Adaptation on rugged fitness landscapes

A Thesis

Submitted For the Degree of

MASTER OF SCIENCE (ENGINEERING)

in the Faculty of Science

by

Sarada S

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

JANUARY 2012

To my family

DECLARATION

I hereby declare that the matter embodied in the thesis entitled " **Adaptation on rugged fitness landscapes** " 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 Dr. 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 S

CERTIFICATE

I hereby certify that the matter embodied in this thesis entitled " **Adaptation on rugged fitness landscapes** " has been carried out by Ms. Sarada S 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.

Dr. Kavita Jain
(Research Supervisor)

Acknowledgements

I express my deep gratitude to my mentor Dr.Kavita Jain for her constant guidance, concern and support. Her patience and enthusiasm were always a source of inspiration. She is always approachable and understanding and it is a joy to work with her.

I am grateful to Prof. M. R. S. Rao, the 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, Dr.Vidhyadhiraja N. S. and Dr. Subir K. Das for giving me the chance to be a part of JNCASR. In these 2 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, Dr. Rama Govindarajan, Dr. Subir Das and Prof. Amitabh Joshi for the classes they have handled for me, the discussion and suggestions regarding my work.

I would like to thank the comp lab members - Ravi, Bikas, Vishnu, Ashish and Sandhya for helping me to resolve the problems I had with my system.

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 also acknowledge the contribution of Akhilesh to the work we did together on greedy walks.

I would also like to thank my friends in JNCASR- Suman, Sutapa, Shaista, Vishwas, Moumita, Shiladitya, Sonia, Jitu, Deb, CD, Sheetal, Antara and Nikhil for helping me with my research and for making my stay in JNC very enjoyable.

I specially thank Sumithra Lal for being a dear friend and responsible for me joining Kavita.

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

Synopsis

When a microbial population is exposed to a hostile environment, as for example on introduction of antibiotics, it might die out or undergo changes that increase its chances of survival. These changes termed *mutations* occur in the genome of the individuals and may get transmitted to the subsequent generations [1]. The beneficial ones lead to adaptation resulting in higher reproductive ability that is indicated by an increase in a numerical value called *fitness*.

In the thesis we study the dynamics of adaptation of asexually reproducing genomic sequences starting from a low fitness value. We have calculated quantities that can experimentally be measured in populations like the number of mutations and the increase in fitness during adaptation. These depend majorly on two factors: the size of the population and the correlation between fitness of the parent and its mutants. The first plays a role in the fixation of a beneficial mutation. When the population size is small, it evolves stochastically since the better mutations can get lost due to random sampling or *genetic drift* whereas in large populations these effects are negligible and adaptation is deterministic. Large populations also produce greater number

of mutations and thus evolve to reach higher fitness values. The second factor, correlation between fitnesses is a reflection of the environment [2] and affects the steady state of the two population sizes attained via adaptation, in a fixed environment.

In the thesis we consider a population of asexually reproducing genomic sequences when the mutation rate is small so that most of the population has same sequence at all times. This sequence is thus the *most populated* or the *dominant sequence* and the properties of the whole population reflect those of this sequence. We study the dynamics of evolution by tracking changes in this sequence which depend on the population size and the path to the fittest sequence. The population size determines the number of mutants produced per generation and when the size is very large, all sequences are populated and a sequence many mutations away from the current most populated sequence can become the next, whereas small populations are localised at the most populated sequence and can access only the neighbouring one mutant sequences which can then replace the current sequence. Therefore the most populated sequence of both the population sizes will reach the highest fitness only when it can be reached via one mutant neighbours starting from any initial sequence. But when this is not the case, only that of large populations can reach the fittest value and small populations will proceed only till fitter mutations are available. Our question of interest is the average number of changes in the identity and the fitness of the most populated sequence for the two population sizes.

In Chapter 1 we introduce some basic terms pertinent to our work and define the quantities that shall be used in later calculations. For the sake of

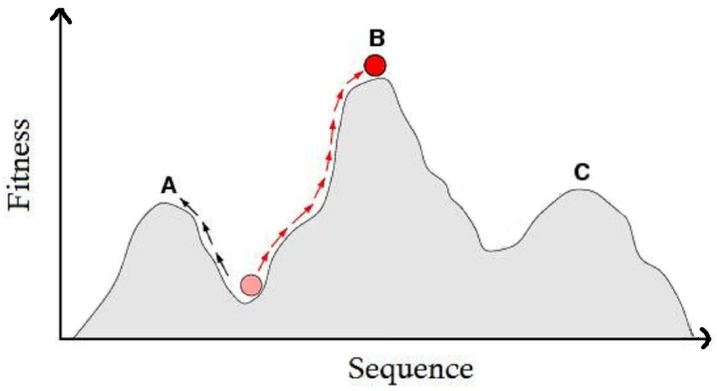


Figure 1: Schematic representation of a rugged fitness landscape with many fitness peaks. Here A and C are local fitness peaks which are fitter than all their nearest neighbours while B is the fittest sequence and hence the global peak. The arrows represent the change in the sequence and fitness of the population. The population (shown by dot) climbs the fitness landscape during adaptation.

simplicity, we work with binary sequences which can *replicate* and *mutate*. The replication rate of each sequence is given by its fitness and all possible sequences along with their associated fitness comprise the *fitness landscape* which encodes information about the environment. For example, when the carbon source of a *E.coli* population is composed of glucose there is only a single metabolism pathway and as expected, the fitness landscape is smooth with a single fitness peak but becomes rugged with many local peaks when the medium is a complex mixture of carbon sources which can be metabolised in multiple ways [3]. The ruggedness of the fitness landscape is determined by the correlations between fitness of the sequences: decreasing correlations produce increasing ruggedness with the completely correlated fitness landscape being smooth. The correlations are introduced in our work using the *Block model* [4] in which a sequence of length L is assumed to be composed

of blocks of equal length L_B with independent and identically distributed (i.i.d.) fitness and the fitness of the sequence is the average of the fitness of the blocks present in the sequence. The length of the block, L_B can be used to tune the correlations between the sequence fitnesses spanning from fully correlated corresponding to $L_B = 1$ to uncorrelated corresponding to $L_B = L$. During adaptation a population climbs the fitness landscape via muations as shown in Fig. 1 and the dynamics of adaptation depend strongly on the supply of beneficial mutations given by the product of population size and mutation probability. If beneficial mutations are easily available as in population of infinite size, the population can reach the global fitness peak on rugged fitness landscapes while a population for which beneficial mutations are rare gets trapped in a local fitness peak. A brief outline of the adaptation models we shall consider in our work is shown in Fig. 2

In Chapter 2 we review earlier works that have dealt with the adaptation of small and infinite population sizes. To study the infinite populations a quasispecies model [5] is used in which a phase transition occurs in the steady state. For most fitness landscapes, this occurs at a critical mutation rate above which the population is uniformly scattered on the fitness landscape and below which it is localised around the fittest sequence surrounded by a suite of mutants [6, 7]. During adaptation the population at each sequence grows and the identity of the dominant sequence may move by many mutations as its population overtakes the current one. Earlier work has focussed on uncorrelated fitness landscapes ($L_B = L$) [8,9] and we discuss their results in this chapter. The simple case of fully correlated fitness landscapes ($L_B = 1$) is also discussed here. Unlike the infinite populations where all

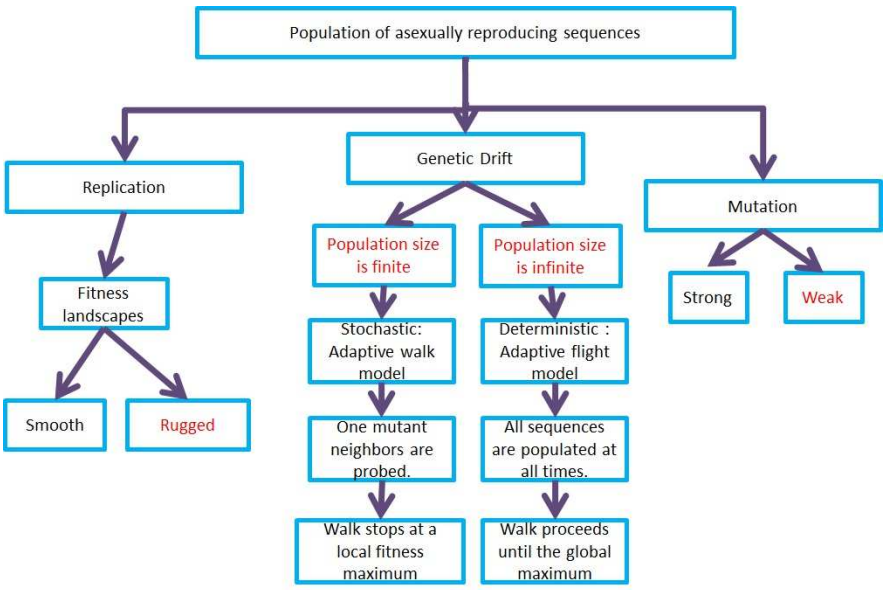


Figure 2: Models of evolution of asexual populations. The parameters we shall consider in our work are highlighted in red.

sequences are populated at all times and the dynamics is due the dominant sequence overtaken by another, small populations are localised at a single sequence and when the mutation rate is low, dynamics involves the whole population moving to fitter one mutant neighbours. The steady state is trivial with the population stabilising at the local fitness peak. An adaptive walk model [1], in which the transition probability determines to which of the better mutant the population would move, is used to study the adaptation of these populations . In this chapter, the known results of the two limiting cases namely greedy walk [10] and random adaptive walk [11] are discussed. In the first case, the transition probability of the population moving to the fittest of the L neighbors is one and in the second, it is equal for all fitter neighbours. However in the biologically relevant situations, the transition

probability of the population moving to a better sequence is proportional to the difference between the fitness of that sequence and the current sequence [1]. The first step of this walk starting with a high initial fitness has been well studied theoretically [12] and experimentally [13] and we discuss some of these results.

In Chapter 3 we describe our results on the dynamics of a quasispecies model [5] when nonzero correlations between fitnesses are present. Starting with a population localised at a sequence at time $t = 0$, the initial population of all other sequences depend on the number of mutations D from the initial sequence. But when the probability of mutations is low, their subsequent growth is exponential with a rate equal to their fitness. The logarithmic population $E(D, t)$ of a sequence with D mutations grows in time as [9]

$$E(D, t) = -D + f(D)t. \tag{1}$$

As illustrated in Fig. 3 the logarthmic population of each sequence grow linearly in time with a slope equalling its logarthmic fitness $f(D)$. The sequence corresponding to the highest $E(D, t)$ is the dominant sequence at that t and at a later time when it is overtaken by another population, its fitness assumes that of the population. We have considered evolution on fitness landscapes with strong correlations by choosing block length $L_B = 2$ and with weak correlations by choosing $L_B = L/2$. In the first case, we have shown exactly that the average number of changes $\bar{J}$ in the dominant sequence is independent of the sequence length but depends on the fitness

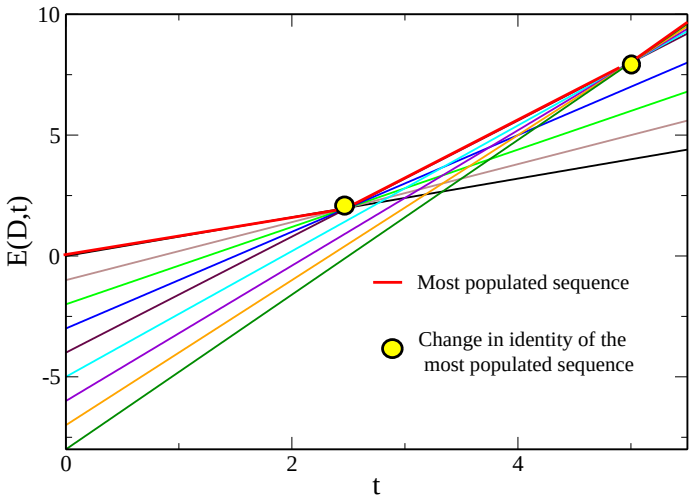


Figure 3: Growth of the logarithmic population $E(D,t)$ at each D with time for $L = 8$ and $L_B = 2$. The initial dominant sequence at $D = 0$ is replaced by the fittest sequence at $D = 4$ which in turn is overtaken by the one at $D = 8$.

distribution $p(f)$ of the blocks. Specifically we find [14]

$$\bar{J} = \begin{cases} \dfrac{23}{72} & (p(f) = e^{-f}) \quad (2) \\ \dfrac{17}{48} & (p(f) = 1) \quad (3) \end{cases}$$

In contrast, when $L_B = L/2$ our simulations show that the average number of changes in the dominant sequence grows as $\sqrt{L}$ with a prefactor that varies with the distribution $p(f)$ [15]. However irrespective of the fitness distribution and the degree of correlation, the temporal distribution of reaching the global fitness maximum has a $1/t^2$ dependence [14].

In Chapter 4 we study the properties of the adaptive walk with low initial

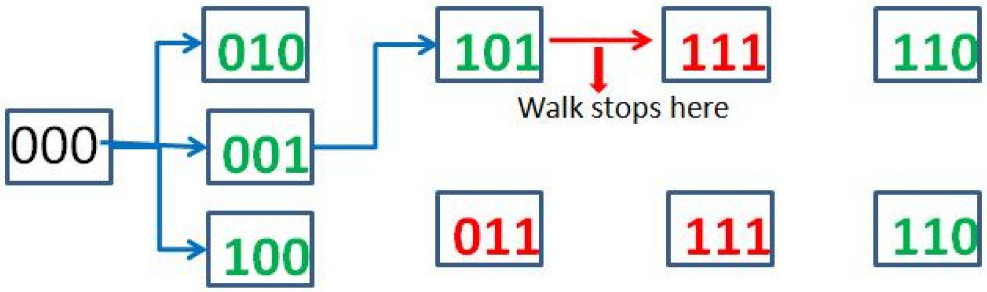


Figure 4: Adaptive walk for $L = 3$. The mutant sequences with better fitness than that of the parent are shown in green and the worse ones in red. The walk stops at the step $J = 2$ since **101** is a local maximum.

fitness. Unlike previous works [12] which deal with the first step here we present results on the entire walk. As shown in Fig. 4 the population scans its L one mutant neighbours and moves to one of the fitter sequences. The walk terminates when no further beneficial mutations are available as shown in Fig.4. For large sequence length L and uncorrelated fitness, the transition probability that the population will move from current fitness h to fitness f in the next step is given by [18]

$$T(f \leftarrow h) = \frac{(f-h)p(f)}{\int_h^u dg(g-h)p(g)} \ , \ f > h. \tag{4}$$

where the fitness are chosen from $p(f)$. If we define the probability of a sequence having a fitness less than the current value h as $q(h) = \int_l^h dgp(g)$, then the probability that not all the L one mutant neighbours of the present sequence have a lower fitness is $1 - q^L(h)$. Now the probability $P_{J+1}(f)$ that

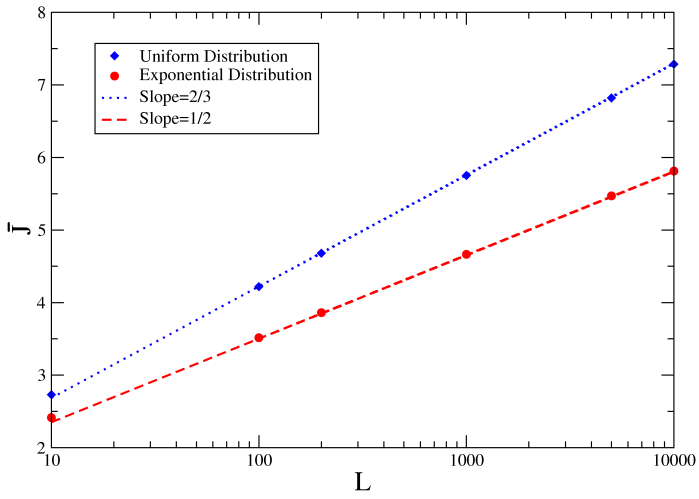


Figure 5: Average number $\bar{J}$ of adaptive steps as a function of sequence length L for uniform and exponential fitness distributions on uncorrelated fitness landscapes.

the walker would take the $(J+1)^{th}$ step and assume fitness f is given by [18]

$$P_{J+1}(f) = \int_l^f dh\ T(f \leftarrow h)\ (1 - q^L(h)) P_J(h)\ ,\ J \geq 0 \tag{5}$$

On solving the above equation analytically within an approximation, we obtain the average number $\bar{J}$ of adaptive steps to reach the local fitness maximum as [18]

$$\bar{J} \approx \begin{cases} \frac{1}{2} \ln L & (p(f) = e^{-f}) \quad (6) \\ \frac{2}{3} \ln L & (p(f) = 1) \quad (7) \end{cases}$$

The above theoretical predictions agree well with the simulation results as shown in Fig. 5. When correlations are introduced using the block model, we have shown that the average number of steps $\bar{J}_B(L)$ in the adaptive walk of sequences of length L with B blocks increases linearly with B [18]:

$$\bar{J}_B(L) = B\bar{J}(L_B), \tag{8}$$

where $\bar{J}(L_B)$ is the average number of steps in the adaptive walk for uncorrelated sequences of length L_B. We get the above solution using the fact that the probability of taking J steps for a correlated sequence can be factorised into product of probabilities of each block taking j_i steps such that $\Sigma_{i=1}^{B} j_i = J$. Our results show that the adaptive walks are short which is consistent with experimental data [19].

Chapter 5 is a brief discussion about the results obtained for the two population sizes. Especially infinite population results and the connection to the adaptive walk of small populations are discussed. The last part of the thesis deals with the open questions in the adaptive walk model that we intend to address in the near future. These include the change in the properties of the walk due to recombination, the fitness advantage conferred by a beneficial mutation and the time taken by it to get fixed in a population.

Bibliography

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

[2] C. R. Miller, P. Joyce, and H.A. Wichman. *Genetics*, 187:185–202, 2010.

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

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

[5] M. Eigen. *Naturwissenchaften*, 58:465 – 523, 1971.

[6] M.A. Nowak. *Evolutionary Dynamics: exploring the equations of life*. Harvard University Press, 2006.

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

[8] K. Jain. *Phys. Rev. E*, 76:031922, 2007.

[9] J. Krug and C. Karl. *Physica A*, 318:137–143, 2003.

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

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

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

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

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

[15] G. Das. *Dynamical properties of a quasispecies model on correlated fitness landscapes*. M.S. thesis, JNCASR, Bangalore, 2010.

[16] S. Seetharaman. *Unpublished*.

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

[18] K. Jain and S. Seetharaman. *Genetics*, doi:10.1534, 2011.

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

Publications

- "Evolutionary dynamics on strongly correlated fitness landscapes", **Sarada Seetharaman** and Kavita Jain, Phys. Rev. E 82:031109, 2010.
- "Nonlinear deterministic equations in biological evolution", Kavita Jain and **Sarada Seetharaman** , J. Nonlin. Math. Phys 18:321–338, 2011.
- "Multiple adaptive substitutions during evolution in novel environments", Kavita Jain and **Sarada Seetharaman** , Genetics 189:1029–1043, 2011.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Biological evolution

Biological evolution refers to changes over time in populations of individuals and may cause varied outcomes like

- **Adaptation**: This makes the organisms better suited to their environment, an example of which is the antibiotic resistance developed by microorganisms.
- **Extinction**: This refers to the disappearance of a whole species. This could occur due to accumulation of deleterious mutations or external factors like change in the climate.
- **Speciation**: This refers to two or more subpopulations descended from a common ancestor, developing into different species during the course of evolution, due to lack of genetic mixing between them and the accumulation of different sets of mutations.

Of these, we are interested in the process of adaptation caused by beneficial mutations as represented schematically in Fig. 1.1 for an asexual population. Recent laboratory experiments on adaptation are measuring the changes in genomes and the corresponding effects on physical traits [1]. Along with the progress in experiments, mathematical models have been developed that not only provide insight into the experimental results, but also make novel predictions that can be experimentally verified [2]. We study the adaptation of a population of asexually reproducing genomic sequences in two population size limits. If the size of the population is infinite, it can be dealt using deterministic equations while the finite sized population requires a stochastic formulation. In this thesis, we discuss analytical solutions to some of the questions related to adaptation.

1.2 Concepts and definitions

Before defining the mathematical models, we explain some basic concepts and definitions which are relevant to the discussion in the thesis [3].

Sequence and sequence space: A sequence $\sigma = \{\sigma_1, ..., \sigma_L\}$ is a string of L letters which are chosen from an alphabet of size a. It represents a protein or a genotype when the letters are amino acids or nucleotides respectively. The total sequence space consists of all possible strings of length L and thus has a size a^L which increases exponentially with L. For computational ease, the alphabet size can be decreased by lumping some of the information in a single letter. In our work we differentiate between genotypes by the absence or presence of a mutation which corresponds to $a = 2$ [4] and hence work with

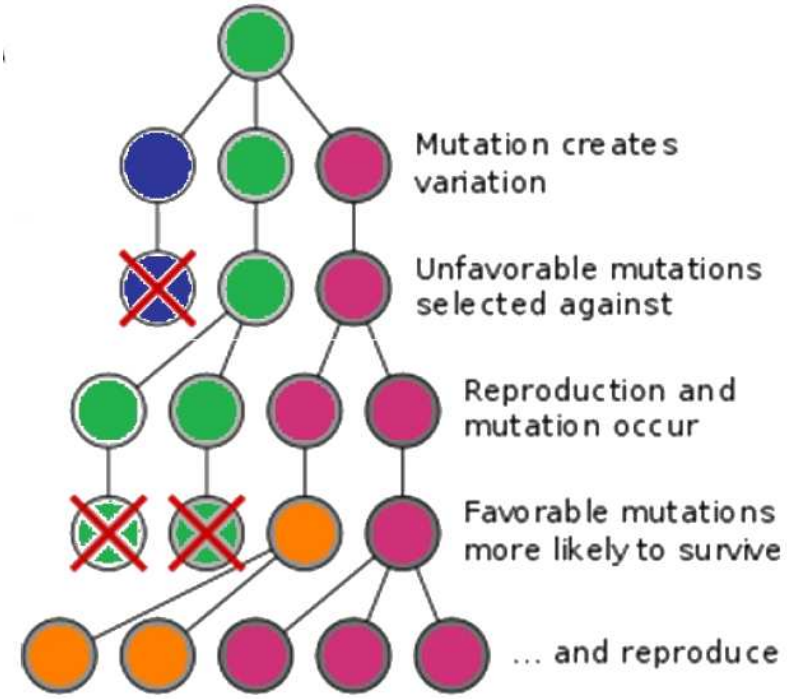
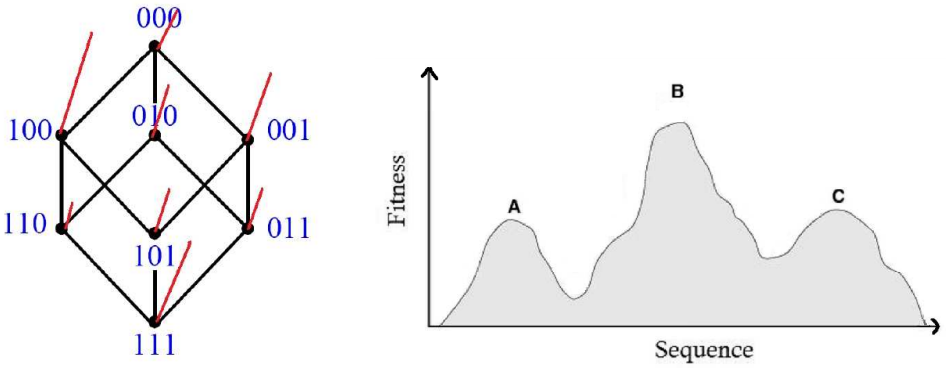


Figure 1.1: A simplified schematic representation of adaptation. Different sequences produced due to mutations in the parent sequence are shown in different colors. If a mutant is fitter than the parent, it may spread within the population.

binary sequences. Such 2^L sequences can be arranged on a L-dimensional sequence space, where the distance $D(\sigma, \sigma')$ between two sequences σ and σ' is equal to the number of loci by which they differ. For a binary sequence in which $\sigma_i = 0$ or 1, one may write

$$D(\sigma, \sigma') = \sum_{i=1}^{L} (\sigma_i - \sigma'_i)^2 \tag{1.1}$$

The sequence space for $L = 3$ is shown in Fig. 1.2(a). Here each sequence is at a vertex and the number of edges between two sequences gives the number



(a) (b)

Figure 1.2: (a) Fitness landscape for sequence length $L = 3$. The sequence space is represented by a cube and the fitness are shown by red lines. (b) Schematic representation of fitness landscape. In the plot, all possible sequences are on the x-axis and their associated fitness on the y-axis. A and C are local fitness peaks while B is the global fitness peak.

of loci by which they differ.

Fitness: The fitness $f(\sigma)$ of a sequence σ is a measure of its reproductive success in a given environment. Its definition is context dependent and may represent the replication rate of a genotype or the functionality of a protein. For sequences of length L, its sequence space along with the fitness of each sequence comprises the $L + 1$ dimensional *fitness landscape* [5] as shown in Fig. 1.2(a) for $L = 3$. For large L, the exact fitness landscape representation is complicated and is shown schematically in two or three dimensional plane. An example of a high-dimensional fitness landscape is shown in Fig. 1.2(b) in which there are several local fitness peaks, defined as sequences that are fitter than their nearest neighbours besides the global fitness peak, which is the fittest of all sequences.

Empirical measurement of fitness landscapes is very hard since the number of sequences increases exponentially with the sequence length L. However several qualitative features particularly the topography of the fitness landscapes has been deduced in experiments on proteins and microbes either by an explicit construction of the fitness landscapes for small $L(\lesssim 5)$ or indirect measurements of relevant quantities. These experiments show that the fitness landscapes can be smooth as evidenced by fast adaptation in some proteins [6] or have multiple peaks as seen in microbial populations that evolve towards different fitness maxima [7–9] and enzymes with short uphill paths to the global fitness peak [10]. Detailed studies in which all or a set of mutants from wild type to an optimum are created and their fitness measured [11] have also indicated the smooth [12] and rugged [4,13] nature of the fitness landscapes. 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].

