln L_B + c_\kappa - \frac{\beta_\kappa f_0}{1-(1+\kappa)f_0} & (4.60) \\ &\approx \bar{J}(L_B|f_0) & (4.61) \end{aligned}$$

while for large f_0, we get

$$\frac{\bar{J}_B(L|f_0)}{B} = \beta_\kappa \ln L_B + c_\kappa - \frac{\beta_\kappa}{\kappa}(\kappa + \ln(-\kappa) + \ln(1+\kappa f_0) + H_{-\frac{1}{\kappa}} - \gamma) \quad (4.62)$$

$$= \bar{J}(L_B|f_0) - \frac{\beta_\kappa}{\kappa}(\kappa + \ln(-\kappa) + H_{-\frac{1}{\kappa}} - \gamma) \quad (4.63)$$

where the Euler-Mascheroni constant $\gamma \approx 0.577$. The walk length expressions above can be succinctly written as

$$\bar{J}_B(L|f_0) = \bar{J}_B(L|0) - \frac{B\beta_\kappa}{\kappa}\ln(1+\kappa f_0) \quad (4.64)$$

and shows that the walk for nonzero fitness is shorter, as one would intuitively expect. For $\kappa = -1$, the equations (4.60) and (4.62) are compared against the numerical results in Fig. 4.2, and we see that the theoretical prediction for the walk length matches the simulation results quite well in the range of initial fitness values where the rate function agrees.

Fréchet class: In this case, the sum distribution (4.47) for large Bf_0 is given by [5]

$$I_B(Bf_0;\alpha) \sim Bg(Bf_0) \quad (4.65)$$

whose tail behavior is the same as that of the fitness distribution $g(f)$. Using this in (4.49), we immediately find

$$\bar{J}_B(L|f_0) = B(\beta_\kappa \ln L_B + c_\kappa) - \frac{\beta_\kappa}{\kappa}(\ln(1+B\kappa f_0) + \kappa(B-1)) \quad (4.66)$$

$$\approx B(\beta_\kappa \ln L_B + c_\kappa) - \frac{\beta_\kappa}{\kappa}(\ln(\kappa f_0) + \ln B + \kappa(B-1)) \quad (4.67)$$

We note that the above answer matches with (4.36) for the two block model discussed in the last subsection. The above equation states that the average walk length decreases logarithmically with initial fitness but, unlike in the Weibull and Gumbel domain, the coefficient of $\ln f_0$ does not scale with the number of blocks. Thus in this case

$$\bar{J}_B(L|f_0) = \bar{J}_B(L|0) - \frac{\beta_\kappa}{\kappa} \ln(1 + B\kappa f_0) \tag{4.68}$$

In Fig. 4.3, the above expression is compared with the simulation data for $\kappa = 2/3$, and we see a good quantitative agreement between the theory and the simulations.

4.3 Walk length in the full model

As mentioned in the previous chapters, the transition probability (2.14) used to calculate the walk length is valid only when the relative fitness difference is small. However, large fitness differences during successive steps in the walk can occur when the initial fitness is small or if the fitness distribution has a fat tail [16]. In such cases, the approximation (2.14) breaks down, and we should consider the full transition probability (2.13). We have not been able to obtain analytical results for this model, and present our simulation results below.

As in the linear model, the walks are long for the full model when the

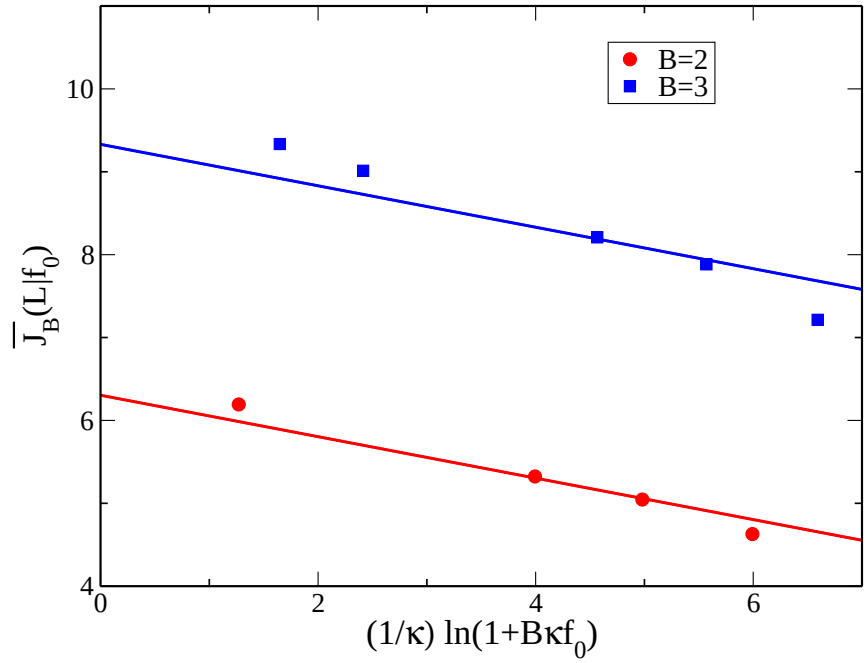


Figure 4.3: Plot shows the variation of the average walk length with the initial fitness for the linear model on correlated fitness landscapes for various B when $\kappa = 2/3$ and $L_B = 1000$. The theoretical prediction (4.66) (lines) is compared against the simulation data (points).

initial fitness is low or when the fitness are correlated [16]. However qualitative difference between the linear and the full model is seen with regard to the walk length dependence on the extreme value domain. As explained in Sec. 4.2.1, the divergence of the denominator on the RHS of (2.14) is responsible for the independence of the walk length on the initial fitness when $\kappa > 1$ in the linear model. However the normalization constant in (2.13) remains finite for all κ and therefore the walk length always decreases with increasing f_0 here. In the full model, for an infinitely long sequence, the

walk goes on forever for all κ [22] but similar to what happens in the linear model, for finite L, the walk terminates at a local fitness peak and the walk length is expected to increase with the sequence length. Here we are unable to analytically calculate the average walk length and present our numerical results in Fig. 4.4. For uncorrelated fitnesses, we find that in all the extreme value domains, the average walk length decreases with increasing f_0 due to decreasing availability of beneficial mutations at higher initial fitnesses. The simulation results also indicate that the average walk length $\tilde{\bar{J}}(L|f_0)$ has a logarithmic dependence on the rank $m_0 = L(1+\kappa f_0)^{-1/\kappa}$ of the initial fitness as given in (4.19). Thus we can write

$$\tilde{\bar{J}}(L|f_0) = \tilde{\beta}_\kappa \ln m_0 + \tilde{c_\kappa} \tag{4.69}$$

where $\tilde{\beta}_\kappa$ and $\tilde{c_\kappa}$ depend on the exponent κ and the block number B. Interestingly, the prefactor $\tilde{\beta}_\kappa$ has a *nonmonotonic* dependence on the exponent κ: with increasing κ, it decreases in the Weibull domain and increases in the Fréchet domain with a minimum occurring in the Gumbel domain. As shown in the inset of Fig. 4.4, on correlated fitness landscapes, the adaptive walks are longer than those on uncorrelated ones since the number of local fitness peaks decrease with increasing correlations [28]. Furthermore the average walk length $\tilde{\bar{J}}_B(L|f_0)$ seems roughly linear in $\ln m_0$ in all the three extreme value domains with a slope that depends nonmonotonically on exponent κ.