Adaptation occurs when the population climbs a fitness landscape via mutations and is determined by the topography of the fitness landscape. If the fitness landscape is smooth with a single peak so that from any sequence the fittest sequence can be reached via fitter neighbours, the population may reach the global fitness maximum. On the other hand, if the fitness landscape is rugged with many local fitness peaks, the path to the fittest sequence from any other may encounter fitness valleys in which case, the population can get

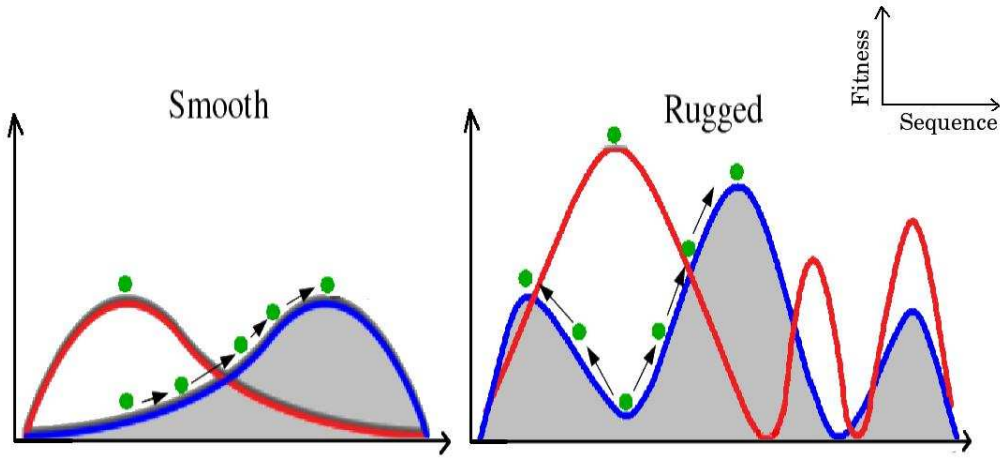


Figure 1.3: Schematic representation of adaptation dynamics on smooth and rugged fitness landscapes. The red curve is the fitness landscape before the change in environment and the blue curve is the new fitness landscape after the change. The adaptation of a population (represented by a dot) depends on the shape of the landscape (see text).

trapped at a local fitness peak as can be seen in Fig. 1.3.

Correlations between sequence fitnesses are a measure of the ruggedness of the fitness landscapes, with decreasing correlations producing increased ruggedness. Experiments suggest an intermediate degree of correlations in fitness landscapes [10,15,16] and many theoretical models like the NK model [17], Mt. Fuji model [18] and block model [10] have been proposed to vary the correlations in fitness landscapes. Of these, we use the *block model* in which a sequence of length L is divided into B blocks of equal length L_B with independent and identically distributed (i.i.d.) fitnesses from a fitness distribution $p(f)$. The fitness of a block may or may not depend on its position in the sequence and in our work we have considered both the cases. The fitness of the sequence is the average of the fitness of the blocks present

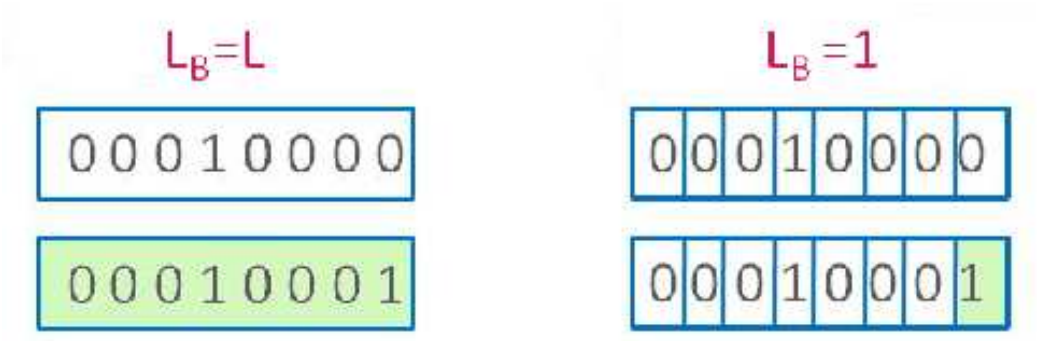


Figure 1.4: Block model. The blocks in the bottom sequence that are different from the top one are shaded green. Left: When block length $L_B = L$, though the sequences vary by only one mutation, they are uncorrelated as each block fitness is an i.i.d. variable. Right: When block length $L_B = 1$, the sequences are correlated due to the presence of common blocks.

in it. The correlations in the fitness landscape can be tuned by varying L_B with the extreme cases $L_B = 1$ and L producing fully correlated and fully uncorrelated landscapes respectively as described in Fig. 1.4.

A fitness landscape can be constant in time if the environment is fixed or it can be time dependent due to a constantly changing environment or if the fitness depends on the population density [3]. In this thesis, we assume that a fitness landscape has changed due to a sudden change in the environment, so that a population previously at a fitness peak drops to a fitness valley. From this point onwards, the fitness landscape is taken to be constant over the time scales considered.

Mutation: Stochastic changes known as mutations may happen in the genome of an individual during replication or when it is exposed to certain external mutagens like radiation. These may insert, delete or change the nucleotides in the genome and thus create a new sequence as shown in Fig.

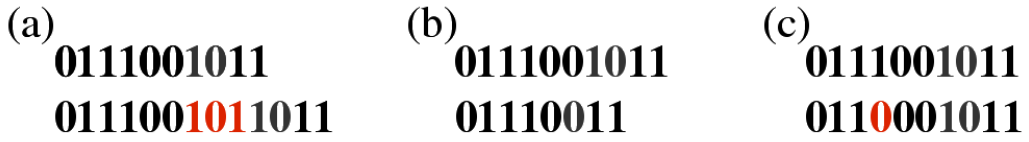


Figure 1.5: (a) Insertion mutation. (b) Deletion mutation. (c) Point mutation. Mutations are highlighted in red.

1.5. In this thesis, we will consider only *point mutations* that change a letter from 0 to 1 and vice versa thus preserving the length of the sequence. If μ is the mutation probability per locus per generation, the sequence σ' of length L will mutate to σ separated from it by D point mutations with a probability

$$p_{\sigma \leftarrow \sigma'} = \mu^D (1-\mu)^{L-D} \tag{1.2}$$

A mutation may increase the fitness of a sequence (beneficial) or reduce it (deleterious) or cause no change in the fitness (neutral) [19]. If the mutation is beneficial, the change may propagate in the population, otherwise it may get eliminated.

In this thesis, we work in the *weak mutation limit*, which corresponds to $N\mu \ll 1$ if the population is of finite size N and $\mu \ll \mu_c$ in the case of infinitely large populations, where μ_c is a critical mutation rate which shall be explained in Chapter 2. In this limit, most of the population is localised at a single sequence and the properties of the whole population is determined by this sequence.

Genetic drift: Real populations are finite in size and as a result, they are subject to stochastic fluctuations due to random sampling termed genetic

drift. To understand this evolutionary force, consider a population of fixed size N with fitness of the wild type 1 and that of the mutant $1+s$. Here s is the *selection coefficient* giving the relative fitness difference between the mutant and the non-mutant 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 γ_i that the mutant will sweep through the population at large times starting with an initial number i is given by [20]

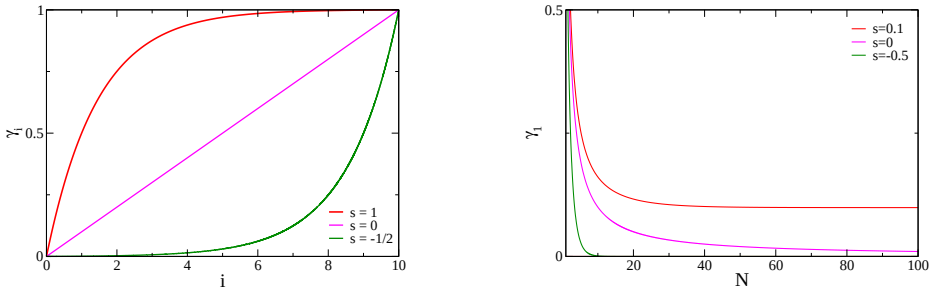
$$\gamma_i = \frac{1-(1+s)^{-i}}{1-(1+s)^{-N}} \tag{1.3}$$

The evolution is this case is stochastic where, beneficial mutations might get lost if the mutation is rare and deleterious mutation might get fixed if i is large as shown in Fig. 1.6(a). Let us now consider the fate of a rare mutation ($i = 1$) when $N \to \infty$, $s \to 0$ such that Ns is finite. The probability of fixation of the mutant can be written as

$$\gamma_1 = \frac{s}{1-e^{-Ns}} \tag{1.4}$$

If $Ns \ll 1$ (weak selection limit), the mutation is nearly neutral and its probability of fixation 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

$$\gamma_1 \approx \begin{cases} s & \text{if } s > 0 \qquad (1.5) \\ 0 & \text{if } s < 0 \qquad (1.6) \end{cases}$$



(a) (b)

Figure 1.6: (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.

The probability if fixation of a rare mutant as a function of population size N is shown in Fig. 1.6(b).

Although the real populations are finite and evolve stochastically, phenomena observed in deterministic setting may survive in the presence of stochasticity as well [21], and deterministic solutions can also be utilised to get insight in the corresponding stochastic problem [22] and to develop stochastic theories [23].

1.3 Overview of the thesis

In this thesis, we study the adaptation of asexual populations on rugged fitness landscapes with many local peaks. The limiting cases of strong and weak fitness correlations are considered. We are mainly interested in the number of beneficial mutations accumulated and the corresponding fitness increase during the course of adaptation. We study the following two models in which the number of mutants produced per generation are infinite (Quasispecies

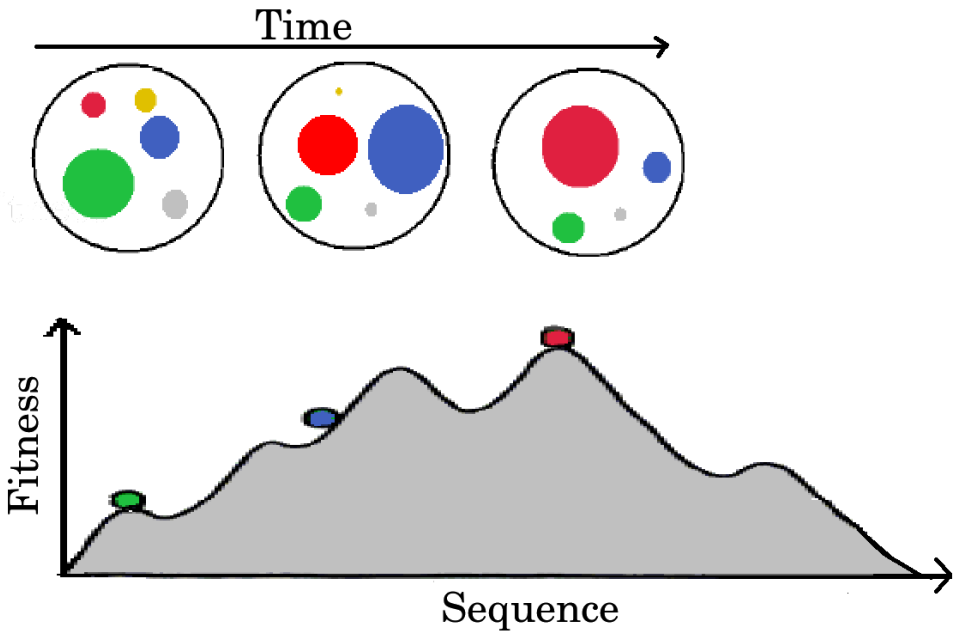


Figure 1.7: Adaptation dynamics of a quasispecies population. Different colors represent different sequences with the area of the circle being proportional to the population size of that sequence and the circle with the largest area is the dominant sequence at that time. The fitness of the whole population is close to that of the dominant sequence and changes every time the dominant sequence changes (as shown by the color of the dot) on the fitness landscape.

model) or less than one (Adaptive walk model).

1. **Quasispecies model:** For infinite populations, all sequences are populated at all times but when the mutation rate $\mu \to 0$, most of the population is localised at a fit sequence and therefore the population behaviour is largely determined by the properties of this sequence termed the *dominant sequence* as shown in Fig. 1.7. Using a quasispecies model [3, 24] of infinitely large populations, the number of times the dominant sequence changes till the population reaches the global fitness maximum is studied. Since the dominant sequence can change

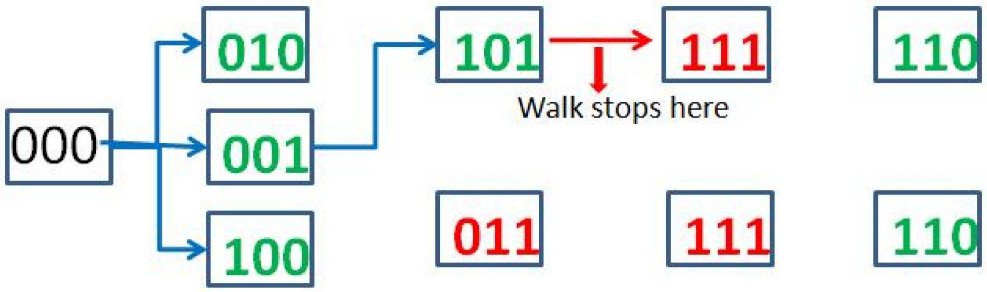


Figure 1.8: An adaptive walk of a population with sequence length $L = 3$. The one mutant sequences fitter than the current sequence are shown in green and the less fit ones are shown in red.

by more than one mutation, this model is also termed the *adaptive flight model*. Another quantity of interest in this model is *records* or sequences whose fitness is greater than all the previous dominant sequences. For strongly correlated fitness landscapes we find that the record probability is independent of the length of the sequence and the underlying fitness distribution. The average number of changes in the dominant sequence is also found to be independent of L but unlike records, it is shown to depend on the fitness distribution.

2. **Adaptive walk model:** If a finite population is subjected to strong selection, then unlike infinite population, it is localised at a single sequence. At low mutation rates, the population can move to fitter sequences one mutation away and can only reach a local fitness maximum as shown in Fig. 1.8. In the adaptive walk model, the population from the present sequence moves to any of the fitter one mutant neighbours with a probability that is proportional to the fitness difference between

the two. Using this model we find that the number of mutations accumulated during adaptation till it reaches a local fitness peak is proportional to the logarithm of the sequence length. Moreover we find the distribution of walk length and fitness at any step in the walk for uniform and exponential underlying fitness distributions.

An outline of adaptation models studied in this thesis is shown in Fig. 1.9.

1.4 Plan of the thesis

In Chapter 2, we review some of the earlier works that have dealt with the adaptation of finite and infinite populations. We discuss the steady state solution of quasispecies models in which a phase transition occurs on most fitness landscapes at a critical mutation rate μ_c, above which the population is uniformly distributed over all sequences. Below this value, the population is concentrated at the fittest sequence surrounded by a suite of mutants. Most theoretical works deal with the dynamics of the quasispecies population in the weak mutation regime, $\mu \to 0$. The analytical results for the dynamics on uncorrelated and fully correlated fitness landscapes is discussed here along with our numerical results for weakly correlated fitness landscape corresponding to block length $L_B = L/2$. For the adaptive walk model, the average walk length for the limiting cases of greedy walk, in which the population moves to the fittest of the one mutant neighbours and random walk, in which the population moves to any of the fitter one mutant neighbours with equal probability are explained. For the biologically relevant adaptive walk model, we discuss the theoretical results regarding the first step and

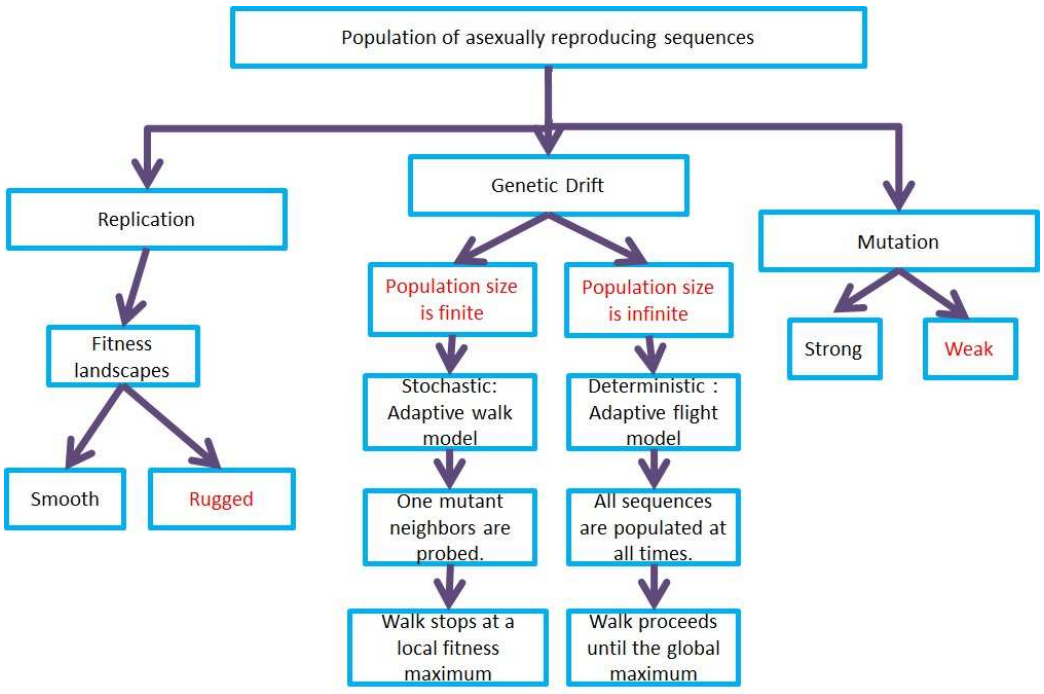


Figure 1.9: Models of evolution of asexual populations. The parameters we shall consider in our work are highlighted in red.

the experimental verification of the same.

In Chapter 3, we discuss our results on the dynamics of an infinite quasispecies population with strong fitness correlations that are introduced using the block model. Our exact calculations for the number of changes in the identity of the dominant sequence is shown when the block length $L_B = 2$. We find that this value is independent of the length of the sequence but varies with the underlying block fitness distributions. Other quantities of interest such as the record probability of a sequence fitness exceeding the fitness of the previous dominant sequences and the change in the fitness of the dominant sequences are also derived.

In Chapter 4 we present our results on the adaptive walk model. We

show that the number of steps in the adaptive walk has a logarithmic dependence on the length of the sequence with a prefactor determined by the underlying fitness distribution. The length of the walk when correlations are introduced using the block model is shown to increase linearly with the number of blocks. Other results pertaining to walk length and fitness distribution are also described in detail.

Finally in Chapter 5, we discuss a connection between the results of the quasispecies model and the adaptive walk model and briefly outline the open questions that we intend to address in the near future.

Bibliography

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

[2] H. A. Orr. *Nat. Rev. Genet.*, 6:119–127, 2005.

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

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

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

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

[7] R. Korona, C. H. Nakatsu, L. J. Forney, and R. E. Lenski. *Proc. Natl. Acad. Sci. USA*, 91:9037–9041, 1994.

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

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

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

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

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

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

[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] C. Carneiro and D.L. Hartl. *Proc. Natl. Acad. Sci. USA*, 107:1747–1751, 2010.

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

[17] S. Kauffman and E. Weinberger. *J. Theor. Biol.*, 141:211–245, 1989.

[18] T. Aita and Y. Husimi. *J. Theor. Biol.*, 182:469–485, 1996.

[19] L. Loewe and W. G. Hill. *Phil. Trans. R. Soc. B*, 365:1153–1167, 2010.

[20] M.A. Nowak. *Evolutionary Dynamics: exploring the equations of life*. Harvard University Press, 2006.

[21] J. Quer, R. Huerta, I. S. Novella, L. Tsimring, E. Domingo, and J.J. Holland. *J. Mol. Biol.*, 264:465–471, 1996.

[22] K. Jain and J. Krug. *Genetics*, 175:1275, 2007.

[23] K. Jain. *Genetics*, 179:2125, 2008.

[24] M. Eigen. *Naturwissenchaften*, 58:465 – 523, 1971.

Chapter 2

Review of previous works

2.1 Introduction

Consider a microbial population evolving in a complex environment that can be modeled by rugged fitness landscapes. At large times, most of the population resides at the globally fittest sequence of the fitness landscape and due to mutations, a suite of mutants is also present. If the population size is infinite, a nonzero population is present at *all* the sequences whereas a finite population produces only a small number of mutants around the present sequence [1] and may acquire a fitter mutation only if it does not get lost due to genetic drift as discussed in Chapter 1.

In this thesis, we examine the adaptation dynamics in these two limits. Before discussing our work, we review earlier results on adaptive dynamics obtained using the quasispecies and adaptive walk models. We also discuss the block model that is used to introduce fitness correlations in our system.

2.2 Block model of tunably correlated fitnesses

The block model was introduced by Perelson and Macken who were motivated by the observation that many biomolecules such as proteins and antibodies are composed of domains or partitions [2]. In this model, sequence of length L is divided into B independent blocks of equal length $L_B = L/B$, $1 \leq L_B \leq L$. Each block configuration is assigned a fitness which may also depend on the position of the block. In this work, we consider both the cases when a block configuration at any location in the sequence carries the same block fitness (position independent block fitness) and when the fitness of a block depends on its location in the sequence (position dependent block fitness) Thus the number of independent random fitnesses are 2^{L_B} or $B2^{L_B}$ respectively. In each case the block fitnesses are chosen *independently* from a common distribution with support on the interval $[l, u]$ where l and u are respectively the lower and upper limits of the block fitness distribution. The sequence fitness is given by the average of the corresponding block fitnesses.

The topographical features such as the number of local maxima depends on L_B. To explain this point, let us consider the model where the block sequence fitnesses are position dependent. For a sequence to be a local maximum, each of its B block sequences must also be a local maximum. Since a sequence is composed of independent blocks and the average number of local optima of a sequence of length L_B with i.i.d. fitness is $2^{L_B}/(L_B + 1)$, it follows that the average number n_{opt} of local maxima of a sequence of length L and block length L_B is given by $(2^{L_B}/(L_B+1))^B$ [2]. Except for $L_B = 1$ for which there is a single local (same as global) fitness peak, n_{opt} increases with

increasing L_B and L (see Fig. 2.1(a)). For example in the case of $L_B = 2$ there are $\approx 1.15^L$ local optima on an average. Arguing as above for local maximum, it can be seen that the globally fittest sequence is composed of identical blocks with the largest block fitness.

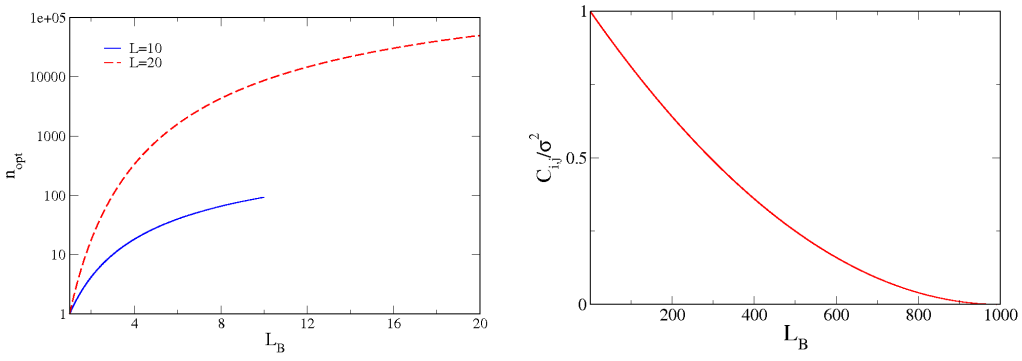
An attractive feature of the block model is that the correlations can be tuned with the block length L_B. As illustrated in Fig. 1.4, when two sequences have at least one common block, their respective fitnesses are correlated. For $L_B = 1$, the sequence fitnesses are maximally correlated while for $L_B = L$, we obtain the model with maximally uncorrelated fitnesses. This statement can be quantified by considering the correlation function $C_{0,j}$ between the fitness $w_0 = w(\sigma^{(0)})$ of a sequence, $\sigma^{(0)}$ with identical position independent blocks and the fitness w_j of a sequence one mutation away from it, which is given by

$$w_j = \frac{(L - L_B)f_0 + L_B f_j}{L} \ , \ j = 0, ..., L_B \tag{2.1}$$

where f_j is the fitness of the block of length L_B with 1 in the jth position. Using the fact that f_j's are i.i.d. random variables, we can write the correlation function as [2]

$$C_{0,j} = \langle w_0 w_j \rangle - \langle w_0 \rangle \langle w_j \rangle = \frac{L - L_B}{L} \sigma^2 \tag{2.2}$$

where σ^2 is the variance of the block fitness distribution $p(f)$. The above correlation function is largest at $L_B = 1$ and vanishes at $L_B = L$. Similarly



(a) (b)

Figure 2.1: (a) Average number $n_{\rm opt}$ of local maxima in block model as a function of L_B for various L. (b) Correlation function between two sequences separated by a single mutation as a function of the block length of the sequences.

the correlation function $C_{i,j}$ amongst the one mutant neighbors given by [3]

$$C_{i,j} = \langle w_i w_j \rangle - \langle w_i \rangle \langle w_j \rangle = \left[\left(\frac{L - L_B}{L} \right)^2 + \delta_{i,j} \left(\frac{L_B}{L} \right)^2 \right] \sigma^2 \qquad (2.3)$$

is a monotonically decreasing function of L_B for $i \neq j$ as can be seen in Fig. 2.1(b).

2.3 Quasispecies model

Let us consider an infinitely large population of binary sequences where a sequence $\sigma \equiv \{\sigma_1, ..., \sigma_L\}$, $\sigma_i = 0, 1$ is a string of L letters. Due to the infinite size of the population, there are no fluctuations in the population frequency of a sequence and one can work with the averages. The population evolves by the elementary processes of replication and mutation. We study adaptation by tracking the evolution of the *dominant sequence* in time. In the

following, we begin with the quasispecies model of biological evolution [4, 5] and proceed to relate it to the shell model [6]. Now if the environment is changed by changing (say) the nutrient medium of the microbial population, the fittest subpopulation before the environment change will be typically maladapted to the new environment and due to the large population size, a small population may be present at the new fittest sequence. The question of interest is how the new global maximum is reached starting with an initial condition in which the whole population is at the sequence that was globally fittest before the environmental change.

If the replication rate $A(\sigma)$ of the sequence σ is defined as the average number of copies produced per generation and $p_{\sigma\leftarrow\sigma'}$ is the probability that a sequence σ' mutates to the sequence σ at a mutational distance $D(\sigma,\sigma')$ given by (1.1), the population fraction $X(\sigma,t)$ of sequence σ at time t evolves according to the following quasispecies equation [4, 7]:

$$X(\sigma,t+1)=\frac{\sum_{\sigma'}p_{\sigma\leftarrow\sigma'}A(\sigma')X(\sigma',t)}{\sum_{\sigma'}A(\sigma')X(\sigma',t)} \tag{2.4}$$

where the denominator on the right hand side ensures the normalisation condition $\sum_\sigma X(\sigma,t)=1$ is satisfied at all times and $p_{\sigma\leftarrow\sigma'}$ is defined in (1.2).

2.3.1 Steady state: Error threshold transition

For certain fitness landscapes, there exists a error threshold, μ_c below which at large times the fittest sequence is maximally populated surrounded by a suite of mutants [8]. If the mutation rate exceeds this value then the

population does not have a dominant sequence and is uniformly scattered on the fitness landscape. This can be shown by considering a sharp peaked fitness landscape given by [7]

$$A(\sigma) = A_0\delta_{\sigma,\mathbf{0}} + (1 - \delta_{\sigma,\mathbf{0}}) \ , \ A_0 > 1 \tag{2.5}$$

where $\mathbf{0} = \{0, 0, ..., 0\}$ is the fittest sequence with all zeros. This sequence has replication rate $A_0 > 1$ and all others have fitness 1. Using this choice for $A(\sigma)$ in (2.4) for the sequence $\mathbf{0}$, we get in the steady state

$$X(\mathbf{0}) = \frac{A_0(1-\mu)^L X(\mathbf{0}) + \sum_{\sigma' \neq \mathbf{0}} M(\sigma \leftarrow \sigma')X(\sigma')}{A_0 X(\mathbf{0}) + 1 - X(\mathbf{0})} \tag{2.6}$$

In the scaling limit $\mu \to 0, L \to \infty$, the terms in the numerator on the right hand side of the equation arising due to mutations to sequence $\mathbf{0}$ vanish and by rearranging the terms one obtains [9]

$$X(\mathbf{0}) = \frac{A_0(1-\mu)^L - 1}{A_0 - 1} = 1 - \frac{\mu}{\mu_c} \tag{2.7}$$

where $\mu_c = \ln A_0/L$ is the critical mutational probability. Thus the master sequence $\mathbf{0}$ supports a finite fraction of population below μ_c. Above the critical probability μ_c, the population is homogeneously distributed over the sequence space.

The error threshold for the two extreme cases of correlations corresponding to $L_B = L$ and $L_B = 1$ in the block model have also been calculated as explained below.

Block model: Uncorrelated fitness landscapes

The error threshold has been calculated on uncorrelated fitness landscapes for sequences of length L, whose fitness is defined as

$$A(\sigma) = e^{\kappa f(\sigma)} \tag{2.8}$$

where $\kappa > 0$ and $f(\sigma)$ is chosen from a Gaussian distribution given by

$$P(f) = \frac{1}{\sqrt{L\pi}} \exp\left(\frac{-f^2}{L}\right) \tag{2.9}$$

It has been shown that the critical mutation probability μ_c in this uncorrelated fitness landscape is given by [10–12]

$$\mu_c = 1 - \exp[\frac{\kappa^2}{4} - \kappa\sqrt{\ln 2}] \tag{2.10}$$

Block model: Fully correlated fitness landscapes

Here a multiplicative (or *Fujiyama*) fitness landscape is considered where the fitness of any sequence is given by

$$A(\sigma) = \prod_{i=1}^{L} e^{\lambda\sigma_i} = \exp[\lambda(L - 2D(\sigma, \sigma'))] \tag{2.11}$$

where $\lambda > 0$ and $\tilde{\sigma} = \{1, 1, 1, ..., 1\}$. It has been shown that a population evolving on this fitness landscape always exists in the localised phase and does not have a phase transition corresponding to the error threshold [13]. The logarithm of $A(\sigma)$ gives the fully correlated fitness landscape of the block model and hence the case, $L_B = 1$ does not have phase transition.

2.3.2 Dynamics: Shell model

The quasispecies equation (2.4) can be transformed by introducing the unnormalised population $Z(\sigma, t)$ defined as $X(\sigma, t) = Z(\sigma, t)/\sum_{\sigma'} Z(\sigma', t)$ [14]. The evolution of $Z(\sigma, t)$ is given as

$$Z(\sigma, t+1) = \sum_{\sigma'} \mu^{D(\sigma,\sigma')}(1-\mu)^{L-D(\sigma,\sigma')} A(\sigma') Z(\sigma', t) \tag{2.12}$$

We take a monomorphic initial population that is localised at the sequence $\sigma^{(0)}$ so that $X(\sigma, 0) = Z(\sigma, 0) = \delta_{\sigma,\sigma^{(0)}}$. In the first step, all sequences are populated depending on the number of mutations, $D(\sigma, \sigma^{(0)})$ that separate them from $\sigma^{(0)}$ as

$$Z(\sigma, 1) \sim \mu^{D(\sigma,\sigma^{(0)})} A(\sigma^{(0)}) \tag{2.13}$$

As the mutation probability $\mu \ll \mu_c$, beyond the first step the growth of the population at sequence σ will be dominated by the replication of the existing population in it than due to the mutation of other sequences. Then beyond the first step, the population $Z(\sigma, t+1) \sim A(\sigma) Z(\sigma, t)$ and earlier works [6, 7, 15] have shown that, the statistical properties of the dominant sequence in the quasispecies model are accurately described by a simplified *shell model* which approximates the solution of (2.12) by

$$Z(\sigma, t) \sim \mu^{D(\sigma,\sigma^{(0)})} A^t(\sigma) \tag{2.14}$$

It has been shown that this result exactly matches that of quasispecies model for the dominant sequence at all times but in case of others, it produces similar results only for highly fit sequences and for very short times [15].

Our quantities of interest involve only the dominant sequence at any time and hence this approximation works well in this case.

Taking the logarithm of both sides in (2.14) and rescaling by $|\ln \mu|$ to get rescaled logarithmic population $E(\sigma, t) \sim \ln Z(\sigma, t)$. This grows linearly in time with a slope $w(\sigma) = \ln A(\sigma)$ as given by the equation

$$E(\sigma, t) = -D(\sigma, \sigma^{(0)}) + w(\sigma)t \tag{2.15}$$

At $t = 0$ corresponding to the first step in evolution of our system, all sequences differing from $\sigma^{(0)}$ by $D(\sigma, \sigma^{(0)}) = D$ number of mutations will have the same initial seeding population. Thus we can consider them as emerging out of a shell of radius D with different growth rates. Of these $\binom{L}{D}$ sequences the fittest one with fitness $w^{(\max)}(D)$ grows the fastest and has the probability of becoming the dominant sequence in the population. Though in this case the fittest sequence at every shell is non-identically distributed, there are also other models, in which it is identically distributed [6]. At every D value we denote this sequence by its shell number and in both cases, its growth is given by

$$E(D, t) = -D + w^{(\max)}(D)t \tag{2.16}$$

The fitness of the dominant sequence changes abruptly whenever its identity changes. This occurs whenever the population of the fittest sequence of shell D equals, in the shortest time, the current dominant sequence in shell D' such that $D > D'$. From (2.16) the time at which this occurs can be calculated

to be

$$T(D, D') = \frac{D - D'}{w^{(\max)}(D) - w^{(\max)}(D')} \tag{2.17}$$

From the above equation we see that for the fittest sequence in shell D to be the dominant sequence, it must be fitter than all sequences in $D' < D$, that is it must be a *record*. However, not every record will become the dominant sequence since for $D > D'$, the record in shell D might equal the current dominant sequence faster than that in shell D'. This abrupt change in fitness of the dominant sequence is termed a *jump*. For example, for $L = 4$, the growth of the logarithmic populations of the fittest sequence at every D is shown in Fig. 2.2 for uncorrelated fitnesses. For the choice of fitnesses made, the sequence at $D = 2$ becomes the dominant sequence after $D = 0$ and it is later replaced as the dominant sequence by the one at $D = 4$ even though the fitnesses at $D = 1$ and $D = 3$ are also records. Thus to become the dominant sequence, not only should a sequence be a record but it should also overtake the current dominant sequence in minimum time.

The quantities of interest are the statistics of the number of records, number of changes in the dominant sequence when the correlations between the fitnesses are varied using the block model as explained in Chapter 1. The fitness and the number of blocks of each kind are represented by f_i and n_i respectively, where i is the decimal equivalent of the block sequence so that $i = 0, 1..., L_B$. The constraint is $\sum_{i=1}^{L_B} n_i = L/L_B$. For the sake of convenience, the initial sequence is chosen to be a string of 0s and has fitness f_0. The known results for the extreme cases $L_B = 1$ and $L_B = L$ and the

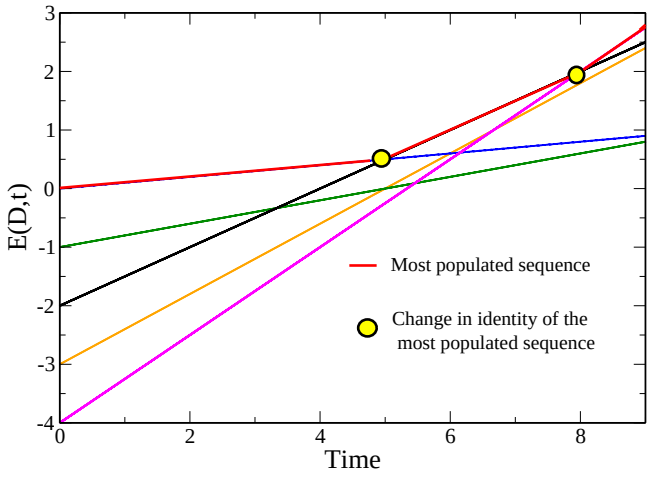


Figure 2.2: The growth of the fittest logarithmic population at each D. The change in dominant sequence is indicated by the circle.

numerical results for $L_B = L/2$ are discussed in this Chapter.

Dynamics of shell model when $L_B = 1$

For the position independent blocks in the fully correlated case, $L_B = 1$, only two kinds of block sequences are possible, namely $\{0\}$ and $\{1\}$ with fitness f_0 and f_1 respectively. At any D, the maximum fitness is given by

$$w^{(\max)}(D) = \frac{n_0 f_0 + (L - n_0) f_1}{L} \tag{2.18}$$

It is obvious that when $f_0 > f_1$, the total number of records and changes in the dominant sequence is 0. But if $f_1 > f_0$, there is a record at every D. So the average number of records is $L/2$. In this case, the time at which the

fittest population at any D will equal the original population is

$$T(D,0) = \frac{D-0}{w^{(\max)}(D) - w^{(\max)}(0)} = \frac{1}{f_1 - f_0} \tag{2.19}$$

As (2.19) is independent of D, all record populations equal the population of the initial dominant sequence at the same time and beyond this the fittest sequence at $D = L$ assumes this identity. So the average number of changes in the identity of the dominant sequence is 0.5 [16]. Both the results are independent of the underlying fitness distribution.

Dynamics of shell model when $L_B = L$

At each D there are $\binom{L}{D}$ sequences with i.i.d. fitnesses. The distribution of the maximum fitness amongst them is

$$P_{max}(f) = \binom{L}{D} p(f) \left(\int_l^f p(f')df' \right)^{\binom{L}{D}-1} \tag{2.20}$$

For large $\binom{L}{D}$, the distribution $P_{max}(f)$ moves towards the tail of the fitness distribution, $p(f)$. From *extreme value statistics*, the asymptotic distribution of $P_{max}(f)$ can be shown to fall into one of the three universal distributions depending on the underlying distribution $p(f)$ [17]

- *Gumbel distribution*: If $p(f) = \gamma f^{\gamma-1} e^{-f^\gamma}$, $\gamma > 0$
- *Fréchet distribution*: If $p(f) = (\delta - 1)(1+f)^{-\delta}$, $\delta > 1$
- *Weibull distribution*: If $p(f) = \nu(1-f)^{\nu-1}$, $\nu > 0, f < 1$

In all three cases, the probability of D being a record is given by

$$P_R(D) = \frac{\binom{L}{D}}{\sum_{i=0}^{D}\binom{L}{i}} \approx 1 - \frac{D}{L-D} \tag{2.21}$$

From the above equation we can show that the average number of records increases linearly with L as $(1-\ln 2)L$. In case of exponential distribution, the number of jumps has been calculated as $\mathcal{J} = \sqrt{L\pi}/2$ and the temporal distribution of the last jump is shown to have $1/t^2$ dependence [6, 7, 15].

Dynamics of shell model when $L_B = L/2$

When a sequence is built of two blocks of length $L_B = L/2$, the total number of possible blocks is 2 2^{L_B} when the block fitnesses are position dependent and 2^{L_B} in case of position independent block fitnesses. We work with exponential fitness distribution and generate the fitness of the extreme cases corresponding to $D' = 0$ and $L/2$ from it. Since it is computationally very expensive to generate $\binom{L/2}{D}$ random variables and choose the maximum amongst them at each D, we produce them for all $1 \leq D < L/2$ from the Gumbel distribution. If $f(j,k)$ is the fittest amongst the $\binom{L_B}{k}$ blocks at position j having k mutations, the maximum at each D is determined by comparing $f(1,i) + f(2, D-i)$ between all $i = 0, .., D$ and choosing the highest. The sequence fitness at D is a record if its fitness is higher than all $D' < D$. Every time the population of the leader is overtaken by another in minimum time, the number of jumps is incremented by 1. The whole process is iterated 10^6 times and the average calculated.

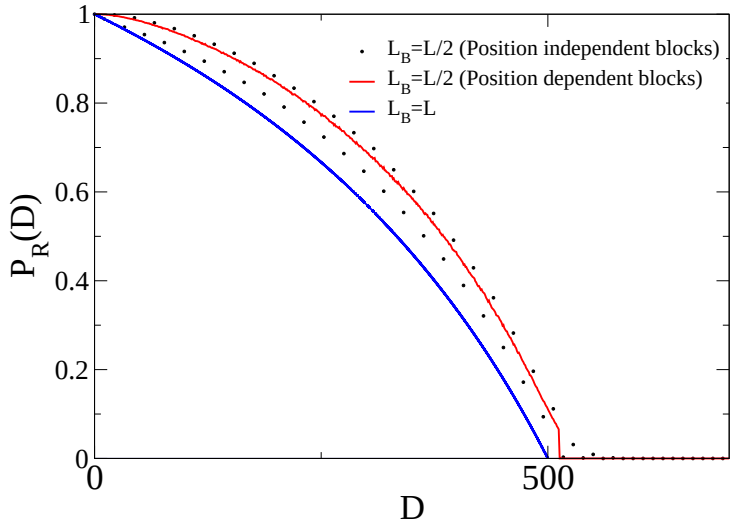


Figure 2.3: The simulation results for the record distribution $P_R(D)$ when $L = 1000$.

The population at $D < L/2$ is a record when [18]

$$f(1,i) + f(2, D-i) = max\{(f(1,i) + f(2, D'-i))\} \tag{2.22}$$

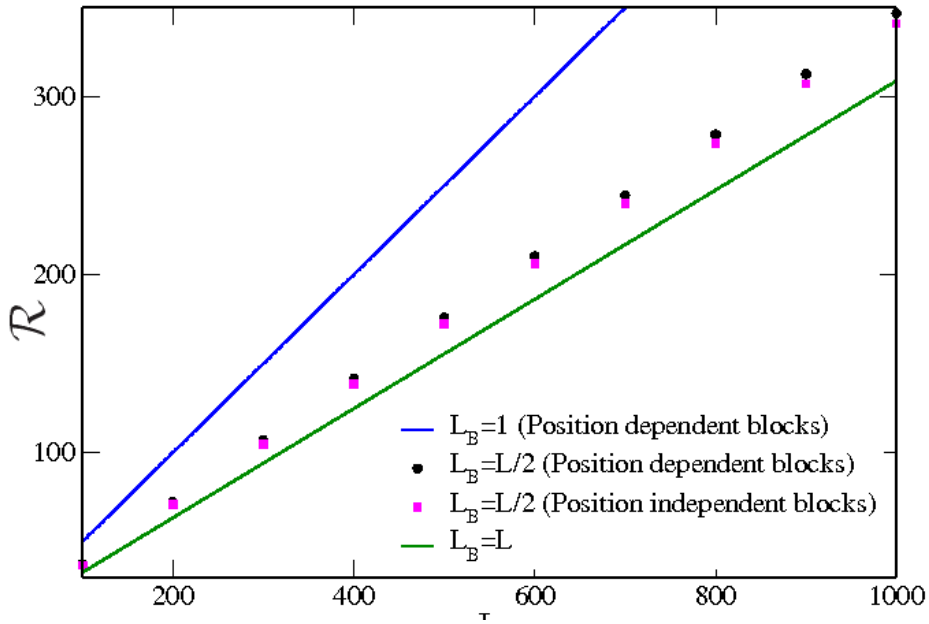
where $D' = 0, .., D$ and $i = 0, .., D$. For the first step this value is given by

$$P_R(1) = \frac{2L_B}{L_B+1} - \left(\frac{L_B}{L_B+1}\right)^2 \tag{2.23}$$

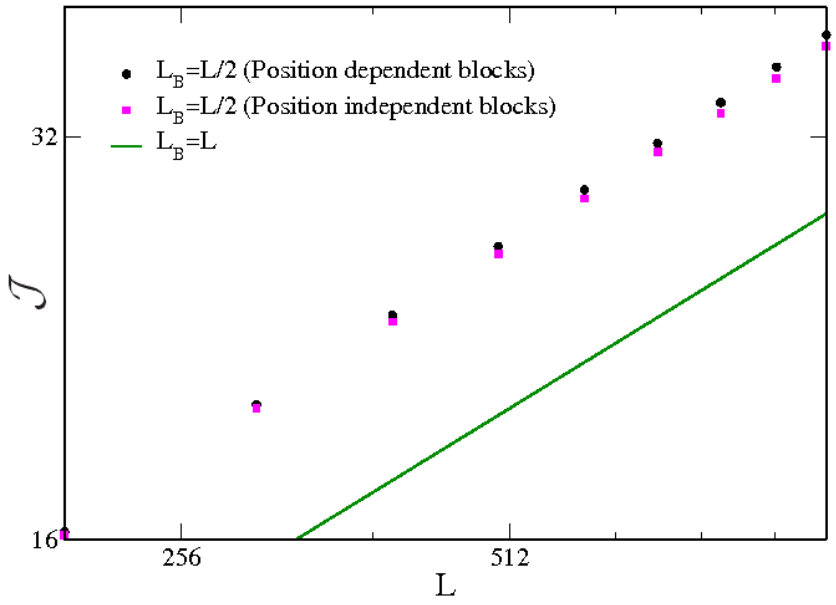
In case of position independent fitness, the position of the block loses its significance and so $f(1,i) = f(2,i)$ and the probability of the first step being a record is trivial with

$$P_R(1) = \frac{L_B}{L_B+1} \tag{2.24}$$

The plot of $P_R(D)$ against D is shown in Fig. 2.3 for both position dependent



(a)



(b)

Figure 2.4: (a) The simulation results for the average number of records when $L_B = L$ and $L_B = L/2$ along with the theoretical expressions of $L_B = L$ and $L_B = 1$. (b) The simulation results for the average number of jumps when $L_B = L/2$ for position independent and position independent exponentially distributed fitnesses along with the theoretical expression for $L_B = L$.

and independent blocks with exponentially distributed fitness. The record probability of position independent model displays a similar pattern as $L_B = 2$ (See Chapter 3) with even D having a higher value than odd D.

The general expression for $P_R(D)$ and the average number of records are yet to be worked out. The simulation results comparing the average number of records between $L_B = L$ and $L_B = L/2$ are shown in Fig. 2.4(a). Moving on to jumps, the average number of jumps scales with B and in the case of exponential distribution it is found to be, $\mathcal{J} \propto \sqrt{L_B}$ as shown in Fig. 2.4(b) [3, 18].

Having dealt with large populations using the quasispecies model, we now shift our focus to small populations and study them using an adaptive walk model.

2.4 Adaptive walk models

We now turn to a discussion of an adaptive walk model for a population of haploid binary sequences of size N [19]. The model is defined in the strong selection, $Ns \gg 1$ and weak mutation $N\mu \ll 1$ regime where s is the selection coefficient and μ is the mutation probability per locus per generation. In this regime, beneficial mutations arise sequentially and fix rapidly [20]. Due to the small number of mutants produced in any generation, it is a good approximation to neglect two or higher mutations and assume that the mutational neighbourhood accessible to a sequence at any time comprises of only its L one-step mutants. Due to the strong selection, one of the better mutants will sweep through the population earlier than the others and thus,

the population can be considered to be localised at a single genotype. Such a monomorphic population performs an adaptive walk by moving uphill on a fitness landscape until no more beneficial mutations can be found.

If the sequence at which the population is located and its one mutant neighbours are ranked according to their fitnesses, with rank 1 accorded to the sequence with the highest fitness, the transition probability that the population will move from the present rank i to a fitter one mutant of rank j is given by

$$T(j \leftarrow i) = \frac{f(j) - f(i)}{\sum_{k=1}^{i-1} f(k) - f(i)} \ , \ 1 \leq j \leq i-1 \tag{2.25}$$

The underlying fitness distribution is chosen from one of these three categories

$$p(f) = \begin{cases} (\delta - 1)(1+f)^{-\delta} & , \delta > 2 \quad \text{(Fréchet)} & (2.26) \\ \gamma f^{\gamma-1} e^{-f^{\gamma}} & , \gamma > 0 \quad \text{(Gumbel)} & (2.27) \\ \nu(1-f)^{\nu-1} & , \nu > 0, f < 1 \quad \text{(Weibull)} & (2.28) \end{cases}$$

since the distribution of the maximum of a large number of values from these will belong to the universal extreme value distributions as discussed in Sec. 2.3.2. The first step in the adaptive walk model has been well studied and we discuss it in Sec. 2.4.3. The two limiting cases for which adaptive walk has been analysed are defined as

- Greedy walk: This is obtained from (2.25) when $\delta \to 1$ in (2.26). The population moves from the present sequence to the fittest one mutant neighbour with probability one so that $T(j \leftarrow i) = \delta_{j,1}$ (see Sec. 2.4.1).

- Random walk: This is obtained from (2.25) when $\nu \to 0$ in (2.28). The population moves to any one of the fitter one mutant neighbours with equal probability so that $T(j \leftarrow i) = \dfrac{1}{i-1}$ (see Sec. 2.4.2).

In the next few subsections assuming uncorrelated fitness landscapes we calculate the average length of the walk, that is, the average number of mutations that occurs till the population reaches a local fitness peak in case of greedy walk and random walk. We show that greedy walk has the shortest walk length and the random walk has a walk length that diverges with L. The average walk length of the adaptive walk is expected to be in between these two values and shall be dealt with in Chapter 4. However, this Chapter shall deal with the properties of the first step of the adaptive walk.

2.4.1 Greedy walk

As mentioned before, in this model the population moves to the fittest of all the L one mutant neighbours of the present sequence [22]. This process stops when all the nearest neighbours have a lower fitness than the present sequence. We are interested in the record probability, $\alpha_J(L)$ defined as the probability of a sequence at step J being fitter than all other sequences in the steps $J' < J$ and the probability, $P_J(L)$ of the walker taking at least J steps from which the average walk length $\bar{J}$ can be calculated. We first discuss the known results for $L \to \infty$ and then present some of our results for finite L.

When $L \gg 1$ we can ignore the fact that a few of the one mutant neighbours have been already sampled in the walk and approximate the number

of novel mutants by L at each step. The probability that the step J is a record is given by

$$\alpha_J(L \to \infty) = \frac{L}{JL} = \frac{1}{J} \tag{2.29}$$

The walk will last at least J steps if at each step, the population encounters a record, the probability of which is given by

$$P_J(L \to \infty) = \prod_{i=1}^{J} \alpha_i = \frac{1}{J!} \tag{2.30}$$

Since the probability of taking exactly J steps is given by

$$Q_J(L \to \infty) = P_J(L \to \infty) - P_{J+1}(L \to \infty) = \frac{J}{(J+1)!} \tag{2.31}$$

we have the average length of the walk as [22]

$$\bar{J}(L \to \infty) = \sum_{i=0}^{\infty} JQ_J = e - 1 \tag{2.32}$$

We note that the above results are independent of the distribution $p(f)$ and thus the statistics is universal.