In the previous chapter, we saw that the behaviour of fitness fixed can be

understood using the small selection coefficient approximation in the Weibull and Gumbel domains. Below we will compare our results in Fig. 4.4 with those obtained in the linear model that assumes that the selective effects are small. In section 4.2, the analytical expressions for average walk length using (2.14) were obtained for both uncorrelated and correlated fitnesses [8,9,22,29] and it was shown that a transition occurs in the behaviour of the walk length at $\kappa = 1$. For $\kappa < 1$ where the mean of the fitness distribution (2.8) is finite, the average walk length calculated using the transition probability (2.14) was found to show a logarithmic dependence on the rank of the initial fitness as given in (4.18). An intuitive understanding of the logarithmic dependence of the walk length in the domain of κ where its variance is finite can be obtained by equating the fitness fixed at the final step $\bar{J}$ to the average fitness (2.10) of a local fitness maximum [14, 22]. Note that β_κ in (4.4) decreases monotonically with the exponent κ. For $\kappa > 1$ where the mean of the fitness distribution (2.8) is undefined, the walk length is found to be independent of the initial fitness rank. Our analytical calculations for the average walk length on correlated fitness landscapes [29] show that the length of the adaptive walk increases with increasing block number B and there is a monotonic decrease in β_κ with increasing κ.

We now exploit the results summarised above to understand the walk length behaviour when the transition probability is given by (2.13). Fig. 4.4 shows that in the Weibull domain, for uncorrelated fitnesses, the simulation data and the expression (4.18) are in good agreement and for correlated

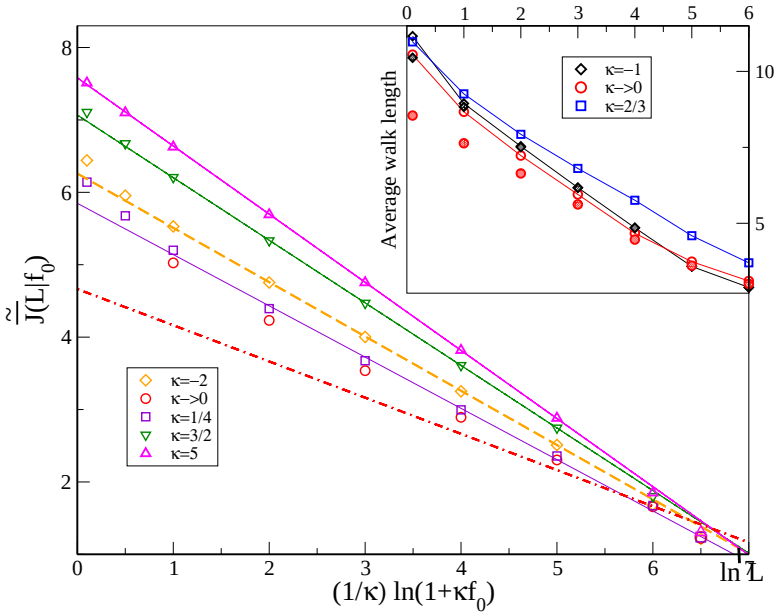


Figure 4.4: The plot shows the variation of the average walk length with initial fitness for various κ on uncorrelated (main) and correlated fitness landscapes with $B = 2$ (inset) for the full model. In the main plot, the broken lines show the result (4.18) with the constants $\tilde{c}_\kappa = 1.08$ and 1.21 for $\kappa = -2$ and $\to 0$ respectively, while the solid lines are the best fit to (4.69) with $\tilde{\beta}_\kappa \approx 0.71, 0.86, 0.94$ for $\kappa = 1/4, 3/2$ and 5 respectively. In the inset, the open symbols give the simulation data points of the average walk length obtained using the transition probability (2.13) while the shaded ones are those obtained using the transition probability (2.14) in the small selection coefficient approximation. In all the simulations, the sequence length $L = 1000$.

fitnesses, when the numerical data obtained using (2.13) is plotted with the ones using (2.14), the two data sets coincide for a wide range of initial fitness. Similar plots in the Gumbel domain for uncorrelated and correlated fitnesses show that the small selection coefficient approximation does not work as well as in the Weibull domain. In the Fréchet domain, the results (4.18) and (4.4)

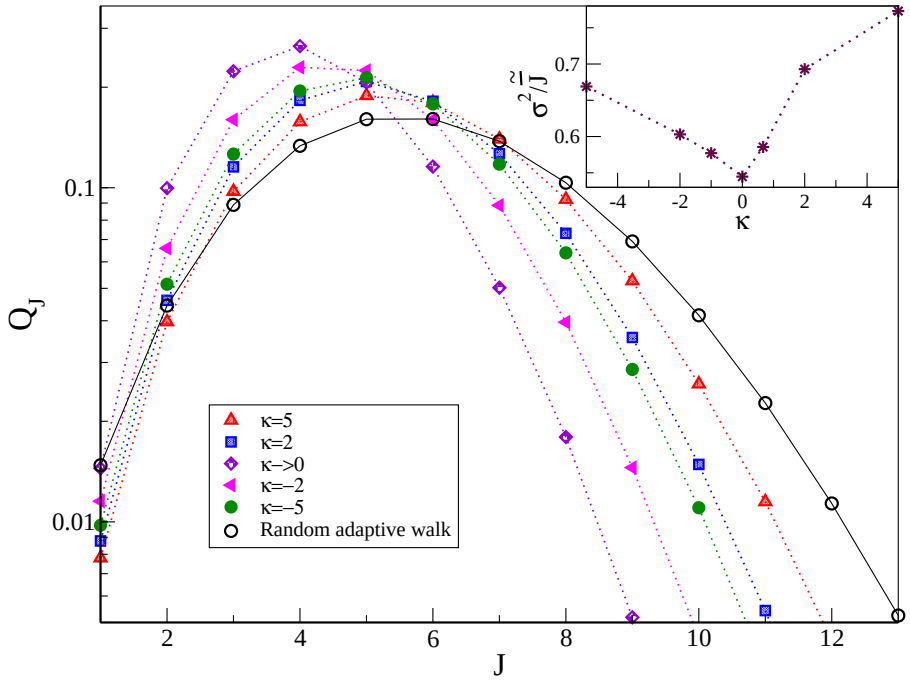


Figure 4.5: Plot to show the simulation data for the walk length distribution (main) and the index of dispersion (inset) of the walk length for various κ when $L = 1000$ using the full model on uncorrelated fitness landscapes. In the inset, $(1/\kappa)\ln(1 + \kappa f_0) = 2$.

obtained by neglecting the large effect mutations predict decreasing walk length with increasing κ, a trend opposite to that seen in Fig. 4.4. That the walk length $\bar{J}$ should be longer than $\tilde{\bar{J}}$ is expected - as the step length (3.45c) is smaller than that given by (3.8), more adaptive steps to a local fitness peak can be taken in the former case. To understand the trend of coefficient $\tilde{\beta}_\kappa$, it is useful to consider the limits $\kappa \to \pm\infty$. It has been shown that the limit $\kappa \to -\infty$ corresponds to a random adaptive walk in which transition to any beneficial mutation occurs with the same probability [7] and in this case,

$\beta_\kappa = 1$ in the expression (4.18) for the walk length [14]. The opposite limit $\kappa \to \infty$ corresponds to a greedy adaptive walk [13, 30] in which the fittest mutant is chosen with probability one if selection coefficient is assumed to be small, but a random adaptive walk when large selective effects are taken into consideration [7]. Thus we arrive at the conclusion that in the two limiting cases, if the selection coefficient is allowed to be large, the prefactor $\tilde{\beta}_\kappa$ in (4.69) must be one. As the prefactor $\tilde{\beta}_\kappa$ decreases with increasing κ due to (4.4) in the Weibull domain, it must increase in the Fréchet domain in order to satisfy the $\kappa \to \infty$ limit. We also mention that the transition in the fitness fixed at $\kappa = 1$ does not seem to affect the walk length.

The inset of Fig. 4.1 shows that the full model is approximated very well by the linear model in the Weibull domain, and is a reasonable approximation in the Gumbel domain. This agreement is explained by the fact that the fitness difference between successive steps are indeed small in these two domains as discussed in [16]. However in the Fréchet domain, the relative fitness differences between the successive steps in the adaptive walk can be as large as hundred [16] thus rendering the linear model invalid. For a fixed initial fitness rank, the inset of Fig. 4.1 shows that in the full model, the walk length increases with increasing κ in the Fréchet domain. Thus the behaviour of the walk length is nonmonotonic in κ with the minimum occurring in the Gumbel domain.

Figure 4.5 shows the distribution of the walk length for various κ and uncorrelated fitnesses, and we observe that as $|\kappa|$ increases, this distribution

approaches the corresponding result for the random adaptive walk where the walk distribution is known to be a Poisson distribution with mean $\ln L$ [14]. A related quantity is the index of dispersion of the walk length which is the ratio of the variance to the mean which is shown in the inset of Fig. 4.5 and displays a nonmonotonic behaviour with the minimum occurring at $\kappa = 0$ and approaching unity for $\kappa \to \pm\infty$. A similar nonmonotonic behavior is seen in the linear model but in that case, the index of dispersion approaches unity when $\kappa \to -\infty$ and one [8].

4.4 Discussion

In this chapter, we studied a model of adaptation in which beneficial mutations sweep the population sequentially as it adapts by climbing up a rugged fitness landscape. The broad question addressed here is regarding the average number of adaptive mutations that occur until the population reaches a local fitness peak. This quantity has been measured in recent experiments on various systems like bacteriophage $\phi X174$ [10], fungus *A. nidulans* [11] and bacteria *E. coli* [12]. Theoretically, the number of adaptive changes have been calculated on uncorrelated fitness landscapes for zero initial fitness [9,22] and high initial rank [1,8,31]. Some studies for correlated fitnesses have also been carried out [22, 28, 32, 33]. Here we have extended the previous works and studied how the length of the adaptive walk depends on the initial fitness, extreme value domains and fitness correlations.

For the linear model that assumes small relative fitness differences in all the extreme value domains, we find that the walk length decreases with increasing initial fitness logarithmically provided the mean of the fitness distribution is finite, otherwise it remains a constant. The walks are found to be shorter for fitness distributions that decay slower - in the limit $\kappa \to \infty$, the walk length approaches the greedy walk limit (4.1) while in the other extreme of $\kappa \to -\infty$, it tends to the random adaptive walk (4.2) [7]. The logarithmic variation with the same dependence on the fitness distribution as here has also been seen in other systems [8]. On correlated fitness landscapes, the previous studies have been largely numerical [28, 32, 33], while here we have presented analytical results. Interestingly, the large deviation theory finds an application in the calculation of the walk length for correlated fitnesses. We find that, as on uncorrelated fitness landscapes, the walk length decreases with increasing initial fitness and GPD exponent κ. But increasing fitness correlations also lengthen the adaptive walk since the population encounters lesser number of local fitness peaks. Our detailed analysis shows that the walk length difference $\bar{J}_B(L|f_0) - \bar{J}_B(L|0)$ scales linearly with the number of blocks (that are a measure of correlations) in the Weibull and Gumbel domains, and shows a weaker logarithmic dependence on the number of blocks in the Fréchet domain. For the sake of completeness, we also performed simulations for $\kappa > 1$ and found that the average walk length in this case shows a linear dependence on the block number (data not shown). These results for the linear model are summarised in Table 4.1.

For the full model that is not restricted to small relative fitness differences, we find that the walk length decreases with initial fitness for all κ, unlike in the linear model. The walk length is however seen to match quantitatively well in the Weibull domain where small fitness differences arise [16]. In contrast, in the Fréchet domain, even the qualitative trends in the two models are opposite: while the walk length decreases with increasing $\kappa(< 1)$ in the linear model, it increases in the full model. Thus in the full model, the walk is shortest in the Gumbel domain. An analytical understanding of these results is however not available.

Experiments show that a moderately sized population reaches a fitness plateau in two to four substitutions [10–12] (although one population has been seen to gain nine beneficial mutations as well [10]) thus indicating that the adaptive walks are generally short. An inverse relationship between the initial fitness and the walk length has been observed in some experiments [10, 12] in agreement with the full model. However a constant walk length independent of initial fitness has been seen in a recent experiment [11]. As described above, the full model predicts the walk length to be a nonmonotonic function of the parameter κ. The adaptive walk is expected to last longer in the experimental set ups in which Weibull [10, 34] or Fréchet [35] domain is observed than in the ones in which the distribution of beneficial mutations has an exponential tail [6, 36–38]. However the walk length has not been measured in these experiments, while in the walk length experiments [10–12], the extreme value domain of the beneficial mutation has not been studied

EVT domain	Dependence on initial rank	Dependence on number of blocks
Weibull, $\kappa < 0$	Logarithmic	Linear
Gumbel $\kappa \to 0$	Logarithmic	Linear
Fréchet, $0 < \kappa < 1$	Logarithmic	Logarithmic
Fréchet, $\kappa > 1$	Independent	Linear

Table 4.1: Table summarizing the dependence of the walk length on extreme value domains, initial fitness and fitness correlations in the linear model.

and therefore presently the theoretical predictions regarding the connection between the extreme value theory and the length of the adaptive walk remains experimentally untested. Although some of the available experimental results are in qualitative agreement with the theoretical predictions described above, a quantitative comparison between the experiments and the theory seems difficult. This is because in experiments measuring the walk length, the walk is assumed to terminate if the fitness remains constant over some time period but that need not imply that the adaptation is over [10]. Besides most experiments [11] cannot measure mutations whose fitness difference is below a threshold value and miss out on mutations conferring slight benefit thus underestimating the walk length. A better understanding of the theoretical results vis-à-vis the experimental ones remains a goal for the future.

Bibliography

[1] J. H. Gillespie. *Theor. Popul. Biol.*, 23:202–215, 1983.

[2] J. H. Gillespie. Oxford University Press, Oxford, 1991.

[3] A. Eyre-Walker and P.D. Keightley. *Nat. Rev. Genet.*, 8:610, 2007.

[4] P. D. Sniegowski and P. J. Gerrish. *Phil. Trans. R. Soc. B*, 365:1255–1263, 2010.

[5] D. Sornette. Springer, Berlin, 2000.

[6] D. R. Rokyta, C. J. Beisel, and P. Joyce. *J Theor Biol.*, 243:114–120, 2006.

[7] P. Joyce, D. R. Rokyta, C. J. Beisel, and H. A. Orr. *Genetics*, 180:1627–1643, 2008.

[8] J. Neidhart and J. Krug. *Phys. Rev. Lett.*, 107:178102, 2011.

[9] K. Jain. *EPL*, 96:58006, 2011.

[10] D. R. Rokyta, Z. Abdo, and H. A. Wichman. *J Mol Evol*, 69:229, 2009.

[11] D. R. Gifford, S. E. Schoustra, and R. Kassen. *Evolution*, 65:3070–3078, 2011.

[12] A. Sousa, S. Magalhães, and I. Gordo. *Mol. Biol. Evol.*, 29:1417–1428, 2012.

[13] H. A. Orr. *J. theor. Biol.*, 220:241–247, 2003.

[14] H. Flyvbjerg and B. Lautrup. *Phys. Rev. A*, 46:6714–6723, 1992.

[15] B. Charlesworth and D. Charlesworth. Roberts and Company Publishers, 2010.

[16] S. Seetharaman and K. Jain. *Evolution*, 68:965–975, 2014.

[17] K. Jain and J. Krug. *J. Stat. Mech.: Theor. Exp.*, page P04008, 2005.

[18] C. Sire, S. Majumdar, and D. S. Dean. *J. Stat. Mech.: Theor. Exp.*, page L07001, 2006.

[19] Universal extremal statistics in a freely expanding jepsen gas. *Phys. Rev. E*, 75:051103, 2007.

[20] S. Sabhapandit, I. Bena, and S.N. Majumdar. *J. Stat. Mech.*, page P05012, 2008.

[21] H. Touchette. *Phys. Rep.*, 478:1–69, 2009.