We now turn to a discussion of greedy walk on uncorrelated landscapes when the sequence length is finite [18]. When sequence length is small, as the number of one mutant neighbours sampled in the previous steps becomes comparable to the number of new neighbours that appear as the walk proceeds, it is necessary to keep track of the sequences and their corresponding fitness. We first attack the problem for small L by enumerating the possible paths that the walker can take starting from a binary sequence comprising of

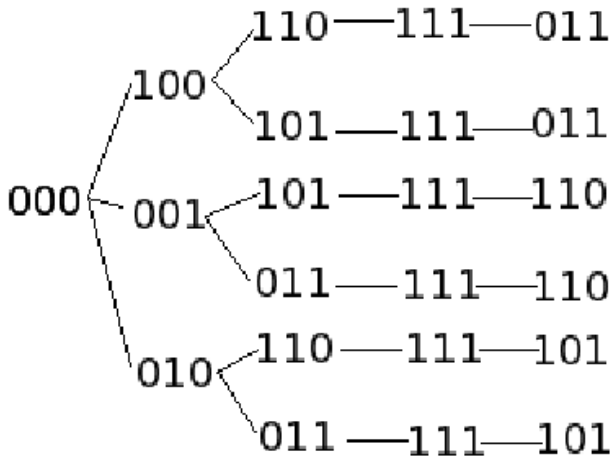


Figure 2.5: All possible path configuration for $L = 3$. The number of adaptive steps is independent of the path taken in the sequence space.

all 0s, and obtaining the probability $\alpha_J(L)$ that the step J has record fitness. We find that

$$\alpha_1(L) = \frac{L}{L+1} \tag{2.33}$$

$$\alpha_2(L) = \frac{L-1}{2L} \tag{2.34}$$

$$\alpha_3(L) = \frac{L-2}{3L-2} \tag{2.35}$$

$$\alpha_4(L) = \frac{L-3}{4L-4} + \frac{1}{4L-4} \tag{2.36}$$

$$\alpha_5(L) = \frac{L-3}{5L-6} + \frac{L-3}{(L-2)(5L-7)} \tag{2.37}$$

$$\alpha_6(L) = \frac{(L-3)^2}{(L-2)(6L-8)} + \frac{(L-3)^2}{(L-2)^2(6L-9)} + \frac{1}{6L-9} \tag{2.38}$$

In the first step, the path has L sequences to choose from and hence the record probability is given by $L/(L+1)$ whereas in the second and the third

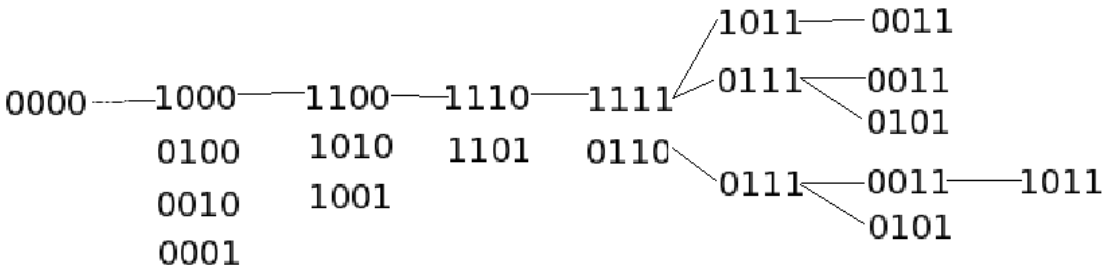


Figure 2.6: Some possible paths for $L = 4$ to illustrate sequence dependence $J = 5$ onwards. The number of adaptive steps depends on the path taken in the sequence space.

steps the number of new sequences encountered decreases by one and hence the values of $\alpha_2(L)$ and $\alpha_3(L)$ are $(L-1)/2L$ and $(L-2)/(3L-2)$ respectively. The paths for $L = 3$ are shown in Fig. 2.5. Here, all the paths are identical in the sense that the number of sequences accessible to a given sequence at some step is the same irrespective of the specific choice of the sequence. Interesting features of the problem emerge for higher L. When we consider $L = 4$, we find that $J = 4$ onwards, the subsequent paths cease to be identical. Depending on which particular sequence is chosen, different number of sequences become accessible for the next step. This history-dependence becomes more pronounced for $L = 5$ and higher, and appears as the absence of symmetry in the path configurations. Thus, the probability $\alpha_J(L)$ become history-specific, and we need to add separately the individual contributions of paths originating from different sequences as can be seen formulae (2.36-2.38). One set of paths is shown in Fig. 2.6 to illustrate the path-specificity of $\alpha_J(4)$. We can see that beyond the fourth step the number of new one mutant sequences depends on the present sequence thus necessitating the summing over different paths to obtain $\alpha_J(L)$. We also note that there is a

	By counting	By simulation
$\alpha_1(4)$	0.800000	0.799990
$\alpha_2(4)$	0.375000	0.375303
$\alpha_3(4)$	0.200000	0.200179
$\alpha_4(4)$	0.166666	0.166695
$\alpha_5(4)$	0.109890	0.110181
$\alpha_6(4)$	0.114582	0.114616
$\alpha_7(4)$	0.015625	0.015453

Table 2.1: Comparison of $\alpha_J(4)$ obtained by numerical simulations with those obtained by counting the possible paths.

$J^*(L)$ such that $\alpha_J(L)$ is zero for all $J \geq J^*(L)$. This occurs when all points in the sequence space accessible to a certain sequence have been explored previously, and hence the walk terminates. For example, when $L = 3$ this occurs at $J^*(3) = 5$ and for $L = 4$, $J^*(4) = 8$. Rosenberg had calculated the same quantities in [23]. However we note that the probabilities computed above are not in agreement with his results for two reasons. One, Rosenberg's formula does not take into account the sequence-specificity that arises at $J = 5$ and persists subsequently. Thus, the results in [23] give only the largest of the $\alpha_J(L)$ terms. Second, the formula does not incorporate the fact that $\alpha_J(L) = 0$ for all $J > J^*(L)$. For example, we find that $J^*(3) = 5$ implying $\alpha_5(3) = 0$ when $L = 3$ whereas Rosenberg's formula gives $\frac{1}{9}$. Thus we see that the analytical formulation of $\alpha_J(L)$ is a nontrivial problem, and we now resort to numerical simulations.

Numerical simulations were done where the probabilities $\alpha_J(L)$ were obtained by considering 10^6 different realisations of the fitness landscape, with fitnesses assigned randomly from an exponential distribution. Table 2.1 compares $\alpha_J(4)$ values obtained above with the numerical simulations.

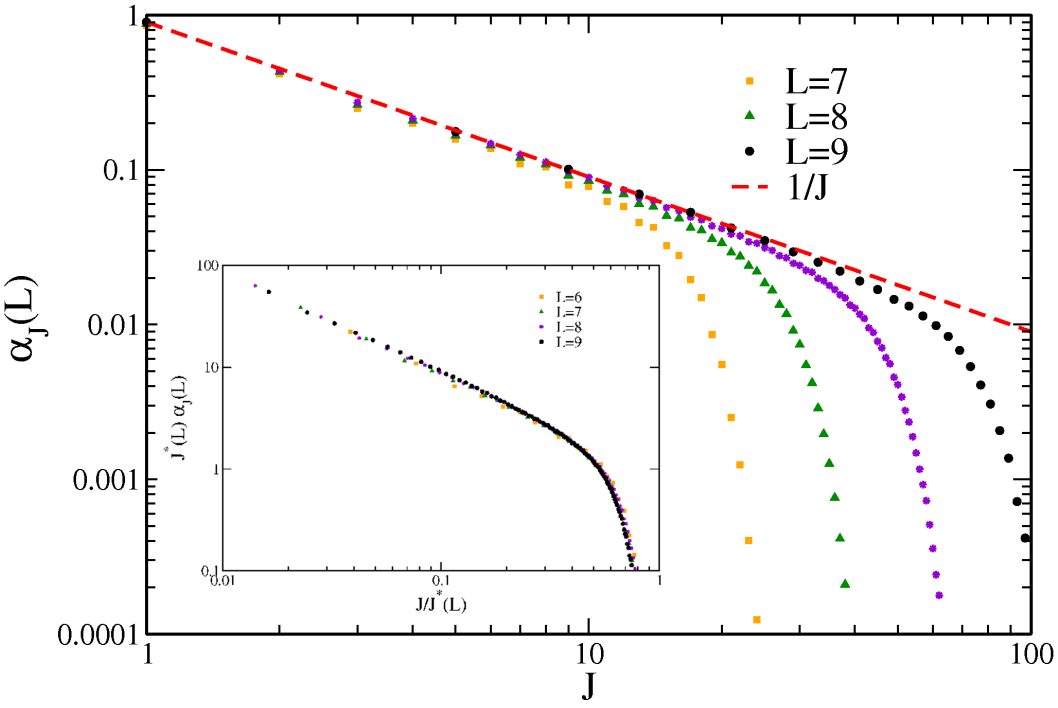


Figure 2.7: The variation of $\alpha_J(L)$ with J for various L and the inset shows the scaling collapse of $\alpha_J(L)$ with $J^*(L)$.

Our simulation results for $\alpha_J(L)$ against J for various L is plotted in Fig. 2.7. We notice that the plot shows a $1/J$ dependence for small J matching the large L limit and drops sharply as the finite effects set in. The plot thus has the following scaling form

$$\alpha_J(L) = \frac{1}{J} F\left(\frac{J}{\tilde{J}}\right) \tag{2.39}$$

where $F(x) = 1$ for $x \ll 1$ and decays for $x \gg 1$. Assuming that there is a single scale in the system, we expect that $\tilde{J} \propto J^*$ and therefore a data collapse can be obtained as shown in the inset of Fig. 2.7.

To find the dependence of the maximum walk length J^* on L, we plot

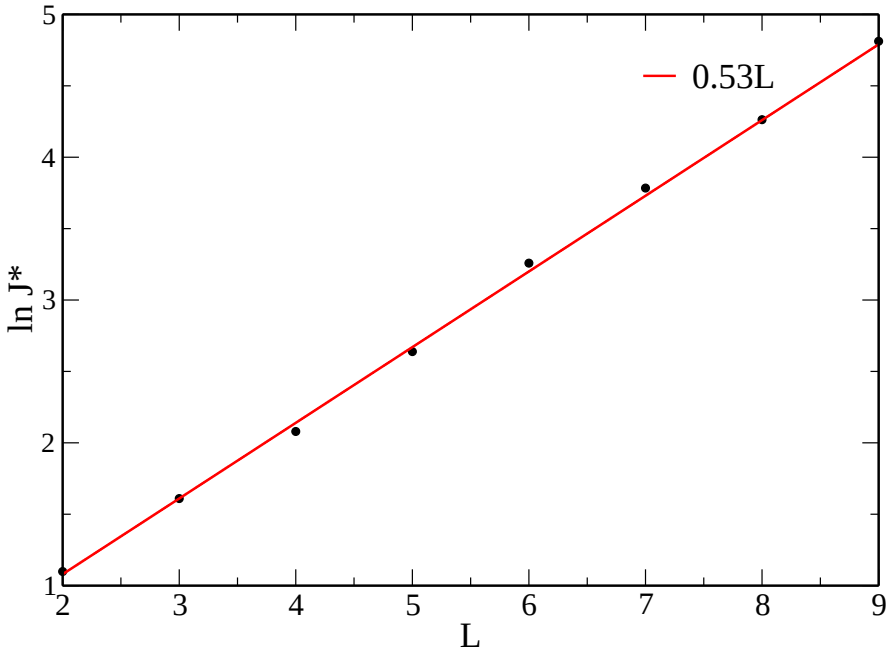


Figure 2.8: Straight line fit for log $J^*(L)$ versus L.

it as a function of L and find that there exists an exponential relationship between $J^*(L)$ and L (see Fig. 2.8). A plot of the mean walk length, $\bar{J}$ versus L is shown in Fig. 2.9. As discussed before, in the limit of large L, mean path length approaches $e-1$. To see how quickly the mean path lengths converge to this value, we plot the difference between this limit and the mean lengths for various L and note that the convergence is fast (inset of Fig. 2.9). Thus due to a fast rise in fitness at every step, the walk length in greedy walks is short and converges to a constant for large L. In the next section, we shall show that the other extreme case of random walk has an average walk length which diverges with L.

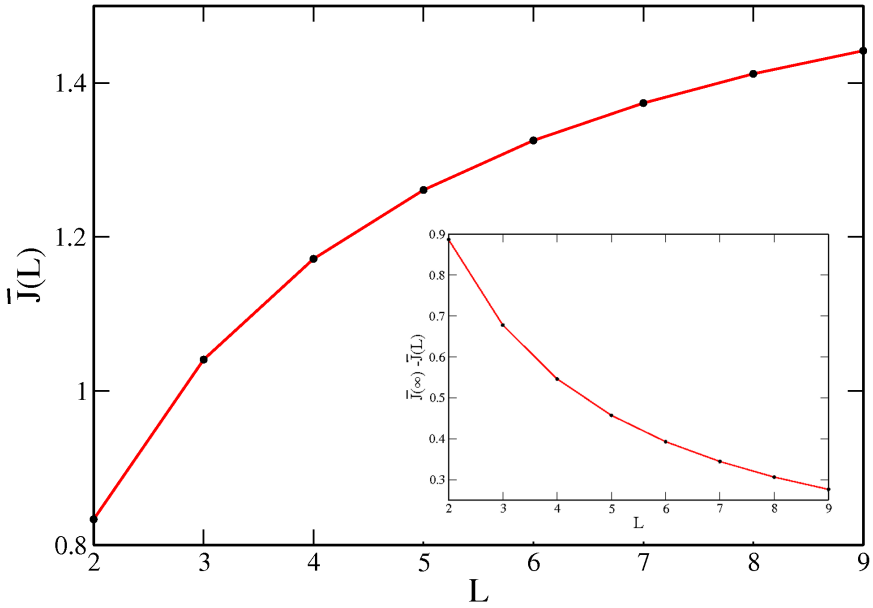


Figure 2.9: The variation of mean walk length with L and the inset shows the convergence of $\bar{J}(L)$ to the infinite limit with increasing L.

2.4.2 Random adaptive walk

Here we briefly review the known results for random adaptive walk in which all better mutants are chosen with equal probability [24, 26, 27]. The probability distribution of the walk lasting at least J steps and assuming fitness f in that step, $P_J(f)$ obeys the following recursion relation [26]:

$$P_{J+1}(f) = \int_l^f dh \, \frac{p(f)}{\int_h^u dg \; p(g)} \; \left[1 - q^L(h)\right] P_J(h) \tag{2.40}$$

where $q(f) = \int_l^f dg \; p(g)$, the probability that a sequence has fitness less than f. A change of variable from the fitness f to the cumulative probability $q(f)$ gives

$$P_{J+1}(q) = \int_0^q dq' \frac{1 - q'^L}{1 - q'} \; P_J(q') \tag{2.41}$$

If we define $P_J = \int_l^u df P_J(f)$, since the walk length distribution for the random adaptive walk obeys $Q_J = P_J - P_{J+1} = \int_l^u dh\ q^L(h) P_J(h)$, we have the probability that the walk lasts exactly J steps as

$$Q_J = \int_l^u dh\ q^L(h) P_J(h) = \int_l^u dq\ q^L P_J(q) \tag{2.42}$$

which shows that Q_J is a universal distribution in that it is independent of the underlying fitness distribution $p(f)$. Differentiating (2.41) with respect to q immediately gives

$$\frac{dP_{J+1}(q)}{dq} = \frac{1-q^L}{1-q}\ P_J(q) = \sum_{n=0}^{L} q^n\ P_J(q) \tag{2.43}$$

The generating function $G(x,q) = \sum_{J=1}^{\infty} x^J P_J(q)$ then obeys the following first order differential equation:

$$G'(x,q) - xP_1'(q) = x\frac{1-q^L}{1-q} G(x,q) \tag{2.44}$$

For the initial condition $P_0(f) = \delta(f)$, we have $P_1(q) = 1$ and due to (2.41), the distribution $P_J(0) = 0$. Solving the above differential equation using these boundary conditions gives $G(x,q) = xe^{xH_L(q)}$ where $H_L(q) = \sum_{k=1}^{L} q^k/k$ and hence the distribution $P_J(q)$ is given by [26]

$$P_J(q) = \frac{H_L^{J-1}(q)}{(J-1)!} \tag{2.45}$$

Since the product $q^L P_J(q)$ in random adaptive walk peaks around $q = 1$, using $H_L(q) \approx \ln L$ for q close to unity for finite but long sequences and

performing the integral in (2.42), we get

$$Q_J \approx e^{-\bar{J}} \frac{\bar{J}^{J-1}}{(J-1)!} \tag{2.46}$$

where $\bar{J} = \ln L$. Thus the walk length distribution is a Poisson distribution (in J) with mean $\bar{J} = \ln L$ [26].

We shall show in Chapter 4 that the average walk length of an adaptive walk lies between the two limiting case results of Greedy walk ($\bar{J} = e-1$) and random walk ($\bar{J} = \ln L$). But before we discuss the entire walk properties, we show in the next section the properties of the first step of the adaptive walk.

2.4.3 Adaptive walk: First step in the walk

We discuss the first step in the adaptive walk where the transition probability given by (2.25) depends on the fitness difference between the sequences [28]. Due to a change in the environment, the rank at which the population resides drops from 1 to i among the $L+1$ sequences where rank 1 corresponds to the fittest sequence, 2 to the second highest and so on. The rank at which the population now resides is assumed to be near the top in fitness (i is small) since a lot of mutations are deleterious and the change in rank cannot be drastic since environmental changes are mostly gradual.

When the number of sequences $L+1$ is large, from Extreme value theory that describes the tail behaviour of most distributions, the average of the

fitness spacing, Δ between the fittest sequences can be shown to be [17]

$$E[\Delta_1] = C_m \qquad E[\Delta_2] = C_m/2 \qquad E[\Delta_3] = C_m/3 \tag{2.47}$$

where Δ_1 is the fitness difference between the fittest sequence and the 2^{nd} fittest sequence, Δ_2 is the difference between the 2^{nd} and 3^{rd} fittest sequences and so on. The value of C_m is determined by the underlying fitness distribution. Using the fitness difference described above in (2.47) the mean transition probability is obtained as [28]

$$E[T(j \leftarrow i)] = \frac{1}{i-1} \sum_{k=j}^{i-1} \frac{1}{k} \tag{2.48}$$

and the average rank to which the population jumps is

$$E[j] = \frac{i+2}{4}. \tag{2.49}$$

When i is large, this value as expected is between the expected rank of the greedy walk where rank 1 is always fixed and the random walk where the mean average rank is $i/2$.

To test the above prediction Rokyta *et al.*, [29] carried out 20 single step adaptations from a single ancestral genotype of an icosahedral, single stranded DNA bacteriophage, ID11. Replicate populations were allowed to fix a single beneficial mutation under strong selection and weak mutation. The identity of the substitution was determined by whole-genome sequencing of each final population and the ranks of the sequences were determined by standard fitness assays [30].

From the adaptive walk model discussed above, the fittest mutation should occur the maximum number of times, the second fittest should be the second most frequent and so forth. But in this experiment, though the fittest mutation was a $G \rightarrow T$ transversion (pyrimidine-purine substitution), the most frequent substitutions were $C \rightarrow T$ transitions (pyrimidine-pyrimidine substitution). The discrepancy is due to the mutational bias since transitions occur at a higher rate than transversions. While in the adaptive walk model the mutation rates are averaged out and so, the mutational bias ignored, in real scenarios they may play a major role and the model is altered to accommodate it, so that the mean transition probability is written as

$$E[T(j \leftarrow i)] = \frac{\mu_j}{\sum_{k=1}^{i-1} \bar{\mu}_k} \sum_{k=j}^{i-1} \frac{1}{k} \tag{2.50}$$

where $\bar{\mu}_k = \frac{1}{k} \sum_{i=1}^{k} \mu_i$. With this correction the results are consistent with the experiment.

In the next Chapters, the progress that we have made and the results that we have obtained by building on these models shall be explained.

Bibliography

[1] K. Jain and J. Krug. *Genetics*, 175:1275, 2007.

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

[3] G. Das. *Dynamical properties of a quasispecies model on correlated fitness landscapes*. M.S. thesis, JNCASR, Bangalore, 2010.

[4] M. Eigen. *Naturwissenchaften*, 58:465 – 523, 1971.

[5] K. Jain and J. Krug. In *Structural Approaches to Sequence Evolution: Molecules, Networks and Populations*, pages 299–340. Springer, Berlin, 2007.

[6] J. Krug and C. Karl. *Physica A*, 318:137–143, 2003.

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

[8] T. Wiehe. *Genet. Res. Camb.*, 69:127–136, 1997.

[9] M. Nowak and P. Schuster. *J. theor. Biol.*, 137:375–395, 1989.

[10] B. Derrida. *Phys. Rev. B*, 24:2613–2626, 1981.

[11] C. Amitrano, L. Peliti, and M. Saber. *J. Mol Evol.*, 29:513, 1989.

[12] S. Franz, L. Peliti, and M. Sellitto. *J. Phys. A: Math. Gen.*, 26:L1195, 1993.

[13] G. Woodcock and P. G. Higgs. *J. theor. Biol.*, 179:61–73, 1996.

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

[15] K. Jain. *Phys. Rev. E*, 76:031922, 2007.

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

[17] E.J. Gumbel. *Statistics of extremes*. Columbia University Press, New York, 1958.

[18] S. Seetharaman. *Unpublished, 2011.*

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

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

[21] H.A. David and H.N. Nagaraja. *Order Statistics*. Wiley, New York, 2003.

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

[23] N.A. Rosenberg. *J. theor. Biol.*, 237:17–22, 2005.

[24] C. A. Macken and A. S. Perelson. *Proc. Natl. Acad. Sci. USA*, 86:6191–6195, 1989.

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

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

[27] S. A. Kauffman. *The Origins of Order*. Oxford University Press, New York, 1993.

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

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

[30] D.R. Rokyta, M.R. Badgett, I.J. Molineux, and J.J. Bull. *Mol. Biol. Evol.*, 19:230–238, 2002.

Chapter 3

Quasispecies model on strongly correlated fitness landscapes

3.1 Introduction

In the last Chapter we reviewed the quasispecies model and related it to the shell model. Also the known results for the shell model on fully correlated and uncorrelated fitness landscapes were discussed. But experiments [1, 2] show that real fitness landscapes have an intermediate degree of correlations. In Chapter 2, our simulation results for weakly correlated fitness landscapes, where $L_B = L/2$, were presented. In this Chapter we study the dynamics of adaptation in the quasispecies model using the shell model approximation, when the correlations amongst the fitnesses are high corresponding to the block length $L_B = 2$. We mainly deal with position independent block model, though a few results of the position dependent model are also discussed [3].

In the position independent block model, all sequences of length L are

built of blocks $\{0,0\}, \{0,1\}, \{1,0\}$ and $\{1,1\}$. The fitness of the four blocks is denoted by f_0, f_1, f_2 and f_3 and the number of blocks of each kind by n_0, n_1, n_2 and n_3 respectively. The initial sequence, $\sigma^{(0)}$ is a string of 0s and so, the constraints on the system is $2(n_0+n_1+n_2+n_3) = L$ and $n_1+n_2+2n_3 = D$. If the values of n_1 and n_2 are known, then from these two equations n_0 and n_3 are calculated as $n_0 = (L-D-n_1-n_2)/2$ and $n_3 = (D-n_1-n_2)/2$. The fitness of any sequence of length L and differing from the initial sequence by D mutations is given by

$$w_{n_1,n_2}(D) = \frac{(L-D-n_1-n_2)f_0 + 2n_1f_1 + 2n_2f_2 + (D-n_1-n_2)f_3}{L} \tag{3.1}$$

The constraints will hold only when for even D, the sum n_1+n_2 is even whereas for odd D, it is odd. Also in order to ensure the non-negativity of n_0 for $D \leq L/2$, the conditions $n_1+n_2 \leq D, n_1 \leq D$ must be satisfied as $n_3 \geq 0$ and for $D > L/2$, $n_1+n_2 \leq L-D, n_1 \leq L-D$ are required.

In the next section, we discuss the properties of the largest fitness at fixed D. In section 3.3 the statistics of records (introduced in Chapter 2) are studied and finally in section3.4.1 we address the statistics of jumps.

3.2 Extreme value statistics

Since our question of interest deals with the identity of the dominant sequence at any point of time, we need to consider only the fittest sequence at every D. If $f_1 > f_2$, then for fixed $n = n_1 + n_2$, the maximum fitness occurs when $n_1 = n$. Now the fitness of $w_{n+k,0}$, $k \neq 0$ for $D \geq 2$ is given by

$w_{n+k,0} = w_{n,0}(D) - (k/L)(f_0 + f_3 - 2f_1)$. When $f_1 > f_2$, the sign of the second term determines the value of n that gives the maximum fitness at each D and is shown to be

$$w^{(\max)}_{n_1,n_2}(D) = \begin{cases} w_{D,0}(D) & \text{if } f_0 - 2f_1 + f_3 < 0 \quad (3.2) \\ w_{1,0}(D) & \text{if } f_0 - 2f_1 + f_3 > 0 \text{ and } D \text{ is odd} \quad (3.3) \\ w_{0,0}(D) & \text{if } f_0 - 2f_1 + f_3 > 0 \text{ and } D \text{ is even} \quad (3.4) \end{cases}$$

For $D > L/2$, the largest possible fitness is obtained on replacing D by $L-D$ and when $f_1 < f_2$ the corresponding conditions are obtained by interchanging fitnesses f_1 and f_2 and the indices n_1 and n_2 in the preceding equations.

Only one of the above three fitnesses can be the maximum at constant D and the cumulative distribution of the maximum being less than w is given by $\mathcal{P}_E(w, D)$. For unbounded underlying distribution $p(f)$ with $f > 0$, we can write

$$\begin{aligned} \mathcal{P}_E(w, D) &= \int_0^\infty df_0 p(f_0) \int_0^\infty df_1 p(f_1) \int_l^u df_2 p(f_2) \\ &\quad \int_0^\infty df_3 p(f_3) \Theta(w - w_{D,0}) \Theta(w - w_{0,D}) \Theta(w - w_{0,0}) \\ &= \int_0^{\frac{w}{1-r}} df_0 \; p(f_0) \int_0^{\frac{w-(1-r)f_0}{r}} df_3 \; p(f_3) \left[\int_0^{\frac{w-(1-2r)f_0}{2r}} df_1 \; p(f_1) \right]^2 \quad (3.5) \end{aligned}$$

where $\Theta(...)$ is the Heaviside step function and $r = D/L < 1/2$. Specifically,

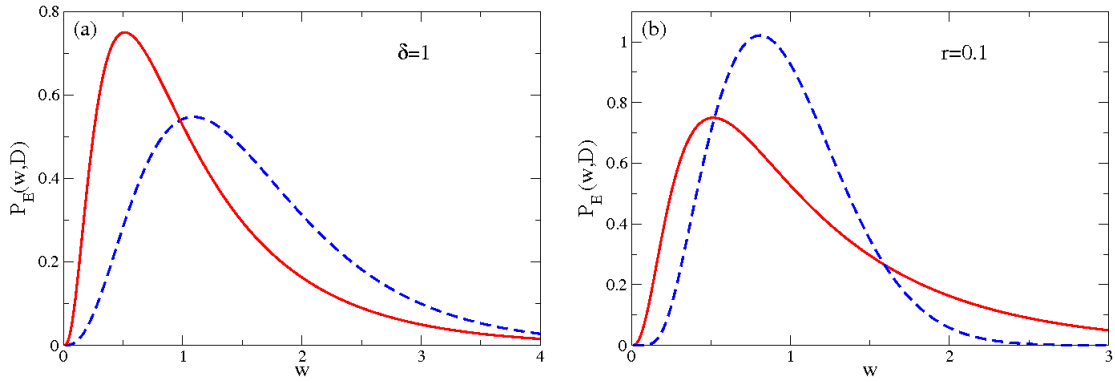


Figure 3.1: Distribution $P_E(w, D)$ of maximum fitness for (a) $r = 0.1$ (solid) and $r = 0.4$ (broken) with $\delta = 1$ and (b) $\delta = 1$ (solid) and $\delta = 2$ (broken) with $r = 0.1$.

for $p(f) = \delta f^{\delta-1} e^{-f^\delta},\ \delta > 0$, we have

$$\begin{aligned}\mathcal{P}_E(w,D) &= \delta \int_0^{\frac{w}{1-r}} df\, f^{\delta-1} e^{-f^\delta} \left[1 - e^{-\left(\frac{w-(1-r)f}{r}\right)^\delta}\right] \left[1 - e^{-\left(\frac{w-(1-2r)f}{2r}\right)^\delta}\right]^2 \\ &= a \int_0^1 df e^{-af} \left[1 - e^{-a\left(\frac{(1-r)(1-f^{1/\delta})}{r}\right)^\delta}\right] \left[1 - e^{-a\left(\frac{(1-r)(1-\frac{1-2r}{1-r}f^{1/\delta})}{2r}\right)^\delta}\right]^2 \end{aligned} \quad (3.6)$$

where $a = (w/(1-r))^\delta$. The probability $P_E(w, D) = d\mathcal{P}_E/dw$ that the largest sequence fitness with D mutations has a value w can be easily computed for $\delta = 1$ and is given by

$$P_E(w,D) = \frac{e^{-\frac{w}{1-r}} - e^{-\frac{2w}{r}} + 2e^{-\frac{3w}{2r}}}{1-2r} - \frac{2e^{-\frac{w}{2r}}}{1-4r} - \frac{\left(e^{-\frac{2w}{1-r}} - e^{-\frac{w}{r}}\right) r}{1-5r+6r^2} + \frac{4e^{-\frac{3w}{2(1-r)}} r}{1-6r+8r^2} \quad (3.7)$$

In the above expression when r is increased the plot shifts to the right as can be seen in Fig. 3.1(a). This is because the value of the maximum fitness

increases with D. Similar trend is also noted for higher δ values as the four fitness values to be chosen tend to stay close to the mean of the distribution (Fig. 3.1(b)). This is in contrast to finding the maximum of a large number of variables, where due to the fast decaying tails, the plot will move to the left for increasing δ values.

3.3 Statistics of record fitnesses

As explained in the previous section, the relation between block fitnesses determines the maximum fitness at any $D \leq L/2$ (see (3.2)-(3.4)). Here, the fitness $w_{D,0}(D)$ can be a record if it exceeds all the fitnesses at constant D as well as the ones with number of mutations $D' < D$. The first condition is met if (3.2) is satisfied. As the conditions in (3.2) are independent of D (barring parity), the largest fitness in a shell with D' mutations is also $w_{D',0}(D'), 1 < D' < D$. Then $w_{D,0}(D) > w_{D',0}(D')$ for all $D' \geq 0$ if $f_1 > f_0$. Thus the probability of $w_{D,0}(D)$ being a record can be written as

$$\begin{aligned} P(w_{D,0} \text{ is a record}) &= \int_l^u \prod_{i=0}^{3} df_i p(f_i) \Theta(f_1 - f_0) \Theta(f_1 - f_2) \Theta(2f_1 - f_0 - f_3) \\ &= \int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_l^{2f_1 - f_0} df_3 p(f_3) \end{aligned} \quad (3.8)$$

For $D > L/2$, the fitness $w_{L-D,0}(D)$ can be record if $w_{L-D,0}(D) > w_{L-D',0}(D')$ for $D' \geq L/2$ and $w_{L-D,0}(D) > w_{D',0}(D')$ for $D' < L/2$ along with the conditions $f_1 > f_2$ and $f_0 - 2f_1 + f_3 < 0$ (see (3.2)). The first two

inequalities are satisfied if $f_3 > f_1$ and $f_0 < f_1$. Thus we can write

$$\begin{aligned} P(w_{L-D,0} \text{ is a record}) &= \int_l^u \prod_{i=0}^{3} df_i p(f_i) \Theta(f_1 - f_0)\Theta(f_1 - f_2) \\ &\quad \times \Theta(f_3 - f_1)\Theta(2f_1 - f_0 - f_3) \\ &= \int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{f_1}^{2f_1 - f_0} df_3 p(f_3) \end{aligned} \quad (3.9)$$

For even D, the fitness $w_{0,0}(D)$ can be a record if $w_{0,0}(D) > w_{0,0}(D')$ for even D' and $w_{0,0}(D) > w_{1,0}(D')$ for odd D' besides satisfying (3.4). If $f_2 < f_1$, the fitness $w_{0,0}(D)$ can be a record if $f_3 > f_0$ and $f_3 > 2f_1 - f_0$. The last two conditions can be split into two cases, namely $f_3 > f_0$ if $f_0 > f_1$ and $f_3 > 2f_1 - f_0$ if $f_0 < f_1$. Similarly, for $f_2 > f_1$, the conditions for $w_{0,0}(D)$ to be a record are obtained by interchanging f_2 and f_1. Combining all the above conditions, we get

$$\begin{aligned} P(w_{0,0} \text{ is a record}) &= 2 \int_l^u \prod_{i=0}^{3} df_i p(f_i) \Theta(f_1 - f_2)\Theta(f_3 - f_0)\Theta(f_0 + f_3 - 2f_1) \\ &= 2[\int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{2f_1 - f_0}^u df_3 p(f_3) \\ &\quad + \int_l^u df_3 p(f_3) \int_l^{f_3} df_0 p(f_0) \int_l^{f_0} df_1 p(f_1) \int_l^{f_1} df_2 p(f_2)] \end{aligned} \quad (3.10)$$

For odd D, the fitness $w_{1,0}(D), D > 1$ can be a record if (3.3) is satisfied, $w_{1,0}(D) > w_{1,0}(D')$ for odd $D' < D$ and $w_{1,0}(D) > w_{0,0}(D')$ for even $D' < D$. The last two conditions are satisfied if $f_0 < f_3$ and $f_0 < f_1$ respectively. Then

the probability of $w_{1,0}(D), D > 1$ being a record is given by

$$\begin{aligned} P(w_{1,0} \text{ is a record}) &= \int_l^u \prod_{i=0}^{3} df_i p(f_i) \Theta(f_1 - f_0)\Theta(f_1 - f_2)\Theta(f_3 - f_0) \\ &\quad \times \Theta(f_0 + f_3 - 2f_1) \\ &= \int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{2f_1 - f_0}^u df_3 p(f_3) \end{aligned} \quad (3.11)$$

The above expression holds for $D = 1$ also as $w_{1,0}(1)$ is a record if $w_{1,0}(1) > w_{0,0}(0)$ which implies $f_0 < f_1$ besides $f_2 < f_1$.

3.3.1 Record occurrence distribution

Using the results derived above, we now calculate the probability $P_R(D)$ that a record occurs in the shell with $D > 0$ mutations given $P_R(0) = 1$. Fig. 3.2 shows that $P_R(D)$ is not a smooth function - the value of $P_R(D)$ depends on whether D is odd or even and whether it is below or above $L/2$. Thus four distinct cases arise due to this character of $P_R(D)$ which we will discuss below. We shall find that the distribution $P_R(D)$ is universal i.e. does not depend on the choice of the underlying distribution of the block fitness. As the global maximum is the last record and the only global maximum for $D > L/2$ occurs with probability $1/4$, we may expect the record occurrence probability for $D > L/2$ to be smaller than that for $D \leq L/2$.

Even D: When D is even, either $w_{D,0}(D)$ or $w_{0,D}(D)$ can be a record for $D \leq L/2$, $w_{L-D,0}(D)$ or $w_{0,L-D}(D)$ for $D > L/2$ or $w_{0,0}(D)$ for any even D

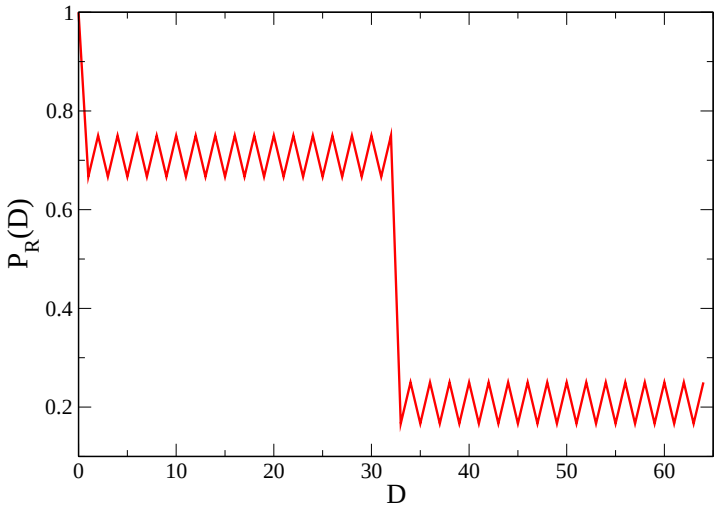


Figure 3.2: Variation of record occurrence probability $P_R(D)$ with the number of mutations D for $L = 64$.

.Thus the probability of even D for $D \leq L/2$ having a record is given by

$$\begin{aligned} P_R(D) &= 2(P(w_{D,0} \text{ is a record}) + P(w_{0,0} \text{ is a record})) \qquad (3.12) \\ &= 2\int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \\ &+ 2\int_l^u df_3 p(f_3) \int_l^{f_3} df_0 p(f_0) \int_l^{f_0} df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \\ &= \frac{2}{3} + \frac{1}{12} = \frac{3}{4}\ ,\ D \leq L/2 \qquad (3.13) \end{aligned}$$

Similarly for $D > L/2$, the record occurrence probability is given by

$$\begin{aligned} P_R(D) &= 2P(w_{L-D,0} \text{ is a record}) + P(w_{0,0} \text{ is a record}) && (3.14) \\ &= 2\int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{f_1}^u df_3 p(f_3) + \frac{1}{12} \\ &= \frac{1}{4}\ ,\ D > L/2 && (3.15) \end{aligned}$$

Odd D: For $w_{D,0}(D), D > 1$ to be a record when D is odd, the same conditions as for even D are required so that (3.8) holds. Thus the probability of a shell with odd $D, 1 < D \leq L/2$ having a record is given by

$$\begin{aligned} P_R(D) &= 2\left[P(w_{D,0} \text{ is a record}) + P(w_{1,0} \text{ is a record})\right] && (3.16) \\ &= 2\int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \\ &= \frac{2}{3}\ ,\ D \leq L/2 && (3.17) \end{aligned}$$

For $D > L/2$, the probability that $w_{L-D,0}(D)$ is a record is given by (3.9) and $w_{1,0}(D)$ is a record by (3.11). Thus the probability of a record occurring for odd $D > L/2$ can be expressed as

$$\begin{aligned} P_R(D) &= 2\left[P(w_{L-D,0} \text{ is a record}) + P(w_{1,0} \text{ is a record})\right] && (3.18) \\ &= 2\int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{f_1}^u df_3 p(f_3) \\ &= \frac{1}{6}\ ,\ D > L/2 && (3.19) \end{aligned}$$

3.3.2 Record value distribution

In this subsection, we calculate the probability $\mathcal{P}_R(w, D)$ that the record value in shell D is smaller than or equal to w. For this purpose, we will need the probability $\mathcal{P}_R(w(D) \leq w)$ that the fitness $w(D)$ in shell D does not exceed w. As the record value distribution is not expected to be universal, we will restrict ourselves to distributions with support on the interval $[0, \infty)$. It can be checked that the cumulative distribution $\mathcal{P}_R(w, D)$ gives the probability $P_R(w)$ obtained in the last subsection when $w \to \infty$. Below we present the expressions for $D \leq L/2$ as the corresponding distributions for $D > L/2$ can be written in an analogous manner.

Even D: As discussed before, the distribution for the record value is a function of the ratio $r = D/L$ for even D. Since either $w_{D,0}(D)$ or $w_{0,0}(D)$ can be a record for even $D \leq L/2$, the cumulative probability $\mathcal{P}_R(w, D) = 2\mathcal{P}(w_{D,0} \leq w) + \mathcal{P}(w_{0,0} \leq w)$ where

$$\begin{aligned}
\mathcal{P}(w_{D,0} \leq w) &= \int_0^\infty \prod_{i=0}^{3} df_i p(f_i)\ \Theta(w - w_{D,0})\Theta(f_1 - f_2) \\
&\quad \times\Theta(2f_1 - f_0 - f_3)\Theta(f_1 - f_0) \\
&= \int_0^w df_0 p(f_0) \int_{f_0}^{\frac{w-f_0}{2r}+f_0} df_1 p(f_1) \int_0^{f_1} df_2 p(f_2) \int_0^{2f_1-f_0} df_3 p(f_3) \qquad (3.20)
\end{aligned}$$

and

$$\begin{aligned}\mathcal{P}(w_{0,0} \leq w) &= 2\int_0^\infty \prod_{i=0}^{3} df_i p(f_i)\ \Theta(w - w_{0,0})\Theta(f_3 - f_0) \\ &\quad \times\Theta(f_1 - f_2)\Theta(f_0 + f_3 - 2f_1) \\ &= 2\int_0^w df_0 p(f_0) \int_0^{f_0} df_1 p(f_1) \int_0^{f_1} df_2 p(f_2) \int_{f_0}^{\frac{w-f_0}{r}+f_0} df_3 p(f_3) \\ &\quad + 2\int_0^w df_0 p(f_0) \int_{f_0}^{\frac{w-f_0}{2r}+f_0} df_1 p(f_1) \int_0^{f_1} df_2 p(f_2) \int_{2f_1-f_0}^{\frac{w-f_0}{r}+f_0} df_3 p(f_3) \quad (3.21)\end{aligned}$$

Using these expressions, it is straightforward to see that

$$\begin{aligned}\mathcal{P}_R(w, D) &= 2\int_0^w df_0 p(f_0) \int_0^{f_0} df_1 p(f_1) \int_0^{f_1} df_2 p(f_2) \int_{f_0}^{\frac{w-f_0}{r}+f_0} df_3 p(f_3) \\ &\quad + 2\int_0^w df_0 p(f_0) \int_{f_0}^{\frac{w-f_0}{2r}+f_0} df_1 p(f_1) \int_0^{f_1} df_2 p(f_2) \int_0^{\frac{w-f_0}{r}+f_0} df_3 p(f_3) \quad (3.22)\end{aligned}$$

Taking the derivative of the last expression with respect to w, we obtain the distribution $P_R(w, D)$ that the record value equals w. For $p(f) = e^{-f}$, the distribution $P_R(w, D)$ is given by

$$P_R(w, D) = \frac{e^{-4w} + 2e^{-\frac{3w}{2r}} - e^{-\frac{2w}{r}}}{1 - 2r} - \frac{2e^{-\frac{w}{2r}}}{1 - 4r} + \frac{e^{-2w}(3 - 8r)}{1 - 6r + 8r^2} + \frac{e^{-\frac{w}{r}} r - e^{-3w}(3 - 8r)}{1 - r(5 - 6r)} \quad (3.23)$$

The above result for the record value distribution is compared with the extreme value distribution $P_E(w, D)$ given by (3.7) in Fig. 3.3 for two values of r. Though the record fitness is also the extreme fitness in shell D, the converse is not true and the distribution $P_R(w, D) < P_E(w, D)$ for all w at a given D. We also note that the most probable record value in shell D is smaller than the corresponding extreme value - this behavior is unlike that

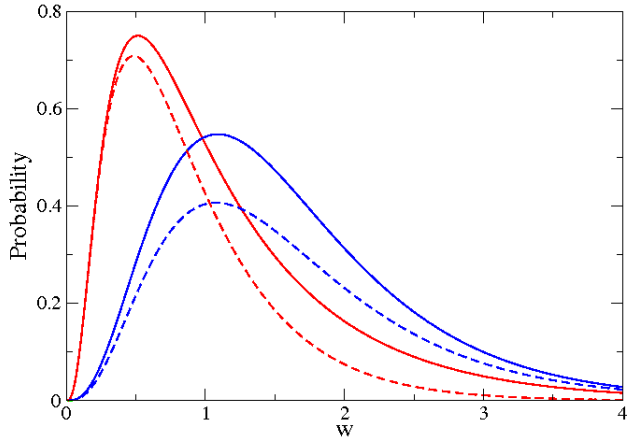


Figure 3.3: The probability distribution of the extreme value (solid lines) given by (3.7) and record value (dashed lines) by (3.23) for $r = 0.1$ (left curves) and $r = 0.4$ (right curves) for $p(f) = e^{-f}$.

for uncorrelated fitnesses for which record is a maximum of a larger set of independent variables.

Odd D: To find the record value distribution for odd D, besides $\mathcal{P}(w_{D,0} \leq w)$, we require the cumulative probability $\mathcal{P}(w_{1,0} \leq w)$ that the fitness $w_{1,0}(D)$ in shell D does not exceed w. The latter can written as

$$\begin{aligned}\mathcal{P}(w_{1,0} \leq w) &= \int_0^\infty \prod_{i=0}^{3} df_i p(f_i)\ \Theta(w - w_{1,0})\Theta(f_1 - f_2) \\ &\quad \times\Theta(f_0 + f_3 - 2f_1)\Theta(f_1 - f_0)\Theta(f_3 - f_0) \\ &= \int_0^w df_0 p(f_0) \int_{f_0}^{\frac{L(w-f_0)}{2D}+f_0} df_1 p(f_1) \int_0^{f_1} df_2 p(f_2) \int_{2f_1-f_0}^{\frac{Lw-(L-D-1)f_0-2f_1}{D-1}} df_3 p(f_3) \qquad (3.24)\end{aligned}$$

which reduces to the second integral in (3.21) for $L \gg 1$. Thus for large L, the cumulative distribution $\mathcal{P}_R(D, w)$ for odd D is also a function of r. However

unlike extreme value distribution for odd D, the distributions for even and odd D do not match for $L \gg 1$ as the expression for the distributions for the distributions $\mathcal{P}(w_{1,0} \leq w)$ and $\mathcal{P}(w_{0,0} \leq w)$ do not coincide.

3.3.3 Distribution of the number of records

To find the probability $N_R(n)$ that the total number of records equals n, we first calculate the record configuration probability $Q(\{w_{n_1,n_2}(D)\})$ defined as the probability that all the elements in the set $\{w_{n_1,n_2}(D)\}$ are records. This distribution depends on the location of the global maximum. If f_0 is the largest block fitness, the global maximum occurs at $D = 0$ and obviously there are no records beyond $D = 0$ in this case.

When f_0 is not a global maximum and $f_1 > f_2$, only four record configurations occur with a nonzero probability. When the fittest block has a fitness f_1, a record cannot occur beyond $D = L/2$ and only the conditions in (3.8) are satisfied since $2f_1 - f_0 - f_3$ must be positive. Thus the fitness $w_{D,0}(D)$ for all $D \leq L/2$ is a record with probability

$$Q(w_{1,0}(1), ..., w_{L/2,0}(L/2)) = \frac{1}{4} \tag{3.25}$$

When the block fitness f_3 is the largest, the records occur until $D = L$ at a spacing of one or two depending on the sign of $f_1 - f_0$ as explained below:

(i) Using the conditions in (3.10), we can see that when $f_2 < f_1 < f_0 < f_3$, a record occurs only in even D shells. As f_i's are independent and identically distributed (i.i.d.) random variables, all 4! block fitness configurations are

equally likely and therefore we get

$$Q(w_{0,0}(2), w_{0,0}(4), ..., w_{0,0}(L)) = \frac{1}{24} \tag{3.26}$$

(ii) If $f_1 > f_0$ (and f_2), the fitness $w_{1,0}(1)$ is a record. The next record depends on the sign of $2f_1 - f_0 - f_3$. From (3.10) and (3.11), it follows that if $2f_1 - f_0 - f_3 < 0$, the fitness $w_{0,0}(D)$ is a record for all even D and $w_{1,0}(D)$ for all odd D with probability

$$\begin{aligned} & Q(w_{1,0}(1), w_{0,0}(2), ..., w_{1,0}(L-1), w_{0,0}(L)) \\ & \quad = \int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{2f_1 - f_0}^u df_3 p(f_3) \end{aligned} \tag{3.27}$$

If $2f_1 - f_0 - f_3 > 0$, due to (3.8) and (3.9), the fitnesses $w_{D,0}(D)$ for all $D \leq L/2$ and $w_{L-D,0}(D)$ for all $D > L/2$ are records. This event occurs with probability

$$\begin{aligned} & Q(w_{1,0}(1), ..., w_{L/2,0}(L/2), w_{L/2-1,0}(L/2+1), ..., w_{0,0}(L)) \\ & \quad = \int_l^u df_0 p(f_0) \int_{f_0}^u df_1 p(f_1) \int_l^{f_1} df_2 p(f_2) \int_{f_1}^{2f_1 - f_0} df_3 p(f_3) \end{aligned} \tag{3.28}$$

From the above discussion, it is evident that the total number of records (ignoring the one at $D = 0$) can be either $L/2$ (due to (3.25) and (3.26)) or L (see (3.27) and (3.28)). The probability $N_R(n)$ of total number n of records is independent of underlying block fitness distribution and is given by

$$N_R(L/2) = 2\left(\frac{1}{4} + \frac{1}{24}\right) = \frac{7}{12}\ ,\ N_R(L) = \frac{2}{12} = \frac{1}{6} \tag{3.29}$$

where we have used that twice the sum of (3.27) and (3.28) equals (3.19). The average number $\mathcal{R}$ of records can be found using $N_R(n)$ or $P_R(D)$ and is given by

$$\mathcal{R} = \sum_{n=1}^{L} nN_R(n) = \sum_{D=1}^{L} P_R(D) = \frac{11L}{24} \approx 0.458L \tag{3.30}$$

for any even L.

3.4 Statistics of the jumps

As discussed in Chapter 2, all records are contenders for being a leader; however only those records for which the overtaking time is minimised qualifies to be a jump [4–6]. Like records, the statistics of jumps depends on the location of the global maximum. If f_0 is the fittest block, the unmutated sequence with fitness $w_{0,0}(0) = f_0$ is the leader throughout.

If $f_1(> f_2)$ is the global maximum, the last record and hence the last jump occurs at $D = L/2$. Since the time of intersection $T(0, D)$ of the population $E(D, t), D \leq L/2$ with the population $E(0, t)$ given by

$$T_1 = T(0, D) = \frac{D}{w_{D,0}(D) - w_{0,0}(0)} = \frac{L}{2(f_1 - f_0)} \ , \ D \leq L/2 \tag{3.31}$$

is independent of D, all the populations overtake the population of the initial sequence at the same point. Thus all the record populations participate in the evolutionary race. But as the population $E(L/2, t)$ has the largest fitness, it becomes the final leader thus leading to a single jump when f_1 (or f_2) is the largest fitness.

If the global maximum is f_3 which occurs at $D = L$, the following cases as discussed in Sec. 3.3.3 arise:

(i) If $f_1 < f_0$, the population with the record fitness $w_{0,0}(D), D \leq L$ overtakes that with the initial fitness $w_{0,0}(0)$ at a time given by

$$T_{3,1} = T(0, D) = \frac{D}{w_{0,0}(D) - w_{0,0}(0)} = \frac{L}{f_3 - f_0} \ , \ D \leq L \tag{3.32}$$

so that all the populations with record fitness $w_{0,0}(D)$ intersect at the same time and the population of the global maximum at $D = L$ takes over in a single jump.

(ii) If $f_1 > f_0$ and $2f_1 - f_0 - f_3 < 0$, the population with fitness $w_{0,0}(D)$ for all even D and $w_{1,0}(D)$ for all odd D intersects $E(0, t)$ at the following intersection time:

$$\begin{aligned} T(0, D) &= \frac{D}{w_{1,0}(D) - w_{0,0}(0)} \\ &= \frac{LD}{(D-1)f_3 + 2f_1 - (D+1)f_0} \ , \text{ for odd } D \qquad (3.33) \\ T(0, D) &= \frac{D}{w_{0,0}(D) - w_{0,0}(0)} = \frac{L}{f_3 - f_0} \ , \text{ for even } D \qquad (3.34) \end{aligned}$$

By virtue of the condition $2f_1 - f_0 - f_3 < 0$, the intersection time for odd D is greater than that for even D. Therefore the current leader at $D = 0$ is overtaken by $D = L$ resulting in a single jump at time $T_{3,2} = L/(f_3 - f_0)$.

If $2f_1 - f_0 - f_3 > 0$, the record fitnesses are $w_{D,0}(D)$ for $D \leq L/2$ and $w_{L-D,0}(D)$ for $D > L/2$. The populations corresponding to these fitnesses

overtake the leader at $D = 0$ at time

$$T(0, D) = \frac{L}{2(f_1 - f_0)} \ , \ D \leq L/2 \tag{3.35}$$

$$T(0, D) = \frac{DL}{(2D - l)f_2 + 2(L - D)f_1 - Lf_0} \ , \ D > L/2 \tag{3.36}$$

As the intersection time for $D \leq L/2$ is minimum amongst the rest and $w_{L/2,0}(L/2)$ is the largest fitness, the first jump occurs when the population of the sequence with fitness $w_{L/2,0}(L/2)$ overtakes $E(0, t)$. The next change in leader occurs at the point of intersection of populations involving the fitness $w_{L-D,0}(D), D > L/2$ with the current leader at a time

$$T_{3,3} = T(L/2, D) = \frac{L}{2(f_3 - f_1)} \ , \ D > L/2 \tag{3.37}$$

which is again D independent. Thus the population $E(L, t)$ is the leader after $E(L/2, t)$ and the global maximum is reached in two jumps.

3.4.1 Distribution of the number of jumps

It is obvious that when any block fitness other than f_0 is the globally largest fitness, there will be at least one jump (corresponding to globally fittest being the final leader) so that the probability of at least one jump equals 3/4. In addition, there can be one more jump when f_3 is the global maximum and $2f_1 - f_0 - f_3 > 0$ (see (3.37)). Due to (3.28), the probability p_2 of the second

jump is given by

$$p_2 = 2\int_l^u df_0 p(f_0)\int_{f_0}^u df_1 p(f_1)\int_l^{f_1} df_2 p(f_2)\int_{f_1}^{2f_1-f_0} df_3 p(f_3)\Theta(u-2f_1+f_0) \tag{3.38}$$

Thus the average number $\mathcal{J}$ of jumps is given by $(3/4)+p_2$. As p_2 is independent of L, the average number of jumps is of order unity for any underlying distribution but the constant p_2 is not universal. For instance, when the block fitnesses are chosen from an exponential probability distribution, $p_2 = 5/72 \approx 0.069$ while for uniform distribution, it equals $5/48 \approx 0.104$.