[22] K. Jain and S. Seetharaman. *Genetics*, 189:1029–1043, 2011.

[23] C.M. Bender and S.A. Orszag. Springer, 1999.

[24] A. S. Perelson and C. A. Macken. *Proc. Natl. Acad. Sci. USA*, 92:9657–9661, 1995.

[25] W. Feller. John Wiley and sons, 2000.

[26] D. E. Knuth. Addison-Wesley Professional, Indianapolis, 1997.

[27] M. Abramowitz and I.A. Stegun. Dover, 1964.

[28] H. A. Orr. *Evolution*, 60:1113, 2006.

[29] S. Seetharaman and K. Jain. *Phys. Rev. E*, 90:32703, 2014.

[30] P. R. A. Campos and F. G. B. Moreira. *Phys. Rev. E*, 71:061921, Jun 2005.

[31] H. A. Orr. *Evolution*, 56:1317–1330, 2002.

[32] J. A. L. Filho, F. G. B. Moreira, P. R. A. Campos, and V. M. Oliveira. *J. Stat. Mech.*, -:P02014, 2012.

[33] J. Neidhart, G. S. Szendro, and J. Krug. *Genetics*, 198:699–721, 2014.

[34] T. Bataillon, T. Zhang, and R. Kassen. *Genetics*, 189:939–949, 2011.

[35] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. *PLoS Genet.*, 8:e1002783, 2012.

[36] R. Sanjuán, A. Moya, and S.F. Elena. *Proc. Natl. Acad. Sci. USA*, 101:8396–8401, 2004.

[37] R. Kassen and T. Bataillon. *Nat. Genet.*, 38:484–488, 2006.

[38] R. C. MacLean, G. G. Perron, and A. Gardner. *Genetics*, 186:1345–1354, 2010.

Chapter 5

Adaptation dynamics in the strong mutation regime

5.1 Introduction

In Chapter 3, we have shown analytically and numerically that qualitatively different patterns occur during the adaptation dynamics of populations in different DBFEs when the number of mutants produced per generation is very small [1–4]. The fitness gain obtained in each fixation event follows the pattern of diminishing returns in Weibull domain, constant returns in Gumbel domain and accelerating returns in Fréchet domain which suggests that this quantity can be used to predict the DBFE. These observations are in strong selection-weak mutation (SSWM) regime where the genetic variation in the population is minimal, that is, only one beneficial mutation is

present in the population in time interval between its appearance and fixation [5]. It is then natural to ask if the relationship between the adaptation dynamics and the DBFE mentioned above holds for large populations as well in which there might be more than one beneficial mutation competing for dominance in the population. To address the above question, we use the Wright-Fisher model explained in Chapter 2 to study the adaptation dynamics of an asexual population when the mutation rate is varied in the three EVT domains of DBFE. If two or more beneficial mutation exist in the population, then they have been experimentally observed to compete with each other for dominance in the population [6–9]. This competition is termed clonal interference that we introduced in Chapter 1. We find that the qualitatively different trends for the fitness difference in the three EVT domains seen in the SSWM regime holds even when the mutation rate is increased and the population experiences clonal interference. We also study the dependence of the rate of adaptation of the population on the number of mutants produced per generation in the population. We observe that the rate of adaptation depends strongly on the number of mutations when the beneficial fitnesses are distributed according to the Fréchet domain, whereas it is nearly independent in the case of Gumbel and Weibull distributions. We suggest that these distinct trends can be used to predict the DBFE from experimental studies on adaptation.

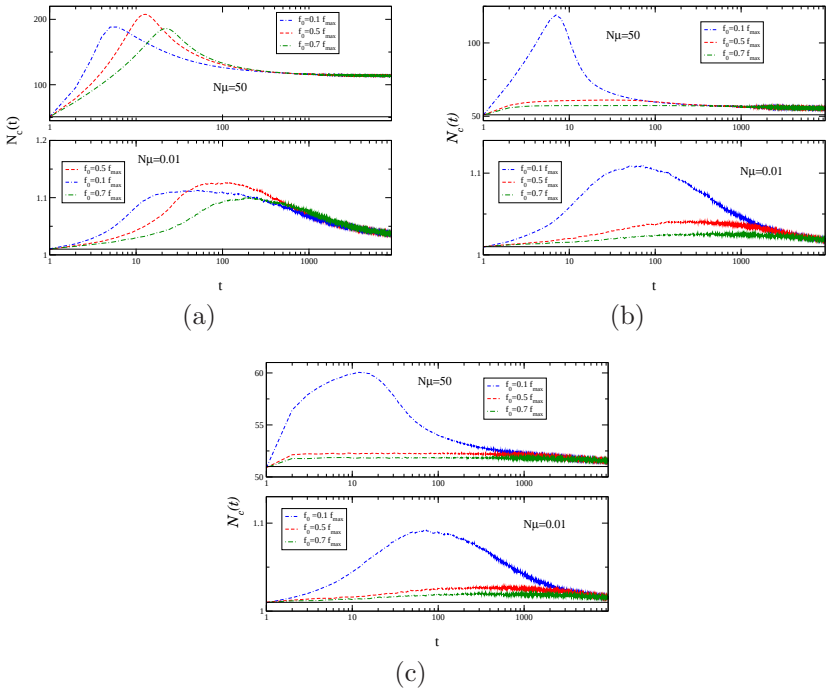


Figure 5.1: The plot shows the average number of classes in the population as function of time for various initial fitnesses. The fitnesses are chosen from (2.8) with (a) $\kappa = -1$ (b) $\kappa \to 0$ and (c) $\kappa = 1/4$. For each κ value, the plot shows $\mathcal{N}_c(t)$ in both the high mutation (top panels) and low mutation (bottom panels) regimes. The straight line in all plots show $N\mu + 1$.

5.2 Results

5.2.1 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 ($\mathcal{N}_c$) present in the population is shown in Fig. 5.1 as a function of time, for all the three domains of DBFE.

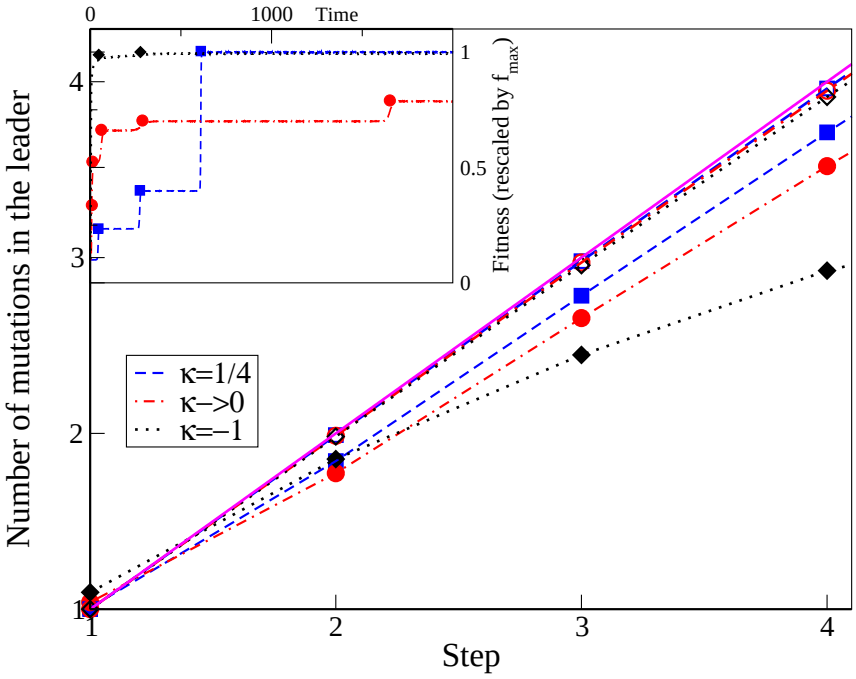


Figure 5.2: The main plot shows the number of mutations in the leader of any step for various κ and mutation rates. The simulation data are represented by points while the broken lines are guide to the eye. The solid line shows $y = x$. In the inset, from a single simulation run, the fitness of the whole 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.