3.4.2 Temporal jump distribution

We are interested in the probability $P(t)$ that the last jump occurs at time $t > 0$ shown in Fig. 3.4 for $p(f) = e^{-f}$. This distribution is a sum of the probability $P_A(t)$ that the last jump occurs at t when f_1 or f_2 is a global maximum and $P_B(t)$ when f_3 is a global maximum. We first consider the cumulative probability $\mathcal{P}_A(t) = \int_0^t dt' P_A(t')$ which on using that f_1 (or f_2) is a global maximum and (3.31) gives

$$\begin{aligned}\mathcal{P}_A(t) &= 2\int_l^u \prod_{i=0}^{3} df_i p(f_i)\Theta(t-T_1)\Theta(f_1-f_0)\Theta(f_1-f_2)\Theta(f_1-f_3)\\ &= 2\int_{l+\frac{L}{2t}}^u df_1 p(f_1)\int_l^{f_1-\frac{L}{2t}} df_0 p(f_0)\int_l^{f_1} df_2 p(f_2)\int_l^{f_1} df_3 p(f_3)\end{aligned} \tag{3.39}$$

Differentiating $\mathcal{P}_A(t)$ with respect to time t yields

$$P_A(t) = \frac{-L}{2t^2}\frac{d\mathcal{P}_A}{d\epsilon} = \frac{L}{t^2}\int_{l+\epsilon}^u df p(f)p(f-\epsilon)\left(\int_l^f dg p(g)\right)^2 \tag{3.40}$$

where we have defined $\epsilon = L/(2t)$. For large times $t \gg L/2$, the integral on the right hand side of the above equation reduces to the probability $G(0)$ that the gap between the globally largest and the second largest in a set of i.i.d. random variables is zero [5]. Thus the probability $P_A(t)$ decays as $\sim LG(0)/t^2$ at large times.

When f_3 is the largest fitness (and $f_1 > f_2$), the last jump can occur at times given by (3.32), (3.34) and (3.37). As $T_{3,1} = T_{3,2}$, the corresponding conditions (discussed in Sec. 3.3.3) on the block fitnesses can be combined to give the following cumulative probability

$$\mathcal{P}_1(t) = \int_{l+2\epsilon}^{u} df_3 p(f_3) \int_{l}^{f_3-2\epsilon} df_0 p(f_0) \int_{l}^{\frac{f_0+f_3}{2}} df_1 p(f_1) \int_{l}^{f_1} df_2 p(f_2) \quad (3.41)$$

and the probability distribution

$$P_1(t) = \frac{L}{t^2} \int_{l+2\epsilon}^{u} df_3 p(f_3) p(f_3 - 2\epsilon) \int_{l}^{f_3-\epsilon} df_1 p(f_1) \int_{l}^{f_1} df_2 p(f_2) \quad (3.42)$$

which also decays as $1/t^2$ at large times. An expression for the distribution for the last jump time $T_{3,3}$ can also be written down in an analogous manner and reads as

$$P_2(t) = \frac{L}{2t^2} \int_{l+\epsilon}^{u} df_1 p(f_1) p(f_1 + \epsilon) \int_{l}^{f_1-\epsilon} df_0 p(f_0) \int_{l}^{f_1} df_2 p(f_2) \stackrel{\epsilon \to 0}{\to} \frac{L}{2t^2} G(0) \quad (3.43)$$

Clearly the distribution $P_B(t) = 2(P_1(t) + P_2(t)) \sim t^{-2}$. Thus the probability distribution $P(t) = P_A(t) + P_B(t)$ obeys the inverse square law for any block fitness distribution.

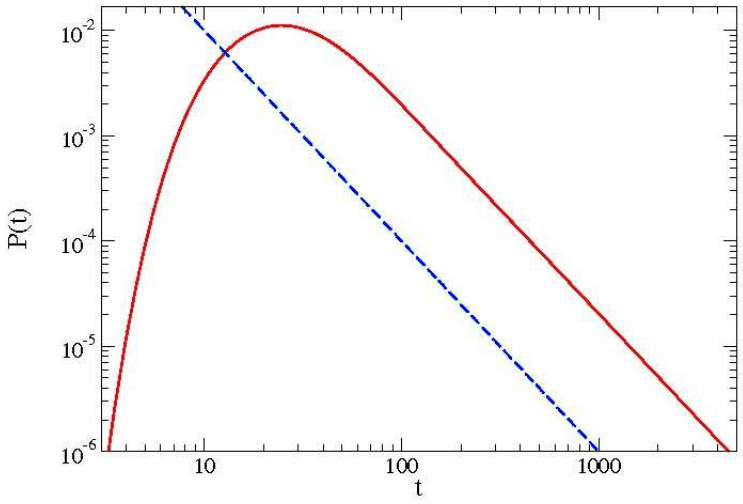


Figure 3.4: Log-log plot of the distribution $P(t)$ of the last jump for $p(f) = e^{-f}$ and $L = 100$. The broken line has a slope -2.

3.5 Discussion

In this Chapter we have studied a deterministic model [4] describing the evolution of a population of self-replicating sequences on a class of strongly correlated fitness landscapes with several fitness peaks [7]. The broad questions addressed have been studied on completely uncorrelated fitness landscapes in previous works [4–6]. Here we are interested in finding how the various evolutionary properties are affected when the sequence fitnesses are strongly correlated.

We are primarily interested in the evolutionary dynamics and in particular, the properties of jumps that occur in the population fitness when the most populated sequence changes. As discussed in Chapter 2, the largest

fitness at a constant mutational distance from the initial sequence only need to be considered for this purpose. This led us to consider the problem of the extreme statistics of correlated random variables [8, 9] which has been much less studied than its uncorrelated counterpart. We found that the extreme value distribution is not of the Gumbel form which is obtained when the random variables are i.i.d. and their distribution decays faster than a power law. In fact, we expect that the universal scaling distributions which depend only on the nature of the tail of the underlying distribution do not exist for such correlated random variables as the number of independent variables namely the block fitnesses is too small.

As the minimum requirement of a sequence to qualify as a leader is that it must be a record, we also studied several record properties of correlated variables. Recently the statistics of record events when the number of observations added at each time step increases either deterministically [10] or stochastically [11] have been studied. The records defined in the shell model are an example of the former category as the number of observations changes as $\binom{L}{D}$ with D. It was shown that when the block fitnesses are position independent the probability for a record to occur in a shell with D mutations is not a continuous function unlike the record distributions for independent random variables [5]; however the universality property that the distribution is independent of block fitness distribution continues to hold. The average number of records was found to increase linearly with L as in the maximally uncorrelated case but with the prefactor given by $(1 - \ln 2) \approx 0.306$ for the latter case which is smaller than in (3.30).

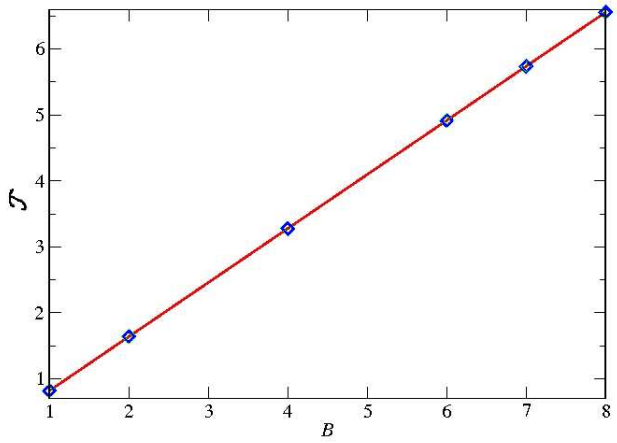


Figure 3.5: Average number of jumps as a function of B for the block model with position dependent block fitness chosen from exponential distribution and fixed $L_B = 2$. The line has a slope given by $3/4 + p_2 = 0.819$.

In the uncorrelated and the weakly correlated fitness models, the L dependence of the average number of jumps was seen to depend on the class of the fitness distribution $p(f)$. For $p(f)$ decaying faster than a power law, the average number of jumps increased as $\sqrt{L}$ [4,5]. In contrast, for the strongly correlated case the average number of jumps was shown to be independent of L for any choice of block fitness distribution $p(f)$ although the value of the constant was found to be nonuniversal. These results suggest that for block fitness distributions decaying faster than a power law, the average number of records increases but the average number of jumps decreases with increasing correlations. But the result of our numerical simulations shows that the average number of jumps increases linearly with the number of blocks when

the fitness of the blocks depend on their position in the sequence. However the prefactor is given by the average number of jumps obtained in the position independent block fitness model namely $(3/4) + p_2$ (see Fig. 3.5). This suggests that the different blocks behave independently in the position dependent block fitness model.

The temporal distribution for the last jump to occur at time t obeys t^{-2} law for infinite (and finite) populations evolving on uncorrelated fitness landscapes [4–6]. Here we have shown that on a class of strongly correlated fitness landscapes, the same law is obeyed. The origin of this power law can be understood using a simple scaling argument when the fitness variables are independent variables [4] but it is not obvious at the outset that such an argument can be used here since the sequence fitnesses are correlated. But it turns out that the jump time involves the i.i.d. block fitnesses and therefore t^{-2} law is obtained here as well.

So far we have studied infinitely large populations but stochastically evolving finite populations will be the focus of our next Chapter and we study their properties till they evolve to reach a local fitness peak.

Bibliography

[1] C. Carneiro and D.L. Hartl. *Proc. Natl. Acad. Sci. USA*, 107:1747–1751, 2010.

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

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

[4] J. Krug and C. Karl. *Physica A*, 318:137–143, 2003.

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

[6] K. Jain. *Phys. Rev. E*, 76:031922, 2007.

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

[8] H.A. David and H.N. Nagaraja. *Order Statistics*. Wiley, New York, 2003.

[9] K. Jain, A. Dasgupta, and G. Das. *J. Stat. Mech.*, page L10001, 2009.

[10] J. Krug. *J. Stat. Mech.:Theor. Exp.*, page P07001, 2007.

[11] I. Eliazar and J. Klafter. *Phys. Rev. E*, 80(6):061117, 2009.

Chapter 4

Adaptive walk on correlated and uncorrelated fitness landscapes

4.1 Introduction

In this Chapter we present our results [1] on the adaptive walk model in which beneficial mutations arise sequentially and fix rapidly [2]. As explained in Chapter 2, if the mutation rate is small and the selection coefficient is large (compared to the inverse population size), it is a good approximation to assume that only the one-step mutants are accessible at any time and the population is localised at a single genotype. Such a monomorphic population performs an adaptive walk by moving uphill on a fitness landscape until no more beneficial mutations can be found.

The results from the limiting cases of choosing the beneficial mutation

in the next step, namely the greedy [3] and the random [4] adaptive walks have been discussed in Chapter 2 . As already mentioned in Chapter 2, adaptive walk is a more realistic model whose first step is well studied [5, 6]. However as the properties of the entire walk are required to design a drug or a biomolecule [7] and as experimental data on multiple adaptive substitutions is becoming available [8, 9], it is important to extend the existing theory to address the statistical properties of the entire walk. With this aim, we present our results on the entire adaptive walk on rugged fitness landscapes with many local fitness optima [1]. An important difference between our work and the previous ones is that we start the adaptive walk with low fitness to describe the adaptation process in novel environments such as when antibiotics are introduced [10, 11] whereas the initial fitness is assumed to be high in other studies [2, 5, 12, 13].

For generic fitness distributions, we argue that the average number of adaptive steps increases logarithmically with sequence length with a prefactor that depends on the choice of fitness distribution. Although our argument does not capture the proportionality constant correctly, the logarithmic dependence is seen to be in excellent agreement with the simulation results. We also present detailed results on the statistical properties of entire walk for exponentially and uniformly distributed fitnesses as these two distributions lend themselves to an analytic treatment and are also consistent with the experiments [14, 15]. Following the approach of [4], we write a recursion relation for the fitness distribution of *fixed* beneficial mutations at an adaptive step which is valid for long sequences and fitness distributions with a finite mean. For the above mentioned distributions, we also find the distribution

of walk length. The average walk length calculated using this approach gives a prefactor consistent with the numerical results.

Although in most part of this Chapter we consider uncorrelated fitnesses and assume that the distribution of the fitness does not change during the course of evolution, the effect of correlations is also discussed. As discussed in Chapter 2, experiments support an intermediate degree of correlations in fitness landscapes [16,17] and changing fitness distributions may be modeled by correlated fitnesses [12], we calculate the average number of steps to an optimum on a fitness landscape generated by the block model of correlated fitnesses in which a sequence is divided into several independent blocks and correlations arise when two sequences share some blocks [18]. The average walk length has been measured using numerical simulations in a block model in [12] and it was speculated that the average number of adaptive steps is independent of the underlying fitness distribution and increases linearly with the number of blocks. We show that while the latter result is roughly correct, the average number of steps to a local optimum is not independent of the fitness distribution which is a consequence of the result discussed above for the uncorrelated fitness landscapes.

4.2 Adaptive walk model for long sequences

As explained in Chapter 2 we work with haploid binary sequences of length L in the strong selection-weak mutation (SSWM) regime [1]. If the fitnesses of the wild type sequence and its L one-mutant neighbors are arranged in a descending order with the best fitness assigned the rank 1, the transition

probability that the population moves from the wild type with fitness rank i and value $f(i)$ to a mutant with rank $j < i$ and value $f(j)$ is proportional to the fixation probability which is well approximated by $2(f(j) - f(i))/f(i)$ in the strong selection limit [2]. The normalised transition probability from fitness $f(i)$ to fitness $f(j)$ is given by

$$T(f(j) \leftarrow f(i)) = \frac{f(j) - f(i)}{\sum_{k=1}^{i-1} f(k) - f(i)} \ , \ 1 \leq j \leq i-1 \tag{4.1}$$

Once the population has moved to a mutant sequence with fitness $f(j)$ with probability $T(f(j) \leftarrow f(i))$, it produces a set of new mutants which are rank ordered and chosen according to (4.1) and the process repeats itself until the population reaches a local optimum whose nearest neighbors are all less fit than itself. Note that the parameters N and μ have dropped out of the picture and the properties of the model depend on the sequence length (or the initial rank) and the distribution of sequence fitnesses.

The model described above has been studied using (4.1) and EVT (Extreme Value Theory)in previous works [2,5,12,13] assuming the initial fitness to be high (small i). In contrast, we start with a low fitness and write a recursion relation for the probability $P_J(f)$ that an adaptive walk has at least J steps and the fitness is f at the Jth step, following [4] who studied this distribution for random adaptive walks as explained in Chapter 2. In the following discussion, it is assumed that the sequence length is large which allows the following two simplifications: first, the events in which a sequence is backtracked can be ignored and second, the transition rates can be written in terms of absolute fitnesses instead of fitness ranks. Consider a population

at the Jth adaptive step and with fitness h. It can proceed to the next step provided at least one fitter mutant is available. If $q(h) = \int_l^h dg\ p(g)$, this event occurs with a probability $1 - q^L(h)$ where it is assumed that at each step in the evolutionary process, L novel mutants are available which have not been encountered before. While this is true at the first step, the number of novel mutants is $L - 1$ at the second step since one of the mutants is the parent sequence itself which is not an allowed descendant as the walk always proceeds uphill. In fact for any $J \geq 2$, some of the mutants have already been probed but the error introduced by ignoring this complication is of the order of $1/L$ which is negligible for large L [4]. Then for long sequences we can write

$$P_{J+1}(f) = \int_l^f dh\ p(f)T(f \leftarrow h)\ (1 - q^L(h))P_J(h)\ ,\ J \geq 0 \tag{4.2}$$

where the underlying fitness of the sequences is chosen from one of following three distributions

$$p(f) = \begin{cases} (\delta - 1)(1 + f)^{-\delta} & , \delta > 2 \quad \text{(Fréchet)} & (4.3) \\ \gamma f^{\gamma - 1} e^{-f^\gamma} & , \gamma > 0 \quad \text{(Gumbel)} & (4.4) \\ \nu(1 - f)^{\nu - 1} & , \nu > 0, f < 1 \quad \text{(Weibull)} & (4.5) \end{cases}$$

In (4.2), $p(f)T(f \leftarrow h)$ gives the probability that a mutant with fitness $f > h$ is chosen. Furthermore for large L, it is a good approximation to replace the sum in the denominator of (4.1) by an integral and we may write

$$T(f \leftarrow h) = \frac{f - h}{\int_h^u dg\ (g - h)\ p(g)}\ ,\ f > h \tag{4.6}$$

Thus we work with absolute fitnesses instead of fitness ranks. Since the transition probability (4.6) is undefined for slowly decaying fitness distributions $p(f) \sim f^{-\delta}$, $\delta \leq 2$, we restrict $\delta > 2$ in (4.3). Using (4.6) in (4.2), we finally obtain

$$P_{J+1}(f) = \int_l^f dh \; \frac{(f-h)p(f)}{\int_h^u dg \; (g-h) \; p(g)} \; (1 - q^L(h)) P_J(h) \; , \; J \geq 0 \qquad (4.7)$$

The above equation is the central equation of this part of the Chapter and we will employ it to obtain various results on the statistical properties of adaptive walks. In the following, we assume the initial condition $P_0(f) = \delta(f)$ corresponding to zero initial fitness. As $P_J(f)$ obeys an integral equation which are harder to analyse, we may try to write a differential equation for $P_J(f)$. Differentiating (4.7) with respect to f, we get:

$$\begin{aligned} P'_{J+1}(f) &= \int_l^f dh \; \frac{(f-h)p'(f) + p(f)}{\int_h^u dg \; (g-h) \; p(g)} \; (1 - q^L(h)) P_J(h) \; , \; J \geq 0 \qquad (4.8) \\ P''_{J+1}(f) &= \int_l^f dh \; \frac{(f-h)p''(f) + 2p'(f)}{\int_h^u dg \; (g-h) \; p(g)} (1 - q^L(h)) P_J(h) \\ &+ \frac{p(f)(1-q^L(f))}{\int_f^u dg \; (g-f) \; p(g)} P_J(f) \; , \; J \geq 1 \qquad (4.9) \end{aligned}$$

where prime denotes a f-derivative. On using (4.7) and (4.8) in (4.9), we find

$$\begin{aligned} P''_{J+1}(f) &= 2\frac{p'(f)}{p(f)} P'_{J+1}(f) + \left[\frac{p''(f)}{p(f)} - 2 \left(\frac{p'(f)}{p(f)} \right)^2 \right] P_{J+1}(f) \\ &+ \frac{p(f)(1-q^L(f))}{\int_f^u dg \; (g-f) \; p(g)} P_J(f) \; , \; J \geq 1 \qquad (4.10) \end{aligned}$$

The first derivative term in the above equation can be eliminated by writing $P_J(f) = p(f)\tilde{P}_J(f)$ which finally yields

$$\tilde{P}''_{J+1}(f) = \frac{p(f)(1-q^L(f))}{\int_f^u dg\ (g-f)\ p(g)}\tilde{P}_J(f)\ ,\ J \geq 1 \tag{4.11}$$

Here we will restrict our attention to exponentially and uniformly distributed fitnesses as these two fitness distributions are consistent with the available empirical data. We show that due to (4.11), a second order ordinary differential equation is obeyed by a generating function of $P_J(f)$ for these two distributions which can be solved within an approximation subject to the following boundary conditions:

$$P_J(f)|_{f=l} = 0\ ,\ J \geq 1 \tag{4.12}$$

$$P'_J(f)|_{f=l} = \frac{p(l)}{\int_l^u dg\ g\ p(g)}\ \delta_{J,1} \tag{4.13}$$

where (4.12) is a direct consequence of (4.7) and the equation (4.13) arises on using the initial condition in (4.8).

Besides $P_J(f)$, we also find the walk length distribution Q_J and the average fitness $\bar{f}_J$ at the Jth step which can be related to $P_J(f)$ as explained below. Integrating over f on both sides of (4.7), we get

$$P_{J+1} = \int_l^u df\ P_{J+1}(f) \tag{4.14}$$

$$= \int_l^u dh \int_h^u df\ \frac{(f-h)p(f)}{\int_h^u dg(g-h)p(g)}\ (1-q^L(h))P_J(h) \tag{4.15}$$

$$= \int_l^u dh\ (1-q^L(h))P_J(h) = P_J - \int_l^u dh\ q^L(h)P_J(h) \tag{4.16}$$

Then the walk length probability Q_J that exactly J steps are taken is given by

$$Q_J = P_J - P_{J+1} = \int_l^u dh\ q^L(h) P_J(h) \tag{4.17}$$

with $Q_0 = 0$ since the initial fitness is zero. The above equation has a simple interpretation: Since $P_J(h)$ is the probability that at least J steps are taken and the fitness at the Jth step is h, exactly J steps will be taken if all the L mutants of the sequence at the Jth step carry a fitness smaller than h from which (4.17) follows. The average walk length $\bar{J} = \sum_{J=0}^{2^L} JQ_J \approx \sum_{J=0}^{\infty} JQ_J$ for large L. The average fitness $\bar{f}_J$ is defined as $\bar{f}_J = \int_l^u df\ fP_J(f)$. Using (4.7), we can write

$$\bar{f}_{J+1} = \int_l^u df f \int_l^f dh\ \frac{(f-h)p(f)}{\int_h^u dg(g-h)p(g)}\ (1-q^L(h))P_J(h) \tag{4.18}$$

$$= \int_l^u dh \frac{(1-q^L(h))P_J(h)}{\int_h^u dg(g-h)p(g)} \int_h^u df f(f-h)p(f) \tag{4.19}$$

Note that neither (4.17) nor (4.19) are closed equations.

Our analytical results are also compared with numerical simulations which were performed using an exact procedure for $L \leq 10$ and an approximate method outlined in [5] for larger L as explained. While simulating short sequences of length $L \leq 10$ and uncorrelated fitnesses, a randomly chosen sequence was assigned a fitness equal to zero. Then the rest of the fitness landscape comprising of $2^L - 1$ fitnesses was generated by drawing random variables independently from a common distribution $p(f)$. The transition probability from the initial sequence to each of the better sequences among

the L nearest neighbors was calculated according to (4.1) and the fixed sequence at the first step in the adaptive walk was chosen. Then the transition probability from the chosen mutant sequence to its better neighbors was calculated and this process was repeated until a fitter sequence was not available. To simulate sequences with length $L \gtrsim 10^2$, we followed an approximate procedure outlined in [5] as the total number of sequences 2^L is prohibitively large for long sequences. Starting with zero fitness, L i.i.d. random variables were generated and a higher fitness f was chosen according to the transition probability (4.1). During the next step in the process, L new i.i.d. random variables were generated and the transition probability from f to a better fitness was calculated. These steps were repeated until the new set of random fitnesses does not exceed the currently fixed fitness. The block model was simulated to generate weakly correlated fitnesses by assigning independent fitnesses to each block sequence. In all the simulations, the data was collected using 10^6 independent realisations of the fitness landscape and the results obtained were used to verify the theoretical claims as explained below.

4.3 Average fitness and walk length for general fitness distributions

For a broad class of fitness distributions, the average fitness for an infinitely long sequence can be computed. Although this limit is biologically unrealistic, it provides a good approximation to the average fitness $\bar{f}_J$ for small

J (see Fig. 4.1) as the population can not sense the finiteness of sequence length far from the local optimum. On taking the limit $L \to \infty$ in (4.19) and denoting the average fitness in this limit by F_J, we obtain

$$F_{J+1} = \int_l^u dh \; \frac{\int_h^u df \; f(f-h)p(f)}{\int_h^u dg \; (g-h)p(g)} \; P_J(h)|_{L\to\infty} \tag{4.20}$$

Algebraically decaying fitness distributions: On substituting (4.3) in (4.20) and performing the integrals involving $p(f)$, we get

$$F_{J+1} = \int_0^\infty dh \; \frac{2+(\delta-1)h}{\delta-3} \; P_J(h)|_{L\to\infty} = \frac{2}{\delta-3} + \frac{\delta-1}{\delta-3} F_J \; , \; \delta > 3 \tag{4.21}$$

where we have used that $P_J|_{L\to\infty} = 1$ due to (4.16) and the initial condition $P_0 = 1$. Repeated iteration with $F_0 = 0$ yields

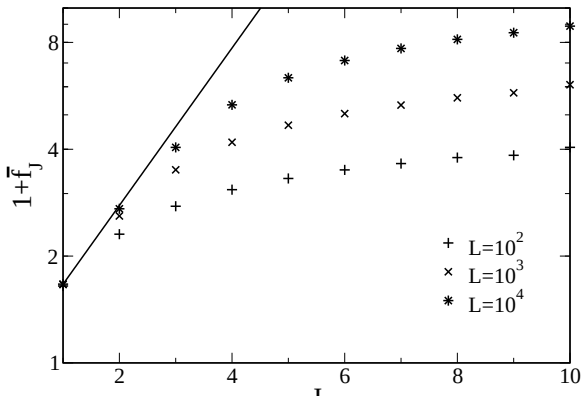
$$F_J = \left(\frac{\delta-1}{\delta-3}\right)^J - 1 \tag{4.22}$$

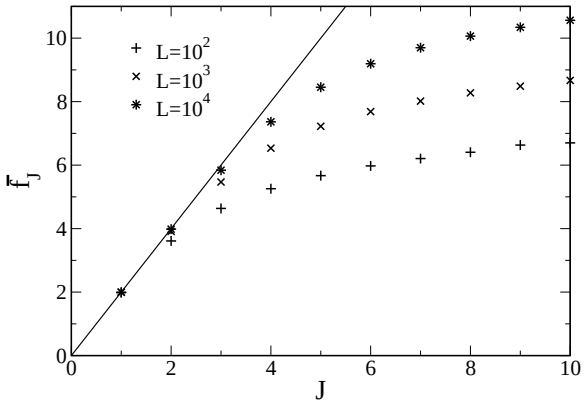
which increases geometrically with J. This result is compared in Fig. 4.1a with the average fitness for finite sequences which shows that the number of steps up to which $\bar{f}_J$ and F_J match increases with L.