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 increase during the initial time steps as more mutant classes accumulate and decrease at later times when the classes of lower fitness are eliminated by the high fit classes. Also, we see that while the maximum of the number of classes in the population at any time is greater for lower initial fitness when $\kappa \geq 0$, in the case of $\kappa = -1$, it

shows a nonmonotonic trend with respect to initial fitness. The maximum number of classes existing in the population for the first case as shown in Fig.5.1(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 and must compete with this fitter class.

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. 5.1. 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, and thus it takes longer time to take over the population as shown in the bottom panel of Fig.2.3(a). This results in a larger number of mutants in Weibull domain which can be observed in the bottom panel of Fig.5.1(a). We can clearly see from the top panels of the figure that not only are number of classes for a higher mutation rate greater than that for lower rates (as expected), but also that at a fixed high mutation rate, the number of classes increases with decreasing κ, as in the low mutation 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, whereas 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.

5.2.2 Number of mutations in the leader

In the low mutation regime, the average number of mutations in the leader is predicted to be equal to the step number since the genetic variation in the population is low and any mutation that escapes drift quickly takes over the population [10]. We verify this point via simulations as depicted in Fig. 5.2. We find that the mutation number equals the step, in all the three EVT domains of 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 differ 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 since 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 least for the fat-tailed distributions and maximum for the truncated ones, as shown in Fig. 5.2. 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 whereas 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.

5.2.3 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

$$\bar{f}_1 = a_\kappa^{(N\mu)} f_0 + b_\kappa^{(N\mu)} \tag{5.1}$$

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 analytically obtain the fitness at the first step, $\bar{f}_1$ as [1, 2]

$$\bar{f}_1 = \int_{f_0}^{u} df\ T(f \leftarrow f_0) f \tag{5.2}$$

where the transition probability is given by (2.13). In this model, from (5.2) the coefficient $a_\kappa^{(N\mu \ll 1)}$ was obtained as $0.33, 1.0$ and 1.6 for $\kappa = -1, 0,$ and $1/4$ respectively, as explained in Section 3.3. The corresponding $b_\kappa^{(N\mu \ll 1)}$ for the aforementioned κ were $0.66,\ 2.0$ and 1.89 [2]. In the high mutation regime where the adaptive walk model is not applicable, we obtained the values for the coefficients in (5.2) 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.

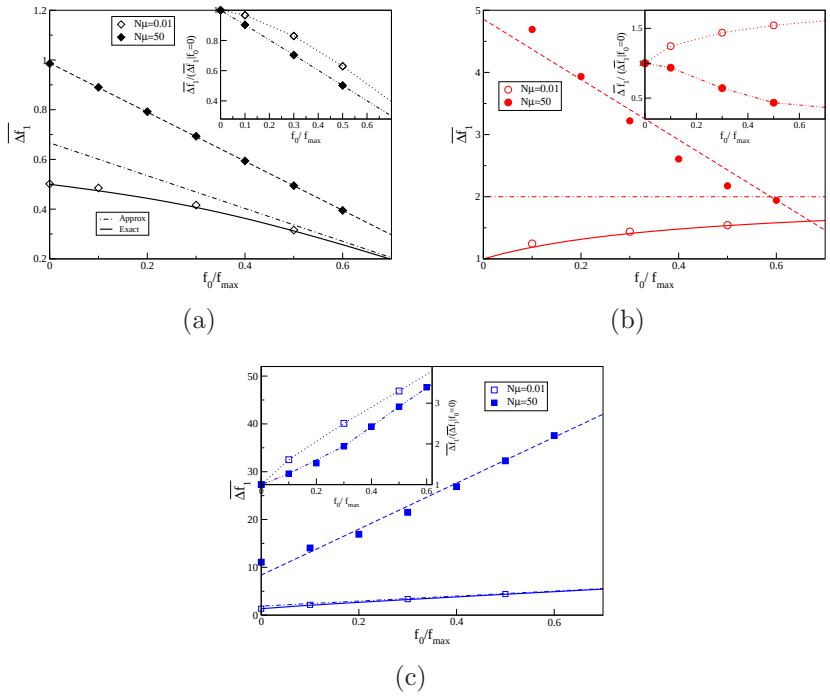


Figure 5.3: The main plot shows the fitness difference at the first step as a function of the initial fitness for various $N\mu$. The fitnesses are chosen from (2.8) with (a) $\kappa = -1$ (b) $\kappa \to 0$ and (c) $\kappa = 1/4$. The solid lines in the main plot are obtained by numerically evaluating the integral given by (5.1), 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 guide to the eye. 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 guide to the eye.

The interesting result from our work is that, irrespective of the number of mutants produced in the population, the difference $\overline{\Delta f_1} = \bar{f}_1 - f_0$ between the fitness of the first step and the initial fitness displays results that are qualitatively different in each EVT domain of DBFE, as shown in Fig. 5.3 and Fig. 5.4. This is clearly seen in the low mutation regime where as the

initial fitness is increased, the fitness difference at the first step increases, approaches a constant and decreases when κ is positive, zero and negative, respectively. In the high mutation regime, though the population is no longer monomorphic, the fitness of the population is almost equal to the fitness of the leader as shown in the inset of Fig. 5.2. More importantly, even in the high mutation regime where, for a fixed initial fitness, the fitness of the first step is greater than the value in the low mutation regime, as can be seen in Fig. 5.4, we find that with increasing initial fitness, the qualitative trend of $\overline{\Delta f_1}$ increasing or decreasing depends on whether the underlying fitness distribution decays as a power law ($\kappa > 0$) or is truncated ($\kappa < 0$). This is because $\overline{\Delta f_1} = (a_\kappa^{(N\mu)} - 1)f_0 + b_\kappa^{(N\mu)}$ and, since $a_\kappa^{(N\mu)}$ is greater than one and less than one for the power law ($\kappa > 0$) and truncated ($\kappa < 0$) distributions respectively. 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. 5.3 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, go further to show that the trends hold irrespective of the number of mutants produced in the population. The qualitatively different trends of the fitness increase increasing, staying a constant and decreasing in the Fréchet, Gumbel and Weibull domain, respectively can be used to distinguish between the EVT

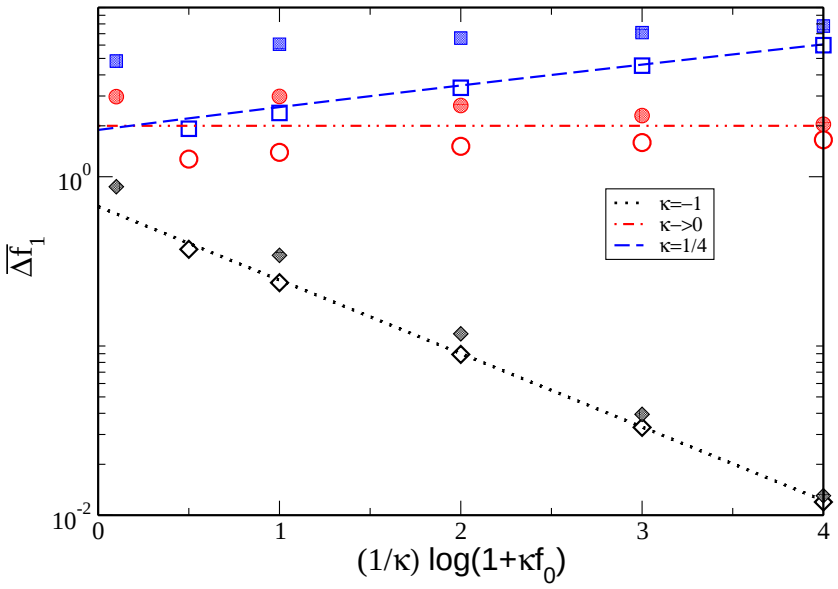


Figure 5.4: The plot shows the fitness difference at the first step as a function of the initial fitness for different κ and two different $N\mu$. The lines give the theoretical values while the open symbols are the simulation output for $N\mu = 0.02$ and the closed symbols are those for $N\mu = 5$.

domains.

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 rescaled by the fitness difference obtained when the initial fitness is zero (insets of Fig.5.3), we see that this increase slows in the high mutation regime compared to the results obtained in the low mutation regime. This indicates that as the mutation

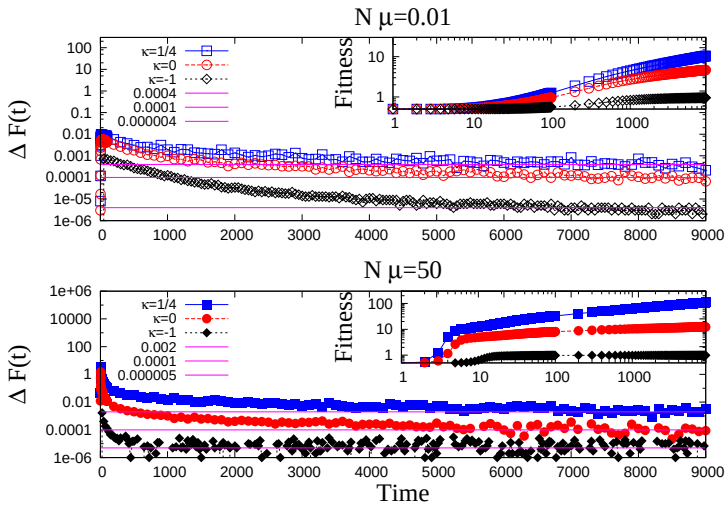


Figure 5.5: Main figure shows the fitness increment in each time step and the inset figure shows the increase in fitness for three different values of κ. In all the cases, the population starts with the same initial fitness $f_0 = 0.5$.

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 is lower for all the fitness distributions.

5.2.4 Rate of adaptation

Besides the fitness increment at a fixed event of leader change, we also measure the fitness as a function of time. We observe 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 transient phase, then it slowly evolves with time and finally it reaches a constant.

The initial transient phase is strongly dependent on the initial condition

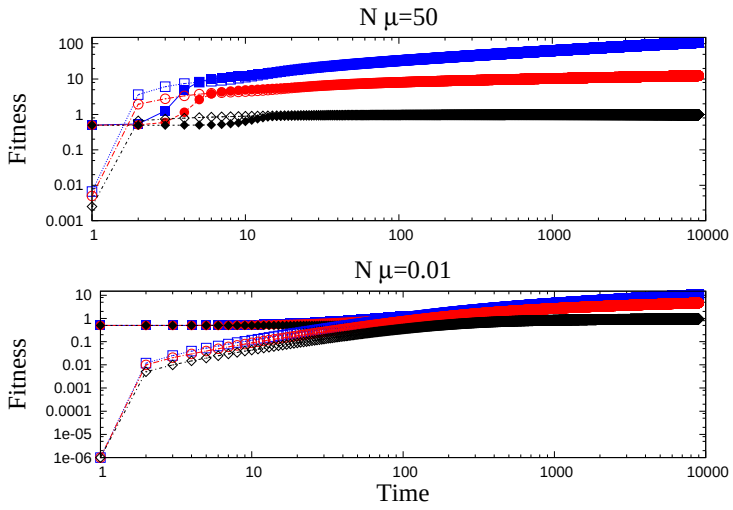


Figure 5.6: The figure shows the average fitness of the population for various κ in both the low and high mutation regimes. Two different initial conditions $f_0 = 0$ (open symbols) and $f_0 = 0.5$ (closed symbols) are considered.

as well as the mutation rate as shown in Fig. 5.6. The increase in fitness is fastest for the lowest initial condition, but it approaches the same fitness values as in the case of higher initial fitness in few generations. The time taken for different initial fitness population 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. We observe that the rate of increase in average fitness ($\bar{\mathcal{F}}(t)$) with time also depends on the mutation rate as shown in the inset of Fig. 5.5. This is because when 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.

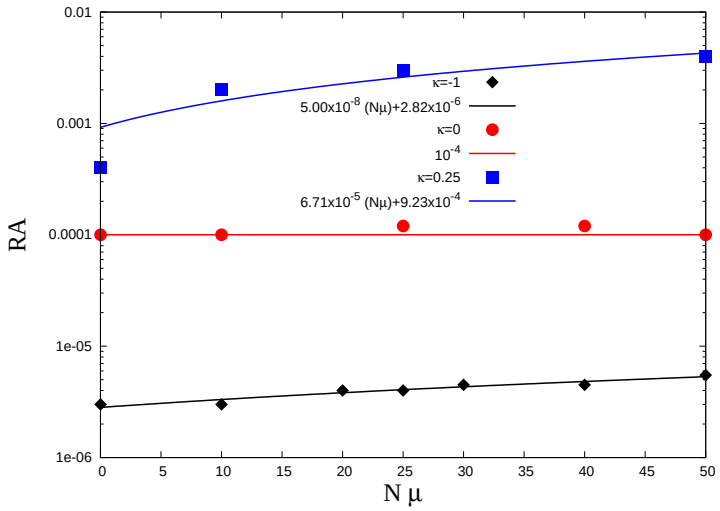


Figure 5.7: The figure shows the rate of adaptation with $N\mu$ for various κ values. The initial fitness is fixed as $f_0 = 0.5$.

After the transient phase, we also measured at the fitness increment defined as

$$\Delta\bar{\mathcal{F}}(t) = \langle \bar{\mathcal{F}}(t+1) - \bar{\mathcal{F}}(t) \rangle \tag{5.3}$$

at each step. The $\Delta\bar{\mathcal{F}}(t)$ initially increases, then slowly decreases and settles down to a constant as shown in Fig. 5.5. This phase is initial condition independent, and the rate at which it approaches the constant adaptation rate is different in different mutation regimes.

Once $\Delta\bar{\mathcal{F}}(t)$ becomes a constant, it is called as the rate of adaptation (RA), which can be calculated using equation 5.3. We measure RA for the three DBFEs with different mutation rates. The values were averaged over 3000 time steps. The fastest rate of adaptation is observed in power law distribution with high mutation rates. Since the distribution is unbounded,

it can produce high fit mutants when the number of mutants produced is large, so maximum advantage is expected to be for this distribution. The rate of adaptation is slowest for bounded distribution, due to limited number of higher fitness mutations available.

Thus dependence of RA on number of mutants for three distributions are significantly different in the three EVT domains of DBFE. There is a clear increase in RA with number of mutants in the case of power law distribution, whereas it is nearly constant in other two distributions as shown in Fig. 5.7. These trends lead us to the conclusion that large number of mutants produced is not particularly advantageous except for fat-tailed distributions where $\kappa > 0$. In all other cases ($\kappa \leq 0$), even though large mutation rate results in very fast fitness increase in initial steps, the rate of adaptation reaches the same value for both low and high mutation regimes.

5.3 Discussion

The main purpose of our work is to determine the quantities which qualitatively show different behaviour for different extreme value domains of the DBFE. The fitness gain at each fixation event shows qualitatively different trends in each DBFE domain, when the number of mutants produced in the population is much less than one at every generation as explained in Chapter 3 [2,4]. The focus of this work is to explore the parameter regime in which the number of mutants produced is much above one. When the mutation rate is

high, the population becomes polymorphic and the better mutants existing in the population compete with each other. In this case as well, we observe that the qualitative trends found in the low mutation regime hold, which shows that the fitness gain at each step in adaptation process is strongly dependent on the DBFE, irrespective of the mutant number produced.