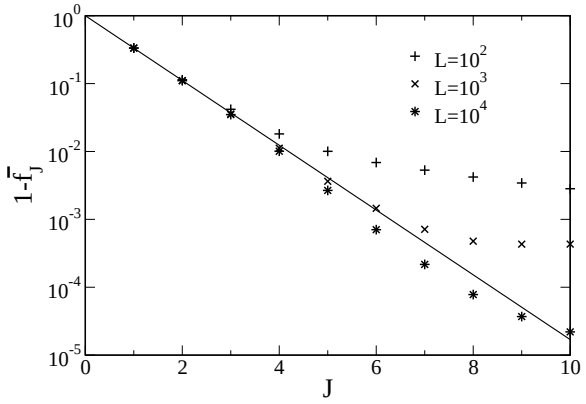
Exponential fitness distribution: For fitness distributions given by (4.4), the equation for F_J does not close except for $\gamma = 1$. For $p(f) = e^{-f}$, we get $F_J = 2 + F_{J-1}$ which gives

$$F_J = 2J \tag{4.23}$$

Fig. 4.1b shows that the rate of increase of fitness $\bar{f}_J$ is slower than a constant at larger J's.







(c)

Figure 4.1: Evolution of average fitness with the number of adaptive steps starting from zero initial fitness obtained numerically (points) and compared with the average fitness in infinite sequence length limit (lines) for (a) power law distributed fitness with $\delta = 6$, equation (4.22) (b) exponentially, equation (4.23) and (c) uniformly distributed fitness, equation (4.25).

Bounded fitness distributions: A calculation similar to above for $p(f)$ in (4.5) gives

$$F_{J+1} = \frac{2 + \nu F_J}{2 + \nu} \tag{4.24}$$

and therefore

$$F_J = 1 - \left(\frac{\nu}{2+\nu}\right)^J \tag{4.25}$$

For uniformly distributed fitness ($\nu = 1$), we find that $1 - F_J = 3^{-J}$ in good agreement with the numerical data in Fig. 4.1(c) for small J.

We now give an argument to estimate the average walk length $\bar{J}$ using the above results for the average fitness F_J and the EVT [4]. We first note that since $P_J|_{L\to\infty} = 1$ for all J, every step in the adaptive walk is definitely taken for infinitely long sequences and hence the average walk length is expected to diverge with L. For a sequence of finite length, the adaptive walk stops when the population has reached a local optimum whose fitness is the largest among $L+1$ i.i.d. random variables. But since the average number of fitnesses with value $\geq f$ is given by $(L+1)(1-q(f))$, at a local optimum we have [19]

$$(L+1)\int_{F_{\bar{J}}}^{u} df\ p(f) = 1 \tag{4.26}$$

where we have approximated $\bar{f}_{\bar{J}}$ by $F_{\bar{J}}$. The above equation yields

$$F_{\bar{J}} \approx \begin{cases} L^{\frac{1}{\delta-1}} - 1 & \text{(Algebraic)} \quad (4.27) \\ \ln L & \text{(Exponential)} \quad (4.28) \\ 1 - L^{-\frac{1}{\nu}} & \text{(Bounded)} \quad (4.29) \end{cases}$$

On matching the expected fitness $F_{\bar{J}}$ with the F_J obtained in the above

discussion for various distributions, we get

$$\bar{J} \approx \begin{cases} \dfrac{1}{\delta-1}\dfrac{\ln L}{\ln(\frac{\delta-1}{\delta-3})} & \text{(Algebraic)} \quad (4.30) \\ \dfrac{1}{2}\ln L & \text{(Exponential)} \quad (4.31) \\ \dfrac{1}{\nu}\dfrac{\ln L}{\ln(\frac{2+\nu}{\nu})} & \text{(Bounded)} \quad (4.32) \end{cases}$$

Thus the above argument shows that for large L,

$$\bar{J} \approx \alpha \ln L \tag{4.33}$$

where the prefactor α depends on $p(f)$. We note that $\alpha_{\text{algebraic}} < \alpha_{\text{exponential}} < \alpha_{\text{bounded}}$ which implies that smaller number of substitutions occur for fat-tailed fitness distributions than the bounded ones. To understand this qualitative trend, consider the transition probability for the first step given by $T(f \leftarrow 0)p(f) \sim fp(f)$. At large f, this probability is higher for slowly decaying distributions and thus a large fitness gain occurs initially. But as the probability to exceed the high fitness achieved at the first step is small, the walk terminates sooner for broad distributions.

The results of our numerical simulations for $\bar{J}$ shown in Fig. 4.2 are in agreement with the logarithmic dependence on L but the value of the prefactor does not match with that obtained above (except for $p(f) = e^{-f}$). The prefactor α is expected to interpolate between the two limiting cases of adaptive walks namely greedy walk in which the best mutant is chosen with probability one and random adaptive walk in which all better mutants are chosen with equal probability. As explained in Chapter 2, the former limit

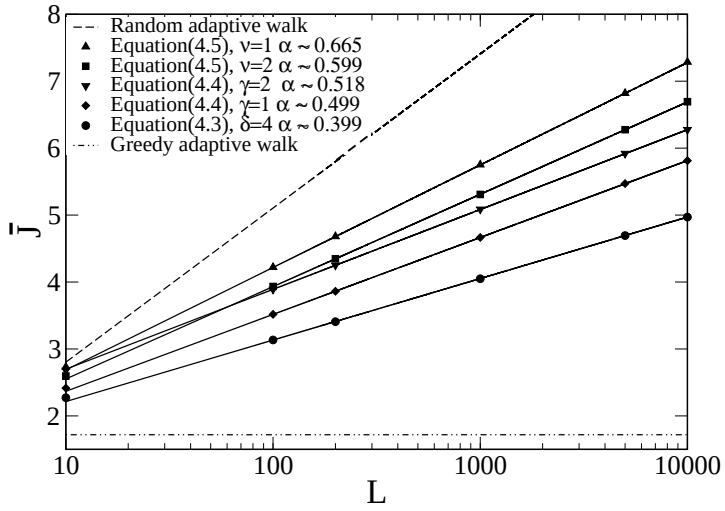


Figure 4.2: Average number $\bar{J}$ of adaptive steps as a function of sequence length L for various fitness distributions when the fitnesses are uncorrelated. The points show the data obtained using numerical simulations and the lines are the best fit to the function $\bar{J} = \alpha \ln L + \beta$. The results for greedy walk and random adaptive walk (up to an additive constant) are also shown. The numerical fit for the prefactor α for exponential and uniform fitness distribution matches well with the analytical results given by (4.46) and (4.56) respectively.

is obtained when $\delta \to 1$ in (4.3) and the latter when $\nu \to 0$ in (4.5) [13]. Since the average walk length for a greedy walker is a finite constant equal to $e-1 \approx 1.718$ for infinitely long sequences [3], the prefactor $\alpha = 0$ while $\alpha = 1$ for random adaptive walk. In the following sections, we find that $\alpha = 1/2$ for exponentially distributed fitness and $2/3$ for the uniform case which are consistent with the results in Fig. 4.2 and the analytical results of [20] which are obtained using a simpler version of the adaptive walk model considered here.

4.4 Distribution of fitness and walk length

We now present our calculation for the distribution of the walk length and the fitness at the J^{th} step of the walk. Our results obtained using an approximation work well as long as the walker is not close to the local fitness optimum. If we consider the whole population to have an initial fitness f_0, using $P_0(f) = \delta(f - f_0)$ in (4.7) we have

$$P_1(f) = \frac{(f - f_0)p(f)(1 - q^L(f_0))}{\int_l^u dg\ gp(g)} \propto (f - f_0)p(f) \tag{4.34}$$

The above fitness distribution at the first step is nonmonotonic for all fitness distributions in (4.3)-(4.5) except for truncated distributions with $\nu \leq 1$. The implications of this result are examined in Sec.4.6. Though the solution for the first step is trivial, it is complicated for $J > 1$ and we have used certain approximations to solve it for uniform and exponential distributions as shall be explained in the next few sections.

4.4.1 Entire walk with exponentially distributed fitness

For $p(f) = e^{-f}$, from (4.11) we obtain

$$\tilde{P}''_{J+1}(f) = (1 - q^L(f))\tilde{P}_J(f)\ ,\ J \geq 1 \tag{4.35}$$

where $q(f) = 1 - e^{-f}$. Due to (4.12) and (4.13), the boundary conditions are $P_J(0) = 0$ and $P'_J(0) = \delta_{J,1}$.

We define a generating function $G(x,f) = \sum_{J=1}^{\infty} \tilde{P}_J(f) x^J \ , \ x < 1$ which obeys the following second order ordinary differential equation:

$$G^{''}(x,f) = x(1-q^L(f))G(x,f) \tag{4.36}$$

To arrive at the above equation, we have used that $\tilde{P}_1(f) = f$ which is obtained on using the initial condition in (4.7). The generating function $G(x,f)$ obeys a Schrödinger equation for the wave function of a particle in a one-dimensional potential $V(f) \sim 1 - q^L(f)$ and energy zero [21]. Since $1 - q^L(f) \approx 1 - e^{-Le^{-f}}$ is close to unity for $f \ll \ln L$ and vanishes for $f \gg \ln L$, the potential $V(f)$ decreases smoothly from one to zero and moves rightwards with increasing L. Similar potentials also arise when two materials with different transport properties are joined together and in such systems, an analytical solution is obtained within a step function potential approximation [22, 23]. We follow this approach here and approximate the distribution $1 - q^L(f)$ by the Heaviside theta function $\Theta(\tilde{f} - f)$ where $\tilde{f} = \ln L$. Within this step distribution approximation, we have

$$G^{''}(x,f) = \begin{cases} xG(x,f) & , \ f < \tilde{f} \\ 0 & , \ f > \tilde{f} \end{cases} \tag{4.37}$$

For $f < \tilde{f}$, the differential equation (4.37) has a solution of the form $G_<(x,f) = a_+ e^{\sqrt{x}f} + a_- e^{-\sqrt{x}f}$ which reduces to $G_<(x,f) = c \sinh(\sqrt{x}f)$ since $G(x,0) = 0$ due to $P_J(0) = 0$. Since the solution for $f < \tilde{f}$ can not depend on $\tilde{f}$, we appeal to the infinite sequence length limit to fix the proportionality

constant c. As noted earlier, the distribution $P_J|_{L\to\infty} = 1$ for all $J \geq 0$ which implies that

$$\int_0^\infty df\ e^{-f} G_<(x,f) = \frac{x}{1-x} \tag{4.38}$$

and therefore

$$G_<(x,f) = \sqrt{x}\sinh(\sqrt{x}f) \tag{4.39}$$

We check that the boundary condition $P'_J(0) = \tilde{P}'_J(0) = \delta_{J,1}$ which is equivalent to $G'(x,0) = x$ is also satisfied by the above solution.

For $f > \tilde{f}$, the solution $G_>(x,f) = af + b$ where the constants of integration a, b can be fixed by matching the solutions $G_<$ and $G_>$ and their first derivative at $f = \tilde{f}$. Thus the constant a and b are determined by the following conditions:

$$G_<(x,\tilde{f}) = G_>(x,\tilde{f}) = a\tilde{f} + b \tag{4.40}$$

$$G'_<(x,f)|_{f=\tilde{f}} = G'_>(x,f)|_{f=\tilde{f}} = a \tag{4.41}$$

A simple algebra shows that

$$G_>(x,f) = x\cosh(\sqrt{x}\tilde{f})(f-\tilde{f}) + \sqrt{x}\sinh(\sqrt{x}\tilde{f}) \tag{4.42}$$

Using the above expressions for $G(x,f)$, the fitness distribution $P_J(f)$ for the fixed beneficial mutations can be calculated. On expanding (4.39) and (4.42) in a power series about $x = 0$ and picking the coefficient of x^J, we

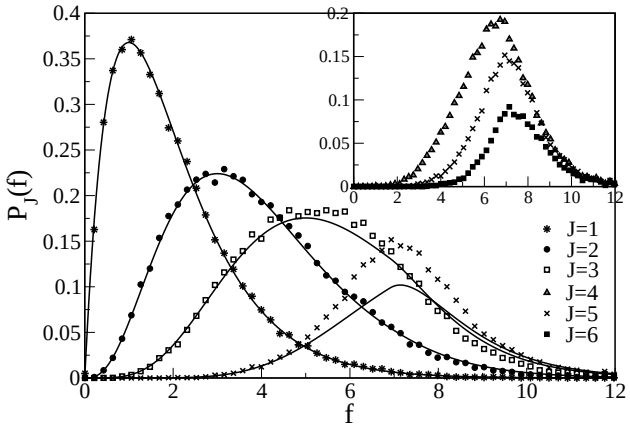


Figure 4.3: Main: Comparison of the distribution $P_J(f)$ for $J = 1, 2, 3, 5$ obtained numerically (points) and analytically (lines) given by (4.43) for exponentially distributed fitness and sequence length $L = 1000$. Inset: Numerical data for $P_J(f)$ for $J = 4, 5, 6$ to show that the fitness distribution does not shift appreciably beyond $\bar{J} \approx 4.6$ as local optimum with average fitness ≈ 7 is approached.

have

$$P_J(f) = \frac{e^{-f} f^{2J-1}}{(2J-1)!} \times \begin{cases} 1 & , \ r \leq 1 \\ \frac{(2J-1)r-(2J-2)}{r^{2J-1}} & , \ r > 1 \end{cases} \tag{4.43}$$

where $r = f/\tilde{f}$. Figure 4.3 shows our numerical results for $P_J(f)$ for the first few adaptive steps. As the walk proceeds, the distribution moves rightwards as expected and its amplitude decreases since the probability $q^L(f)$ that the walker can not find a better neighbor approaches unity with increasing f.

Our analytical result (4.43) is also shown in Fig. 4.3 for comparison. For $L = 10^3$, the step distribution approximation used to find (4.43) gives $1 - q^L(f) \approx 1$ for $f < \ln L = 6.9$ and zero otherwise. However as the probability $1 - q^L(f)$ stays close to unity for $f \leq 5$ and decreases gradually to zero when $f \approx 12$, the distribution (4.43) in the region $5 < f < 12$

does not match well with the simulation results but outside this crossover region, we see a good quantitative agreement. We also note that the fitness distribution does not move appreciably for $J \geq 4$ and is centred around $f \approx 7$ (see inset of Fig. 4.3). This is because the average walk length for $L = 10^3$ is about 4.6 steps (refer Fig. 4.2) and as the local optimum is approached, the fitness distribution of fixed beneficial mutation remains centred close to the typical fitness of the local optimum given by (4.26) which is $\ln L \approx 6.9$. This also explains the initial linear rise in the average fitness followed by a slower increase in Fig. 4.3.

We next calculate the walk length distribution Q_J defined by (4.17). Since $q^L(f) = \Theta(f - \tilde{f})$ within the step distribution approximation discussed above, (4.17) reduces to

$$Q_J = \int_{\tilde{f}}^{\infty} df \; P_J(f) \tag{4.44}$$

On integrating $P_J(f)$ given in (4.43), we get

$$Q_J = e^{-\ln L} \left[\frac{(\ln L)^{2J-2}}{(2J-2)!} + \frac{(\ln L)^{2J-1}}{(2J-1)!} \right] \; , \; J > 0 \tag{4.45}$$

This expression is compared with numerical results in Fig. 4.4 and shows a reasonable agreement. The average number of adaptive steps calculated using (4.45) is given by

$$\bar{J} = \sum_{J=1}^{\infty} J Q_J \approx \frac{1}{2} \ln L \tag{4.46}$$

which is in good agreement with the simulation result in Fig. 4.2. The width of the distribution Q_J measured using the variance $\sigma^2 = \bar{J^2} - \bar{J}^2 \approx \ln L/4$

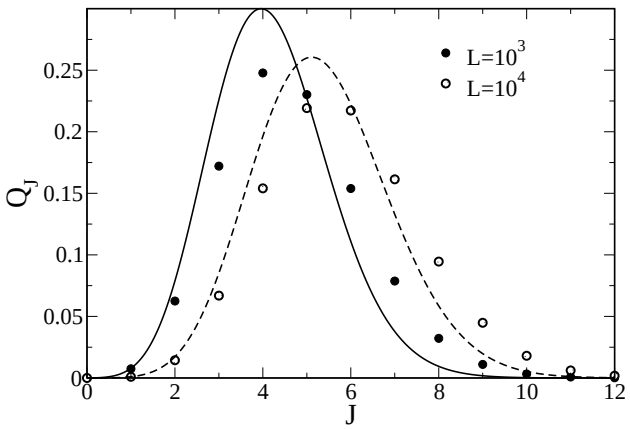


Figure 4.4: Walk length distribution Q_J for $p(f) = e^{-f}$ comparing numerical (points) and analytical result (lines) given by (4.45).

also increases with L.

4.4.2 Entire walk with uniformly distributed fitness

For $p(f) = 1$, since $P_J(f) = \tilde{P}_J(f)$, the differential equation (4.11) reduces to

$$P''_{J+1}(f) = \frac{1 - f^L}{\int_f^1 dg\ (g - f)}\ P_J(f) = \frac{2(1 - f^L)}{(1 - f)^2} P_J(f)\ ,\ J \geq 1 \qquad (4.47)$$

with boundary conditions $P_J(0) = 0$ and $P'_J(0) = 2\delta_{J,1}$. As before, we define a generating function $G(x, f) = \sum_{J=2}^{\infty} x^{J-2} P_J(f)$ which obeys the following second order ordinary differential equation:

$$G''(x, f) = \frac{2(1 - f^L)}{(1 - f)^2}\ (xG(x, f) + 2f) \qquad (4.48)$$

where we have used that $P_1(f) = 2f$. We treat this case also within the step distribution approximation discussed earlier. Since the probability $1 - f^L \approx 1 - e^{-L(1-f)}$, we approximate it by a step function $\Theta(\tilde{f} - f)$ where $\tilde{f} = (L-1)/L$. For $f < \tilde{f}$, we obtain an inhomogeneous second order ordinary differential equation with variable coefficients:

$$G''_<(x,f) = \frac{2x}{(1-f)^2} G_<(x,f) + \frac{4f}{(1-f)^2} \tag{4.49}$$

This equation can be solved by standard methods (as detailed in Appendix A) to yield

$$G_<(x,f) = a_+(1-f)^{\alpha_+} + a_-(1-f)^{\alpha_-} + u_+(f)(1-f)^{\alpha_+} + u_-(f)(1-f)^{\alpha_-} \tag{4.50}$$

where the exponents

$$\alpha_\pm = \frac{1 \pm \sqrt{1+8x}}{2} \tag{4.51}$$

The first two terms on the right hand side give the solution of the homogeneous equation and the last two terms are the particular integral involving the variational parameters $u_\pm(f)$ given in Appendix A. The constants of integration $a_\pm$ can be obtained using the boundary conditions $G(x,0) = 0$ and $\int_0^1 df\ G_<(x,f) = (1-x)^{-1}$. After some straightforward algebra, we find that

$$G_<(x,f) = \frac{-2}{x}\left[\frac{(1-f)^{\alpha_+} - (1-f)^{\alpha_-}}{\alpha_+ - \alpha_-} + f\right] \tag{4.52}$$

We verify that the condition $P'_J(0) = 0$ for $J > 1$ which amounts to $G'(x,0) = 0$ is also satisfied. For $f > \tilde{f}$, as $G''_>(x,f) = 0$, the solution $G_>(x,f) = af + b$

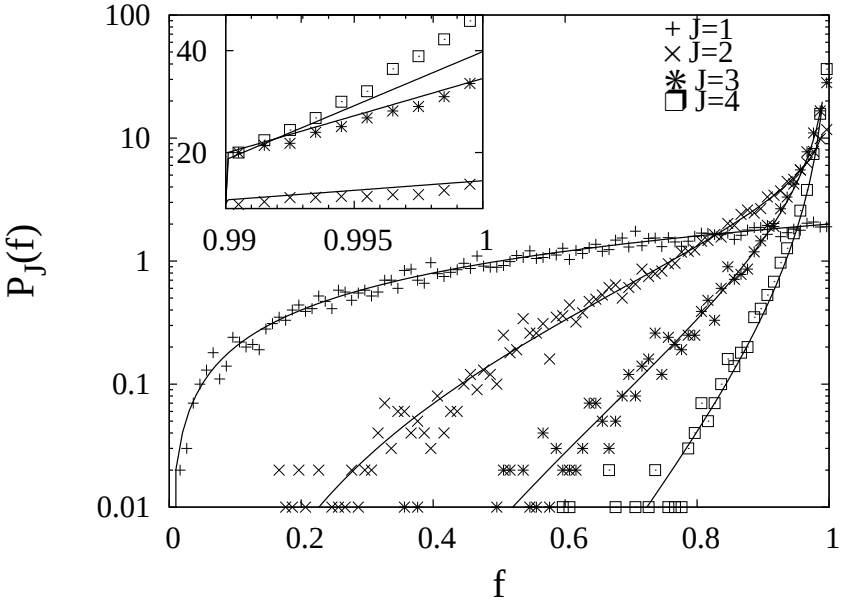


Figure 4.5: Comparison of the distribution $P_J(f)$ for $J = 1, 2, 3, 4$ obtained numerically (points) and analytically (lines) given by (A.6)-(A.9) for uniformly distributed fitness and sequence length $L = 100$. The distribution for $f \leq \tilde{f}$ is shown in the main plot and for $f > \tilde{f}$ in the inset.

where a, b can be determined using (4.40) and (4.41) to give

$$\begin{aligned} G_>(x,f) &= \frac{-2}{x}\left[\frac{\alpha_-(1-\tilde{f})^{\alpha_- -1} - \alpha_+(1-\tilde{f})^{\alpha_+ -1}}{\alpha_+ - \alpha_-} + 1\right] f \\ &- \frac{2}{x}\left[\frac{(1-\tilde{f})^{\alpha_+} - (1-\tilde{f})^{\alpha_-} - \alpha_-\tilde{f}(1-\tilde{f})^{\alpha_- -1} + \alpha_+\tilde{f}(1-\tilde{f})^{\alpha_+ -1}}{\alpha_+ - \alpha_-}\right] \end{aligned} \tag{4.53}$$

Explicit expressions for $P_J(f)$ for first few adaptive steps are given in Appendix A and a comparison between the analytical and the simulation results is shown in Fig. 4.5.

To find the walk length distribution $Q_J = \int_{\tilde{f}}^{1} df\ P_J(f)$, we define

$$H(x) = \sum_{J=1}^{\infty} x^J Q_J = xQ_1 + x^2 \int_{\tilde{f}}^{1} df\ G_>(x,f) \tag{4.54}$$

$$= \frac{x(1-\tilde{f})}{\alpha_- - \alpha_+}\left[(2-\alpha_+)(1-\tilde{f})^{\alpha_+} - (2-\alpha_-)(1-\tilde{f})^{\alpha_-}\right] \tag{4.55}$$

As an explicit expression for Q_J is rather unwieldy, its derivation and the expression itself are given in Appendix A and a comparison with the simulations is shown in Fig. 4.6. The average number of steps is given by

$$\bar{J} = \left.\frac{dH(x)}{dx}\right|_{x=1} = \frac{-6\ln(1-\tilde{f})}{9} \tag{4.56}$$

which shows that for large L, the number of adaptive steps grows as $(2/3)\ln L$ in agreement with the numerical results shown in Fig. 4.2. The higher moments can also be found straightforwardly and we find that the variance $\bar{J^2} - \bar{J}^2 \approx (10/27)\ln L$ and the skewness of the distribution decays slowly as $(\ln L)^{-1/2}$.

4.5 Effect of correlations on the number of adaptive steps

We now turn to a discussion of adaptive walk properties when the fitnesses are correlated and given by a block model introduced in Chapter 1. We compute the average number $\bar{J}_B(L)$ of adaptive steps given by $\sum_{J=1}^{\infty} JQ_J(L,B)$ where $Q_J(L,B)$ is the probability that exactly J adaptive mutations occur when a

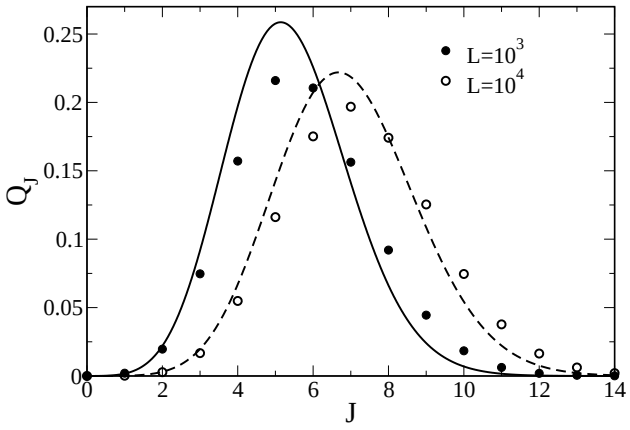


Figure 4.6: Walk length distribution Q_J for uniformly distributed fitnesses comparing simulation (points) and analytical result (lines) in (A.27).

sequence of length L is divided in B blocks.

Consider the distribution $\mathcal{Q}(m_1, ..., m_B)$ which gives the joint probability that the ith block of length L_B in a sequence of length L carries m_i adaptive mutations where $i = 1, ..., B$. An important property of the block model is that this joint distribution factorises, that is [18]

$$\mathcal{Q}(m_1, ..., m_B) = \prod_{b=1}^{B} Q_{m_b}(L_B, 1) \tag{4.57}$$

where $Q_J(L_B, 1) \equiv Q_J(L_B)$ is the walk length probability when the fitnesses are uncorrelated and the sequence length is L_B. The above equation expresses the fact that the block fitnesses evolve independently. As only one mutation occurs in the sequence at any step so that all but one block sequence remains unchanged and since the block fitnesses are i.i.d. random variables, (4.57) holds.

Since the distribution $Q_J(L,B)$ is given by

$$Q_J(L,B) = \sum_{m_1,...,m_B=0}^{J} \mathcal{Q}(m_1,...,m_B)\delta(m_1+...+m_B-J) \tag{4.58}$$

it follows that

$$\begin{aligned}
\bar{J}_B(L) &= \sum_{J=1}^{\infty} J \sum_{m_B=0}^{J} Q_{m_B}(L_B) \sum_{m_1,...,m_{B-1}=0}^{J-m_B} \prod_{b=1}^{B-1} Q_{m_b}(L_B)\delta(\sum_{b=1}^{B-1} m_b - (J-m_B)) \\
&= \sum_{J=1}^{\infty} J \sum_{m_B=0}^{J} Q_{m_B}(L_B) Q_{J-m_B}(L-L_B, B-1) \\
&= \sum_{m=0}^{\infty} Q_m(L-L_B,B-1) \sum_{n=0}^{\infty}(n+m)Q_n(L_B) \\
&= \bar{J}(L_B) + \sum_{m=1}^{\infty} m Q_m(L-L_B,B-1) \\
&= \bar{J}(L_B) + \bar{J}_{B-1}(L-L_B) \\
&= B\bar{J}(L_B)
\end{aligned} \tag{4.59}$$

where we have used that $\sum_{J=0}^{\infty} Q_J(L,B) = 1$ and $\bar{J}$ is the average number of steps in the adaptive walk for uncorrelated fitnesses. Figure 4.7 shows the results of our numerical simulations for average walk length when the block length $L_B = L/B$ is kept fixed and the block fitnesses are exponentially and uniformly distributed. For fixed L_B, (4.59) predicts that $\bar{J}_B$ increases linearly with B which is in excellent agreement with the numerical data.