Thus, an important quantity that can be used to predict the DBFE is the fitness difference between the mutations that spread in the population. From our simulations, we see that as the initial fitness is increased the fitness difference at the first step given by $\overline{\Delta f_1}$ reduces, approaches a constant or increases in the Weibull, Gumbel and Fréchet domains, respectively. We can understand these increasing and decreasing trends by the following heuristic reasoning. In both the low and high mutation regimes, for large f_0, the fitness at the first step increases linearly with the initial fitness as given in (5.1) and so, we can write the selection coefficient defined as the relative fitness difference, at the first step as

$$s = \frac{\bar{f}_1 - f_0}{f_0} = \frac{(a_\kappa^{(N\mu)} - 1) f_0}{f_0} + \frac{b_\kappa^{(N\mu)}}{f_0}, \quad \forall \quad \kappa \ , N\mu \tag{5.4}$$

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 (5.4) 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 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:

$$\bar{f}_1 - f_0 < u - f_0, \tag{5.5}$$

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.

Another important measure in adaptation is the rate at which it occurs. Most of the previous studies which measured the adaptation rate only considered exponentially distributed fitness distributions [11–15]. In this work, we measured the rate of adaptation in all the three EVT domains of DBFE and studied its dependence on the mutation rate. We observed a clear increase in RA with the number of mutants produced in the population in the case of Fréchet distribution, whereas it is nearly constant in Gumbel and Weibull distributions as shown in Fig. 5.7. This pattern can be used as another measure in predicting the underlying DBFE. A previous study measuring RA with exponentially distributed beneficial fitness effect observed that after the initial few generations of high rate of adaptation, the rate of

adaptations settles down to a constant value which is independent of the mutation rate [16] which is consistent with our observation.

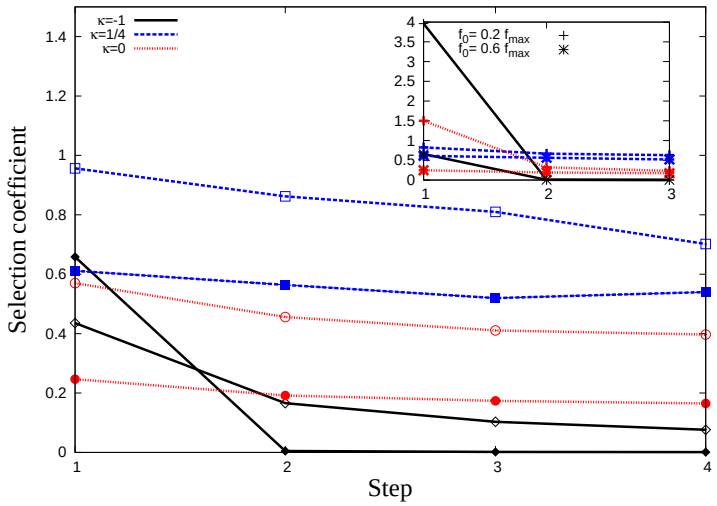


Figure 5.8: The main figure shows the selection coefficient as a function of step for all three κ values with two different $N\mu$ where open symbols and closed symbols are for $N\mu = 0.01$ and $N\mu = 50$, respectively. The inset shows the selection coefficient of various steps for two different the initial fitnesses $f_0 = 0.2 f_{max}$ and $f_0 = 0.6 f_{max}$, where f_{max} is calculated using (2.25) in the high mutation regime.

Experimentally, the distribution of beneficial fitness effects can be inferred by two methods. In the first, mutations are introduced in the wild type 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 [5,17–25]. In contrast, here we focus on learning about DBFE via adaptation dynamics. Though many works have tracked the dynamics of the population during adaptation [5, 26–29], in most of them only the selection coefficient of the mutant

fixed was measured. But our simulations in both the high and low mutation regimes and also, the previous works [1, 2] in the SSWM regime find that irrespective of the EVT domain of the DBFE, the selection coefficient as given by (5.4) always decreases, with the increasing initial fitness or increasing steps as shown in Fig. 5.8 and hence this quantity, is not useful to distinguish between the EVT domains, while the fitness difference between steps show different pattern depending on the EVT domain of the DBFE. In this work, we numerically show that the fitness returns in each EVT domain is very robust and holds even when the number of mutations produced is large. Fitness difference can be measured in experiments as, for example, shown in [20]. 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 the first mutation when the initial fitness is varied. Such experimental studies are desirable.

5.4 Future work

In this thesis, one important process that we ignored is *recombination* which leads to exchange of genetic material between sequences. In microbial populations such as bacteria, this happens mostly by means of horizontal gene transfer such as *conjugation* in which two sequences swap genetic material. The advantage that the population experiences because of this is that, beneficial mutations in different organisms can come together to produce a fitter

organism than either. This process may thus effectively eliminates competition between beneficial mutations since they can spread together. So in the long term, recombination may provide fitness advantage to the population [30]. In our work, we have also assumed the effect of each mutation to be independent of the effect of others. However often one mutation affects the effect of other mutations and this effect is termed *epistasis* [31].

In future, we intend to study the adaptation dynamics of microbial populations by varying the recombination rate. We also plan to introduce epistatic interaction between mutations and check if the qualitative trends of the fitness difference between successive mutations still exhibit the distinct trends of decreasing, staying a constant and increasing in the Weibull, Gumbel and Fréchet domains hold even in these cases. We also hope to get analytical expressions for the numerical results discussed in this chapter.

Bibliography

[1] K. Jain and S. Seetharaman. *Genetics*, 189:1029–1043, 2011.

[2] S. Seetharaman and K. Jain. *Evolution*, 68:965–975, 2014.

[3] S. Seetharaman and K. Jain. *Phys. Rev. E*, 90:32703, 2014.

[4] S. Seetharaman. M.S. thesis, JNCASR, Bangalore, 2011.

[5] D.R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. *Nat. Genet.*, 37:441–444, 2005.

[6] J.A.G.M. de Visser and D. E. Rozen. *Genetics*, 172:2093–2100, 2006.

[7] J.A.G.M. de Visser, C.W. Zeyl, P.J. Gerrish, J.L. Blanchard, and R.E. Lenski. *Science*, 283:404–406, 1999.

[8] R. Miralles, P. J. Gerrish, A. Moya, and S.F. Elena. *Science*, 285:813–815, 1999.

[9] D.E. Rozen, J.A.G.M. de Visser, and P. J. Gerrish. *Curr Biol.*, 12:1040–1045, 2002.

[10] J. H. Gillespie. *Theor. Popul. Biol.*, 23:202–215, 1983.

[11] P. J. Gerrish and R. E. Lenski. *Genetica*, 102:127–144, 1998.

[12] S.-C. Park and J. Krug. *PNAS*, 104:18135–18140, 2007.

[13] M.M. Desai and D.S. Fisher. *Genetics*, 176:1759–1798, 2007.

[14] S.-C. Park, D. Simon, and J. Krug. *J. Stat. Phys.*, 138:381–410, 2010.

[15] P.R.A. Campos and L. M. Wahl. *Evolution*, 64(7):1973–1983, 2010.

[16] P.R.A. Campos and V.M. de Oliveira. *Evolution*, 58(5):932–937, 2004.

[17] R. Sanjuán, A. Moya, and S.F. Elena. *PNAS*, 101:8396–8401, 2004.

[18] R. Kassen and T. Bataillon. *Nat. Genet.*, 38:484–488, 2006.

[19] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman. *J Mol Evol*, 69:229, 2008.

[20] R. C. MacLean and A. Buckling. *PLoS Genetics*, 5:e1000406, 2009.

[21] T. Bataillon, T. Zhang, and R. Kassen. *Genetics*, 189:939–949, 2011.

[22] M. F. Schenk, I. G. Szendro, J. Krug, and J. A. G. M. de Visser. *PLoS Genet.*, 8:e1002783, 2012.

[23] 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. *PLoS Genet*, 10(2), 2014.

[24] C. Bank, T. H. Ryan, D. J. Jeffrey, and N.A.B. Daniel. *Mol. Biol. Evol.*, 2014.

[25] D. R. Rokyta, Z. Abdo, and H. A. Wichman. *J Mol Evol*, 69:229, 2009.

[26] S.E. Schoustra, T. Bataillon, D.R. Gifford, and R. Kassen. *PLoS Biol*, 7 (11):e1000250, 2009.

[27] R. C. MacLean, G. G. Perron, and A. Gardner. *Genetics*, 186:1345–1354, 2010.

[28] D. R. Gifford, S. E. Schoustra, and R. Kassen. *Evolution*, 65:3070–3078, 2011.

[29] A. Sousa, S. Magalhães, and I. Gordo. *Mol. Biol. Evol.*, 29:1417–1428, 2012.

[30] W. R. Rice. *Nat. Rev. Genet.*, 3:241–251, 2002.

[31] D. R. Rokyta, Z. Abdo, and H. A. Wichman. *PLoS Genetics*, 7:229e1002075, 2011.

Appendix A

Solution of the generating function equation (4.6)

The probability distribution $\mathcal{P}_J(f|f_0)$ obeys the recursion equation given by (2.16) [1] in which $T(f \leftarrow h)$ is given by (2.14). The above equation simply means that the population moves from fitness h to a higher fitness f at the next step with probability $T(f \leftarrow h)$ provided at least one fitter mutant is available, the probability of whose is given by $1 - q^L(h)$. For monomorphic initial condition with fixed fitness f_0, we have the boundary conditions given in (2.19) For the linear model with transition probability (2.14), the integral equation (2.16) can be recast as a second order differential equation (2.17).

For infinitely long sequences, the cumulative probability distribution $q^L(h) \rightarrow$

0 and the differential equation (4.6) for the generating function $G(x, f)$ reduces to

$$G''(x, f) = \frac{x(1-\kappa)}{(1+\kappa f)^2} G(x, f) \tag{A.1}$$

From (2.18) and (2.19), we have

$$G(x, f_0) = 0 \tag{A.2}$$

$$G'(x, f_0) = \frac{x}{\int_{f_0}^{u} dg\ (g - f_0)\ p(g)} \tag{A.3}$$

The solution of (A.1) subject to above initial conditions is given by [2]

$$G(x, f) = \frac{x(1-\kappa)(1+\kappa f_0)^{1/\kappa}}{\sqrt{\kappa^2 + 4x(1-\kappa)}} \left[\left(\frac{1+\kappa f}{1+\kappa f_0} \right)^{\alpha_+} - \left(\frac{1+\kappa f}{1+\kappa f_0} \right)^{\alpha_-} \right] \tag{A.4}$$

where

$$\alpha_\pm = \frac{1}{2} \left(1 \pm \sqrt{1 + \frac{4x(1-\kappa)}{\kappa^2}} \right) \tag{A.5}$$

The functions a_1, a_2 appearing in (4.14a) and (4.14b) can be calculated explicitly using the above result. In terms of z defined in (4.8), the solution (A.4) for $\kappa \neq 0$ can be written as

$$G(x, z) = \frac{x(1-\kappa)(1+\kappa f_0)^{1/\kappa}}{\sqrt{\kappa^2 + 4x(1-\kappa)}} \left[\left(z \frac{1+\kappa \tilde{f}}{1+\kappa f_0} \right)^{\alpha_+} - \left(z \frac{1+\kappa \tilde{f}}{1+\kappa f_0} \right)^{\alpha_-} \right] \tag{A.6}$$

Comparing the above equation with (4.9a), we get

$$a_1 = \frac{x(1-\kappa)(1+\kappa f_0)^{1/\kappa}}{\sqrt{\kappa^2+4x(1-\kappa)}}\left(\frac{1+\kappa\tilde{f}}{1+\kappa f_0}\right)^{\alpha_+} \tag{A.7a}$$

$$a_2 = -\frac{x(1-\kappa)(1+\kappa f_0)^{1/\kappa}}{\sqrt{\kappa^2+4x(1-\kappa)}}\left(\frac{1+\kappa\tilde{f}}{1+\kappa f_0}\right)^{\alpha_-} \tag{A.7b}$$

For exponentially distributed fitnesses, taking the limit $\kappa \to 0$ in (A.4) and using (4.8), we find that

$$G(x,z) = \frac{\sqrt{x}e^{f_0}}{2}\left(e^{z\sqrt{x}}e^{(\tilde{f}-f_0)\sqrt{x}} - e^{-z\sqrt{x}}e^{-(\tilde{f}-f_0)\sqrt{x}}\right) \tag{A.8}$$

from which we obtain

$$a_1 = \frac{\sqrt{x}e^{f_0}}{2}e^{(\tilde{f}-f_0)\sqrt{x}} \tag{A.9a}$$

$$a_2 = -\frac{\sqrt{x}e^{f_0}}{2}e^{-(\tilde{f}-f_0)\sqrt{x}} \tag{A.9b}$$

Appendix B

Walk length using Gaussian approximation for exponentially distributed fitnesses

By virtue of central limit theorem [3], the distribution $\mathcal{N}_B(X)$ of the sum of B i.i.d. random variables is given by

$$\mathcal{N}_B(X) = \frac{1}{\sqrt{2\pi B\sigma^2}} \exp\left[-\frac{(X - B\bar{f})^2}{2B\sigma^2}\right] \tag{B.1}$$

provided the mean $\bar{f}$ and the variance σ^2 of the parent distribution $p(f)$ exist. Since the Gaussian distribution is a good approximation to the exact distribution of the sum when $X \sim B\bar{f} \pm \sqrt{2B\sigma^2}$, we expect that it will provide a good estimate of the walk length when $f \sim B(f_0 - \bar{f}) \pm \sqrt{2B\sigma^2}$ in the integrand in (4.28). With increasing B, as the core of the distribution $\mathcal{N}_B(X)$ moves rightwards while the factor fe^{-f} in the integrand peaks around one, the overlap is significant when $f_0 \sim 1 \mp \sqrt{2\sigma^2/B}$. Thus the Gaussian approximation for the sum distribution is likely to work well in the neighborhood of initial fitness one. This can be seen more explicitly as follows: using (B.1) in the integral appearing in (4.28), we get

$$I_{clt} = \int_0^{Bf_0} df\ fe^{-f}\mathcal{N}_{B-1}(Bf_0 - f) \tag{B.2}$$

$$= \frac{ae^{a^2}e^{-2ab}}{\sqrt{\pi}}\left[e^{-(a-b)^2} - e^{-4a^2} + \sqrt{\pi}(a-b)(\mathrm{erf}(a-b) - \mathrm{erf}(2a))\right] \tag{B.3}$$

where $a = \sqrt{(B-1)/2}$ and $b = (Bf_0 - B + 1)/(2a)$. For large B, using the asymptotic expansion of error function, we get

$$I_{clt} \approx \frac{ae^{a^2}e^{-2ab}e^{-(a-b)^2}}{2\sqrt{\pi}(a-b)^2} \tag{B.4}$$

The expression (4.28) for the average walk length then gives

$$\bar{J}_B(L|f_0) = B(\beta_0 \ln L_B + c_0) - \beta_0 B\frac{e^{-f_0(f_0-1)}}{(2-f_0)^2} \tag{B.5}$$

$$\approx B(\beta_0 \ln L_B + c_0) - B\beta_0 f_0\ ,\ f_0 \to 1 \tag{B.6}$$

Thus it is only when the initial fitness is close to unity, the Gaussian approximation captures the linear relationship between $\bar{J}$ and f_0 correctly.

Bibliography

[1] K. Jain and S. Seetharaman, Genetics **189**, 1029 (2011).

[2] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers* (Springer, 1999).

[3] D. Sornette, *Critical Phenomena in Natural Sciences* (Springer, Berlin, 2000).