For large L, due to (4.33) we have

$$\bar{J}_B(L) \approx \alpha B \ln(L/B) + B\beta \tag{4.60}$$

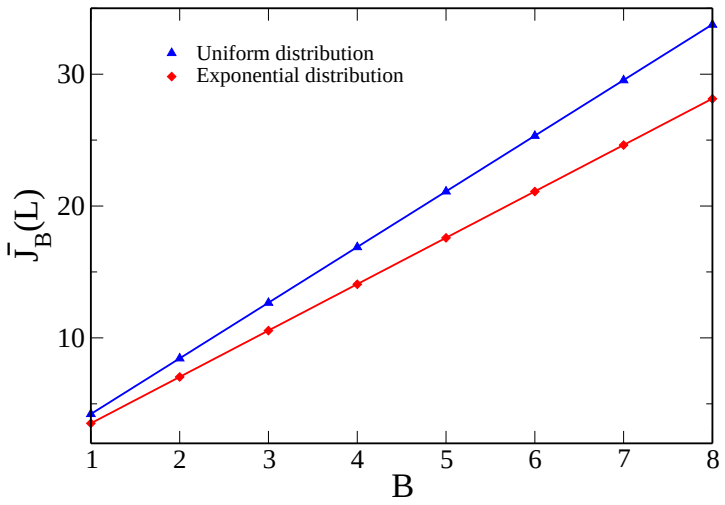


Figure 4.7: Average number $\bar{J}_B$ of adaptive steps as a function of block number B for fixed $L/B = 100$. The numerical data is in excellent agreement with (4.59) shown by solid line.

where β is a constant in Fig. 4.7. For small B, a linear rise in the average number of steps with the number of blocks has been seen numerically for exponential-like distributions and it was inferred that the mean walk length is independent of underlying fitness distributions [12]. However as discussed in the previous sections, the average number $\bar{J}$ depends on the fitness distribution $p(f)$ and therefore the average $\bar{J}_B$ is also nonuniversal.

4.6 Discussion

In the last few years, several analytical results have been obtained for the adaptive walk model [2]. However many of these results deal with the first step in the adaptation process [5, 12, 13] and an extension of the theory to

full adaptive walk is necessary. Previous studies also assume that the process of adaptation starts from a highly fit sequence which is not applicable to situations in which the population is subjected to high stress and hence has a very low initial fitness [10,11]. In this Chapter, we have obtained results for the entire adaptive walk starting from a low initial fitness but as discussed below, we expect some of these results to hold for moderately high initial fitness also.

4.6.1 Walk length distribution and average walk length

In previous works, the walk length distribution for the greedy walk and the random adaptive walk have been studied and found to be universal in that they are independent of the underlying fitness distribution $p(f)$. The origin of this universality property is clear in the light of the results of [13] who pointed out that these two models can be obtained as a limit of (4.1) which defines the adaptive walk model. For the random adaptive walk, the distribution Q_J for infinitely long sequence vanishes and the average walk length diverges with sequence length. In contrast, for greedy walk, the walk length distribution in the $L \to \infty$ limit decreases exponentially fast with J for the greedy walk as a result of which the average number of steps turns out to be a constant [3,24].

In this Chapter, we have calculated the walk length distribution for exponentially and uniformly distributed fitnesses and found the average walk length for general fitness distributions. An important conclusion of our study is that the average number of adaptive steps increases logarithmically with the sequence length with a prefactor smaller than unity if the walk starts

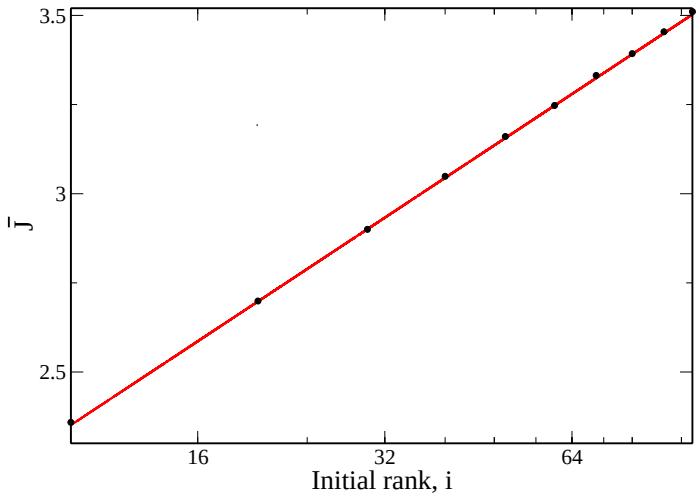


Figure 4.8: Average walk length $\bar{J}$ as a function of the rank i of the initial sequence for $L = 100$ and exponential distribution. The numerical results for $L = 100$ (black circles) is plotted along with the theoretical prediction given by $\bar{J} = \frac{1}{2} \ln i$ (red line).

from zero fitness. Our simulations also indicate that if the initial rank is of order L, the average number of steps increases logarithmically with the rank and with the same proportionality constant as that for the zero initial fitness case as shown in Fig. 4.8 for $L = 100$. Thus for a wild type sequence with initial rank (or L) of the order 100, the number of substitutions are expected to be less than 5. Although short adaptive walks have been observed in experiments [8, 9], more detailed experimental studies testing the logarithmic dependence would be desirable. Although a test of the L-dependence of the average walk length may not be experimentally viable, it should be possible to study the average walk length as a function of the initial rank.

Besides the sequence length, the number of steps to a local optimum depend on the underlying fitness distribution and the fitness correlations also. If the fitnesses are uncorrelated, as the numerical data in Fig. 4.2 shows, the prefactor α in (4.33) depends on the shape of the fitness distribution and therefore a rather detailed knowledge of the full fitness distribution (how fast it decays) is required to test this which is presently unavailable. However one can discern a trend in the value of α: it decreases as the fitness distribution broadens. This suggests that systems with fitness distribution in the Gumbel class [6,10,25–27] will register shorter walks than those in the Weibull domain [15]. As shown here in the block model of correlated fitnesses, the average number of adaptive steps increases as the number of blocks (and hence fitness correlations) increase. This is in accordance with the expectation that on a smooth correlated fitness landscape, as the local optima are less common [18], there is a less chance to get trapped and therefore uphill walk can last longer [12, 28, 29].

4.6.2 Distribution of fixed beneficial mutations during the walk

The fitness distribution $P_J(f)$ has not been studied in previous theoretical studies of adaptive walks in the SSWM limit and here we have computed this fitness distribution analytically using the recursion relation (4.7). The fitness distribution at the first step given by (4.34) can give a qualitative idea about the shape of $p(f)$. For most fitness distributions, $P_1(f)$ is expected to be nonmonotonic but for bounded distributions which diverge at the upper

limit or the uniform distribution, $P_1(f)$ increases monotonically towards the upper bound. An inspection of the experimental data of [6] shows the fitness distribution at the first step to be nonmonotonic which is consistent with their assumption of exponentially decreasing distribution of beneficial effects. It would be interesting to check if the distribution $P_1(f)$ in [15] is monotonic as the data in this study is consistent with a uniformly distributed fitness. The above behavior of $P_1(f)$ is expected to be robust in the presence of correlations as at the first step in evolution, the population has not sensed the correlations in the fitness landscape [12].

For the fitness distribution for the entire walk, we presented an analysis for two distributions namely exponential and uniform which are consistent with the available experimental data. The distribution $P_J(f)$ is obtained within a step distribution approximation which captures the shape of the fitness distribution correctly for the first few steps and leads to an accurate estimate of the number of average steps. Our approximation consists of replacing the probability $1-q^L(f)$ by a step function $\Theta(\tilde{f}-f)$ where $\tilde{f}$ is given by (4.28) for exponentially and (4.29) for uniformly distributed fitnesses. For $f \ll \tilde{f}$ and $f \gg \tilde{f}$, our approximate solution matches the simulation results well for any J. With increasing J, the distribution $P_J(f)$ shifts towards higher fitnesses and peaks about $\tilde{f}$ for larger J's. As explained earlier, the fitness $\tilde{f}$ is reached when J is close to $\bar{J} \propto \ln L$ and therefore we expect our approximation to work well for $J \ll \ln L$.

When the underlying fitness distribution is exponential, we find that the fitness distribution of the fixed beneficial mutation also has an exponential

tail (see (4.43)). The robustness of this result *i.e.* whether any fitness distribution in the Gumbel class exhibits exponential tail for $P_J(f)$ is however not clear. For uniformly distributed fitnesses, as the width of the distribution $1 - q^L(f)$ decreases with increasing L, the step distribution approximation works better in this case than in the exponential case where the width is a constant (compare Figs. 4.4 and 4.6).

In the next and the last Chapter of this thesis, we shall present a few preliminary results on other statistical properties of adaptive walk and a brief outlook of the problems we hope to address in the near future.

Bibliography

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

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

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

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

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

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

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

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

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

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

[11] M.J. McDonald, T. F. Cooper, H. J. E. Beaumont, and P. B. Rainey. *Biol. Lett.*, 7:98–100, 2010.

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

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

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

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

[16] C. Carneiro and D.L. Hartl. *Proc. Natl. Acad. Sci. USA*, 107:1747–1751, 2010.

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

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

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

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

[21] J. Mathews and R. L. Walker. *Mathematical methods of physics*. Pearson Education Limited, 1970.

[22] G. E. Blonder, M. Tinkham, and T. M. Klapwijk. *Phys. Rev. B*, 25:4515, 1982.

[23] van B. Schaeybroeck and A. Lazarides. *Phys. Rev. A*, 79:053612, 2009.

[24] N.A. Rosenberg. *J. theor. Biol.*, 237:17–22, 2005.

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

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

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

[28] E. D. Weinberger. *Phys. Rev. A*, 44:6399–6413, 1991.

[29] S. A. Kauffman. *The Origins of Order*. Oxford University Press, New York, 1993.

Chapter 5

Summary and outlook

5.1 Summary of the results

In our work, we studied the adaptation process using a quasispecies model for infinite populations and an adaptive walk model for finite populations. A block model was used to introduce fitness correlations in both the cases. In this last Chapter, we give a brief summary of our results and the relation between the models that we have used. The final part of the Chapter discusses the questions that we hope to address in the near future.

We first compare the deterministically evolving populations of infinite size studied here vis-a-vis finite populations that are subject to stochastic fluctuations on multi-peaked fitness landscapes. The basic difference between a finite and an infinite population is that while the former has a finite mutational spread in the sequence space [1], all the mutants are available at all times in the deterministic case. In infinite population, a transition to a higher fitness peak takes place by *overtaking* the less fitter populations

as explained in Chapter 3. Also the most populated sequence involved in the jump event is not necessarily a local maximum (for any correlation) for infinite populations. To see this, consider the fittest sequence with fitness $w^{(max)}(D)$ at a fixed number of mutations D from the initial sequence $\sigma^{(0)}$. Barring the initial sequence, all the one-mutant neighbors of sequence with fitness $w^{(max)}(1)$ are at mutational distance two from the initial sequence. Consider the scenario when the sequence with fitness $w^{(max)}(2)$ is a nearest neighbor of sequence with fitness $w^{(max)}(1)$. Then the fittest sequence at distance unity from the initial sequence can be a jump if at least $w^{(max)}(1) > w(\sigma^{(0)})$ and the minimum intersection time condition $(w^{(max)}(1) - w(\sigma^{(0)}))^{-1} < 2(w^{(max)}(2) - w(\sigma^{(0)}))^{-1}$ is obeyed. Clearly the latter condition rewritten as $w^{(max)}(2) - w^{(max)}(1) < w^{(max)}(1) - w(\sigma^{(0)})$ can be satisfied even when $w^{(max)}(1)$ is not a local maximum. Thus the number of jump events are not related to the number of local optima for an infinite population. In contrast, on rugged fitness landscapes, a finite population can get trapped at a local optimum from which it can escape by *tunneling* through a fitness valley [2]. In fact at late times, most of the population passes exclusively through the local fitness peaks and thus such sequences are the most populated ones when the population size is finite.

In the study of quasispecies model, our main concern is how the fitness correlations affect the dynamics. By varying the fitness correlations using the block model from strongly correlated fitnesses ($L_B = 1, 2$) to weakly correlated fitnesses ($L_B = L, L/2$), we found that the temporal distribution of the last jump has $1/t^2$ dependence which suggests that this property maybe universal, in that it is independent of the fitness correlations. The average

number of records was found to increase linearly with sequence length L, with a prefactor that increases with correlations. The results are however independent of the underlying fitness distribution, $p(f)$. On the other hand, the average number of jumps was found to increase with the sequence length as $\sqrt{L}$ for weakly correlated fitnesses but linearly with L for strongly correlated ones. Also this number depends on the underlying fitness distribution.

In the adaptive walk model, the effect of correlations on the length of the adaptive walk was studied. We showed that the walk length is dependent on the logarithmic length of their genomic sequence and as in the case of the quasispecies model, here also for fixed L, this number increases linearly with the number of blocks and hence correlations.

5.2 Relation between the quasispecies and adaptive walk models

In the adaptive walk model used in Chapter 4, at every step L in the walk new one mutant fitnesses are produced and the population moves to one of them according to the transition probability defined in (4.1). We found the relation between the walk length and sequence length for two choices of $p(f)$. Recently, this has been generalised to a larger class of fitness distributions without using the step distribution approximation [3] (see Chapter 4).

A simpler adaptive walk model, in which the walk proceeds in a fixed neighbourhood has also been studied. Here the population performs an adaptive walk on a space of L mutually accessible alleles till it reaches the fittest

sequence [4]. For various fitness distributions, the mean walk length was calculated analytically [5] in this model by averaging the time taken to reach the fittest sequence with respect to the fitness distribution and the result was found to be identical to the full model considered here. It was noted in [5] that the transition probability from the present sequence to a one mutant neighbour in the adaptive walk model and the rate at which the dominant sequence jumps from one sequence to another in quasispecies model depend linearly on the fitness difference between the two. As a result the number of jumps for i.i.d fitnesses in quasispecies model [6, 7] and the walk length distribution in the adaptive walk model are found to be identical.

5.3 Future work

The properties of multiple steps in an adaptive walk have been studied in some recent experiments [8, 9]. In Schoustra et al., [9] 118 replicate evolving populations of a fungus *Aspergillus nidulans* were studied and the number and fitness effects of new mutations in each lineage was measured. They find a negative correlation between the mean fitness of an evolving population and the fraction of beneficial mutations available at each step, thus indicating that the supply of beneficial mutations is depleted as the fitness of the population increases. Also, it was observed in this experiment that the increase in fitness becomes smaller with successive mutations indicating reduction in the selection coefficient values.

We intend to calculate in our model the distribution $P(s_J)$ of the selection coefficient $s_J = (f_J - f_{J-1})/f_{J-1}$ at the Jth step in the adaptive walk. As

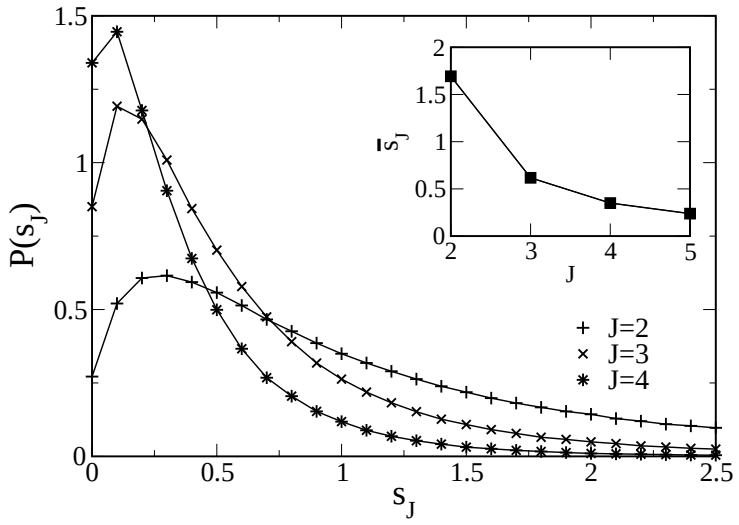


Figure 5.1: Distribution $P(s_J)$ of selection coefficient s_J for $L = 1000$ and $p(f) = e^{-f}$. The inset shows the decay in average selection coefficient $\bar{s}_J$ as a function of J. The points are joined by line to guide the eye.

we start with zero fitness, the selection coefficient is defined for $J \geq 2$. Our preliminary numerical results for $P(s_J)$ are shown in Fig. 5.1 for the first few steps in the walk and we observe that the typical selection coefficient decreases as the walk proceeds. This behavior matches qualitatively with the experimental results of [9]. A theoretical analysis of the distribution $P(s_J)$ requires the joint distribution of the fitness at step $J-1$ and J and we hope to address this question in a future work. We are also interested in the time taken for a fitter mutant to get fixed in a population and how this is related to the fitness difference between the present sequence and the mutant one.

Bibliography

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

[2] K. Jain and J. Krug. *Genetics*, 175:1275, 2007.

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

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

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

[6] J. Krug and C. Karl. *Physica A*, 318:137–143, 2003.

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

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

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

Appendix A

Adaptive walk model for uniformly distributed fitness

Solution of differential equation 4.49 The generating function $G_<(x, f)$ obeys the following inhomogeneous second order differential equation:

$$G''(x, f) - \frac{2x}{(1-f)^2} G(x, f) = \frac{4f}{(1-f)^2} \tag{A.1}$$

where we have dropped the subscript for brevity. The general solution of such differential equations is a linear combination of the general solution $G_H(x, f)$ of the homogeneous equation obtained by setting the right hand side equal to zero and the particular solution G_P of the inhomogeneous equation [1]. The homogeneous solution is of the form

$$G_H(x, f) = a_+(1-f)^{\alpha_+} + a_-(1-f)^{\alpha_-} \tag{A.2}$$

where $\alpha_\pm$ are the solutions of the quadratic equation $\alpha^2 - \alpha - 2x = 0$ and given

by (4.51). The particular solution is found using the method of variation of parameters and is of the form $G_P(x,f) = u_+(x)(1-f)^{\alpha_+} + u_-(x)(1-f)^{\alpha_-}$ where the functions $u_\pm(f)$ obey the following first order differential equations [1]:

$$u'_+(f)(1-f)^{\alpha_+} + u'_-(f)(1-f)^{\alpha_-} = 0 \qquad \text{(A.3)}$$

$$\alpha_+ u'_+(f)(1-f)^{\alpha_+-1} + \alpha_- u'_-(f)(1-f)^{\alpha_--1} = \frac{4f}{(1-f)^2} \qquad \text{(A.4)}$$

On solving the above equations, we obtain

$$G_P(x,f) = \frac{4}{\alpha_+\alpha_-} - \frac{4(1-f)}{(1-\alpha_+)(1-\alpha_-)} = \frac{-2f}{x} \qquad \text{(A.5)}$$

Finally using the boundary conditions in the general solution $G_<(x,f) = G_P(x,f) + G_H(x,f)$, the desired result (4.52) is obtained.

Distribution of fixed beneficial mutations: The fitness distribution found using (4.52) and (4.53) is given below for the first few adaptive steps:

$$P_1(f) = 2f \quad , \ f \le 1 \qquad \text{(A.6)}$$

$$P_2(f) = \begin{cases} -8f + 4(f-2)\ln(1-f) & , \ f \le \tilde{f} \\ \frac{4\tilde{f}(f+\tilde{f}-2)}{1-\tilde{f}} + 4(f-2)\ln(1-\tilde{f}) & , \ f > \tilde{f} \end{cases} \qquad \text{(A.7)}$$

$$P_3(f) = 4\begin{cases} 12f + \ln(1-f)(12-6f+f\ln(1-f)) & , \ f \le \tilde{f} \\ \frac{1}{1-\tilde{f}}\Big[6\tilde{f}(2-f-\tilde{f}) + 2(6-(6-\tilde{f})\tilde{f} - f(3-2\tilde{f}))\ln(1-\tilde{f}) & \\ \quad + f(1-\tilde{f})\ln^2(1-\tilde{f})\Big] & , \ f > \tilde{f} \end{cases} \qquad \text{(A.8)}$$

$$P_4(f) = \frac{-8}{3} \begin{cases} 120f + 60(2-f)\ln(1-f) + 12f\ln^2(1-f) + (2-f)\ln^3(1-f)\ , & f \le \tilde{f} \\ \frac{1}{(1-\tilde{f})}\Big[60\tilde{f}(2-f-\tilde{f}) - 12(f(5-3\tilde{f}) - 2(5-(5-\tilde{f})\tilde{f}))\ln(1-\tilde{f}) & \\ + 3(f(2-3\tilde{f}) + (2-\tilde{f})\tilde{f})\ln^2(1-\tilde{f}) + (2-f)(1-\tilde{f})\ln^3(1-\tilde{f})\Big]\ , & f > \tilde{f} \end{cases} \tag{A.9}$$

Walk length distribution: On matching powers of x^J on both sides in (4.55), we get

$$Q_1 = e^{-2\ell}(-1 + 2e^{\ell}) \tag{A.10}$$

$$Q_2 = 2e^{-2\ell}(3 + \ell + (-3 + 2\ell)e^{\ell}) \tag{A.11}$$

$$Q_3 = e^{-2\ell}\left[-2(18 + 8\ell + \ell^2) + 4e^{\ell}(9 - 5\ell + \ell^2)\right] \tag{A.12}$$

$$Q_4 = \frac{4e^{-2\ell}}{3}\left[180 + 84\ell + 15\ell^2 + \ell^3 + e^{\ell}(-180 + 96\ell - 21\ell^2 + 2\ell^3)\right] \tag{A.13}$$

where $\ell = \ln L$. A general solution of Q_J by this method does not seem possible but an approximate analytic expression for Q_J can be obtained as explained below.

From the definition of the generating function $H(x)$ in (4.55), it follows that

$$Q_J = \frac{1}{J!}\frac{d^J H(x)}{dx^J}\bigg|_{x=0} \tag{A.14}$$

By the residue theorem for complex variables, we have [1]

$$\frac{1}{2\pi i}\int_C dz\ f(z) = \frac{1}{n!}\frac{d^n}{dz^n}\left((z - z_0)^{n+1} f(z)\right)\bigg|_{z=z_0} \tag{A.15}$$

where z_0 is a pole of order $n + 1$ of the function $f(z)$ and the contour C

encloses the singularities of $f(z)$. From (A.14) and (A.15), we can write

$$Q_J = \frac{1}{2\pi i}\int_C dz\ \frac{H(z)}{z^{J+1}} = \frac{1}{2\pi i}\int_C dz\ e^{K(z)} \tag{A.16}$$

where $K(z) = \ln H(z) - (J+1)\ln z$. We solve this integral by the method of steepest descent which for large J gives [1]

$$Q_J \approx \sqrt{\frac{1}{2\pi K''(z_s)}}e^{K(z_s)} = \sqrt{\frac{1}{2\pi K''(z_s)}}\frac{H(z_s)}{z_s^{J+1}} \tag{A.17}$$

where prime refers to derivative with respect to z. In the above equation, z_s is a solution of the equation

$$\frac{H'(z_s)}{H(z_s)} = \frac{J}{z_s} \tag{A.18}$$

and

$$K''(z_s) = \left(\frac{H'(z)}{H(z)}\right)'\Big|_{z=z_s} + \frac{J}{z_s^2} \tag{A.19}$$

$$= \left(\frac{H'(z)}{H(z)}\right)'\Big|_{z=z_s} + \frac{1}{z_s}\frac{H'(z_s)}{H(z_s)} \tag{A.20}$$

where prime denotes a derivative with respect to z. Since $\alpha_+ > 0$, neglecting the exponentially small term in $(1-\tilde{f})^{\alpha_+}$ in (4.55), we get

$$H(z) \approx \frac{e^{-3\ell/2}e^{\ell y/2}(3+y)(y^2-1)}{16y} \tag{A.21}$$

where $y = \sqrt{1+8z}$. Differentiating $H(z)$ once with respect to z gives

$$\frac{H'(z)}{H(z)} \approx \frac{8(y+3)+4(2y+3)(y^2-1)+2y(y+3)(y^2-1)\ell}{y^2(y^2-1)(y+3)} \tag{A.22}$$

Using the above expression in (A.18) for large y, we get $y_s \approx 4J/\ell$ and therefore

$$z_s \approx \frac{2J^2}{\ell^2} \tag{A.23}$$

On differentiating (A.22) once, we have

$$\left(\frac{H'(z)}{H(z)}\right)' \approx \frac{4}{y}\left[\frac{4}{3(y+3)^2}+\frac{4}{(1+y)^2}-\frac{4+6\ell}{3y^2}+\frac{8}{y^3}-\frac{4}{(1-y)^2}\right] \tag{A.24}$$

Using (A.22) and (A.24) in (A.20), we obtain

$$K''(z_s) \approx \frac{8\left[-36+6y_s(y_s^2-3)+y_s(y_s+3)^2(1+y_s)^2\ell\right]}{y_s^4(y_s+3)^2(y_s^2-1)} \tag{A.25}$$

$$\approx \frac{8\ell}{y_s^3} = \frac{a^4}{8J^3} \tag{A.26}$$

Thus we have

$$Q_J \approx \frac{2J^{3/2}}{\sqrt{\pi}\ell^2} \times \frac{2-\alpha_-(z_s)}{\alpha_+(z_s)-\alpha_-(z_s)} \times \frac{(1-\tilde{f})^{1+\alpha_-(z_s)}}{z_s^J} \tag{A.27}$$

where $\alpha_\pm$ is given by (4.51).

Bibliography

[1] J. Mathews and R. L. Walker. *Mathematical methods of physics*. Pearson Education Limited, 1970.